

5-1-2003

## Vol. 2, No. 1 (Full Issue)

JMASM Editors

Follow this and additional works at: <http://digitalcommons.wayne.edu/jmasm>

---

### Recommended Citation

Editors, JMASM (2003) "Vol. 2, No. 1 (Full Issue)," *Journal of Modern Applied Statistical Methods*: Vol. 2 : Iss. 1 , Article 32.

DOI: 10.22237/jmasm/1051747200

Available at: <http://digitalcommons.wayne.edu/jmasm/vol2/iss1/32>

This Full Issue is brought to you for free and open access by the Open Access Journals at DigitalCommons@WayneState. It has been accepted for inclusion in Journal of Modern Applied Statistical Methods by an authorized editor of DigitalCommons@WayneState.



# Lahey/Fujitsu Fortran

The standard for Fortran programming  
from the leader in Fortran language systems

**SOFTWARE SOLUTIONS**  
for Science & Engineering

## LF95 Fortran for Linux and Windows

Full Fortran 95/90/77 support  
Unsurpassed diagnostics  
Intel and AMD optimizations

IMSL compatible  
Fujitsu SSL2 math library  
Wisk graphics package

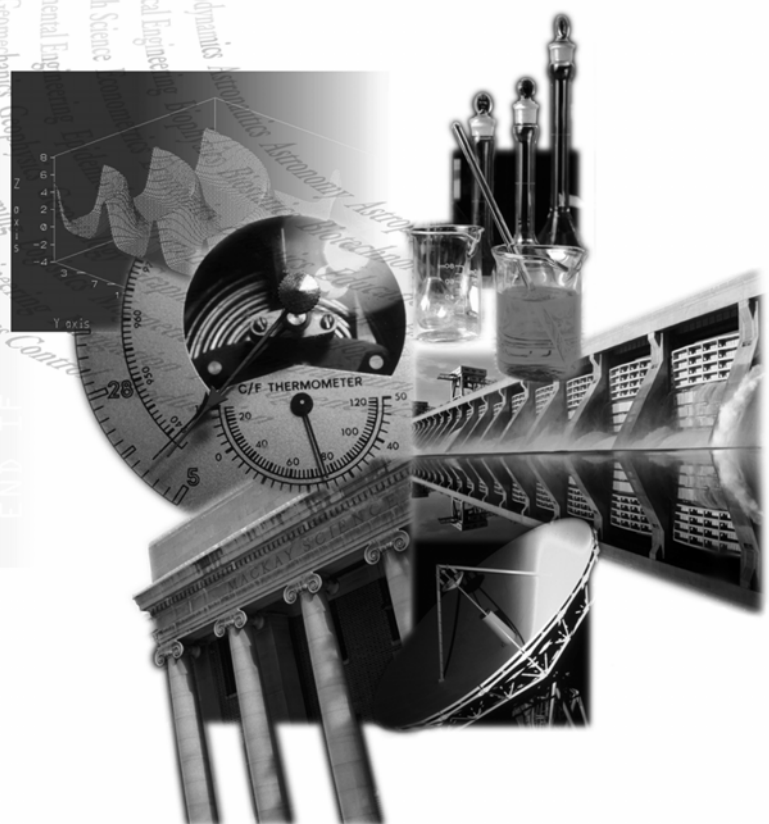
## LF Fortran for the Microsoft® .NET Framework - Coming Soon !

Visual Studio integration  
Windows / Web Forms designer  
Project and code templates

On-line integrated help  
XML Web services  
ADO.NET support

Visit [www.lahey.com](http://www.lahey.com) for more information

```
ELSE
  poly_coef
END IF
ELSE
  poly_coef
END IF
END FUNCTION poly_c
SUBROUTINE poly_ini
TYPE(poly), INTENT
REAL(fpkind), INTE
IF ( .NOT. PRESENT
  NULLIFY ( p%coef
ELSE
  m = UBOUND(v,i)
  IF ( max_degree
  ALLOCATE ( p%
  p%coeffs
ELSE
  ALLOC
  p%coeffs
END IF
```



Lahey Computer Systems, Inc.  
865 Tahoe Blvd - P.O. Box 6091  
Incline Village, NV 89450 USA  
1-775-831-2500  
[www.lahey.com](http://www.lahey.com)

## Journal Of Modern Applied Statistical Methods

### *Invited Articles*

- |         |                                       |   |
|---------|---------------------------------------|---|
| 2 – 13  | <b>George Marsaglia</b>               | Random Number Generators                                  |
| 14 – 15 | <b>Walt Brainerd</b>                  | The Importance Of Fortran In The 21 <sup>st</sup> Century |
| 16 – 26 | <b>Tom Richards,<br/>Lyn Richards</b> | The Way Ahead In Qualitative Computing                    |

### *Regular Articles*

- |           |   |  |
|-----------|---|--|
| 27 – 49   | <b>J. D. Opdyke</b>                                       | Fast Permutation Tests That Maximize Power Under Conventional Monte Carlo Sampling For Pairwise And Multiple Comparisons |
| 50 – 72   | <b>James Algina,<br/>H. J. Keselman,<br/>A. R. Othman</b> | Analyzing Group By Time Effects In Longitudinal Two-Group Randomized Trial Designs With Missing Data                     |
| 73 – 79   | <b>Wim Van den Noortgate,<br/>Patrick Onghena</b>         | A Parametric Bootstrap Version Of Hedges' Homogeneity Test   |
| 80 – 86   | <b>Vance W. Berger,<br/>Costas A. Christophi</b>          | Randomization Technique, Allocation Concealment, Masking, And Susceptibility Of Trials To Selection Bias                 |
| 87 – 107  | <b>H. Evangelaras,<br/>C. Koukouvinos</b>                 | Screening Properties And Design Selection Of Certain Two-Level Designs   |
| 108 – 127 | <b>Christopher W. T. Chiu,<br/>Ronald S. Fecso</b>        | Incorporating Sampling Weights Into The Generalizability Theory For Large-Scale Analyses                                 |
| 128 – 132 | <b>Shlomo S. Sawilowsky</b>                               | A Different Future For Social And Behavioral Science Research  |
| 133 – 151 | <b>Anthony J. Onwuegbuzie,<br/>Joel R. Levin</b>          | Supporting Statistical Evidence  |
| 152 – 160 | <b>Scott J. Richter,<br/>Mark E. Payton</b>               | Performing Two-Way Analysis Of Variance Under Variance Heterogeneity   |
| 161 – 167 | <b>Mourad Tighiouart</b>                                  | Modeling Correlated Time-Varying Covariate Effects In A Cox-Type Regression Model  |
| 168 – 176 | <b>B. Sango Otieno,<br/>C. M. Anderson-Cook</b>           | A More Efficient Way Of Obtaining A Unique Median Estimate For Circular Data   |

177 – 188	<b>Chao-Ying Joanne Peng, Rebecca Naegle Nichols</b>	Using Multinomial Logistic Models To Predict Adolescent Behavioral Risk
189 – 194	<b>Dominique Haughton, Nguyen Phong</b>	Bayesian Analysis Of Poverty Rates: The Case Of Vietnamese Provinces
195 – 201	<b>Daniel X. Wang</b>	Comparisons Of Estimates Of Proprietary And Syndicated Methods In Auto Industry Surveys
202 – 209	<b>Kailash C. Madan, Walid Abu-Dayyeh, Firas Tayyan</b>	Steady State Analysis Of An M/D/2 Queue With Bernoulli Schedule Server Vacations
210 – 217	<b>Ricardo Ocaña-Riola, Emilio Sanchez-Cantalejo, Carmen Martinez-Garcia</b>	Homogeneous Markov Processes For Breast Cancer Analysis

### *Invited Debate*

218 – 225	<b>Shlomo S. Sawilowsky</b>	<i>Target Article:</i> You Think You've Got Trivials?
226 – 230	<b>J. Kyle Roberts, Robin K. Henson</b>	<i>Response:</i> Not All Effects Are Created Equal
231 – 236	<b>Joel R. Levin, Daniel H. Robinson</b>	<i>Comment:</i> The Trouble With Interpreting Statistically Nonsignificant Effect Sizes In Single-Study Investigations
237 – 241	<b>Thomas R. Knapp</b>	<i>Comment:</i> Was Monte Carlo Necessary?
242 – 246	<b>Shlomo S. Sawilowsky</b>	<i>Rejoinder:</i> Trivials: The Birth, Sale, And Final Production Of Meta-Analysis

### *Early Scholars*

247 – 255	<b>Laura K. Miller, Ping Sa</b>	Improved Multiple Comparisons With The Best In Response Surface Methodology
256 – 267	<b>Jianhua Hu, Guosheng Yin</b>	A Semiparametric Regression Model For Oligonucleotide Arrays

### *JMASM Algorithms & Code*

268 – 271	<b>Todd C. Headrick</b>	JMASM6: An Algorithm For Generating Exact Critical Values For The Kruskal-Wallis One-Way ANOVA (Fortran 77)
272 – 278	<b>Joseph McCarthy, Robert DiSario, Hakan Saraoglu</b>	JMASM7: A Recursive Algorithm For Fractionally Differencing Long Data Series (SAS)

## Journal Of Modern Applied Statistical Methods

*JMASM* is an independent print and electronic journal (<http://tbf.coe.wayne.edu/jmasm>) designed to provide an outlet for the scholarly works of applied nonparametric or parametric statisticians, data analysts, researchers, classical or modern psychometricians, quantitative or qualitative evaluators, and methodologists. Work appearing in *Regular Articles*, *Brief Reports*, and *Early Scholars* are externally peer reviewed, with input from the Editorial Board; in *Statistical Software Applications and Review* and *JMASM Algorithms and Code* are internally reviewed by the Editorial Board.

Three areas are appropriate for *JMASM*: (1) development or study of new statistical tests or procedures, or the comparison of existing statistical tests or procedures, using computer-intensive Monte Carlo, bootstrap, jackknife, or resampling methods, (2) development or study of nonparametric, robust, permutation, exact, and approximate randomization methods, and (3) applications of computer programming, preferably in Fortran (all other programming environments are welcome), related to statistical algorithms, pseudo-random number generators, simulation techniques, and self-contained executable code to carry out new or interesting statistical methods. Elegant derivations, as well as articles with no take-home message to practitioners, have low priority. Articles based on Monte Carlo (and other computer-intensive) methods designed to evaluate new or existing techniques or practices, particularly as they relate to novel applications of modern methods to everyday data analysis problems, have high priority.

Problems may arise from applied statistics and data analysis; experimental and nonexperimental research design; psychometry, testing, and measurement; and quantitative or qualitative evaluation. They should relate to the social and behavioral sciences, especially education and psychology. Applications from other traditions, such as actuarial statistics, biometrics or biostatistics, chemometrics, econometrics, environmetrics, jurimetrics, quality control, and sociometrics are welcome. Applied methods from other disciplines (e.g., astronomy, business, engineering, genetics, logic, nursing, marketing, medicine, oceanography, pharmacy, physics, political science) are acceptable if the demonstration holds promise for the social and behavioral sciences.

---

Editorial Assistant  
**Holly Atkins**

Professional Staff  
**Bruce Fay,**  
Business Manager  
**Joe Musial,**  
Marketing Director

Production Staff  
**Holly Atkins**  
**Christina Gase**  
**Bulent Ozkan**  
**Sarah Sawilowsky**  
**Jack Sawilowsky**

Internet Sponsor  
**College of Education,**  
**Wayne State University**

Entire Reproductions and Imaging Solutions Internet: <a href="http://www.entire-repro.com">www.entire-repro.com</a>	248.299.8900 (Phone) 248.299.8916 (Fax)	e-mail: <a href="mailto:sales@entire-repro.com">sales@entire-repro.com</a>
--	--	---

## Editorial Board of Journal of Modern Applied Statistical Methods

Subhash Chandra Bagui  
Department of Mathematics & Statistics  
University of West Florida

Chris Barker  
MEDTAP International  
Redwood City, CA

J. Jackson Barnette  
Community and Behavioral Health  
University of Iowa

Vincent A. R. Camara  
Department of Mathematics  
University of South Florida

Ling Chen  
Department of Statistics  
Florida International University

Christopher W. Chiu  
Test Development & Psychometric Rsch  
Law School Admission Council, PA

Jai Won Choi  
National Center for Health Statistics  
Hyattsville, MD

Rahul Dhanda  
Forest Pharmaceuticals  
New York, NY

John N. Dyer  
Dept. of Information System & Logistics  
Georgia Southern University

Matthew E. Elam  
Dept. of Industrial Engineering  
University of Alabama

Mohammed A. El-Saidi  
Accounting, Finance, Economics &  
Statistics, Ferris State University

Carol J. Etzel  
University of Texas M. D.  
Anderson Cancer Center

Felix Famoye  
Department of Mathematics  
Central Michigan University

Barbara Foster  
Academic Computing Services, UT  
Southwestern Medical Center, Dallas

Shiva Gautam  
Department of Preventive Medicine  
Vanderbilt University

Dominique Haughton  
Mathematical Sciences Department  
Bentley College

Scott L. Hershberger  
Department of Psychology  
California State University, Long Beach

Joseph Hilbe  
Departments of Statistics/ Sociology  
Arizona State University

Peng Huang  
Dept. of Biometry & Epidemiology  
Medical University of South Carolina

Sin-Ho Jung  
Dept. of Biostatistics & Bioinformatics  
Duke University

Jong-Min Kim  
Statistics, Division of Science & Math  
University of Minnesota

Harry Khamis  
Statistical Consulting Center  
Wright State University

Kallappa M. Koti  
Food and Drug Administration  
Rockville, MD

Tomasz J. Kozubowski  
Department of Mathematics  
University of Nevada

Kwan R. Lee  
GlaxoSmithKline Pharmaceuticals  
Collegeville, PA

Hee-Jeong Lim  
Dept. of Math & Computer Science  
Northern Kentucky University

Devan V. Mehrotra  
Merck Research Laboratories  
Blue Bell, PA

Prem Narain  
Freelance Researcher  
Farmington Hills, MI

Balgobin Nandram  
Department of Mathematical Sciences  
Worcester Polytechnic Institute

J. Sunil Rao  
Dept. of Epidemiology & Biostatistics  
Case Western Reserve University

Brent Jay Shelton  
Department of Biostatistics  
University of Alabama at Birmingham

Karan P. Singh  
University of North Texas Health  
Science Center, Fort Worth

Jianguo (Tony) Sun  
Department of Statistics  
University of Missouri, Columbia

Joshua M. Tebb  
Department of Statistics  
Oklahoma State University

Dimitrios D. Thomakos  
Department of Economics  
Florida International University

Justin Tobias  
Department of Economics  
University of California-Irvine

Jeffrey E. Vaks  
Beckman Coulter  
Brea, CA

Dawn M. VanLeeuwen  
Agricultural & Extension Education  
New Mexico State University

David Walker  
Educational Tech, Rsrch, & Assessment  
Northern Illinois University

J. J. Wang  
Dept. of Advanced Educational Studies  
California State University, Bakersfield

Dongfeng Wu  
Dept. of Mathematics & Statistics  
Mississippi State University

Chengjie Xiong  
Division of Biostatistics  
Washington University in St. Louis

Andrei Yakovlev  
Biostatistics and Computational Biology  
University of Rochester

Heping Zhang  
Dept. of Epidemiology & Public Health  
Yale University

**International**  
Mohammed Ibrahim Ali Ageel  
Department of Mathematics  
King Khalid University, Saudi Arabia

Mohammad Fraiwan Al-Saleh  
Department of Statistics  
Yarmouk University, Irbid-Jordan

Keumhee Chough (K.C.) Carriere  
Mathematical & Statistical Sciences  
University of Alberta, Canada

Debasis Kundu  
Department of Mathematics  
Indian Institute of Technology, India

Christos Koukouvinos  
Department of Mathematics  
National Technical University, Greece

Lisa M. Lix  
Dept. of Community Health Sciences  
University of Manitoba, Canada

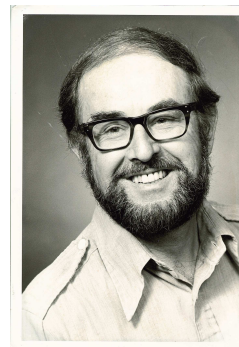
Takis Papaioannou  
Statistics and Insurance Science  
University of Piraeus, Greece

Mohammad Z. Raqab  
Department of Mathematics  
University of Jordan, Jordan

Nasrollah Saebi  
School of Mathematics  
Kingston University, UK

## Random Number Generators

George Marsaglia  
Professor Emeritus of Statistics  
Florida State University



---

The author discusses some promising new random number generators, as well as formulates the mathematical basis that makes them random variables in the same sense as more familiar ones in probability and statistics, emphasizing his view that randomness exists only in the sense of mathematics. He discusses the need for adequate seeds that provide the axioms for that mathematical basis, and gives examples from Law and Gaming, where inadequacies have led to difficulties. He also describes new versions of the widely used Diehard Battery of Tests of Randomness.

Key words: Random number generator, Diehard Test

---

### Introduction

In 1985 I was invited to give the keynote address “A current view of random number generators” at Statistics and Computer Science: XVI Symposium on the Interface. An article based on that address was published in the Proceedings of that conference,[5]. Judging from newsgroups and citations, the article seems to have been widely

---

George Marsaglia is Professor Emeritus of Statistics at Florida State University & Professor Emeritus of Pure and Applied Mathematics & Computer Science at Washington State University. His PhD was in Mathematics under H.B. Mann at Ohio State Univ., 1950, and he was a Fulbright Scholar under M. S. Bartlett and Alan Turing at Univ. of Manchester in 1949-50, then Associate under Harold Hotelling at Univ. N. Carolina. He was Professor and Director of the School of Computer Science at McGill Univ. 1970-78. He has published articles in over fifty math, computer science, statistics, physics, medicine and law journals, & is probably best known for work on random numbers, generating non-uniform variates and testing for randomness. His email address is [geo@stat.fsu.edu](mailto:geo@stat.fsu.edu)

read, although such proceedings are often difficult to access. Availability of the file `keynote.ps` in the CDROM [6], [stat.fsu.edu/pub/diehard](http://stat.fsu.edu/pub/diehard), may have made the article easier to get. Two other postscript files in that CDROM provide more detail on topics of the present article: `mwc1.ps` and `monkey.ps`.

In this article I will update that “current” view, dwelling at some length on what I see as more important kinds of RNGs, particularly Multiply-With-Carry (MWC) and Complimentary-Multiply-With-Carry (CMWC), because they have simple implementations, are very fast, can have incredibly long periods and pass tests randomness at least as well as, and often better than, other kinds of RNGs.

But first I will provide a summary discussion of *congruential* RNGs, because they remain the most common kind, and of *xorshift* RNGs, because they are as fast and simple as congruential but better behaved in tests of randomness. I will list all 648 of the full-period, 32-bit xorshift RNGs. There will also be a short description of *lagged Fibonacci* RNGs. These have diminished in importance, because MWC and CMWC RNGs provide far far longer periods for the same effort, and have better performance on tests.

But one kind is still important because it can provide floating point uniforms directly, without the usual floating of integers.

I will also dwell on the problem of seeds and their relation to randomness, and on the need for an adequate number of seeds that has arisen in Law and Gaming. Finally, I will discuss the latest version of my *DIEHARD Battery of Tests of Randomness* [6], which includes some new, difficult-to-pass tests.

#### Random Number Generators (RNGs)

The mathematics of random number generators requires a set  $\mathcal{Z}$ , an invertible function  $f$  over  $\mathcal{Z}$ , and, for a random choice of a seed  $z$  from  $\mathcal{Z}$ , the sequence of random values in  $\mathcal{Z}$  produced by iterating the function  $f$ :

$$f(z), f^2(z), f^3(z), \dots,$$

where  $f^2(z)$  means  $f(f(z))$ ,  $f^3(z)$  means  $f(f^2(z))$ , etc. Sometimes—in fact, most often—the set  $\mathcal{Z}$  is just the set of integers represented by 32-bit computer words, but for RNGs that meet more stringent requirements, the set  $\mathcal{Z}$  might be the set of all  $m$ -tuples  $(x_1, x_2, \dots, x_m)$  of 32-bit integers, and  $f$  a function that converts one such  $m$ -tuple into another.

If  $f$  is a one-to-one function over  $\mathcal{Z}$ , then for any seed  $z$  chosen uniformly from  $\mathcal{Z}$ , the random variable  $f(z)$  is also uniformly distributed over  $\mathcal{Z}$ . (Just as if you randomly choose a digit  $d$  from  $\mathcal{Z} = \{0, 1, 2, \dots, 8, 9\}$  and I instead choose  $3d + 5 \bmod 10$ , my choice has the same uniform distribution as yours, since  $f(x) = 3x + 5 \bmod 10$  is one-to-one over  $\mathcal{Z}$ .) For the general case with seed set  $\mathcal{Z}$ , the choice of a random seed  $z$  from  $\mathcal{Z}$  will provide, through  $f(z), f^2(z), \dots$  a long sequence of uniform random choices from  $\mathcal{Z}$ . They will not be independent random choices, but for many purposes they may behave as though they were, allowing us, with the minimal effort of choosing a random seed from  $\mathcal{Z}$ , to provide the huge samples that many simulation studies call for.

Note that when  $\mathcal{Z}$  is  $\{(x_1, x_2, \dots, x_m)\}$ ,

a set of  $m$ -tuples, and a random seed  $z$  from  $\mathcal{Z}$  leads to a sequence of uniform but not-independent random choices from  $\mathcal{Z}$ :  $f(z), f^2(z), \dots$ , then the elements of each  $(x_1, x_2, \dots, x_m)$  may themselves be uniform over their range, and furthermore, substrings such as  $x_1, x_2, x_3$  may be quite close to uniform and independent over their joint product set. This suggests, as experience shows, that RNGs with seed sets  $\mathcal{Z}$  made up of  $m$ -tuples  $(x_1, x_2, \dots, x_m)$ , may be more desirable, although they may require that the user provide many more than the usual single random integer seed. Perhaps the axiom: “You get what you pay for” applies.

#### Congruential RNGs

Given a suitable modulus  $m$ , multiplier  $a$ , additive constant  $k$  and initial random seed  $x_0$ , use of the sequence  $x_n = ax_{n-1} + k \bmod m$  is probably the oldest and most common method of producing random integers. If  $a$  is a primitive root of the prime  $p$ , and  $x_0$  is a random seed from

$$\mathcal{Z} = \{1, 2, \dots, p-1\},$$

then the sequence generated by  $x_n = ax_{n-1} \bmod p$  will be strictly periodic, with period  $p-1$ , and each element of that sequence will be a uniform random variable on the set  $\mathcal{Z}$ , but of course they will not be independent.

Getting  $ax \bmod p$  for a prime  $p$  is usually much more difficult than getting  $ax \bmod 2^{32}$ , as the latter is virtually automatic in most CPUs. Thus sequences such as  $x_n = ax_{n-1} + k \bmod 2^{32}$ , with  $k$  odd and  $a = \pm 3 \bmod 8$  have dominated, since, given a random seed

$$x_0 \in \mathcal{Z} = \{0, 1, \dots, 2^{32}-1\},$$

each element in the sequence will be uniformly distributed over  $\mathcal{Z}$ , and the sequence will have period  $2^{32}$ .

Congruential RNGs have the flaw of “falling mainly in the planes”, [2]. For example, if  $x, y, z$  are any three successive integers produced by a congruential RNG with multiplier  $a$ , then the point  $(x, y, z)$  falls on the lattice of points generated by all linear combinations,



with integer coefficients, of the three points  $(1, a, a^2), (0, m, 0), (0, 0, m)$ ; any point  $(x, y)$  lies on the lattice generated by  $(1, a), (0, m)$ ; any four consecutive outputs provide a point  $(x, y, z, w)$  in 4-space that must fall on the lattice of integer combinations of the four points  $(1, a, a^2, a^3), (0, m, 0, 0), (0, 0, m, 0), (0, 0, 0, m)$ , etc.. See [2,3,4]. The last reference describes a simple way to characterize the lattice of a congruential RNG, in terms of ratios of edges of a unit cell—nearly cubic is better than long and thin.

Partly because it is one of the few exact measures for congruential RNGs, a widely used assessment is that of Knuth's 'Spectral test', [1] which dwells on the lattice structure only tangentially. The spectral test amounts to characterizing a lattice by the minimum distance between its hyperplanes. Although I discovered the lattice structure of congruential RNGs, I have never found it a very useful measure of their goodness or badness, but it remains a measure that is frequently taken, because of its exact, interesting mathematical underpinnings.

### Cracking a Congruential RNG

Because congruential RNGs are so common, and are often a system RNG, it may be worth pointing out a simple method for determining whether an unknown RNG is congruential, and if so, how to determine its modulus and multiplier. I have used this method for over thirty years, and it is implicit in references [2,3], but has not been stated explicitly in a journal before this. Suppose the rule for the RNG is  $x_n = ax_{n-1} + k \pmod m$ . Suppose  $\alpha, \beta, \gamma$  are any three points in the plane with coordinates successive integers produced by that congruential RNG. Then the determinant of the  $2 \times 2$  matrix with rows  $\beta - \alpha$  and  $\gamma - \alpha$  is the volume of the parallelepiped determined by the three points, and must be an integer multiple of  $m$ , the unit-cell volume of the lattice. Thus the gcd of five or six such determinants will usually provide  $m$ , from which  $a$  and  $k$  may be found.

For example, a certain simple RNG produces integers  
 308,785,930,695,864,237,1006,819,204,777,378,  
 495,376,357,70,747,356,...,  
 leading to points  $\alpha_1 = (308, 785), \alpha_2 = (785, 930),$   
 $\alpha_3 = (930, 695), \dots$

Then the parallelepiped determined by  $\alpha_1, \alpha_2, \alpha_3$  has volume 133120, that determined by  $\alpha_2, \alpha_3, \alpha_4$  has volume 30720, etc. The sequence of volumes determined by  $\alpha_i, \alpha_{i+1}, \alpha_{i+2}$  is 133120,30720,118784,263168,474112,..., and the gcd of the first two, then the first three, then the first four,..., leads to the sequence 10240,2048,1024,1024,1024,1024,..., and thus to the inference that  $m = 1024$ . Solving  $\{308a + k = 785, 785a + k = 930\} \pmod{1024}$  yields  $a = 69, k = 13$ . Thus, with  $x_0 = 308$ , the sequence  $x_n = 69x_{n-1} + 13 \pmod{1024}$  produces the above output of the RNG.

Query: Which congruential RNG produced 768,54,747,221,321,48,225,669,414,163,260, 723,127,119,420,685,809,630?

### Xorshift RNGs

Theory behind these RNGs is based on viewing a 32- (or 64-) bit integer as an element of a vector space with components in the field mod 2. For such, addition of two vectors can be implemented with the exclusive-or (xor) operation. That, combined with the shift operation, can be used to create certain linear transformations over that vector space. Here the seed set  $\mathcal{Z}$  is the set of all non-zero  $1 \times 32$  binary vectors and  $f$  is a linear transformation on  $\mathcal{Z}$ , represented by a  $32 \times 32$  binary matrix  $T$ , nonsingular. Then for a random seed  $y \in \mathcal{Z}$ , the sequence is  $yT, yT^2, yT^3 \dots$ . If, and only if, the order of  $T$  is  $2^{32} - 1$  in the group of  $32 \times 32$  nonsingular binary matrices, then sequence  $yT, yT^2, yT^3, \dots$  will have period  $2^{32} - 1$ .

Applications require a simple and fast way to form the matrix product  $yT$ , and that can be done if, say,  $T = (I + L^a)(I + R^b)(I + L^c)$ , where  $L$  is the matrix that effects a left shift of one (in  $C, y^\wedge = (y < < 1)$ ), so that  $yL^a$  in  $C$  is

$y^{\wedge} = (y \ll a)$ . The matrix  $R$ , the transpose of  $L$ , effects a right shift of one. Thus, for  $T = (I + L^a)(I + R^b)(I + L^c)$ , for a random 32-bit seed  $y$  from  $\mathcal{Z}$ , each new  $y$  in the sequence  $yT, yT^2, yT^3, \dots$  can be produced in C by successive application of the three instructions

$$y^{\wedge} = y \ll 13; \quad y^{\wedge} = y \gg 17; \quad y^{\wedge} = y \ll 5;$$

Such xorshift sequences are among the most desirable of simple RNGs: quick and easy, with seemingly better performance than congruential on tests of randomness.

For 32- (or 64-) bit binary vectors, there are no two-shift matrices  $T = (I + L^a)(I + R^b)$  that have full period, and certainly no one-shift, so 3-shift  $T$ 's are needed. There are 81 triples  $[a, b, c]$ ,  $a < c$ , for which the  $32 \times 32$  binary matrix  $T = (I + L^a)(I + R^b)(I + R^c)$  has order  $2^{32} - 1$ , listed in four columns:

1, 3, 10	1, 5, 16	1, 5, 19	1, 9, 29
1, 11, 6	1, 11, 16	1, 19, 3	1, 21, 20
1, 27, 27	2, 5, 15	2, 5, 21	2, 7, 7
2, 7, 9	2, 7, 25	2, 9, 15	2, 15, 17
2, 15, 25	2, 21, 9	3, 1, 14	3, 3, 26
3, 3, 28	3, 3, 29	3, 5, 20	3, 5, 22
3, 5, 25	3, 7, 29	3, 13, 7	3, 23, 25
3, 25, 24	3, 27, 11	4, 3, 17	4, 3, 27
4, 5, 15	5, 3, 21	5, 7, 22	5, 9, 7
5, 9, 28	5, 9, 31	5, 13, 6	5, 15, 17
5, 17, 13	5, 21, 12	5, 27, 8	5, 27, 21
5, 27, 25	5, 27, 28	6, 1, 11	6, 3, 17
6, 17, 9	6, 21, 7	6, 21, 13	7, 1, 9
7, 1, 18	7, 1, 25	7, 13, 25	7, 17, 21
7, 25, 12	7, 25, 20	8, 7, 23	8, 9, 23
9, 5, 1	9, 5, 25	9, 11, 19	9, 21, 16
10, 9, 21	10, 9, 25	11, 7, 12	11, 7, 16
11, 17, 13	11, 21, 13	12, 9, 23	13, 3, 17
13, 3, 27	13, 5, 19	13, 17, 15	14, 1, 15
14, 13, 15	15, 1, 29	17, 15, 20	17, 15, 23
17, 15, 26			

If  $T = (I + L^a)(I + R^b)(I + L^c)$  has full period, then so does  $(I + L^c)(I + R^b)(I + L^a)$ , and so does  $(I + L^a)(I + L^c)(I + R^b)$ , leading to  $4 \times 81$   $T$ 's with order  $2^{32} - 1$ . But then the transpose of each also has full period. That provides  $8 \times 81 = 648$  matrices. Any  $[a, b, c]$  in

the above table of 81 yields eight lines of C code:

```

y^=y<<a; y^=y>>b; y^=y<<c;
y^=y<<c; y^=y>>b; y^=y<<a;
y^=y>>a; y^=y<<b; y^=y>>c;
y^=y>>c; y^=y<<b; y^=y>>a;
y^=y<<a; y^=y<<c; y^=y>>b;
y^=y<<c; y^=y<<a; y^=y>>b;
y^=y>>a; y^=y>>c; y^=y<<b;
y^=y>>c; y^=y>>a; y^=y<<b;

```

In summary: for each of the above 81 triples  $[a, b, c]$  with  $a < c$ , any one of those eight lines of C can provide the instructions for a 32-bit RNG with period  $2^{32} - 1$ .

For 64-bit integers, there are  $8 \times 275$  or 2200 such xorshift  $T$ 's with periods of  $2^{64} - 1$ . A list is available from the author, as well as a fast C program for finding all full period xorshift  $T$ 's

#### Lagged Fibonacci RNGs

The basic recurrence for a lagged Fibonacci RNG is  $x_n = x_{n-r} \bullet x_{n-s}$ , for 'lags'  $r$  and  $s$ , with  $r > s$ . Here  $\bullet$  is a binary relation for pairs of elements in some set  $\mathcal{X}$  and the seed set  $\mathcal{Z}$  is the set of  $r$ -tuples  $(x_1, x_2, \dots, x_r)$  with the  $x$ 's in  $\mathcal{X}$ . Usually,  $\mathcal{X}$  is the set of 32-bit integers and  $\bullet$  is addition or subtraction mod  $2^{32}$  or addition of binary vectors (exclusive-or:  $\oplus$ ). A promising choice has  $\mathcal{X}$  the set of odd integers and  $\bullet$  multiplication mod  $2^{32}$ . Theory for the latter may be based on expressing elements  $x, y$  from  $\mathcal{X}$  in the form  $x = \pm 3^a, y = \pm 3^b \pmod{2^{32}}$  so that  $x \bullet y = \pm 3^{(a+b \pmod{2^{30}})}$  and the recurrence rules for addition mod  $2^{30}$  apply.

The notation  $F(r, s, \bullet)$  is used for a lagged Fibonacci RNG. For proper choice of the lags  $r, s$ , the period of  $F(r, s, \pm \pmod{2^{32}})$  can be  $2^{32+r}$ , while that of  $F(r, s, \oplus)$  will at best be  $2^r$ , whatever the word size. For proper choice  $r > s$ , the period of  $F(r, s, * \text{ on odds } \pmod{2^{32}})$  is  $2^{30}$ .

As with other RNGs, formal definition of lagged Fibonacci RNGs requires a seed set  $\mathcal{Z}$  and a function  $f$  on  $\mathcal{Z}$ . Here,  $\mathcal{Z}$  is the set  $\{(x_1, x_2, \dots, x_r)\}$ , with  $x$ 's in the set  $\mathcal{X}$  on which we have the binary relation, and the function  $f$ :

$$f([x_1, x_2, \dots, x_r]) = [x_2, \dots, x_r, x_1 \bullet x_{r-s+1}].$$

Implementing lagged Fibonacci sequences with lags  $r > s$  requires keeping a table of the  $r$  most recent values. Their periods, around  $2^{32+r}$ , are far short of the possible  $2^{32r}$  that is attainable with certain RNGs that also keep a table of the  $r$  most recent values, discussed in the next two sections. One of the most useful applications of lagged Fibonacci RNGs is in the generation of floating point uniform [0,1) variates directly, without the usual floating of random integers. For example, suppose we want to generate 64-bit (C's double or Fortran's double precision) uniform [0,1) random variables using the IEEE 754 standard: 1 sign bit, 11 exponent bits, 52 fraction bits, with the implied 1 leading the fraction part. For our binary relation  $x \bullet y$  we use the rule:

If  $x \geq y$  then  $x - y$ , else  $x - y + 1$ .

If  $x$  and  $y$  are floating point representations of rationals  $a/2^{53}$  and  $b/2^{53}$ , then  $x \bullet y$  will produce the (exact) floating point version of  $c/2^{53}$ , with  $c = x - y \bmod 2^{53}$ .

A single precision version of this is in the widely used 'Universal' generator [9], while a double precision version is the RNG in Matlab and described in the new DVD version of [6]. Both combine an  $F(99, 33, -)$  sequence with a simple Weyl sequence  $y_n = y_{n-1} + d$ , with  $d$  a constant and the  $y$ 's double representations of rationals of the form  $j/2^k$ , with  $k = 23$  or 53. Then the double precision operation

if  $x < y$  then  $x - y$  else  $x - y + p/2^k$  produces rationals with denominators  $2^k$  and numerators the difference modulo the largest prime  $p < 2^k$ .

#### Multiply-With-Carry (MWC) RNGs

An early description of MWC RNGs is in the file `mwc1.ps` of [6]. For another, suppose we extend the example of the second paragraph: You randomly choose a number from 1 to 58 as a pair  ${}^c x$ —that is, 23 is represented as  ${}^2 3$ , 49 as  ${}^4 9$ , etc. Your seed set  $\mathcal{Z}$  is the 58

pairs  ${}^c x$ ,  $0 \leq c < 6$ ,  $0 \leq x < 10$ , excluding  ${}^0 0$  and  ${}^5 9$ . I convert your choice  ${}^c x$  into a new pair by means of the function  $f({}^c x) = {}^{c'} x'$ , with  $c' = \lfloor (6x + c)/10 \rfloor$  and  $x' = 6x + c \bmod 10$ . Thus  $f({}^2 5) = {}^3 2$ ,  $f({}^5 1) = {}^1 1$ , etc. For each uniform choice of  $z \in \mathcal{Z}$ , my result  $f(z)$  will be uniform in  $\mathcal{Z}$ , and the sequence  $f(z), f^2(z), \dots$  will be a sequence of uniform choices from  $\mathcal{Z}$ . If you randomly choose, say,  $z = {}^3 5$ , the result of  $f(z), f(z^2) \dots$  is a period-58 sequence that will contain every element of  $\mathcal{Z}$ :

${}^3 3, {}^2 1, {}^0 8, {}^4 8, {}^5 2, \dots, {}^1 9, {}^5 5, {}^3 5, {}^3 3, {}^2 1, \dots$ , each of them the realization of a true random variable in the mathematics sense, uniformly distributed over the finite set  $\mathcal{Z}$ .

If I use the  $x$ -component of each pair in the first cycle, I get a small sample of 58 random digits:

3, 1, 8, 8, 2, 7, 3, ..., 2, 2, 3, 9, 5, 5.

and if I take, in reverse order, the  $x$ 's from the full cycle, then attach a decimal point, I get the decimal expansion  $\frac{33}{59} = .5593220338983050847457627118644067796610169491525423728813\ 5593 \dots$ .

Now take an eminently practical example, (used as one of the components in the KISS RNG below): Let  $a = 698769069$ ,  $b = 2^{32}$ . You randomly choose one of the  $ab - 2$  seeds from the set  $\mathcal{Z}$  of pairs  $[c, x]$ ,  $0 \leq c < a$ ,  $0 \leq x < b$ , excluding  $[0, 0]$  and  $[a-1, b-1]$ . For each choice of seed  $z$ , form the sequence  $f(z), f^2(z), f^3(z), \dots$ , where

$$f([a, c]) = [\lfloor (ax + c)/b \rfloor, (ax + c) \bmod b].$$

The resulting sequence will have period  $ab - 2$ , about  $2^{60.4}$  or  $10^{18.2}$ . The  $x$  components of each element of that sequence of pairs will pass tests of randomness at-least-as-well-as, and usually better-than, most commonly used RNGs that produce 32-bit integers, and with a period far greater than the  $\approx 2^{32}$  of most RNGs. But you must pay a little more for that longer period: two random seeds, the  $c$  in  $0 \leq c < a$  and the  $x$  in  $0 \leq x < b$ . (The forbidden seeds in the examples:  $[0, 0]$  and  $[a - 1, b - 1]$ , have the property that  $f(z) = z$  and thus produce sequences with

period 1. This nuisance restriction is overcome in the next section on CMWC.)

Another feature of this example, and true in general: the generated  $x$ 's will form, in reverse order, the base  $b = 2^{32}$  'digits' of the expansion of  $j/(ab-1)$  for some integer  $0 < j < ab-1$ , while the forbidden seeds  $[0, 0]$  and  $[a-1, b-1]$  will provide base- $b$  expansions of  $0/(ab-1)$  and  $(ab-1)/(ab-1)$ .

And still another feature of the MWC sequence generated on pairs  $[c, x]$  by means of  $f([a, c]) = [ \lfloor (ax+c) \rfloor, (ax+c) \bmod b ]$  is that **the resulting  $x$ 's are just the elements of the congruential sequence  $y_n = ay_{n-1} \bmod (ab-1)$ , reduced mod  $b$** . For example, with seed  $z = [123, 456789]$  in that last example, the sequence of  $x$ 's becomes

939722732,3858638025,3534982343,  
2658951225,1839178858,1673917006...

while with seed  $y_0 = 123b + 45678$ , the congruential sequence  $y_n = ay_{n-1} \bmod (ab-1)$  produces 319190024259564, 656649178557850825, 2696296900490136775, 2470136321377329209... and that sequence, taken mod  $2^{32}$ , yields the  $x$ 's of the MWC sequence.

The above MWC sequences may be described by  $x_n = ax_{n-1} + \text{carry} \bmod b$ , with the 'carry'  $c$  being the number of  $b$ 's dropped in the modular reduction that produced the new  $x$ :  $c = \lfloor (ax_{n-1} + c)/b \rfloor$ . These are lag-1 MWCs. For lag- $r$  MWCs, as with any RNG, we need a collection  $\mathcal{Z}$  of seeds and an invertible function  $f$ . In this case,  $\mathcal{Z}$  is the set of  $(r+1)$ -tuples

$$\mathcal{Z} = \{ [c; x_0, x_1, \dots, x_{r-1}] \},$$

with  $0 \leq c < a$ ,  $0 \leq x < b$ , except for  $[0; 0, \dots, 0]$  and  $[a-1; b-1, \dots, b-1]$ .

Then the function  $f$  is

$$f([c; x_0, x_1, \dots, x_{r-1}]) = [ (ax_0 + c)/b; \\ x_1, x_2; \dots, x_{r-1}, ax_0 + c \bmod b ].$$

For example, with  $a = 5$ ,  $b = 10$  and  $r = 6$ , the lag-6 MWC generator  $x_n = 5x_{n-6} + c \bmod b$ , starting with seed  $z = [4; 2, 3, 5, 3, 9, 4]$ , will produce this sequence of  $z$ 's:

$[1; 3, 5, 3, 9, 4]$ ,  $[1; 5, 3, 9, 4, 4]$ ,  $[2; 3, 9, 4, 4, 6]$ ,

$[1; 9, 4, 4, 6, 6]$ ,  $[4; 4, 4, 6, 6, 7]$ , ... with output the sequence of  $x$ 's: 4,4,6,6,7,...

The period of the sequence is the order of 10 for the prime  $p = ab^6 - 1 = 5999999$ , which is  $(p-1)/2 = 2,499,999$ .

Here is an example of a C program to compute the sequence through a little more than a full period, and to provide basis for comments on programming the general lag- $r$  MWC RNG:

```
int main(void){
  unsigned long i,t,x0=2,x1=3,
  x2=5,x3=3,x4=9,x5=4,c=4;
  for(i=1;i<2500006;i++){
    {t=5*x0+c;c=t/10;x0=x1;x1=x2;
    x2=x3;x3=x4;x4=x5;x5=t%10;
    if(i<7 || i>2499993)printf{
      "%7d,%d;%d,%d,%d,%d,%d\n",
      i,c,x0,x1,x2,x3,x4,x5}; } }
```

The output of that C program will give the first six, then the last six  $z$ 's in the cycle of length 2,499,999, as well as confirming that the first six of the second cycle match those of the first cycle.

As with the lag-1 MWCs, the more general lag- $r$  MWC:

$$x_n = ax_{n-r} + \text{carry} \bmod b$$

will produce a sequence of  $x$ 's that are, in reverse order, the digits in the base- $b$  expansion of  $j/(ab^r - 1)$ , with  $0 < j < ab^r - 1$ . For example, from the above C program, the sequence of  $x$ 's in reverse order are 935328987...8322467664, and, sure enough,

$$\frac{4676644}{5999999} = .9353289870657974131594 \dots \\ 16916123383224676644 \ 935328 \dots,$$

the trailing digits of which can be determined by expanding  $(10^{499979} \times 4676644 \bmod 5999999)$  to 30 places. To find which  $j$  provides the expansion of  $j/p$ , just put a decimal point in front of the reversed  $x$ 's that end a cycle—for the above case, .9353289... and find that 4676644 is the integer closest to .935289 $p$ .

A possibly simpler way is to use the inverse function of  $f$ , say

$$f^{-1}(z) = g([c; x_0, x_1, \dots, x_{r-1}]) =$$

$$[bc + x_{r-1} \bmod a; \\ \lfloor (bc + x_{r-1})/a \rfloor, x_0, x_1, \dots, x_{r-2}].$$

Then, with  $z = [4; 2, 3, 5, 3, 9, 4]$ , the rightmost  $x$ 's in the sequence  $z, g(z), g^2(z), \dots$  will generate, in order, the digits of the base- $b$  expansion of  $4676644/p$ , just as the rightmost  $x$ 's in  $f(z), f^2(z), f^3(z), \dots$  will generate those digits in reverse order.

For computer implementation, we often choose  $b = 2^{32}$ , and then it is clear that the  $f$  sequence is more practical than the  $g$  sequence, as the integer operations  $t = ax + c; c = \lfloor t/b \rfloor; x = t \bmod b$  are built into most CPUs. One merely forms  $t = ax + c$  in 64 bits, then  $c$  is the top-32 and  $x$  the bottom-32 bits of  $t$ .

In the C program above, for lag-6, it is just barely feasible to keep the last six  $x$ 's by means of promotions:  $x_0 = x_1; x_1 = x_2; \dots$  and so on. Keeping a (circular) table of the  $r$  most recent  $x$ 's in an array  $Q[ ]$ , and a pointer that rotates through the elements provides simple and very fast MWC RNG's. An example is this C procedure that provides 32-bit random integers with period greater than  $2^{33245} \approx 10^{10007}$ :

```
static unsigned long Q[1038],c=123;
unsigned long MWC1038(void){
static unsigned long i=1037;
unsigned long long t,a=611373678LL;
t=a*Q[i]+c; c=(t>>32);
if(--i) return(Q[i]=t);
i=1037; return(Q[0]=t); }
```

You need to assign random 32-bit seeds to the static array  $Q[1038]$ .

Note: Unlike simple MWC RNGs

$$x_n = ax_{n-1} + \text{carry} \bmod m,$$

which can be expressed as the reduction, mod  $b$ , of the congruential sequence

$$y_n = ay_{n-1} \bmod ab-1,$$

there seems to be no such simple relation between lag- $r$  MWCs

$$x_n = ax_{n-r} + \text{carry} \bmod b$$

and the congruential sequence

$$y_n = ay_{n-r} \bmod ab^r - 1.$$

Complimentary-Multiply-With-Carry (CMWC) RNGs

A few nagging problems come with MWC RNGs  $x_n = ax_{n-r} + c \bmod b$  when  $b = 2^{32}$  is chosen for computer implementation: the period is the order of  $b$  for the modulus  $m = ab^r - 1$ , but even when  $p = ab^r - 1$  is a prime, the period cannot be  $p - 1$  because  $b = 2^{32}$  is a square. Thus, as in the above example, MWC1038(), even though  $p = ab^{1038} - 1$  is prime, (as is  $(p - 1)/2$ ), the generated 32-bit integers will have period  $(p - 1)/2$ , and they will, in reverse order, form either the base- $2^{32}$  digits of the expansion of  $j/p$  for some  $j$  in the subgroup  $\{b, b^2, b^3, \dots, b^{(p-1)/2} \bmod p\}$ , or else for some  $j$  in the coset  $\{hb, hb^2, hb^3, \dots, hb^{(p-1)/2} \bmod p\}$ , where  $h$  is some group element not in the cyclic subgroup generated by  $b$ .

Thus, strictly speaking, we do not have a seed set  $\mathcal{Z}$  until we choose the seed  $[c; x_0, x_1, \dots, x_{1037}]$ . Half of the choices will lead to the digits in the expansion of  $j/p$  for  $j$  in the group, half for  $j$  in the coset. (An interesting sidelight: if  $[c, x_0, x_1, \dots, x_{r-1}]$  is a seed whose subsequent  $x$ 's form the reversed digits in  $j/p$ , with  $j$  in the subgroup, then the seed  $[a-1-c; b-1-x_0, b-1-x_1, \dots, b-1-x_{r-1}]$  will form the reversed digits in the expansion of  $k/p$ , with  $k$  in the coset—indeed,  $k = p - j$ .)

Another nuisance feature of MWC RNGs is that the two seeds  $[a - 1; b - 1, \dots, b - 1]$  and  $[0; 0, \dots, 0]$  must be avoided, as they have the property that  $f(z) = z$ , so that their periods are 1 (with reversed digits corresponding to the base- $b$  expansions of  $0/p$  and  $p/p$ , as, in base 10,  $23/23 = .999999\dots$ ).

Complimentary-multiply-with-carry RNGs (CMWC) permit us to avoid both of those difficulties. By making  $b = 2^{32} - 1$ , we can still exploit the way that integer arithmetic is carried out in modern CPUs (with a little fiddling for reductions mod  $2^{32} - 1$  rather than mod  $2^{32}$ ). For this, we seek primes of the form  $p = ab^r + 1$  with  $b = 2^{32} - 1$  a primitive root of  $p$ . Then the CMWC

recursion is  $x_n = (b - 1) - [ax_{n-r} + c \bmod b]$ , where rather than the  $x$  of MWC, we return the  $(b-1)$ -complement of that  $x$ . The period will be  $p - 1 = ab^r$ .

Formally, if  $p = ab^r + 1$  is a prime for which  $b$  is a primitive root, then the seed set

$$\mathcal{Z} = \{[c; x_0, x_1, \dots, x_{r-1}]\}, \\ 0 \leq c < a, 0 \leq x < b,$$

has  $ab^r$  elements, and for any  $z \in \mathcal{Z}$ , (including all 0's and  $c = a - 1$  with all  $x_i = b - 1$ ), the sequence  $f(z), f^2(z), \dots$  will have period  $ab^r$ . Here

$$f(z) = f([c; x_0, x_1, \dots, x_{r-1}]) = \\ [(ax_0 + c)/b]; x_1, x_2; \dots, x_{r-1}, \\ (b-1) - (ax_0 + c \bmod b).$$

Furthermore, the sequence of trailing  $x$ 's in the sequence  $f(z), f^2(z), \dots$  will, in reverse order, form the base- $b$  digits in the expansion of  $j/p$  for some  $0 < j < p$ .

Example:  $b = 10$  is a primitive root of the prime  $p = 7b^2 + 1 = 701$ . The seed set is the 700 elements  $\mathcal{Z} = \{[c; x, y]\}$  with  $0 \leq c < 7$ ,  $0 \leq x < 10$ ,  $0 \leq y < 10$ . The iterating function is  $f([c; x, y]) = [(7x + c)/10]; y, 9 - (7x + c \bmod 10)$ . Starting with seed  $z = [2; 3, 4]$ , the sequence  $f(z), f^2(z), \dots$  produces the 700 elements of  $\mathcal{Z}$ , then repeats:

[2; 4, 6], [3; 6, 9], [4; 9, 4], [6; 4, 2], [3; 2, 5], ..., [4; 6, 6], [4; 6, 3], [4; 3, 3], [2; 3, 4], [2; 4, 6], [3; 6, 9], ... The trailing  $x$ 's, taken in reverse order from the end of a cycle, are 4336..., and  $.4336p = 303.9536$ , so we expect  $j = 304$ , and so it is:

$$\frac{304}{701} = .433666191155492154 \dots \\ 932952924393723252496 4336661912 \dots$$

provides the output from the CMWC  $x_n = 9 - [7x_{n-2} + c \bmod 10]$  starting with  $c = 2, x_0 = 3, x_1 = 6$  and put in reverse order.

Those digits could be produced in direct order with the sequence  $z, g(z), g^2(z), \dots$  where  $g$  is the inverse of  $f$ :

$$g([c; x, y]) = [10c + y \bmod 7; \lfloor (10c + y)/7 \rfloor, x].$$

Then the sequence  $z, g(z), g^2(z), \dots$  becomes

[2; 3, 4], [4; 3, 3], [4; 6, 3], [4; 6, 6], [1; 6, 6], ..., and the trailing components, 43366... form the digits in the expansion of  $304/701$ .

The digits in the base- $b$  expansion of  $j/p$  for a large prime  $p$  are likely to serve quite well as random integers from 0 to  $b - 1$ , whether in direct or reverse order. But for computer implementation, with  $b = 2^{32}$  or  $b = 2^{32} - 1$ , the arithmetic in  $g(z)$  is much less well suited to computer operations than that in  $f(z)$ .

With period exceeding  $2^{131086} \approx 10^{39461}$ , here is a C procedure that produces the CMWC sequence

$x_n = (b-1) - [ax_{n-r} + \text{carry} \bmod b]$ ,  
with  $b = 2^{32} - 1, a = 18782$ :

```
static unsigned long Q[4096],c=123;
unsigned long CMWC(void){
  unsigned long long t, a=18782LL;
  static unsigned long i=4095;
  unsigned long x,m=0xffffffff;
  i=(i+1)&4095; t=a*Q[i]+c;
  c=(t>>32); x=t+c; if(x<c){x++;c++;}
  return(Q[i]=m-x); }
```

The static array  $Q[ ]$  must be filled with random 32-bit integers for different runs. Rather than keeping the most recent 4096  $x$ 's in an array  $Q$ , smaller sizes 2048, 1024, 512, ... can be used. (Choice of array size  $2^k$  simplifies incrementing the array index). Different choices of  $r, a$  require slight changes to the above C code for CMWC: Make  $Q[ ]$  have size  $r$ , change multiplier  $a$  and change the two 4095's to (decimal)  $r-1$ .

Here are a few good choices for  $r$  and  $a$ :

$r$	$a$	$r$	$a$
2048	1030770	64	987651206
2048	1047570	64	987657110
1024	5555698	32	987655670
1024	987769338	32	987655878
512	123462658	16	987651178
512	123484214	16	987651182
256	987662290	8	987651386
256	987665442	8	987651670
128	987688302	4	987654366
128	987689614	4	987654978

The results will be CMWC RNGs that seem to pass tests of randomness as well as any I know of, are simple and extremely fast, and have periods  $ab^r$ , with  $b = 2^{32} - 1$ , roughly  $2^{32r+30}$ .

Choice of  $r$  and  $a$  have little effect on speed—about 18 nanoseconds on a 1.2MHz PC, or better than fifty million random numbers per second. Those wanting even more pairs  $r, a$  will need to find primes of the form  $p = ab^r + 1$  for which  $b = 2^{32} - 1$  is a primitive root.

#### Randomness and Choice of Seeds

Just as in geometry, where existence of and conclusions about complicated objects are premised on the existence of fundamentals such as points, lines, planes, etc., most distribution theory in probability is premised on the existence of more fundamental random variables. In particular, the distribution of elements in the RNG sequence  $f(z), f^2(z), f^3(z), \dots$  is premised on the existence of a random selection  $z$  from the seed set  $\mathcal{Z}$ , and it has as firm a basis in mathematics as do results in geometry, number theory and the like.

In my view, use of the name pseudo random number generators (PRNGs) is not appropriate. The qualifier *pseudo* can have several implications, most commonly: unreal, false, pretended, spurious, sham.

Use of *pseudo*, in the sense of *unreal* implies that there is **real** randomness, when the only kind we are sure of is in the sense of mathematics—exactly the sense that applies in our use of  $f(z), f^2(z), \dots$

Use of *pseudo* in the sense of *false* is not appropriate either. If  $x$  and  $y$  are independent standard normal variates then we say that  $x^2 + y^2$  is a chi-square variate; we do not call it a pseudo chi-square variate. Its properties may be deduced from that of its defining variates, just as are those of elements of the sequence  $f(z), f^2(z), f^3(z), \dots$ —both considerations a real consequence of assumptions in the mathematical model.

Use of *pseudo* in the sense of *pretended* might be considered the least objectionable, for it might seem that we are pretending that our sequence  $f(z), f^2(z), f^3(z), \dots$  produces truly ran-

dom numbers. But to many, (including me), true randomness exists only in the sense of mathematics—whether or not we understand it, the Universe is unfolding as it must. So the *pretending* is that there is such a thing as true randomness.

And finally, use of *pseudo* in the sense of *spurious* or *sham* is worst of all. Few would argue over the usefulness of RNGs for the past fifty years, and considerable effort has gone into studying their mathematics. Unfortunately, joint distribution theory for elements in the sequence  $f(z), f^2(z), f^3(z), \dots$  is not readily determined other than through simulation. Thus, except in cases such as the lattice structure of congruential RNGs, use of number or matrix theory to establish the periods, presence or absence of various  $m$ -tuples, relation to base- $b$  decimal expansions, most of what we know about RNGs has been determined from extensive use, and those are far from *spurious* or *sham* endeavors.

If you must use ‘pseudo’, it would be more appropriate to say that our random-in-the-sense-of-mathematics numbers are *pseudo* independent, for we **are** pretending that they are independent—our random variables are identically distributed (id) uniform, but not independent identically distributed (iid) uniform.

#### Choice of Seeds

It is often convenient to choose just one or two random integers as seeds, even when several hundred may be needed to specify an element  $z$  of the seed set  $\mathcal{Z}$ . The other parts of the seed  $z$  may already have been assigned by default or previous use. Use of a RNG sequence can be likened to randomly choosing a starting position on a huge wheel of numbers (the RNG’s cycle), then using them sequentially from that selected starting point. Even though that wheel might contain over  $10^{10000}$  numbers, a single 32-bit integer can provide over four billion potential starting points, and the features you may want to study are likely to be consistent with the underlying probability theory

at all but a few of those starting points.

But there are some applications where the seed selection procedure must be able to provide every element in the seed set  $\mathcal{Z}$ . Such requirements arise in Law. Many states have laws that permit use of computers (and hence a RNG) to select a jury venire—a panel of citizens selected to serve on juries. For example, Flor.Stat.ch.40.225(2000) authorizes use of computers to select jury venires, if such drawing *is by lot and at random by a method approved by The Florida Supreme Court*.

I was retained by that court to see if methods used in the various court districts were meeting this statutory requirement. It turned out that they were not,[8]. In most counties, a seed of perhaps ten digits is chosen by the staff or by the computer clock for the RNG of some proprietary administrative system. Suppose the task were to choose 80 potential jurors from a list of 200 eligibles. There are  $\binom{200}{80} > 10^{57}$  ways to choose such a panel, and the  $\approx 10^{10}$  ways of selecting a single seed, (or worse, the 65,536 possible 16-bit integers from a computer clock) can not come close to providing the necessary number of choices.

The requirement that selection be *by lot and at random* means that a litigant should be entitled to any one of the possible venires; in this case, a RNG requiring at least eight 32-bit random integers to determine the element  $z$  of the seed set  $\mathcal{Z}$  would be required. The Florida Supreme Court has implemented recommended procedures for choosing the necessary number of random seeds, from publicly available data—for example, from a coming week's stock market—which is unpredictable yet verifiable after the fact, see [7,8].

Another place where the need for adequate seeds arises is in the gaming industry. For example, The Michigan Game Control Board received an application to license a computer poker game that would permit the player to play as many as fifty games of poker at a time. The application was initially rejected because the Board ruled that a player was entitled to the

chance that his fifty hands would all be straight flushes, and the RNG used by the machine had far too meager a seed set  $\mathcal{Z}$ . The company was presumably able to get a license after I advised them on overcoming the problem by extending the seed set  $\mathcal{Z}$  for the version of my KISS RNG that they were using (without permission, as published mathematics or algorithms are not protected by patent or copyright law).

#### Combination RNGs

A RNG produces a random sequence of uniform selections from the seed set  $\mathcal{Z}$ . They are used to provide a sequence of integers  $x_1, x_2, \dots$ , each  $x$  coming from one of the random elements in  $\mathcal{Z}$ , either, for simple RNGs, as the element  $x_n = f^n(z)$  itself, or as one of the  $x$ -components of an  $r$ -tuple of  $x$ 's that make up each  $z \in \mathcal{Z}$ . When, as is the most common case, the RNG merely produces a sequence of integers with period  $2^{32}$ , then it can produce only  $1/2^{32}$  of the possible pairs  $(x, y)$ , only  $1/2^{64}$  of the possible triples  $(x, y, z)$ , etc. If your simulation concerns certain properties of triples  $(x, y, z)$  that are adequately represented in the limited supply that the RNG provides, then fine. But with such a limited supply, it is likely that there will be many simulations for which such a RNG is not suitable.

There are ways to overcome this difficulty. One of them is to use lag- $r$  MWC or CMWC RNGs, from which every possible  $r$ -tuple of  $x$ 's can appear. Another method is to combine simple, short period generators. One of the first examples of this was Super Duper, which combined a congruential and a 2-shift xorshift RNG. One of the most widely used is the KISS RNG, which I named during the early days of the Clinton administration when "Keep It Simple Stupid" was a buzz phrase relating to economic policy. The KISS RNG combines three simple RNGs: congruential, xorshift and the lag-1 MWC described above. It has had wide use, because it seems to pass all tests of randomness and yet uses simple computer instructions and needs no tables.



Here is a C version:

```
unsigned long KISS(){
static unsigned long
x=123456789,y=362436000,
z=521288629,c=7654321;
unsigned long long t;
x=69069*x+12345;
y^=y<<13; y^=y>>17; y^=y<<5;
t=698769069LL*z+c; c=t>>32;
return x+y+(z=t); }
```

Seeds  $x, y, z, c$  may be changed from the default values. With a period  $> 2^{124}$ , KISS() may be more suitable than single-seed RNGs for applications. It is used in the gaming industry in North America and Australia. The speed may be only 20+ million per second compared to the 50+ million or so for MWC or CMWC, but a RNGs speed is usually not important in the overall time of a simulation or in game processing. For Fortran, which has no easy way to access the 64-bit product of two 32-bit integers, versions of KISS adjoin two lag-1 MWC's on 16 bits.

For a really powerful combination, I would recommend  $\text{CMWC}() + \text{KISS}()$ , for the rare chance that the 4096-tuples provided by CMWC might benefit from tweaking with KISS. (The "+" can be ordinary addition mod  $2^{32}$ ).

#### Testing RNGs

For some 25 years, in graduate math/cpt.sci/stat courses, I had discussed using a battery of tests of randomness for RNGs, and in one of the classes a Chinese student, who knew the word 'battery' only from pervasive TV ads for Sears car batteries, used the term Diehard in referring to the tests we were discussing—much to the amusement of the class, but the name stuck. Subsequently, under a grant from NSF, I developed

*The Marsaglia Random Number CDROM  
with*

*The Diehard Battery of Tests of Randomness*

The CDROM contained 600 megabytes of random bits produced by combining the output of good RNGs with the output of physical devices

purporting to provide random bits. Some 1000 free copies of that CDROM were distributed to researchers worldwide. The CDROM also contained Fortran and C code for what I called The Diehard Battery of Tests of Randomness. Although the free CDROMs were soon gone, presence of two of them at sites on the web made it readily available, and it seems to be in wide use.

A new version of those tests is now available at [www.csis.hku.hk/~diehard](http://www.csis.hku.hk/~diehard) and will be included in a DVD version of [6], with more extensive files of random bits. The new Diehard tests contain several new 'tough' tests that relate to use of random integers in computational number theory and cryptography.

Some of the most effective tests for randomness are based on what I call Monkey Tests. The idea is that some part of each random number can be used to specify a 'keystroke' that produces a letter from an alphabet. Few images invoke the mysteries and ultimate certainties of a random sequence as well as that of the proverbial monkey at a typewriter. A discussion of such 'monkey tests' is in the file `monkey.ps` of [6].

The most effective monkey tests are those for which counts are maintained for the number of appearances of each  $k$ -letter word in a long string of random letters, and it turns out that if 
$$Q_k = \sum (\text{observed} - \text{expected})^2 / \text{expected}$$
 is the naive Pearson statistic for  $k$ -letter words, then  $Q_k - Q_{k-1}$  is the quadratic form in the weak inverse of the covariance matrix of the  $k$ -letter word counts, and is asymptotically chisquare distributed with  $L^k - L^{k-1}$  degrees of freedom, when the alphabet has  $L$  letters.

It usually happens that the number of possible  $k$ -letter words is too great to maintain a count for each word. In that case, the number of  $k$ -letter words missing from a long string of  $N$  random letters is used. That number will be close to normally distributed with mean  $L^k e^{-N/L^k}$ . Except for smaller cases  $k = 2, 3, 4$ , the variance must be estimated by simulation.

One of the tests in the new version of

Diehard is called the ‘Gorilla test’, in the sense of a strong monkey test. Many RNGs fail it. The Gorilla test counts the number of bit strings of length 26 that are missing from a sequence of  $2^{26} + 25$  bits. Such a string is formed by specifying the bit position for each 32-bit word, then taking that bit from each of  $2^{26} + 25$  calls to the RNG. The Gorilla test reports a p-value for the number of missing 26-bit strings for each of the 32 bits.

Another new test is the gcd test. That test uses two successive 32-bit integers  $u, v$  produced by the RNG, then finds  $k$ , the number of steps needed to find the gcd of  $u, v$  by Euclid’s algorithm, and  $x$ , the resulting gcd. It tests to see if a sample of ten million such  $k$ ’s and  $x$ ’s have distributions consistent with underlying theory. All congruential RNGs—even those with prime modulus—fail the gcd test for distribution of  $k$ ’s, and many as well for distribution of  $x$ ’s.

The new tests include a stronger version of my ‘birthday spacings’ test, and others. These, with the gorilla and gcd tests, all relate to the suitability of numbers as random integers, an area of increasing importance in cryptography and computational number theory.

What might be called more conventional tests are mainly concerned with the performance of UNIs, that is, the uniform  $[0,1)$  variables that result from floating the RNG’s integers. In a way, it may be reassuring that most RNGs pass such tests, because most real-life applications of RNGs seem concerned with the UNIs that result, not with observed non-uniformity in the bits of the integers that produce the UNIs (although poor performance of leading bits often portends bad sets of UNIs).

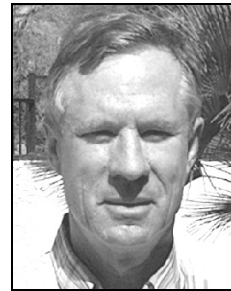
The availability of transfers through the Internet makes it easy to get the new version of the Diehard Battery of Tests at [www.csis.hku.hk/~diehard/](http://www.csis.hku.hk/~diehard/), which interested readers are invited to try for themselves.

#### References

- [1] Knuth, D. E.(1998). *The Art of Computer Programming, Volume II*, 3rd Ed., Addison Wesley, Reading, Mass.
- [2] Marsaglia, G. (1968). Random numbers fall mainly in the planes. *Proceedings of the National Academy of Sciences*, **61**, 25–28.
- [3] Marsaglia, G. (1970). Regularities in congruential random number generators. *Numerische Mathematik* **16**, 8–10.
- [4] Marsaglia, G. (1972). The structure of linear congruential sequences. In *Applications of Number Theory to Numerical Analysis*, Z. K. Zaremba, ed., Academic Press, 249–285.
- [5] Marsaglia, G. (1985). A current view of random number generators. Keynote Address, Statistics and Computer Science: XVI Symposium on the Interface, Atlanta, *Proceedings*, Elsevier.
- [6] Marsaglia G. (1995). *The Marsaglia Random Number CDROM, with The Diehard Battery of Tests of Randomness*, produced at Florida State University under a grant from The National Science Foundation. Available at sites:  
[www.stat.fsu.edu/pub/diehard](http://www.stat.fsu.edu/pub/diehard)  
[www.csis.hku.hk/~diehard/cdrom](http://www.csis.hku.hk/~diehard/cdrom).
- [7] Marsaglia, G. (2001). Problems with the use of computers for selecting jury panels. *Jurimetrics* **41** No. 4, 425–427.
- [8] Marsaglia, G. (2003). Seeds for random number generators. *Communications ACM* May 2003.
- [9] Marsaglia, G., Tsang, W. W. and Zaman, A. (1989). Toward a universal random number generator. *Statistics and Probability Letters* **8**, No. 5.

## The Importance Of Fortran In The 21<sup>st</sup> Century

Walt Brainerd  
The Fortran Company  
Tucson, AZ



---

A brief discussion on the history and purpose of Fortran for scientific and engineering computing is given. This leads to the role Fortran, in its various environments, will likely play well into the 21<sup>st</sup> century.

Key words: Fortran 95, Fortran 2000, F, high performance computing.

---

### Introduction

Let us start with a bold assertion: Fortran is still the best programming language for numerical/scientific computing. The reasons could be discussed and debated extensively, but they include:

- There is a large investment in scientific software written in Fortran, including extensive libraries.
- There is a large investment in the training and experience of scientists that do programming.
- The language is more straightforward to learn and use than most "modern" languages.
- Fortran produces efficient code.

- Fortran is very portable: source code compiles on many platforms with little need for conditional compilation and results are consistent, particularly when executed on standard floating point hardware.

The reason I make this statement is because it means that the continued development and implementation of Fortran will be important in the twenty-first century for the same reasons (listed above) that it has been important in the twentieth century.

However, the computing environment is continually changing. What is being done to ensure that Fortran will remain an outstanding tool for scientists and engineers? That is what will be discussed in the remainder of this article.

---

Walt Brainerd holds one of the first PhDs in computer science awarded in the USA. He was a leader in the development of Fortran 90 and the co-author of several books, including *The Fortran 95 Handbook*. He is one of the originators of the F programming language and maintains the F compiler. Email him at [walt@fortran.com](mailto:walt@fortran.com).

### Language Development and Standardization

In the last quarter century, most of the innovation in the Fortran programming language has come through the group responsible for its standardization. This work is done by the American standards committee J3 under the direction of the International Standards Organization committee WG5. Their web site is: <http://www.nag.co.uk/sc22wg5>.

This work continues, and the next standard, being called Fortran 2003 informally is in the process of approval and publication. (See <ftp://ftp.j3-fortran.org/j3/doc/standing/007>). The major new features include help with interoperation with C programs and enhanced object-oriented facilities, including inheritance and polymorphism, which enhance the excellent data abstraction features of Fortran 95.

#### Modern Development Environments

One of the significant changes for Fortran programmers in recent years has been the availability of modern graphical user interfaces for editing, compiling, executing, and analyzing Fortran programs. Many Fortran implementations include such an environment, in addition to traditional command-line execution and tools.

Unfortunately, these environments are different for almost every compiler. There is hope that there will be an open tool for at least Linux/Unix environments and there does seem to be some convergence by the vendors of Windows compilers to the Microsoft Visual Studio .NET environment. The url: <http://msdn.microsoft.com/vstudio/productinfo/overview/default.asp>.

#### High Performance Computing

In recent years, several tools have been made available to Fortran (and other programming language) programmers to assist them to take advantage of special high performance computer architectures, such as vector processors, distributed memory multiprocessors, and shared memory multiprocessors. These tools include High Performance Fortran (HPF), OpenMP, and MPI. It is reasonable to expect that these tools will continue to be developed as the new versions of Fortran are implemented.

#### Free and Open Source Compilers

Unfortunately, due to the smaller number of compilers a vendor may expect to sell, spreading the development costs means that Fortran compilers are moderately expensive. The only open source Fortran compilers are g77 (but unless you need to compile only legacy codes, who wants to use the quarter-century-old version of Fortran?), and Open64, (<http://open64.sourceforge.net>), a compiler that works only on the Itanium architecture under

Linux. Intel's Linux compiler (<http://www.intel.com/software/products/compiler/s/index.htm>) is available for free, but only for non-commercial use.

There is a g95 project under way to develop a GNU Fortran 95 compiler (<http://g95.sourceforge.net>). It will probably not be available until after Fortran 2003 compilers come out, so it will again be one step behind. Most vendors of Fortran compilers offer academic discounts.

#### F

F is a subset of Fortran 95 consisting of its modern features and excluding the error-prone older features (<http://www.fortran.com/F>). Numerical Algorithms Group (the originators of the first Fortran 90 compiler) has made their compiler technology available for this software, which is maintained by the Fortran Company. It is free, but it is not open source.

It is a compiler that can be used to develop production software, because anything that is compiled by the F compiler will also be compiled by any Fortran 95 compiler. It also provides free software for use by academic institutions that want to expose their students to Fortran programming.

#### Conclusion

The Fortran language, Fortran compilers, Fortran environments, and Fortran tools continue to advance, along with other computing environments. It looks like Fortran will still be the premier programming language to be used in the twenty-first century when serious numerical and scientific computing needs to be done. Additional information about Fortran can be found at <http://www.fortran.com>.

## The Way Ahead In Qualitative Computing



Tom Richards

QSR International  
Melbourne, Australia



Lyn Richards

---

Specialized computer programs for Qualitative Research in social sciences have greatly changed ways of doing QR, the reliability and comprehensiveness of results, the ability to inspect and challenge a researcher's working, and the relationship with quantitative methods in social research. This article explores these claims in the context of N6 (NUD\*IST) and NVivo, the two programs designed by the authors; and considers possible future developments in the field.

Key words: NUD\*IST, NVivo, qualitative research, qualitative computing

---

### Introduction

Qualitative Research (QR) has always centered on the analysis of conversational interviews, field notes and recorded conversations. Its raw data are people talking, and the people can be the researchers with their field notes and conversational turns, as much as the interviewees or subjects. Interviews may be one-on-one, or in groups, the records may be live transcripts or historical recollections. Questionnaires may be used, but mainly as topic prompts expecting prose responses not ticked boxes.

---

Tom Richards is Chief Scientist at QSR International, and designer of NUD\*IST and NVivo. He has a D. Phil. in Logic from Oxford University, and many publications on logic, computer science and methodology. Lyn Richards is founder and Director of Research Services at QSR. She has published books and papers on family sociology, qualitative research and QSR's software. This article is based on a presentation given to the American Educational Research Association, SIG Professors of Educational Research, Chicago, April 21, 2003.

The methods and techniques of doing QR are often corralled into a number of schools, Ethnography, Grounded Theory, Phenomenology, and others. From our point of view these are seen as laying stress on different parts of the research process, and the aim of a developer of software for QR is to ensure there are enough tools to keep them all happy. Their actual practices, viewed as tool-users, have much in common: they just prefer to make different products or build them in different ways because they have different research goals.

QR was done manually until about twenty years ago with the rise of the word processor. Preferred techniques involved typing up the interviews or other raw data, and coding or flagging passages about topics of interest with the goal of gathering together all the passages on a given topic. Coding was done by making marginal notes, or photocopying into file folders, or making notes on system cards. This usually required a messy desk or a large living-room floor as a sorting ground. Needless to say these practices were rickety: clerical and management processes were onerous and scarcely fail-safe. Whilst you might do your initial coding thoroughly, it becomes hard to be sure, for example, that you'd

compared thoroughly how a particular viewpoint is presented by people with different demographics or sets of opinions – just because sorting the data into multiple such groups, often cross-cutting, then trying to do side-by-side comparisons, is so hard. Even trying to find vaguely remembered passages about this or that was a matter of luck. These and many other such difficulties we could call the access problem.

Moreover there is the revision problem. Revising your coding in the light of experience was virtually impossible because of the rigidity of handling coding imposed by paper records and coding management. Using manual methods also meant it was impossible to link the data systematically with quantitative research. Demographic data about respondents, or ticked response boxes, could be analyzed in SPSS; but studying interesting qualitative issues arising in conversational interviews with the respondents, in a way that sorted and compared those discussions using the demographic data, was very difficult. Only simple relations could be effectively investigated. Call this the qual-quant problem.

All of this meant that effective QR was best done with small data sets (by no means a bad thing,  $n$  is not often an important parameter in QR), or conclusions were impressionistic and bolstered by “juicy quotes” rather than dispassionate analyses. Checkability and the reaching of agreement suffered too: disputes over the conclusions reached by a researcher were hard to resolve since there was no way of reviewing the analysis steps. It was more a matter of starting again with the raw data.

### The Rise of Qualitative Computing

If the above characterized QR without computers, how did computing help? Early experiments with electronic files in a word processor improved on the manual situation. Codes could be inserted [like this] in the text, and word search would find all the instances of a code, enabling inspection of their passages. This greatly ameliorated the access problem, but clerical organization of codes, and their comparison, remained elusive. These problems led to the rise of the early dedicated QR programs, which basically provided tools for coding text documents, storing the coding references (usually to lines), and using them to find and display all passages referred to by

a given code such as ‘playground bullying’. From the first dedicated QR programs, simple Boolean searches were supported, thus you could find all passages coded by both ‘playground bullying’ and ‘fear of going to school’. These features were much prized, because researchers could explore, with confidence of completeness, hunches about relationships between different situations or concerns or attitudes; and that is the way qualitative theories are built and tested.

This process came to be known as code-and-retrieve, and because it was computationally simple to program, became the hallmark of computer-based QR. As we shall see however, this was a somewhat limiting approach to QR. For one thing, researchers couldn’t edit the text of their data any more, because to do so would invalidate the coding references made to the text passages; yet flexibility of amending, adding to, fleshing out, the text was a desirable tool for qualitative researchers that word-processing had provided.

Nevertheless code-and-retrieve has formed the core of all QR programs to date. Many of the current software offerings however provide much more than that. This, and the future, is what the rest of this paper will look at, in the context of QSR’s two QR programs.

### Methodology

QSR has two products for qualitative researchers, NVivo and NUD\*IST (Non-numerical Unstructured Data Indexing Searching and Theorizing, a name given when it was being programmed by one of us (TJR) for sole use by the other of us). Its latest version is known as N6 in an attempt to suppress a name, which, however memorable, definitely should not be searched for using a Web search engine! NUD\*IST was first used by LR in the early 1980s, and went commercial in 1986 with the sale of one license (on a university mainframe and with scroll-mode display!). NVivo was launched in 2000. These are very different products, and aimed to support different work practices, as will be described below. Right now however, our aim is to set out how these products both go beyond the code-and-retrieve paradigm just described.

### Edit-While-You-Code

We pointed out above that a restriction imposed by code-and-retrieve was that you

couldn't edit a document – it was frozen. The reason: editing would, by adding or removing text, invalidate the references made by coding to passages in the document. Add a hundred characters at a given point and every reference by every code to passages later in the document will now pick up text a hundred characters before what it used to. Back in the days of paper the problem was different and not so bad. If you coded by photocopying passages to folders of codes, then if you altered the original the coded copy in the folder was unaltered, but might no longer be faithful to the altered original.

Researchers do want to make corrections to interview transcripts, to do partial transcription and flesh it out later as the direction of research indicates, to edit out privacy-infringing material, to add clarifications and greater detail to field notes. Researchers also want to code while they are typing up the transcription, because that's often when they have their best thoughts about what the text is saying and implying and hinting and suggesting. The restriction that all your data documents must be complete and final before you dare to add one code, is a strait-jacketing QR cannot accept.

Aside from the ability to add text at the end of document, which doesn't upset any existing coding, N4 and onwards has provided the ability to edit individual lines or paragraphs – the text units that are the smallest chunks of text that can be coded. NVivo however codes all the way down to individual characters, and moreover supports rich text documents, not just plain text as in N6. Despite this, NVivo supports full editability. Its Document Browsers, where you look at the text of a document, have full editing controls plus controls over the “richness” of the text – font, letter style and color and size, etc. And using the text editor does not in any way invalidate existing coding: NVivo's way of recording coding keeps up with editing changes. So for the first time ever, researchers can feel completely free to modify their documents, and to code them while writing them up.

Nodes – Going Beyond Code-and-Retrieve

The world of QR, including QR computing, talks of codes as the labels attached to and describing the contents of, passages of text. The process of coding is the labeling of the text,

and retrieval of a code involves presenting, somehow or other, the passages referenced by the code.

But both of QSR's products store coding at nodes. These are containers for topics, ideas, places, people, and attitudes, indeed anything that may be relevant to the QR project at hand. There may, for example, be a node 'Schools' which has under it sub-nodes for the schools in the project 'Valley High' and 'Hilltop Primary' for example. 'Valley High' might contain just a memo written by the researcher describing the school and its problems, and 'Schools' contain nothing – it's there just as a generic locator for the nodes for individual schools (this demonstrates why we chose the word 'node' for these entities, and why the two programs can organize nodes in a tree-structured hierarchy like a library catalog or a taxonomy).

Many nodes will however contain coding. If an interviewee talked about Hilltop Primary, it's appropriate to code that passage at the 'Hilltop Primary' node. And of course some nodes are intended primarily for coding, such as 'angry' (marking where interviewees displayed anger) or 'reports of bullying'.

Nodes can also be used to mark cases. If we have ten interviewees, who got interviewed individually a couple of times then in groups, it is useful to collect everything each individual said in one place. This gives rise to case nodes 'Mary', 'Joe', etc., instances of the case type 'Interviewee'.

It's a small step beyond that to use trees of nodes to represent demographic data – called base data trees in N6. (NVivo represents demographic data in tables of so-called attributes). Thus we can have a 'Religion' node, with sub-nodes 'Christian', 'Hindu', 'Jewish', etc. Then if Joe is Jewish, we copy all the coding at the 'Joe' case node to the 'Jewish' node. And the same goes for any other Jewish interviewee case. Why do this? Because now, using the ability to make Boolean combinations of coding at nodes, we can immediately find everything said by Jewish interviewees. And if we have coding at 'Hilltop Primary' and 'reports of bullying' we can find all reports by Jewish interviewees about bullying at Hilltop Primary.

Both N6 and NVivo support importing and exporting such demographic data as tables. For example an SPSS table, whose rows are the

Interviewee cases ‘Joe’ etc, and whose columns are variables such as ‘Religion’ etc., can be imported into N6 to create and code up an entire base data tree. Conversely such a tree (which may be created inside N6 rather than imported) can be exported as a table to any table-handling program. NVivo does this more directly with its attribute tables; but in either program a researcher might create a base data type of tree that records research results, perhaps various categories of social, political or educational opinion the researcher has labeled the interviewees with as a result of careful analysis of what they’ve said. An example would be, for parents or teachers discussing ideal curricula: ‘Curriculum priority/vocational’, ‘Curriculum priority/all-rounder’ and ‘Curriculum priority/none’. The exported table would record which case (interviewee) belongs to each category.

Nodes with coding represent views onto the textual data of a project that are orthogonal to that provided by documents. Any QR program will let you view the contents of a document, e.g. the first interview with Joe. In NVivo and N6, a Document Browser, like an edit window in Microsoft Word, shows you all the text in that document. A node by contrast refers to all passages that have been coded at it. How do you see such passages? In both N6 and NVivo, and unique to these programs, you can view everything coded at a node (in a browser window) in just the same way as you can view a document. This contrasts with being taken to each document in turn with the coded passage highlighted, or a series of cards holding the different passages. In the Node Browser, you can ask to see not only the passages coded, but as much of the context of those passages as you wish, if that helps to understand them.

Now when you’re browsing a document, you can of course code it. Both products provide comprehensive tools for making, viewing and modifying coding in their Document Browsers. But uniquely, they also provide exactly the same coding facilities in their Node Browsers. Since Node Browsers are the place to find and compare nuances in what the node is about, and the place to find what people are or are not talking about in the context of the topic of that node; the Node Browser is the place to code up those nuances and found topics – leading to lots of rich and deep analysis that might well be unrealizable otherwise.

This process is called Coding On, and is made possible by “live” Node Browsers that display their text in context and support coding.

Difficult in the days of paper, the advent of the live Node Browser has made Coding On a simple and universally available tool for qualitative researchers, who are still exploring the power it gives them.

#### Linking: Making the Web of Associations

Edit-while-you-code and the live Node Browser are, in the end, ways of removing fetters from coding. Now we will look at a bunch of tools that are not about coding at all, although the Node system and coding can certainly interact with them. These tools are about making links or associations, involving documents, nodes and other things.

#### Memos and Links

Most qualitative researchers want to keep notes, commonly called memos, about their data and idea. If you have a one-on-one interview with Joe, you may want to have a memo about how Joe behaved in the interview, your thoughts about Joe, and the like. Most QR programs will support writing such a memo, attaching it to Joe’s interview, and adding to it and revising it later.

Such memos can be a valuable source, or indeed explicit repository, of research insights – where the researcher records their evolving thinking about aspects of the project, for instance the rise of a climate of fear and the many ways it interferes with self-esteem. In such a case it seems obvious to link such a memo not to some interview, but to the nodes on fear and self-esteem. N6 and NVivo support that. But more importantly, there is a felt need to code such memos, at anything of research importance they may say. N6 supports this by allowing a memo to be turned into a data document where it can be coded. Obtaining first-class status, if you like. In NVivo, all memos have first-class status anyway. They are no different from any interview document – except that they are called memos.

A memo can be linked to several nodes and documents, so that when you are browsing them you can see they have memos and you can open them in new Browsers. In addition a memo can be linked to any point in the text of a document where it may be relevant, so you see a



little link icon in the text and can access it from there. These in-text links, and others we will be talking about, are all visible in Node Browsers too, and can be accessed from there. And if a memo is rather general in nature, such as a research plan or summary, it needn't be linked anywhere at all, but will still be listed along with all other documents in NVivo's user interface. After all, it is a document. And since it's a document, it can contain links of its own. In this way we can build up a web of links between documents and documents, memos and other documents, nodes and memos or documents. For many researchers, these provide a new way, different from coding, for associating and exploring ideas, topics and themes.

Links can also be to nodes, which provides a sort of converse of coding. If a passage in an interview refers to Joe's peculiar views about sport in the school curriculum, we can insert a link right there in the text, called an extract, to the passage in Joe's interview where he expresses those views. When you set up an extract, the passage being extracted gets put into a node, the extract node, which is what the link in the text jumps you to.

#### Hyperlinking to Other Data

So now in NVivo we can put into the text a fabric of links, joining documents (whether memos or not) to each other, nodes to each other, and between documents and nodes. In addition to such links, marked by little icons in the text, there are more standard hyperlinks to short comments or, significantly, to computer files and web documents. This means that material of any sort at all can be referenced at any place in a document – pictures, web pages, spreadsheets, movies, ... and opened there in its appropriate program. This provides the ability to code such linked items as wholes, by simply coding the hyperlink in the document. In the case of audio and video files, by judicious use of programs that will “snip up” such files, you can attach just the relevant part of a video, for example where Joe is getting worked up about school sport, to the hyperlink. For many researchers, this way of handling the coding of videos is preferable to coding the video file directly. Moreover such links, like document and node links, are always visible and live when presented in Node Browsers, not just in the

editable Document Browser. The ability to make associations, and to link the web of associations with nodes and coding, is now comprehensive.

#### Beyond Retrieval: Asking Questions

Retrieving the text coded at a node may be interesting and illuminating, and lead to a lot of valuable coding-on; but it doesn't show you anything new – you did all that coding. But finding simple Boolean combinations of coding does offer new knowledge. Simple intersection (and) is particularly effective: Given a demographic code such as ‘gender/male’ and a “thematic” node such as ‘bullying’ intersection will show us everything the males have said about bullying. We can by the same procedure put that result alongside what the female interviewees have said about bullying – a contrast likely to be productive of insights to code-on.

A couple of thematic nodes such as ‘playground’ and ‘bullying’ can be intersected to see what's said about bullying in the playground, and a similar search will lead to a contrast with bullying in the classroom. Using the Node Browser facility to see retrievals in context will counteract the way that intersection narrows down its finds.

Several nodes need to be intersected to answer some questions, such as “What do Jewish fathers have to say about playground bullying?” (intersecting four nodes).

N6 and NVivo handle all these searches using a facility called a Search Tool. This supports all other Boolean search operators, so that for example you can ask “What is said about bullying that is not in the playground?” A large range of proximity searches are also provided so that you can ask questions like “Amongst the people who talk about bullying, what do they say about fear of attending school?” – a simple example. The fact that nodes can be organized hierarchically for cataloguing purposes is not forgotten either. So if the ‘curriculum priority’ node has sub-nodes ‘vocational’, ‘all-rounder’ and ‘none’, you can ask to retrieve all the curriculum priority views (nodes below the ‘curriculum priority’ node) and see them together.

Well, where, how, do you see them together? From the earliest versions of NUD\*IST, and in NVivo, the results of any search for any combination of coding has always been stored at a

node. This sort of reflexivity, where results of analyses get stored as new data, is called system closure. It allows the researcher to view the results in a Node Browser, and hence code on – a very fruitful activity with the results of interesting searches. It also allows new questions to be asked involving the search results at any later date. For instance, having stored the answer to “What do Jewish fathers have to say about playground bullying?” as a node, you might ask “Amongst the Jewish fathers who spoke out on playground bullying, what do they have to say about other forms of bullying?” – a proximity search. Such questions are crucial in QR, but how would you get answers to them in a paper-and-file-cabinet research project?

System closure can have significant effects. Consider text search for example, which is supported in comprehensive ways in both products; involving pattern specifications as well as search strings, and in the case of NVivo, approximation searches to allow for misspellings and the like. Given system closure, text search is presented not merely as a way of displaying the next match in the next document; but as a way of collecting all the finds together, optionally in their sentence or paragraph context, and storing them at a node. This not only allows for the sort of coding-on described above, but also means that text search can be brought into the sort of combinatorial searching just described, since the node holding text search results can be input to a Boolean, proximity or other search. It also means that text search can be used as the first rough pass for coding. You make a node holding the passages found by searching for ‘Napoleon’ and ‘Bonaparte’, then add to that by coding when you find indirect references to him.

When you have such comprehensive search tools available, enabling you to ask just about any question expressed in terms of nodes in the project, the task of designing a coding system becomes very much easier, and the resulting system far more flexible. Without such tools, you’d need to ensure you’ve got all the different responses well catalogued by coding ‘Joe on playground bullying’, ‘Joe on classroom bullying’, ‘Henry on playground bullying’, ‘Jewish fathers on classroom bullying’; and so on repetitively to create a morass of combinations of topics and demographics. And all you could do in the end

would be to retrieve them, and asking novel questions like “Do older parents have different views on the effect of teacher discipline on the control of playground bullying than younger parents?” would not be possible. You’d simply have to go back to your documentary data and code for that from the beginning.

Whereas, aware of the power of the search tools, you need to code only for some demographics amongst parents interviewed, and for ‘parent’, ‘bullying’, ‘discipline’, ‘teacher’, ‘playground’, ‘classroom’; then you can ask the questions in the previous paragraph and many others. For instance you find everything said on playground bullying by intersecting ‘playground’, holding everything said about playgrounds and happenings in them, with ‘bullying’, holding everything said about bullying. This makes for simple-to-code, clean, easily organizable node systems that lend themselves to powerful searching, and the crucial ability to make unforeseen searches.

A final but very powerful feature of searches needs to be mentioned. Most of the search operators such as intersection can be applied to not just a pair of nodes (to find their intersection) but to two groups of nodes to create a table or qualitative matrix of their pairwise intersections. For example, to find the views of parents of different religious persuasion on the different curriculum priorities, you take the node ‘Religion’, (below which are ‘Jewish’ etc.) and the node ‘Curriculum priority’ (below which are ‘vocational’ etc.) and you get a table whose cells show what everyone of each religion has said about each curriculum priority. The matrix is stored as coded data, with each cell effectively a node that can be viewed in a Node Browser, where it can be coded-on, used as input to some other search, and so on. A table of numerical data on the cells such as amount of text coded, can be exported to a table-handling program such as SPSS (if the researcher thinks that might be statistically useful data!).

The above outline gives an insight into what N6 and NVivo can do, and a taste of what it’s like to work in such a program. There is a great deal more that can be said, but these are complex and powerful programs, and it would be best to visit the literature on them.

How Do N6 and NVivo Differ? Two Worlds of Work.

In spite of all the common features and tools described above, N6 and NVivo are two rather different products that address two rather different ways of working. One simple example has been mentioned: demographics in N6 are handled in base data trees, but in NVivo in tables of attributes of documents or nodes.

The best way to sum up the differences between the two is that N6 and its forebears are designed for rapid access to textual data via coding, whereas NVivo can handle very complex data with a large variety of tools. Think of NVivo as flexible and subtle, suited for deep analyses as in a typical university PhD project; and N6 as containing a single workmanlike tool that nevertheless provides powerful analyses. Let us spell these out.

N6 requires its data documents to be plain ANSI text, whereas NVivo handles rich text in any font at all. Rich text is more attractive than plain, and is needed of course to display hyperlinks and link icons, but aside from that it gives the user the opportunity to mix languages in a document, to do “visual coding” by highlighting, and to use up to nine level of heading to divide a document into nested subsections. While this presents more opportunities to the researcher, it comes at the price of increased storage demands and slower text handling. In large volumes that can matter, whereas the plain text of N6 makes minimal demands.

For purposes of coding, N6 requires all text to be divided sequentially into text units, which the user can define as sentences, lines or paragraphs. These are the smallest passages of text that can be coded, whereas NVivo supports coding right down to the character level. Fine coding presents better opportunities for researchers interested in the details of what people say, enabling them to pick up words, phrases, and stylistic quirks. At the other extreme, coding at the paragraph level in N6 means it is easy to provide coarse coding economically to enormous volumes of text; and typically for large projects with thousands of interviewees, paragraphs are quite small enough thank you.

The above two features, combined with its ability to automate data handling (see below on Command File scripting), mean that N6 can

handle enormously large projects, limited in general only by the computer’s speed and storage. We know of projects containing tens of thousands of multi-page interview documents, handled well by N6. NVivo would slow down unacceptably on such a big dataset – the recommended maximum is hundreds of documents if they are large and coded to a reasonable level. Of course, even that is no small project.

N6 essentially uses one data type, nodes and their coding, to handle everything – aside from having documents of course. And there’s only one analysis tool, the combinatorial Search Tool described above. NVivo on the other hand has not just nodes and their coding, but comprehensive links as described earlier. It also has sets for grouping documents or nodes in any ways at all. And as mentioned, NVivo avoids the need to use nodes and node trees to handle base data by having a comprehensive attribute/value data type. This is used to set up attributes for documents or nodes, and to assign values to individual documents and nodes – string, Boolean, numeric or date.

As to analysis tools in NVivo, sets have a very comprehensive filtering editor, and attributes have live table displays. In addition there is a Show Tool, for finding lists of related items – all the documents with a particular attribute value for example, or all the nodes coding a given document. And there’s an Assay Tool for looking at the numbers of documents or nodes in a set that have any selected feature – all presented in tabular format with marginals, ready for export to SPSS or other table-handling package.

Moreover NVivo’s Search Tool is more complex than N6’s. Information can be located not just by coding at a node, but in values of attributes, and of course in text search finds. So NVivo’s Search Tool supports Boolean, proximity and Matrix searches, as in N6, but can take as input attribute values and text-search patterns as well as nodes with their coding. In N6 a question like ‘What do Jewish fathers say about classroom bullying’ is framed as intersecting three or four nodes (depending on whether you’ve coded ‘classroom bullying’ as one node or preferably two inviting intersection). In NVivo the intersection would be of attribute-values ‘Religion=Jewish’, ‘Role=father’, and nodes for classroom bullying as before. And if you want to

find where parents talk about the curriculum you don't have to do a 'curriculum' text search first, save the node then intersect with a node or attribute for being a parent. You just intersect the latter with paragraphs containing the word 'curriculum'.

N6 has a scripting tool called command files, allied with a Command Assistant that helps researchers construct complex series of commands to handle large jobs. These can be used over and over again (with editing if need be to change parameters) to cover repetitive work – in the one project or in a series of essentially similar projects. This provided great speedups for many parts of project work. It can even be used to analyze the comparative performance of many coders in a collaborative or multi-site project. NVivo has no scripting, but provides more interactive tools to assist with some complex routines.

NVivo contains a graphical tool for visual exploration of a project's data and their relations. Nodes, documents, sets, attributes and their values can be placed in layers of a graphical model, each being live to its contents (click on a node in a model to open its browser). In addition to the links and groupings a researcher might draw in a model, links can automatically be added to show which nodes code a given document, which documents have a particular attribute value, and the like. Social scientists use "box-and-line" drawings to display a theory or some process or organization in the world, and the graphical modeler is designed to give them great freedom in preparing such diagrams, live to the underlying data. It makes for a great presentation tool! The workmanlike N6 contains no such graphics.

Both products have an associated "Merge" program designed for combining two separate but essentially similar projects into one. They look, for example to see if two same-named documents in the projects are in fact the same in which case their coding can be combined, otherwise treat them as different and change the name of one on merging. N6 treats this merging as essentially a hands-off "batch" process. You set the parameters and let it run. Merge for NVivo however works far more interactively. Before the merging you are taken through an interactive alignment process of examining all potential clashes (like same-named documents) and deciding what to do. At the end of alignment you can stop, having "sorted out" the

two parallel projects so they compare correctly in their document, node and attribute systems, or you can proceed to merge the two of them.

N6, then, is simpler – in its plain text, in its types of data (nodes only) and in its tools and displays. However people working on deep subtle projects, usually in a university research environment, find that compared to N6, NVivo really helps them to soar. It is exhilarating in its richness and flexibility and ways of comparing and showing information. People with simpler needs prefer N6 – there is less to learn and the power remains great. N6 is also the product to use for large projects, which are becoming quite common especially in government or semi-government research organizations, where they are closely allied with extensive quantitative surveys.

These two types of work – simple but powerful and scalable versus complex, flexible and subtle, do effectively divide the field so that most people moving into QR computing recognize which program suits their needs best. These two ways of working, two types of project, are so different that it is unsatisfactory to try to provide one program that handles both excellently – instead you end up with a lowest-common-denominator program.

#### How Qualitative Research is Changing

One of the privileges of being the designers of these programs is to travel the world visiting universities and institutions in very many countries to conduct workshops with users and consult on their projects. This gives a unique insight into how qualitative research is changing under the impact of these computing tools over the last decade. Here are some of the headline changes we have observed since about 1990.

The areas employing QR, especially by computer, as a fundamental tool are broadening. Initially projects and people seemed to come from sociology, educational research, and (intriguingly) areas of engineering. Now there is far more qualitative research in business and organizational studies and consulting, demographics-oriented disciplines such as epidemiology, health sciences (which itself has been a burgeoning discipline over this period), and business-based survey research e.g. market research. Interestingly, history and literary studies remain somewhat aloof.

QR by computer (especially if you're using NVivo) is used to handle a research project end-to-end, not just to analyze filed notes or interviews. All project documentation – project plans, progress summaries, and importantly the research summaries and reports and presentations – are kept inside the NVivo project. This reflects the murky dividing line between data and analysis, and the value of using the linking and coding tools in particular to relate “research” to “data”.

Size of project has increased enormously. Whilst the median size, perhaps a hundred documents at most, remains unchanged, there is a growing tail of huge projects driven by the desire to provide some sort of qualitative analysis of studies with very large  $n$ , and to inform quantitative analysis with such data. Common fields here are government studies, epidemiology and population-wide health studies, global studies by international organizations, and the like. Some specific examples are learning-effectiveness studies of students of a given age across all schools in a state or country, district-by-district analysis of the effects of a new country-wide public safety system, and customer feedback worldwide (where customers are governments) of utilization of major infrastructural capital goods. These may not excite the NVivo-using sociologist who uses QR to develop a theory of social behavior, but their importance, and their need for QR, is great and usually of immediate relevance to communities. They are also all projects suited for N6.

Qualitative-quantitative wars have largely been replaced by collaboration. Some recalcitrant pockets remain, but the change has been remarkable. Of course for many projects on either “side”, there is no need for collaboration; but the incidence of collaborative or mixed-method projects as they are being called, is increasing sharply. The presence of software, particularly the NUD\*IST line over the years, seems to have been quite instrumental here. Two reasons. One, the qual-quant problem mentioned at the start of this paper has been considerably ameliorated by table-handling facilities within qualitative packages combined with table import/export facilities. Thus for example intriguing numerical patterns arising from a matrix search pitching some demographic attribute-values (themselves imported from survey data) against a range of viewpoints elucidated in

interview conversations, can be investigated statistically to test significance or to graph a correspondence analysis.

Reliability is being taken seriously. When the access problem loomed large, researchers tended to erect a number of “monster-barring” defenses here – it's a matter of insight and experience irreplaceable by mere machines, for example. Some defenses by qualitative researchers were quite correct though, for example QR doesn't require a large  $n$  to give it reliability or validity (though some funding committees still think so). After all a biography ( $n=1$ ) can provide tremendous insight into a personality type, a period of history, or a social situation. What matters more is that a QR project carried out in say N6 provides far more auditing of the conclusions a researcher makes. The use of the Search Tool to find the insightful “core” concepts that give an understanding of the problem at hand, can be traced as the results are preserved as nodes. Another researcher can get into the same project and use the Search Tool on the very nodes the original researcher built, to find counter-examples or problematic cases that challenge the original conclusions. Coding patterns can be studied quite directly to see how even they are across the data, and N6's Command Assistant can even produce a script which will compare the coders in a team to find similarities and differences in their coding patterns.

Analysis is going far deeper. Even with smallish projects, the access problem and the clerical time consumed used to put a close limit on the results discoverable and on their exploration. Now however there is little time cost in exploring a large number of hunches and approaches, of combining them and extending them in many ways; in short in encouraging serendipity then putting the discoveries through rigorous analysis.

#### Readings

There is a surprisingly small literature on computational QR, given its ubiquity and the effect it has had to change methods. The series of conferences since 1999 at the University of London, Institute of Education on doing research with QSR's software have led to a special journal issue: *International Journal of Social Research Methodology* (2002a), 5:3.

Amongst its many articles there is a most important discussion of mixed methods by Bazeley (2002a). The evolution of NUD\*IST and NVivo is described in T. Richards (2002). An examination of the effect computing has had on QR methods is set out in L. Richards (2002a; 2002b, 1998). Mixed methods are also discussed in Bazeley (2002b). Bazeley & Richards (2000), Morse & Richards (2002), and (Gibbs, 2002) are three books about how to do QR by computer. The first has a gentle mentoring approach for someone new to qualitative computing; the second is more methodology-oriented (ethnography, phenomenology and so on), while the third takes a more standard text-book approach to the subject.

There are no recent survey books of this fast-changing field. The latest Alexa and Zuell (1999). For a much more comprehensive bibliography of books and articles in the field, visit the following url:  
<http://www.qsrinternational.com/resources/literature/reading.htm>

### Conclusion

The world of computing and software is notoriously unpredictable, which is probably why it has such a huge number of gurus doing the predicting. What shape qualitative research programs will take in ten, or even five years' time is very indeterminate. Arrival in the market by a large established software vendor, or the development by some genius of an unforeseen way of doing QR with computers, can upset any prediction. After all the development of new ways of working has been the hallmark of computing in QR in the past, so why not in the future?

On the other hand, the pressures that might shape QR program revisions in the nearer future can be spelled out. Here are some:

*The rise of mixed methods, the demand for better qual-quant interaction.* This is unlikely to lead to a program that does both, but will lead to innovative thinking on how qualitative programs can better hold up their half of mixed methods. The shape this will take is unforeseeable – something new in research methods may well arise here.

*The application of “intelligent” heuristics.* Using natural language semantics automatically to code documents to the level of intelligence of a

trained researcher, can safely be said to be a very long way off. But there are plenty of more modest artificial intelligence and statistical routines that can be applied to find inductive relationships, to find various sorts of associations between the coding of nodes, and to do data mining as a way of suggesting new and fruitful nodes.

*The pressure to handle large projects with large datasets.* These can be projects where one person or a small local team is studying huge amounts of data. Or there can be multiple researchers gathering data, or joint projects running in several different sites requiring a unified organization and various levels of comparison of site data.

*Handling repetitive or multiple similar projects.* Particularly in the business-driven research world, a successful project will modeled for re-application in similar situations, and hence require the easy definition of its model “skeleton” then easy fleshing-out to the new projects. This can include aggregating the repetitions to an “overall” project.

*Exploiting the Internet.* This is not just finding project data in emails and web pages. The Internet provides a ready-made remote networking and data storage system for people collaborating on projects from multiple sites, and for providing remote and special or customized processing of project data.

*New modes of user interaction.* QR famously makes huge demands on organizing and displaying data, having huge amounts of disorganized data. Early versions of NUD\*IST relied on the scrolling 24 x 80 character display of “glass Teletypes” – which still held sway only 20 years ago. High-resolution color graphics screens, windowing and mousing are all quite recent arrivals, and certainly by no means the last word in user interaction and control. When the next breakthrough arrives it is likely to desert the desktop-and-paper metaphor that the current windowing interface is based on, and provide unforeseen opportunities for novel organization and display of qualitative data.

Given all this, the last word is that we have not yet reached the last word.

## References

- Alexa, M., & Zuell C. (1999). *A review of software for text analysis*. Mannheim: ZUMA.
- Bazeley, P. (2002a). The evolution of a project involving an integrated analysis of structured qualitative and quantitative data: from N3 to NVivo. *International Journal of Social Research Methodology*, 5(3), 229-243.
- Bazeley, P. (2002b). Computerized data analysis for mixed methods research. In A. Tashakkori, & C. Teddlie (Eds.), *Handbook of mixed methods for the social and behavioural sciences*. Thousand Oaks, CA: Sage.
- Bazeley, P. & Richards, L. (2002). *The NVivo qualitative project book*. London & Thousand Oaks CA, Sage.
- Gibbs, G. (2002). *Qualitative data analysis: explorations with NVivo*. Buckingham, Open University Press.
- Morse, J., M., Richards, L. (2002). *Readme first for a user's guide to qualitative methods*. Thousand Oaks & London, Sage.
- Richards L. (1998). Closeness to data: The changing goals of qualitative data handling. *Qualitative Health Research*, (8),3, 319-328.
- Richards, L. (2002a). Rigorous, rapid, reliable and qualitative? Computing in qualitative method. *American Journal of Health Behavior*, 26(6), 425-430.
- Richards, L. (2002b). Qualitative computing – a methods revolution? *International Journal of Social Research Methodology*, 5(3), 263-276.
- Richards, T. (2002). An intellectual history of NUD\*IST and NVivo. *International Journal of Social Research Methodology*, 5(3), 199-214.

## REGULAR ARTICLES

# Fast Permutation Tests that Maximize Power Under Conventional Monte Carlo Sampling for Pairwise and Multiple Comparisons

J.D. Opdyke  
DataMineIt  
Marblehead, MA

---

While the distribution-free nature of permutation tests makes them the most appropriate method for hypothesis testing under a wide range of conditions, their computational demands can be runtime prohibitive, especially if samples are not very small and/or many tests must be conducted (e.g. all pairwise comparisons). This paper presents statistical code that performs continuous-data permutation tests under such conditions very quickly – often more than an order of magnitude faster than widely available commercial alternatives when many tests must be performed and some of the sample pairs contain a large sample. Also presented is an efficient method for obtaining a set of permutation samples containing no duplicates, thus maximizing the power of a pairwise permutation test under a conventional Monte Carlo approach with negligible runtime cost (well under 1% when runtimes are greatest). For multiple comparisons, the code is structured to provide an additional speed premium, making permutation-style p-value adjustments practical to use with permutation test p-values (although for relatively few comparisons at a time). “No-replacement” sampling also provides a power gain for such multiple comparisons, with similarly negligible runtime cost.

Key words: Permutation test, Monte Carlo, multiple comparisons, variance reduction, multiple testing procedures, permutation-style p-value adjustments, oversampling, no-replacement sampling

---

### Introduction

Permutation tests are as old as modern statistics (see Fisher (1935)), and their statistical properties are well understood and thoroughly documented in the statistics literature (see Pesarin (2001) and Mielke and Berry (2001) for extensive bibliographies). Though not always as powerful as their parametric counterparts that rely on asymptotic theory, they sometimes have equal or even greater power (see Andersen and Legendre (1999) for just one example). In addition to their utility when asymptotic theory falls short (e.g. small samples and the Central Limit Theorem), permutation tests are unbiased, and when fully

enumerated, they provide gratifyingly exact results. Most important, however, is that with few exceptions, valid permutation tests rely on no distributional assumptions – only the requirement that the data satisfies the condition of exchangeability (i.e. distributional invariance under the null hypothesis to permutations of the subscripts of the data points). This gives permutation tests a very broad range of application.

Until recently the major drawback of permutation tests has been their high computational demands. Even when sampling from the permutation sample space, as is typically done, rather than fully enumerating it, computer runtimes still have been prohibitive, especially if samples are not very small. Recent advances in computing speed and capacity increasingly have relaxed this constraint, but the continual development of new and computationally intensive statistical methods is easily keeping pace with such advances.

---

J.D. Opdyke is President of DataMineIt, a statistical data mining consultancy (jdopdyke@datamineit.com, www.datamineit.com). I owe special thanks to Geri S. Costanza, M.S., for a number of valuable insights. Any errors are my own.



For example, Westfall and Young (1993) convincingly demonstrated, under a broad range of real-world data conditions, the need for resampling-based multiple testing procedures. However, if the unadjusted p-values themselves are derived from resampling methods, such as permutation tests, the multiple comparisons p-value adjustment requires a computationally intensive nested loop, where a large number (thousands) of additional permutation tests must be performed for each original permutation test to properly adjust its p-value. Obviously, even if each permutation test requires just a few seconds, runtimes quickly become prohibitive if there are many p-values that need to be adjusted.

Similarly, power estimation of tests based on resampling methods require the same intensive nested loop structure (see Boos and Zhang (2000) for a useful computation reduction technique), while power estimation of the multiple comparisons adjustment procedure mentioned above requires an additional (third) loop.

Such examples clearly demonstrate the ongoing need to develop faster code and algorithms that are also increasingly statistically efficient, since variance reduction lessens sampling requirements which, all else equal, increases speed. The goal of the methods described below is to contribute to these efforts.

#### Widely Available Permutation Sampling Procedures

Three procedures in SAS<sup>®</sup> v8.2 – PROC NPARIWAY, PROC MULTTEST, and PROC PLAN – and one procedure in Cytel's Proc StatXact<sup>®</sup> v5.0 – PROC TWOSAMPL – can be used to perform two-sample nonparametric permutation tests. All but PROC PLAN sample the input dataset itself, while PROC PLAN generates a record-by-record list, each record containing a number identifying the corresponding record on the input dataset to include in the “permutation” samples. This list subsequently must be merged with the original data to obtain the corresponding data points, something PROC MULTTEST does automatically by directly generating all the “permutation” samples it uses for permutation-style p-value adjustments (these

samples, however, can be used instead as the samples for the actual permutation tests). In contrast, both PROC NPARIWAY and PROC TWOSAMPL actually conduct the permutation test and provide a p-value, whereas the samples from both PROC MULTTEST and PROC PLAN must be manipulated “by hand” to calculate the value of the test statistic associated with the original sample pair, and then compare it to all those associated with each of the “permutation” samples to obtain a p-value.

Nonetheless, effective use of PROC PLAN, as shown in benchmarks in the Results section below, is much faster than these other procedures – often more than an order of magnitude faster when one of the samples is large. The only potential problem with using PROC PLAN is that it has a sample size constraint – the product of the sum of the two sample sizes ( $n_1 + n_2$ ) and the number of “permutation” samples being drawn (T) cannot exceed  $2^{31}$  (about 2.1 billion, the largest representable integer in SAS) or the procedure terminates. However, this can be circumvented by inserting calls to PROC PLAN in a loop which cycles  $\text{roundup}((n_1 + n_2) * T / 2^{31})$  times, each loop drawing  $T * \lceil \text{roundup}((n_1 + n_2) * T / 2^{31}) \rceil^{-1}$  samples until T samples have been drawn (see code in Appendix C). This looping in and of itself does not slow execution of the procedure.

All of the abovementioned procedures can perform conventional Monte Carlo sampling without replacement *within a sample*, as required of all but a few stylized permutation tests, but none can avoid the possibility of drawing the same sample more than once. In other words, when drawing the sample of “permutation” samples, these procedures can only draw from the sample space of samples (conditional on the data) *with replacement* (WR). This problem of drawing duplicate samples, its effect on the statistical power of the permutation test, and a proposed solution that maximizes power under conventional Monte Carlo sampling for both pairwise and multiple comparisons are discussed in the Methodology section below. First, the background issues of determining the number of “permutation” samples to draw, and sampling approaches other than conventional Monte Carlo, are addressed

below.

#### Determining the Number of Permutation Samples

When drawing samples from the permutation sample space, one must determine how many samples should be drawn. Obtaining an exact p-value from a permutation test via full enumeration – i.e. by generating all possible sample combinations by reshuffling the data points of the samples at hand – quickly becomes infeasible as sample sizes increase. As shown in (1), the number of possible sample combinations becomes very large even for relatively small sample sizes (two samples of 29 observations each, for example, have 30,067,266,499,541,000 possible sample combinations).

$$\# \text{ of two-sample combinations} = {}_n C_{n_1} = \frac{(n_1 + n_2)!}{n_1! n_2!} \quad (1)$$

where  $n_1$  = sample one's size,  $n_2$  = sample two's size, and  $n = n_1 + n_2$

Network algorithms (see Mehta and Patel (1983)) expand the sample size range over which exact p-values realistically may be obtained, but the rapid combinatorial expansion of the “permutation” sample space – defined as conditional on the data in (1) – still limits the full enumeration of continuous data samples to relatively small sample sizes.

Sampling from the permutation sample space, however, can provide an estimate of the exact p-value via a conventional Monte Carlo approach, whereby the probability of drawing any particular sample is equal to one divided by the number of possible sample combinations, as in (2) below:

$$\Pr(S = s) = \frac{1}{{}_n C_{n_1}} \quad (2)$$

(Note that permutations of the same sample do not affect this probability.) A (one-sided) permutation test p-value is simply the number of test statistic values, each corresponding to a “permutation” sample, at least as large as that based on the observed data samples; therefore, the estimated p-value based on conventional Monte Carlo sampling is simply an estimated proportion

distributed binomially. The normal approximation to the binomial distribution allows one easily to obtain specified levels of precision for this estimate, based either on the standard error (se) or the coefficient of variation (cv), as a function of  $T$  = the number of samples drawn. This is done by straightforward solutions of (3) and (4) respectively (see Brown et al. (2001) for descriptions of the “Agresti-Coull” and “Wilson” intervals – superior, if slightly more complex, alternatives to the commonly used Wald approximation shown in (3)).

$$se \approx \sqrt{\frac{p - value(1 - (p - value))}{T}}, \text{ and} \quad (3)$$

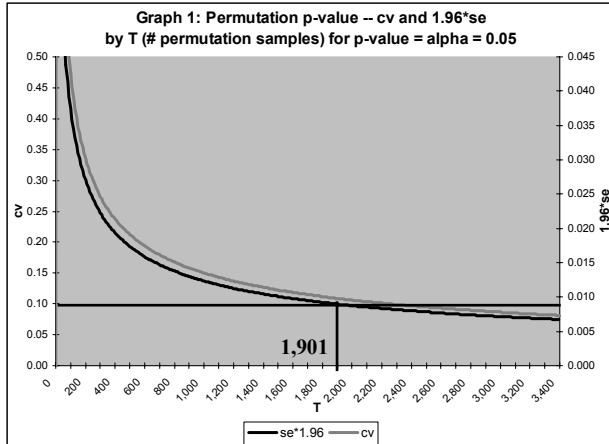
$$95\% \text{ ci} \approx p - value \pm (1.96 \times se)$$

$$cv = \frac{se}{p - value} \quad (4)$$

$$cv = \frac{\sqrt{0.05(1-0.05)}}{0.05} = 0.10 \Rightarrow T = 1,900$$

and for  $cv < 0.10$   $T = 1,901$ .

For example, if  $cv < 0.10$  is needed, one would solve for  $T$  in (4) using the most relevant p-value ( $p\text{-value} = \alpha$ ) and adding one to the solution so that the inequality holds (see Efron and Tibshirani, 1993, pp. 208-211 for an identical calculation). If  $\alpha = 0.05$ , then  $T = 1,901$ , which also yields an approximate 95% confidence interval, based on (3), of just under 0.01 on either side of  $p\text{-value} = \alpha = 0.05$ . While this may be sufficiently precise for many applications, increased precision is obtainable with larger  $T$ , though as shown in Graph 1, marginal gains in precision decrease rapidly in  $T$ . (Note that the normal approximation to the binomial distribution easily satisfies the strictest criteria in the statistical literature for  $T = 1,901$  and  $p\text{-value} = 0.05$  (see Cochran (1977), p. 58, and Evans, et al. (1993), p. 39)).



An efficient alternative to a fixed level of precision, however, especially when conducting many permutation tests, is increasing  $T$  only when the confidence interval of a specific test includes the critical value. Selectively tightening the confidence interval in this way avoids wasteful sampling when p-values are nowhere near the critical value of the test.

#### Other Sampling Methods

The level of precision a method provides for a given number of samples is its efficiency. The efficiency, as well as speed, of conventional Monte Carlo sampling as described above typically are inferior to other sampling methods, such as various forms of importance sampling, which recently have received considerable attention and development (see Owen (2000) for a current survey and recent developments). The idea is that samples are selected not with a uniform probability over the entire sample space, but rather, based on their “importance” for reducing the variance of the estimated p-value. While these and similar variance reduction methods are extremely effective under a wide and growing range of conditions, this paper focuses on conventional Monte Carlo sampling for several reasons: first, some conditions remain under which such methods cannot (yet) be implemented reliably, and results based on quickly implemented conventional Monte Carlo should serve at least as an important verification of the validity of these more efficient methods when their results are suspect; secondly, to date there is little research on the use of such methods in resampling-based

multiple testing procedures (see Naiman and Priebe (2001) and Ortiz and Kaelbling (2000) for related work in this area); and lastly, the sampling procedures in most statistical software packages utilize conventional Monte Carlo, making it much easier to implement when applying resampling methods to stylized statistical tests.

Thus, this paper addresses the need for fast statistical code that quickly performs permutation tests based on conventional Monte Carlo sampling for pairwise and multiple comparisons. It also proposes a simple modification to how most researchers implement conventional Monte Carlo permutation tests: it proposes sampling from the permutation sample space without replacement rather than with replacement which, by definition of conventional Monte Carlo, maximizes power under this sampling approach through variance reduction. The proposed method (“oversampling”) can utilize any “with-replacement” (WR) sampling procedure to accomplish this, in effect efficiently converting any WR sampling procedure into a “no-replacement” (NR) sampling procedure. Before describing “oversampling,” however, the power differential between WR sampling and NR sampling is examined below.

#### Methodology

##### Duplicate Permutation Samples and Power

As mentioned above, all of the procedures examined in this study – PROC PLAN, PROC MULTTEST, PROC NPAR1WAY, and PROC TWOSAMPL – can perform conventional Monte Carlo sampling without replacement *within a sample*, as is required of almost all permutation tests (see Pesarin (2001), Ch. 10, for a notable exception). In other words, no duplicates of the same data point exist within a single sample. This reference to sampling “without replacement” is distinct from drawing an entire set of “permutation” samples that contains no entire sample more than once; this is referred to below as no-replacement (NR) sampling, while generating a set of “permutation” samples that may contain duplicate samples is referred to as “with replacement” (WR) sampling.

*No-replacement (NR) Sampling and Pairwise Comparisons*

Regardless of the number of permutation samples drawn (T), a single pairwise permutation will lose statistical power if there are duplicate samples among the T samples drawn. Intuitively, this makes sense because the fewer duplicates contained in the sample of “permutation” samples, the better represented is the empirical distribution function, and more information almost always implies greater power. In other words, if a difference between population distributions truly exists, more information (i.e. fewer duplicates), on average, should allow us to more readily detect it. And drawing a sample that contains *no* duplicates will yield the greatest power attainable under conventional Monte Carlo.

Statistically, the greater power attributable to NR sampling over WR sampling is due to variance reduction in the estimated p-value ((5.1) – (5.5)). Any permutation test relying on sampling rather than full enumeration will yield an actual significance level (asl) larger than  $\alpha$  due to Monte Carlo error (see Berry & Mielke (1983)). This (one-sided) sampling-based asl is simply the probability under the null hypothesis that the value of the test statistic, based on the “permutation” samples, is equal to or greater than that corresponding to the critical value of the test conditional on the true p-value (the conditional nature of this probability requires summing over all possible values of p, as in (5.8) and (5.9)). The asl under NR sampling is smaller than the asl under WR sampling because the abovementioned conditional distribution of the former is based on the hypergeometric distribution: this has smaller variance than the conditional distribution of WR sampling, which is based on the binomial distribution ((5.6) and (5.7)). This means that once the critical p-values are adjusted to account for  $asl > \alpha$  (the Monte Carlo error), the adjusted critical value for NR sampling will be larger than that of WR sampling ((5.10) – (5.13)). This gives permutation tests based on NR sampling greater power.

$$\sigma_{hyp}^2 < \sigma_{bin}^2 \quad (5.1)$$

$$\Rightarrow \sigma_{NR}^2 < \sigma_{WR}^2 \quad (5.2)$$

$$\Rightarrow asl_{NR} < asl_{WR} \quad (5.3)$$

$$\Rightarrow c_{\alpha_{NR}}^* > c_{\alpha_{WR}}^* \quad (5.4)$$

$$\Rightarrow power_{NR} > power_{WR} \quad (5.5)$$

where

$$\sigma_{bin}^2 = n_p p(1-p) \quad (5.6)$$

$$\sigma_{hyp}^2 = n_p p(1-p) \binom{n C_{n_1} - n_p}{n C_{n_1} - 1} \quad (5.7)$$

where

$n_p$  = number of permutation samples drawn,

$$(5.8)$$

$$asl_{WR} = \Pr(S \leq (n_p \alpha) | p) = \frac{1}{n_p} \sum_{i=0}^{n_p} \sum_{k=0}^{\lfloor n_p \alpha \rfloor} \binom{n_p}{i} \left( \frac{i}{n_p} \right)^k \left( 1 - \frac{i}{n_p} \right)^{n_p - k} \quad (5.9)$$

$$asl_{NR} = \Pr(S \leq (n_p \alpha) | p) = \frac{1}{n_p} \sum_{S=0}^{N \text{ by } n_p} \sum_{k=0}^{\lfloor n_p \alpha \rfloor} \frac{\binom{S}{k} \binom{N-S}{n_p - k}}{\binom{N}{n_p}}$$

where

$S$  = number of “successes” (number of “permutation” sample test statistic values  $\geq$  observed sample test statistic value) among  $n_p$  permutation samples drawn,

$\frac{N}{n_p}$  is an integer, and

$c_{\alpha}^*$  = the critical value adjusted for Monte Carlo error.

(Note that above, the critical p-value of the test is adjusted, rather than the p-values themselves, solely for heuristic and computational purposes when demonstrating the power differential between NR and WR sampling in (5.1)-(5.5). In practice, it is the p-values themselves which should be adjusted for ease of interpretation of the test results. Both adjustments yield identical

results statistically.) The discreteness of both the binomial and hypergeometric distributions prevent the attainment of adjusted critical p-values yielding  $asl = \alpha$  exactly. However, interpolation between  $\alpha$  and the largest p-value yielding  $asl < \alpha$ , based on the percentage change in the corresponding  $asl$ 's, provides a reasonable approximation of the critical p-values that would yield  $asl = \alpha$  if the distributions were continuous. Although this interpolation was used when calculating the asymptotic power differential between NR sampling and WR sampling ((6.2) vs. (6.3) and Table 2), a convenient shorthand provides similar results. If  $(asl / \alpha)$  is assumed to be constant for p-values close to  $\alpha$ , then

$$c_{\alpha}^* \left( \frac{asl}{\alpha} \right) \approx \alpha \quad (5.10)$$

so

$$c_{\alpha}^* \approx \frac{\alpha^2}{asl} \quad (5.11)$$

and

$$c_{\alpha_{NR}}^* \approx \frac{\alpha^2}{asl_{NR}} \quad (5.12)$$

$$c_{\alpha_{WR}}^* \approx \frac{\alpha^2}{asl_{WR}} \quad (5.13)$$

The power differential resulting from use of the two different critical values can be obtained by simulation. An asymptotic approximation, however, provides, as a lower bound, a good idea of its order of magnitude, as well as a useful benchmark against which simulations based on different distributions can be compared to demonstrate relative rates of convergence (efficient use of Boos and Zhang (2000) to perform these simulations is the subject of continuing research).

By the Central Limit Theorem, we know that asymptotically,

$$power = 1 - \Phi \left( z_{\alpha} - \frac{\delta \sqrt{n}}{\sigma} \right) \quad (6.1)$$

where

$$n = n_1 + n_2$$

$\delta$  = size of effect (a location shift)

$\sigma$  = population variance

(see Pesarin (2001), p. 65)

Therefore

$$power_{NR} \approx 1 - \Phi \left( z_{c_{\alpha_{NR}}^*} - \frac{\delta \sqrt{n}}{\sigma} \right) \quad (6.2)$$

$$power_{WR} \approx 1 - \Phi \left( z_{c_{\alpha_{WR}}^*} - \frac{\delta \sqrt{n}}{\sigma} \right) \quad (6.3)$$

(Note that knowledge of  $\sigma$  is unnecessary if  $\delta$  is expressed in terms of  $\sigma$ .) The empirical results of this asymptotic analysis, which are lower bounds for the actual power gains provided by NR sampling, are included in the Results section below in Table 2 (the derivations shown in (5.1) – (6.3) were first presented in Opdyke (2002b)).

### NR Sampling and Multiple Comparisons

The above rationale for the power gains of NR sampling applies to multiple comparisons as well. However, for permutation-style p-value adjustments of permutation test p-values, there are two sources of power gain: a) a stochastically larger distribution of the minimum p-value under NR sampling, and b) smaller original p-values of the permutation tests themselves, after adjustment for Monte Carlo error as described above (note that here, the p-values themselves are adjusted, rather than the critical p-values).

Take the single step multiple testing adjustment procedure described by Westfall and Young (1993) (Algorithm 2.5, pp. 46-48). If we have, say, a family of ten permutation test p-values that need adjustment, we need to generate, under the complete null hypothesis, a vector of ten new p-values by the same process (permutation test) some large number of times, and for each original p-value count the number of times the minimum p-value of each vector is smaller than or equal to that original p-value. Dividing each of these ten counts by the number of times the simulation is run yields ten proportions, which are the ten adjusted p-values.

a) Note that since each p-value in each vector is simply another permutation test, NR sampling will yield a smaller variance for each of these p-values compared to WR sampling, as described in the previous section ((5.1) – (5.2), (5.6) – (5.7)). As a consequence, the minimum p-value will be

stochastically larger when the p-values in each vector are generated using NR sampling than when using WR sampling (7.1). Therefore, the probability that the minimum p-value will be smaller than a given original p-value will be smaller for NR sampling than for WR sampling (7.2). This makes the corresponding numerator (the count) of the adjusted p-value smaller on average, and the adjusted p-value itself smaller on average (7.3), giving the p-value adjustment under NR sampling more power (7.4).

$$\min_{1 \leq j \leq k} P_{j_{NR}}^* \text{ is stochastically larger than } \min_{1 \leq j \leq k} P_{j_{WR}}^* \quad (7.1)$$

$$(7.2)$$

$$\Rightarrow \Pr\left(\min_{1 \leq j \leq k} P_{j_{NR}} \leq p_i \mid H_0^C\right) < \Pr\left(\min_{1 \leq j \leq k} P_{j_{WR}} \leq p_i \mid H_0^C\right)$$

$$\Rightarrow \text{power}_{NR(a)} > \text{power}_{WR(a)} \quad (7.3)$$

$$\Rightarrow \tilde{p}_{i_{NR(a)}} < \tilde{p}_{i_{WR(a)}} \quad (7.4)$$

where

$P_i$  = original p-value

$P_j^*$  = data-based p-value vector of  $j$  p-values

$P_j$  = joint random variable of  $j$  p-values

$H_0^C$  = the complete null hypothesis, i.e. assuming that all null hypotheses included in the family of multiple comparisons are true

$\tilde{p}_{i_{NR}}$  = the adjusted p-value of  $p_i$

b) Another source of power gain from NR sampling is the smaller p-values of the original permutation tests themselves, after adjustment for Monte Carlo error as described in the previous section. Assume that none of the “simulated” p-values in each vector are generated using NR sampling, but that the original p-values are generated, and then Monte Carlo-error adjusted, using NR sampling instead of WR sampling. Because the p-values of the former are smaller (8.1), the probability of the same minimum p-value being less than or equal to the original p-value is smaller for NR sampling (8.2). This means the corresponding numerator (the count) of

the adjusted p-value will be smaller on average, and the adjusted p-value itself will be smaller on average (8.3), giving the p-value adjustment under NR sampling more power (8.4).

$$P_{i_{NR}} < P_{i_{WR}} \quad (8.1)$$

$$(8.2)$$

$$\Rightarrow \Pr\left(\min_{1 \leq j \leq k} P_j \leq P_{i_{NR}} \mid H_0^C\right) < \Pr\left(\min_{1 \leq j \leq k} P_j \leq P_{i_{WR}} \mid H_0^C\right)$$

$$\Rightarrow \tilde{p}_{i_{NR(b)}} < \tilde{p}_{i_{WR(b)}} \quad (8.3)$$

$$\Rightarrow \text{power}_{NR(b)} > \text{power}_{WR(b)} \quad (8.4)$$

Therefore, to maximize NR sampling power gains when using permutation-style p-value adjustments in multiple comparisons of permutation test p-values, combine both a) and b) – use NR sampling to generate both the original Monte Carlo-error adjusted p-values, as well as the “simulated” p-value vectors when making the multiple comparisons adjustment ((9.1) – (9.3)).

$$(9.1)$$

$$\Pr\left(\min_{1 \leq j \leq k} p_{j_{NR}} \leq p_{i_{NR}} \mid H_0^C\right) < \Pr\left(\min_{1 \leq j \leq k} p_{j_{WR}} \leq p_{i_{WR}} \mid H_0^C\right)$$

$$\Rightarrow \tilde{p}_{i_{NR}} < \tilde{p}_{i_{WR}} \quad (9.2)$$

$$\Rightarrow \text{power}_{NR} > \text{power}_{WR} \quad (9.3)$$

The same rationale applies to stepwise multiple comparisons adjustments. Whenever NR sampling is used to generate either or both the minimum p-value and the original Monte Carlo error-adjusted p-values, its variance reduction will yield greater power (these derivations, (7.1)-(9.3), were first presented in Opdyke (2002b)).

Efficient simulation of the power differential shown in (9.1) – (9.3), which requires a computationally intensive nested loop with three levels, is the topic of continuing research. However, its magnitude may very well be larger than that of a single pairwise comparison since variance reduction is achieved from two sources – both a) and b) above – rather than from b) alone.

Before presenting the asymptotic power calculations for a single pairwise comparison, the

next section derives and presents an efficient method for performing NR sampling based on any procedure which uses WR sampling, as do all the “permutation” sampling procedures examined in this paper and known to this author. “Oversampling,” in effect, efficiently converts any WR sampling procedure into an NR sampling procedure, as shown below.

#### “Oversampling” to Avoid Duplicate Samples

“Oversampling” involves simply drawing more than the desired  $T$  samples (say,  $r$  samples), deleting any duplicate samples, and then randomly selecting  $T$  samples from the remaining set (this method, and its results in Table 1, were first presented in Opdyke (2002a)). This approach does not alter the probability of drawing any particular sample (see (2)), so “oversampling” is a statistically valid approach for obtaining  $T$  distinct samples.

The next question to address is, what is the optimal size of  $(r-T)$ ? The goal is to minimize expected runtime, which is a function of  $(r-T)$ , or simply  $r$ , and the size of  $r$  involves the following runtime tradeoff: larger  $r$  will contribute to longer runtimes due to the extra time required to generate more samples, but also will diminish the probability that fewer than  $T$  unique samples will be drawn, which would require another draw of  $r$  samples and increase overall runtime; smaller  $r$  will require less time to generate fewer samples, but at the price of an increased probability of being left with fewer than  $T$  unique samples and having to redraw the samples all over again. Expected runtime is simply the product of a) the expected number of times  $r$  samples need to be drawn to obtain at least  $T$  unique samples, and b) the time it takes to draw  $r$  samples. So if expected runtime =  $g(r, x, y, \dots)$ , we seek  $r$  such that  $\partial g/\partial r = 0$  (and  $\partial^2 g/\partial r > 0$ ).

#### Minimizing Expected Runtime

a) The number of times  $r$  samples must be drawn before obtaining at least  $T$  unique samples is a random variable that follows the geometric distribution, which identifies the number of events occurring before the first success:

$$\Pr(S = s) = p(1-p)^{(s-1)} \quad (10)$$

where  $p$  indicates the probability of success (of obtaining at least  $T$  unique samples) for each event (each call to PROC PLAN, or whichever WR sampling procedure is being used). The expected value of the geometric distribution is  $E[S] = 1/p$ , and  $p$  is derived from a general form of the familiar (coupon or baseball card) collector’s problem. This problem asks the question, “How many card packets must one purchase to collect a complete set of baseball cards?” or equivalently, “How many samples must one draw, when sampling with replacement (because the sample size is so large), to obtain a complete set of all samples from the sampling distribution?” The more general problem, which is the relevant one for this analysis, is “How many samples are required, when sampling with replacement, to obtain  $T$  distinct samples from the sampling distribution?” The number of samples “required” follows a probability mass function (11) which is the sum of geometric random variables.

(11)

$$\Pr(\# \text{ unique samples} = j) = \frac{{}^n C_{n_1}}{j!({}^n C_{n_1} - j)!} \sum_{i=0}^j \frac{(-1)^i j!(j-i)^r}{i!(j-i)!({}^n C_{n_1})^r}$$

where  $r = \#$  of samples drawn and  $j \leq r$

However, we are interested in the probability of obtaining at least  $T$  unique samples, which is simply the cumulative probability of obtaining  $T$ ,  $T+1$ ,  $T+2$ ,  $\dots$ ,  $r-1$ , and  $r$  unique samples, as shown below:

(12)

$$p = \Pr(j \geq T) = \sum_{j=T}^r \left[ \frac{{}^n C_{n_1}}{j!({}^n C_{n_1} - j)!} \sum_{i=0}^j \frac{(-1)^i j!(j-i)^r}{i!(j-i)!({}^n C_{n_1})^r} \right]$$

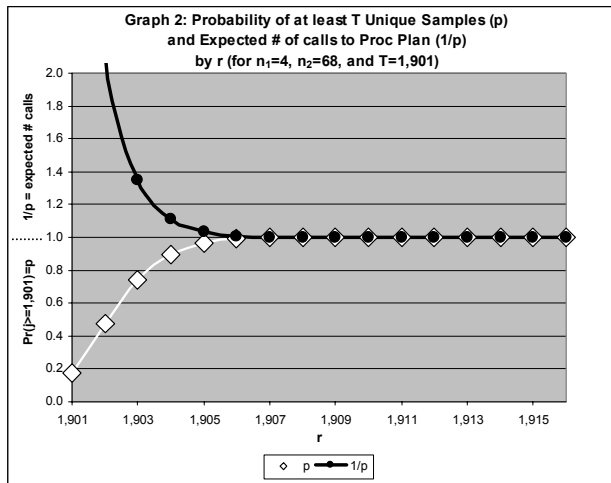
where  $T \leq r$ .

Thus, the expected number of times  $r$  samples must be drawn to obtain at least  $T$  unique samples is a function of the number of possible sample combinations and  $r$ , as shown in (13) below:

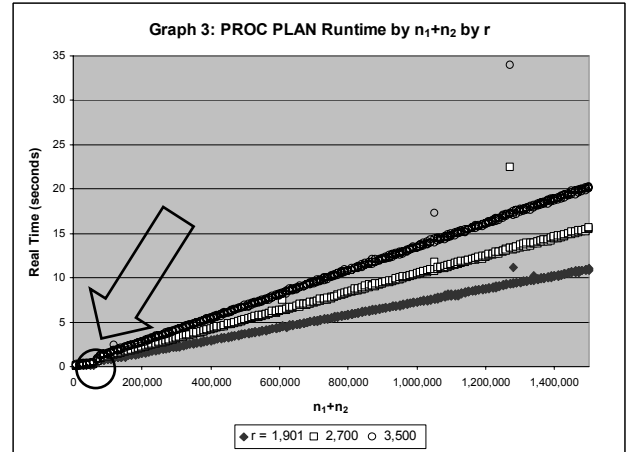
expected # of calls to PROC PLAN = (13)  
 $CTPP(n C_{n_1}, r, T) =$

$$\left( \frac{1}{p} \right) = \left( \sum_{j=T}^r \left[ \frac{n C_{n_1}}{j!(n C_{n_1} - j)!} \sum_{i=0}^j \frac{(-1)^i j!(j-i)^r}{i!(j-i)!(n C_{n_1})^r} \right] \right)^{-1}$$

Graph 2 illustrates the functional relationship between p, 1/p, and r for  $n_1 = 68, n_2 = 4,$  and  $T = 1,901$ :



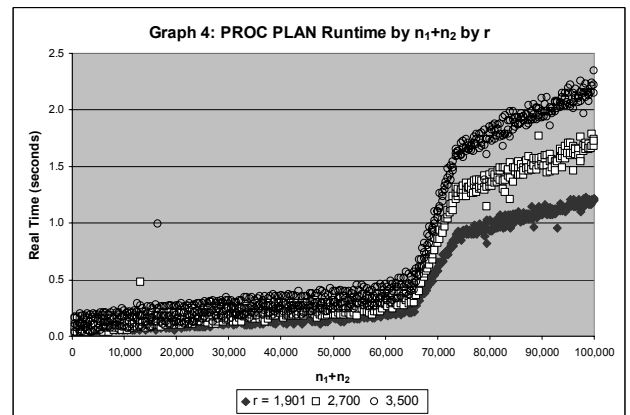
b) Now to return to the other factor determining expected sampling runtime – the time it takes PROC PLAN to draw a sample of r samples. This is simply the runtime of PROC PLAN as a function of, interestingly, not the number of possible two-sample combinations, but rather the sum of the two sample sizes ( $n_1 + n_2$ ), as well as the number of samples drawn, r. This is shown in Graph 3 (see Appendix A for simulation details). Obviously, r and ( $n_1 + n_2$ ) are correlated, but runtime is very well predicted (adj  $R^2 = 0.9884$ ) by the simple ordinary least squares multivariate regression equation in (14):



(14)

$$\text{PROC PLAN Runtime} = \text{PPRT}(n_1, n_2, r) = \beta_0 + \beta_1*(n_1 + n_2) + \beta_2*r + \beta_3*(n_1 + n_2)*r$$

Nonlinearity at about ( $n_1 + n_2$ ) = 65,500 and ( $n_1 + n_2$ ) = 73,500 prompted the inclusion of dummy and interaction terms, leading to the near perfect prediction (adjusted  $R^2 = 0.9927$ ) for  $\text{PPRT}(n_1, n_2, r)$  presented in Appendix B (see Graph 4, which is simply a magnification of Graph 3 up to ( $n_1 + n_2$ )=100,000).

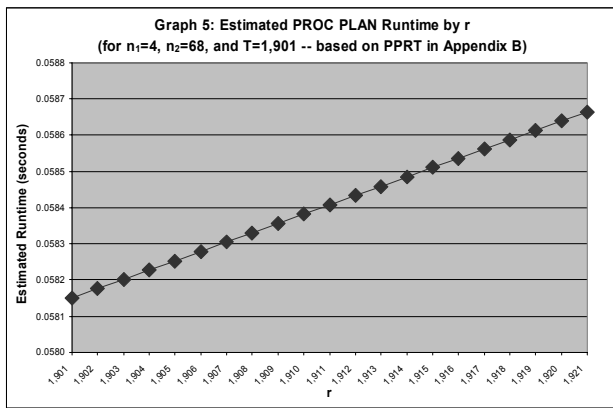


Thus, expected runtime  $g(n_1, n_2, r, T)$  is the product of PROC PLAN Runtime and the expected number of calls to PROC PLAN:

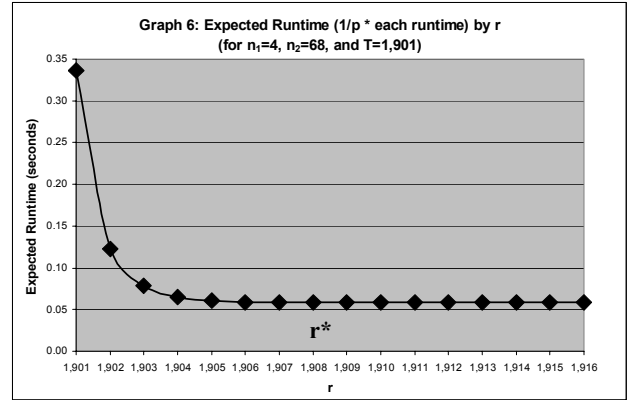


$$\begin{aligned}
 & \text{expected runtime} = g(n_1, n_2, r, T) = (14) \times (13) = \\
 & \text{PPRT}(n_1, n_2, r) * \text{CTPP}(n_1, r, T) = \\
 & \left[ \beta_0 + \beta_1(n_1 + n_2) + \beta_2*r + \beta_3*(n_1 + n_2)*r \right. \\
 & + d_1*\beta_4 + d_1*\beta_5*(n_1+n_2) + d_1*\beta_6*r + d_1*\beta_7*(n_1+n_2)*r \\
 & + d_2*\beta_8 + d_2*\beta_9*(n_1+n_2) + d_2*\beta_{10}*r + d_2*\beta_{11}*(n_1+n_2)*r \\
 & \left. \right] * \\
 & \left( \sum_{j=T}^r \left[ \frac{{}_n C_{n_1}}{j!({}_n C_{n_1} - j)!} \sum_{i=0}^j \frac{(-1)^i j!(j-i)^r}{i!(j-i)!({}_n C_{n_1})^r} \right] \right)^{-1}
 \end{aligned}
 \tag{15}$$

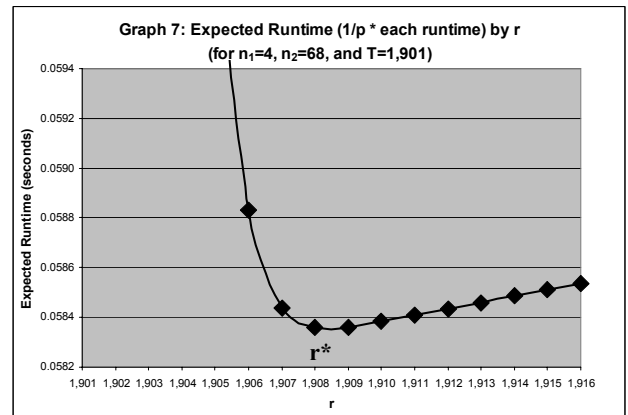
To get an intuitive feel for  $r$  as a function of  $n_1$  and  $n_2$  (for a given  $T$ ), note again that the second term of (15) is a combinatorial function of the sample sizes while the first term is merely a linear function of the sample sizes (see Graph 5).



The combinatorial terms in the second term of (15) end up dominating as sample sizes increase, asymptotically converging to 1.0 (one call to PROC PLAN) faster than the first term (each PROC PLAN runtime) diverges. Hence, for all but very small sample sizes, an optimal  $r$  in terms of expected runtime (where  $\partial g/\partial r = 0$ ) will be fairly close to  $T$ . Graphs 6 and 7 below present  $g(n_1, n_2, r, T)$  – the product of  $1/p$  in Graph 2 and PPRT in Graph 5 above – and demonstrate an optimal  $r$ ,  $r^* = 1,908$ , for  $T = 1,901$ ,  $n_1 = 4$ , and  $n_2 = 68$  (and  ${}_n C_{n_1} = C = 1,028,790$ ).



Graph 7 magnifies the relevant expected runtime range.



Unfortunately, the high level of precision needed to calculate numeric solutions for  $r^*$  based on (15), for different sample sizes and different values of  $T$ , requires use of a symbolic programming language (the Mathematica<sup>®</sup> v4.1 code used to obtain the exact probabilities in Table 1 is available from the author upon request). Thus, exact solutions cannot be implemented “on the fly” in SAS, or any statistical software package, for encountered values of  $n_1$  and  $n_2$ . Good approximations to the probability mass function of the collector’s problem, however, do exist (see Kuonen (2000) and Read (1998), as well as Lindsay (1992) for a unique approach to the problem), but whether using exact or approximate probabilities, for all practical purposes  $r^*$  need not be calculated for each and every combination of values of  $n_1$  and  $n_2$ . Nearly optimal  $r$  can be calculated for ranges of  $C$  because, as shown in Graph 7, the marginal runtime cost of drawing  $r$  slightly larger than  $r^*$  is negligible (though the

marginal runtime cost of drawing  $r$  smaller than  $r^*$  is relatively large). Thus, if we define appropriate ranges of  $C$ , and for the lower bound of each range identify  $r^*$ , these “low-end”  $r^*$ s always will be larger than any other  $r^*$  corresponding to any of the sample pairs within their respective ranges. In other words, though not optimal for every combination of sample sizes within its range, the “low-end”  $r^*$  will be nearly optimal because it will be slightly larger (never smaller) than all other  $r^*$  for sample size pairs within its range, and the marginal runtime cost of being slightly larger than  $r^*$  is negligible.

Table 1 below shows the values of  $r$  used in the permutation test program – the “low-end”  $r^*$ s – for ranges of  $C$ . Although  $g(n_1, n_2, r, T)$  is a function of both  $C$  and  $n_1 + n_2$ , and  $n_1 + n_2$  does vary for (essentially) constant  $C$ , the effect of this can be ignored since, as an empirical matter, it never affects the calculation of each of the “low-end”  $r^*$ s. In other words, CTPP (13) strongly dominates PPRT (14) because  $1/p$  converges to one so quickly.

The code in Appendix C proposes an efficient method for generalizing the results from Table 1, i.e. for obtaining estimates of the optimal “low-end”  $r^*$ s for any value of  $T$ . This method is very fast, perhaps even faster than Kuonen (2000), although it provides only estimates to the exact solution. It first utilizes optimal “low-end”  $r^*$ s already calculated for a particular value of  $T$  (as in Table 1) as the basis for conservative estimates of the distance (standard deviations) between a new  $T$  and the mean of the collector’s problem mass function. Different  $r^*$ s are tested via any of several straightforward convergence algorithms (false position converges more quickly than bisection and, surprisingly, Newton-Raphson in this context) to find those  $r^*$ s yielding distances arbitrarily close to the original conservative distance estimates, typically within just several iterations. The method performs well in practice because of the shape of the runtime function (Graph 7): as long as the original distance estimates are conservative, i.e. slightly larger than necessary, the corresponding estimates of the optimal “low-end”  $r^*$ s also will be slightly larger than necessary, causing only negligible runtime increases over use of the true optimal “low-end”

$r^*$ s.

TABLE 1.  
Nearly Optimal  $r$  (“low-end”  $r^*$ ),  
Probability ( $p$ ) of  $T \geq 1,901$  Unique Samples,  
and Expected # of Calls to PROC PLAN ( $1/p$ )  
by Ranges of # of Sample Combinations,  $C$

$C =_n C_{n_1}$	“low-end” $r^*$	$p$ (lower bound)	$1/p$ (lower bound)
$C < 10,626$	$C$	1.0 (assuming $C \geq T$ )	1.0
$10,626 \leq C < 52,360$	2,138	0.9979293 20330667	1.00207497 6280530
$52,360 \leq C < 101,270$	1,956	0.9990583 42955471	1.00094254 4598290
$101,270 \leq C < 521,855$	1,934	0.9994297 17692296	1.00057060 7715190
$521,855 \leq C < 1,028,790$	1,912	0.9997265 55240808	1.00027351 9551680
$1,028,790 \leq C < 10,009,125$	1,908	0.9995128 39120371	1.00048739 8321020
$10,009,125 \leq C < 25,637,001$	1,904	0.9999615 94180711	1.00003840 7294350
$25,637,001 \leq C < 100,290,905$	1,903	0.9999446 15376581	1.00005538 7691050
$100,290,905 \leq C < 5,031,771,045$	1,902	0.9998396 91379204	1.00016033 4323770
$5,031,771,045 \leq C$	1,901	0.9996411 54940541	1.00035897 3875460

It is worth noting that, for  $T = 1,901$ , the largest value of  $C$  for which one has to actually “oversample” (although one must still check for duplicate samples and redraw if necessary) is relatively small – about  $5 \times 10^9$ . This corresponds to sample sizes of only  $n_1 = 17$  and  $n_2 = 18$  for small  $n = n_1 + n_2$ , and  $n_1 = 2$  and  $n_2 = 100,000$  for large  $n$ . This is due, of course, to the fantastic combinatorial growth of  $C$ , which causes  $1/p$ ’s rapid convergence to one. This convergence

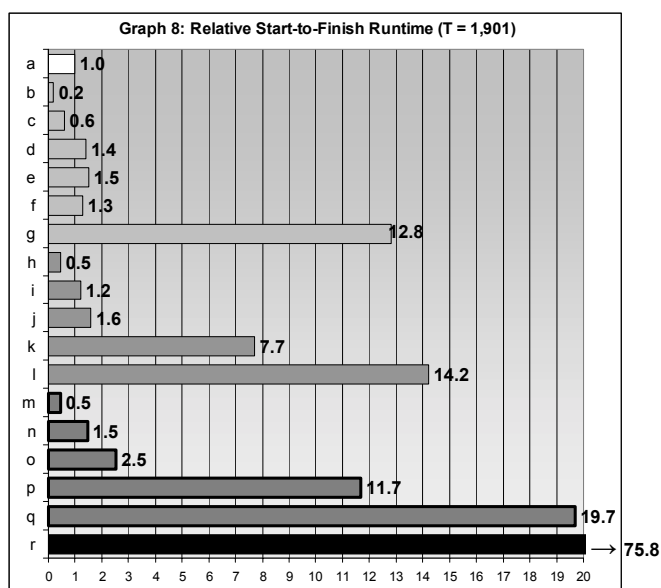
indicates that using “oversampling” as outlined above to perform NR sampling should be applicable to any WR sampling procedure, even if its runtime function, unlike (13), is not linear in  $n$  (i.e. even if it is convex and steep in  $n$ ).

Results

How Fast Is It?

Relative Speed – Some Benchmarks

The start-to-finish runtime of the permutation test program using “oversampling” with PROC PLAN to perform NR sampling is fast relative to other programs and WR procedures, as shown below:



- a = PROC PLAN with “oversampling”
- b = TWOSAMPL,  $(n_1+n_2) < 10,000$ , R = 1
- c = TWOSAMPL,  $(n_1+n_2) < 10,000$ , R > 1
- d = TWOSAMPL,  $10,000 < (n_1+n_2) < 100,000$ , R = 1
- e = TWOSAMPL,  $100,000 < (n_1+n_2) < 150,000$ , R = 1
- f = TWOSAMPL,  $1M < (n_1+n_2) < 1.5M$ , R = 1
- g = TWOSAMPL,  $1M < (n_1+n_2) < 1.5M$ , R > 1
- h = NPARIWAY,  $(n_1+n_2) < 10,000$ , R = 1
- i = NPARIWAY,  $(n_1+n_2) < 10,000$ , R > 1
- j = NPARIWAY,  $10,000 < (n_1+n_2) < 100,000$ , R = 1
- k = NPARIWAY,  $100,000 < (n_1+n_2) < 150,000$ , R = 1
- l = NPARIWAY,  $1M < (n_1+n_2) < 1.5M$ , R = 1
- m = MULTTEST,  $(n_1+n_2) < 10,000$ , R = 1
- n = MULTTEST,  $(n_1+n_2) < 10,000$ , R > 1

- o = MULTTEST,  $10,000 < (n_1+n_2) < 100,000$ , R = 1
- p = MULTTEST,  $100,000 < (n_1+n_2) < 150,000$ , R = 1
- q = MULTTEST,  $1M < (n_1+n_2) < 1.5M$ , R = 1
- r = looping in SAS,  $1 < (n_1+n_2) < 1.5M$ , R > 1

where  $R = \# \text{ Study Groups} / \# \text{ Control Groups}$

(For r above, see Jackson (1998). Beware, however, that this code enters an infinite loop if the number of possible sample combinations for a given sample pair is less than T. Also note that the code, unlike the standard definition of a permutation test which includes “ties” in the numerator of the p-value, splits ties at the boundary after assuming exactly one tie at the boundary (apparently with the intent of making the test less statistically conservative)). The only procedures or programs faster than PROC PLAN with “oversampling” are PROC MULTTEST and PROC NPARIWAY with small samples and one study group per control group, as well as PROC TWOSAMPL with small samples, regardless of the study-control group ratio. For larger samples, the relative speed of PROC PLAN with “oversampling” over MULTTEST and NPARIWAY increases rapidly and nonlinearly, even with a study-control group ratio of one. The relative speeds for large samples and larger study-control group ratios (not shown in Graph 8) are many times larger still (note that MULTTEST runtimes reflect only the time required for sample generation, not p-value calculation, which would increase relative runtime by an additional several multiples for larger samples). The relative speed advantage over TWOSAMPL is only pronounced when one sample is large and the study-control group ratio exceeds one.

On the one hand, smaller samples are where one is most likely to need permutation tests. However, this is where the speed differential matters the least in absolute terms – even when performing two hundred permutation tests with these smaller sample sizes and a study/control group ratio equal to one, none of the other three procedures was ever more than five minutes faster than PROC PLAN with “oversampling.” So the tradeoff in this case is several minutes per run with MULTTEST, NPARIWAY, or TWOSAMPL, versus maximum power with PROC PLAN with “oversampling.”

In contrast, when samples are larger, relative runtimes matter most because even small differences become large in absolute terms. These are precisely the conditions under which PROC PLAN with “oversampling” maintains a very large relative speed advantage over MULTTEST and NPARIWAY, as well as TWOSAMPL when the study-control group ratio exceeds one.

In addition to the speed of PROC PLAN itself, a number of factors contribute to the speed of the entire SAS program used to perform permutation tests with PROC PLAN and “oversampling,” including:

- Use of PROC APPEND to “SET” two large datasets together (one on top of the other) whenever possible.
- Judicious use of multiple PROC TRANSPOSE’s to evaluate the summarized results of the permutation sampling.
- Most test statistics can be constructed based on just one of the two samples in a pair and, if necessary, the pooled summary statistics of the pair. Thus, when conducting permutation sampling, sample only the smaller of the two samples, but keep track of which sample is used (study or control) when constructing the test statistics based on the permutation samples.
- To quickly SET together the potentially large and numerous output dataset lists from PROC PLAN (one set of T samples for every permutation test), use a looping macro that returns all the dataset names into a single SET statement (see code in Appendix C). Alternately, looping on the SET statement and SETting the datasets together cumulatively, one at a time, is extremely inefficient and runtime costly.
- If the dataset is large and contains a large percentage of records with the same response variable value (say, zero), delete these records to avoid sorting and later merging them with the PROC PLAN output. After merging the remaining data with the PROC PLAN output

and retaining all PROC PLAN records in the merge, reassign this value to the response variable when it is missing (i.e. when that record did not merge with the PROC PLAN output because it had been deleted).

- Most importantly, if the data contains multiple study groups per control group, there is no need to output control group records multiple times, once for each corresponding study group, when using PROC PLAN with “oversampling.” The original data simply can be divided into two datasets – one for control group(s) and one for study groups – and each merged separately to the PROC PLAN output (then (PROC) APPENDED together after the merges). Unless one constructs a separate dataset for each permutation test, PROC MULTTEST, PROC NPARIWAY, and PROC TWOSAMPL require control group records duplicated in the input dataset for each study group against which they are being compared. This is what gives PROC PLAN with “oversampling” an additional speed premium in these situations, and similarly, for multiple comparisons. To test a complete null hypothesis under a multiple testing framework, the number of pairwise comparisons required is  $s(s-1)/2$ , where  $s$  is the number of samples. This means that for the other three procedures, a much larger number of observations (16) must be output and sorted compared to the number similarly processed by PROC PLAN with “oversampling” (17).

$$3 \text{ PROCs \#obs} = (s-1) \sum_{i=1}^s n^{(i)} \quad (16)$$

$$\text{PROC PLAN \#obs} = n^{(s)} (s-1) \sum_{i=2}^s n^{(i-1)} (i-1) \quad (17)$$

where

$s$  = the number of samples, and  $n^{(i)}$  = the number of observations in the sample with the  $i$ th largest number of observations

If many permutation tests must be conducted and at least some of these contain large

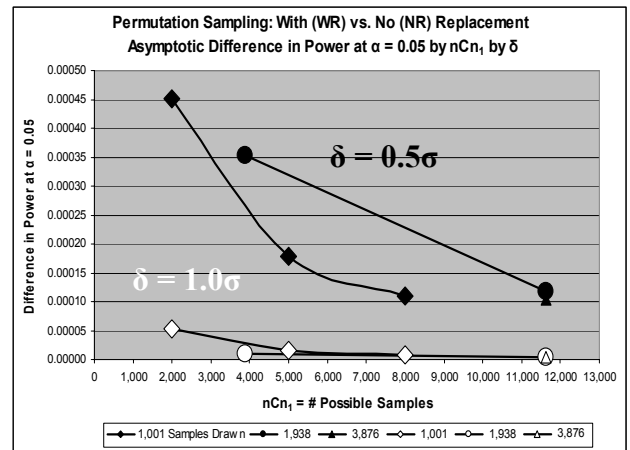
samples, the runtime advantage of (17) over (16) can be extremely large, as seen in Graph 8. However, (17) does not assume the code exploits the fact that with multiple comparisons, the same groups of observations are being used repeatedly in different comparisons. Although the other sampling procedures examined in this study cannot take advantage of this, code based on PROC PLAN can, allowing the researcher to achieve computational efficiencies even beyond those gained by (17) over (16).

*Absolute Speed*

When run on data containing 220 sample pairs where the smaller sample was less than 30 observations but the larger sample was sometimes as large as 64,000 observations, the runtime of the program was 7 minutes, 45 seconds on a desktop PC with two gigabytes of random-access memory and a two gigahertz Pentium® processor. For data containing 6,682 sample pairs where the smaller sample was less than 30 observations but the larger was sometimes over 5,000,000 observations, the runtime was 8 hours, 36 minutes. The former example obviously is more typical of the contexts in which permutation tests are used, but the latter is instructive for demonstrating the limits of the methods and software being relied upon. This study shows that the runtime of PROC PLAN with “oversampling” is not prohibitive even when applied to sample sizes as large (if not far larger) than would ever be used with permutation tests. The same cannot be said for the four alternate methods. (One notable and widespread example of the current application of permutation tests to sample pairs where one sample can be quite large is the telecommunications regulatory arena. Incumbent local exchange carriers have been required by a number of state public service commissions to perform permutation tests on performance measurement data if one sample (typically the CLEC sample) is small, even if the other (typically the ILEC sample) contains many millions of observations.)

NR Sampling – How Much Power Gain?

The asymptotic approximation of the power differential between NR sampling and WR sampling for a single pairwise comparison is calculated below (Table 2 and Graph 9) based on the Central Limit Theorem ((6.2 – (6.3)). There are two notable findings: first, the power gains from using NR sampling over WR sampling are small, even for small values of  $\delta$  (the location difference) and  $nC_{n_1}$ , and even taking into consideration that these asymptotic power differences represent lower bounds for the actual power differences. Secondly, these gains decrease rapidly in  $nC_{n_1}$ . Why is this the case? Recall that the only difference between NR sampling and WR sampling is the variance of the estimated p-value; the former is based on the hypergeometric distribution (5.6) and the latter is based on the binomial distribution (5.7).



$$\sigma_{bin}^2 = n_p p(1 - p) \tag{5.6}$$

$$\sigma_{hyp}^2 = n_p p(1 - p) \frac{(nC_{n_1} - n_p)(nC_{n_1} - 1)}{nC_{n_1}} \tag{5.7}$$

These variances differ only by the finite population correction factor (fpc) of  $(nC_{n_1} - n_p)/(nC_{n_1} - 1)$ . As  $n_1$  and  $n_2$  increase,  $nC_{n_1}$  increases dramatically, causing the rapid convergence of the fpc to one and thus, the practical equivalence of NR sampling and WR sampling. Intuitively, this makes sense as it is clear that the probability of drawing any of a few thousand samples ( $n_p$ ) more than once quickly approaches zero as the number of possible samples

from which to randomly draw rapidly surpasses trillions and quadrillions of possibilities (the exact probability is given by one minus  $(12)^T$  when  $T=r$ ). Therefore, if sample sizes are not very small, it is fair to say that such small power gains would only make NR sampling worth considering if there was little or no runtime cost associated with its implementation. Otherwise, unless the cost of Type II error is astronomically high, NR sampling may not be worth the trouble (however, note NR sampling's more obvious benefit of shorter confidence intervals on the permutation p-values themselves compared to (3), which is based on WR sampling).

#### NR Sampling – Power Gains at What Cost?

A good metric for evaluating the runtime cost of employing NR sampling via “oversampling” is its start-to-finish runtime compared to that associated with WR sampling – i.e. just drawing  $T$  samples and ignoring the duplicate sample problem. This difference is a function of the number of tests performed and their sample sizes. When only two hundred permutation tests were conducted on small sample pairs (both less than 30 observations), NR sampling was 20%-30% slower than WR sampling. However, in absolute terms, this was less than two minutes. When 1,862 tests were conducted, including some sample pairs with one large sample, the runtime cost was always under 2%; for all 6,682 tests, the runtime cost was always well below 1%. Maximizing power via NR sampling arguably is worth this relatively small increase in runtime.

#### Conclusion

This study provides a) statistical code that performs fast continuous-data permutation tests even if one sample is large, and which often is more than an order of magnitude faster than widely available commercial alternatives under these conditions, and b) an answer to the question: does drawing a set of permutation samples containing no duplicate samples increase the power of the permutation test for a single pairwise comparison? If so, by how much, and are there also power gains for multiple comparisons? It is analytically shown that “no-replacement” (NR) sampling of the permutation sample space provides a small power gain over the usual method of “with-replacement” (WR) sampling when using a conventional Monte Carlo approach (this power gain attains, by definition, maximum power under conventional Monte Carlo). This finding holds for pairwise comparisons, as well as for multiple comparisons – specifically, permutation-style p-value adjustments of permutation test p-values – which are made runtime feasible by an additional speed premium built into the code. The power gain for such multiple comparisons, however, may be larger in absolute terms because these procedures achieve variance reduction from two sources rather than just one. Simulating these gains is the focus of ongoing research. The power gains of both pairwise and multiple comparisons, however, quickly diminish as sample sizes increase. This is due to the rapid convergence of the conditional variance of the estimated permutation p-value (based on the hypergeometric distribution) to that of WR sampling (based on the binomial distribution). However, the runtime cost of implementing NR sampling via the proposed method of “oversampling” is negligible – less than 1% of runtime when many tests are conducted and at least some of the sample pairs contain one large sample (which is when runtime matters most in absolute terms). So under a conventional Monte Carlo approach, if the cost of Type II error is not negligible and even if the power gains of NR sampling may be small, there seems to be no reason not to use this straightforward and readily applied method in order to maximize power.

TABLE 2. Asymptotic Approximation of Power Difference Between NR Sampling vs. WR Sampling for a Pairwise Permutation Test

$n_p$	1,001	1,001	1,001	1,938	1,938	3,876	
${}_n C_{n_1}$	2,002	5,005	8,008	3,876	11,628	11,628	
$({}_n C_{n_1} - n_p) / ({}_n C_{n_1} - 1)$	0.5002	0.8002	0.8751	0.5001	0.8334	0.6667	
$asl_{NR}$	0.0511734	0.0513080	0.0513416	0.0501677	0.0502451	0.0501290	
$asl_{WR}$	0.0513977	0.0513977	0.0513977	0.0502838	0.0502838	0.0501677	
$C_{\alpha_{NR}}$	0.0493837	0.0493128	0.0492950	0.0498490	0.0497793	0.0498968	
$C_{\alpha_{WR}}$	0.0492655	0.0492655	0.0492655	0.0497445	0.0497445	0.0498658	
Power <sub>NR</sub>	$\bar{\delta} = 0.5$	0.5870526	0.6121541	0.6361821	0.7030284	0.7027937	0.7031889
	$\bar{\delta} = 1.0$	0.9817270	0.9868391	0.9905697	0.9966619	0.9966550	0.9966665
Power <sub>WR</sub>	$\bar{\delta} = 0.5$	0.5866014	0.6119764	0.6360733	0.7026762	0.7026762	0.7030848
	$\bar{\delta} = 1.0$	0.9816750	0.9868234	0.9905623	0.9966516	0.9966516	0.9966635
Power difference	$\bar{\delta} = 0.5$	0.0004512	0.0001777	0.0001089	0.0003522	0.0001175	0.0001041
	$\bar{\delta} = 1.0$	0.0000520	0.0000157	0.0000073	0.0000103	0.0000034	0.0000030

References

Andersen, M.J. & P. Legendre (1999), An empirical comparison of permutation methods for tests of partial regression coefficients in a linear model, *Journal of Statistical Computation & Simulation*, Vol. 62, No. 3.

Berry, K. & P. Mielke (1983), Moment approximations as an alternative to the f test in analysis of variance, *British Journal of Mathematical & Statistical Psychology*, 36: pp.202-206.

Boos, D., & J. Zhang (June 2000), Monte carlo evaluation of resampling-based hypothesis tests, *Journal of the American Statistical Association*, Vol. 95, No. 450.

Brown, L., T. Cai, & A. DasGupta, Interval estimation for a binomial proportion, *Statistical Science*, Vol. 16, No. 2: pp.101-133.

Cochran, W. (1977), *Sampling techniques*, 2<sup>nd</sup> ed., New York: John Wiley & Sons.

Evans, M., N. Hastings, & B. Peacock (1993), *Statistical distributions*, 2<sup>nd</sup> ed., New York: John Wiley & Sons.

Fisher, Sir R.A. (1935), *Design of experiments*, Edinburgh, Oliver & Boyd.

Efron, Bradley & Robert Tibshirani (1993), *An introduction to the bootstrap*, Chapman & Hall, London & New York.

Affidavit of John Jackson, On Behalf of MCI-Worldcom, Before the Michigan Public Service Commission, Case No. U-11830, November 18, 1998, ATTACHMENT A, "Using Permutation Tests to Evaluate the Significance of CLEC vs. ILEC Service Quality Differentials"

Kuonen, D. (August 2000), A saddlepoint approximation for the collector's problem, *The American Statistician*, Vol. 54, No. 3.

Lindsay, J.D. (1992), A new solution for the probability of completing sets in random sampling: definition of the 'two-dimensional factorial', *The Mathematical Scientist*, 17: 101-110.

Mehta, C., & N. Patel (June 1983), A network algorithm for performing fisher's exact test in  $r \times c$  contingency tables, *Journal of the American Statistical Association*, Vol. 78, No. 382.

Mielke, P. & K. Berry (2001), *Permutation methods: a distance function approach*, Springer-Verlag, New York.

Naiman, D. & C. Priebe (2001), Computing scan statistic p values using importance sampling, with applications to genetics and medical image analysis, *Journal of Computational and Graphical Statistics*, Vol. 10, No. 2.

Opdyke, J.D. (May 5-8, 2002a), PharmaSUG2002: Conference of the Pharmaceutical SAS Users' Group, Salt Lake City, UT, [http://www.pharmasug.org/psug2002/bp2002/psug2002\\_html](http://www.pharmasug.org/psug2002/bp2002/psug2002_html)

Opdyke, J.D. (August 5-7, 2002b), MCP 2002: The 3<sup>rd</sup> International Conference on Multiple Comparisons, Bethesda, Maryland, <http://www.ba.ttu.edu/isqs/westfall/Program.htm>

Ortiz, L. & L. Kaelbling (2000), Sampling methods for action selection in influence diagrams, Proceedings of the Seventeenth National Conference on Artificial Intelligence.

Owen, A. & Y. Zhou (March 2000), Safe & effective importance sampling, *Journal of the American Statistical Association*, Vol. 95, No. 449.

Pesarin, F. (2001), *Multivariate permutation tests with applications in biostatistics*, John Wiley & Sons, Ltd., New York.

Read, K.L.Q. (May 1998), A lognormal approximation for the collector's problem, *The American Statistician*, Vol. 52, No. 2.

Westfall, P., & S. Young (1993), *Resampling-based multiple testing – examples & methods for p-value adjustment*, New York, John Wiley & Sons, Inc.

Appendix A

To estimate PROC PLAN real runtime, SAS® v.8.2 was used on a desktop PC with 2GB RAM and a 2GHz Pentium processor. Sample sizes were generated by assigning values of 3, 16, and 27 to the smaller of the two samples, and, beginning at 100, assigning values by 100 increments to the larger sample up to 100,000, after which point increments of 10,000 were used up to 1.5 million (though the program has been run on sample pairs as large as 29 and 5,000,029). Three values of r were used: 1,901, 2,700, and 3,500.

Appendix B

PROC PLAN RunTime, PPR( $n_1, n_2, r$ ), regression results:  
Left hand side variable = real runtime seconds  
adjusted  $R^2 = 0.9927$

Variable Key	Variable
A	Intercept
B	$(n_1 + n_2)$
C	r
D	$(n_1 + n_2) * r$
E	$[(n_1 + n_2) < 65.5K]$
F	$[(n_1 + n_2) < 65.5K] * (n_1 + n_2)$
G	$[(n_1 + n_2) < 65.5K] * r$
H	$[(n_1 + n_2) < 65.5K] * (n_1 + n_2) * r$
I	$[65.5K \leq (n_1 + n_2) \leq 73.5K]$
J	$[65.5K \leq (n_1 + n_2) \leq 73.5K] * (n_1 + n_2)$
K	$[65.5K \leq (n_1 + n_2) \leq 73.5K] * r$
L	$[65.5K \leq (n_1 + n_2) \leq 73.5K] * (n_1 + n_2) * r$

Variable Key	Parameter Estimate	t value
A	0.0432387277000000	1.80
B	-0.0000001298032000	-2.88
C	0.0000838185000000	9.68
D	0.0000000038095955	234.72
E	-0.0340413560000000	-0.89
F	0.0000004543242500	0.58
G	-0.0000581740000000	-4.24
H	-0.0000000024994500	-8.86
I	-0.4873557050000000	-0.38
J	0.0000071862352000	0.39
K	-0.0016941670000000	-3.70
L	0.0000000228154240	3.47



## Appendix C

```

options = nomprint nomlogic nomrecall;

%MACRO RUN_PRG;

*** the By Variables and npermsampT normally
would be passed in the main macro (RUN_PRG).;

%let byvars=byvar1 byvar2 byvar3 byvar4
byvar5;

*** npermsampT = # of permutation samples;
%let npermsampT=1901;

*** count the number of byvars for parsing;
%let byvars=%cmpres(&byvars);
%let num_byvars=
  %eval(%length(&byvars)-
    %length(%cmpres(&byvars))+1);

*** summarized data (SUMDINPT) contains study
group identifier (study), control group
identifier (cntl), # study group obs, #
control group obs, and any By Variables.;

%let noconverge=0;
data sumdinpt(keep=combins nsamp minrcomb
  minof3 bigcomb ncalls2pp
  topdraws lastdraw smaller
  nobsmallr studynobs contrlnobs
  sumofnobs study cntl &byvars);
set sumdinpt;

*** create variables to be passed to CREATSMP,
which generates the permutation samples
corresponding to each record on SUMDINPT;

if "&npermsampT"="1901" then
  maxcombins=5031771045;
else maxcombins=9*10**16;

*** for versions of SAS v6.12 and older,
comb(,) terminates for results of
approximately 10E70 and higher, so use the
loop below instead;

if ("%sysver"*1)<8 then do;
  combins=1;
  minnobs=min(studynobs,contrlnobs);
  bothnobs=sum(studynobs,contrlnobs);
  do j=minnobs to 1 by -1;
    combins=combins*(bothnobs-j+1)/j;
    if combins>maxcombins then goto enufcomb;
  end;
  enufcomb: combins=round(combins);
end;
else do;
  combins=comb(sum(studynobs,contrlnobs),
    min(studynobs,contrlnobs));
*** if still too large, assign large number;
  if combins=. then combins=maxcombins;
end;

```

\*\*\* The 'table' below was calculated based on the exact probabilities of the Collectors Problem distribution and presents the optimal "low-end" sample sizes by ranges of nCn1 (p.7 above) only for npermsampT = 1901.;

```

IF "&npermsampT" = "1901" THEN DO;
  if combins<&npermsampT then
    nsamp=&npermsampT;
  else if combins<10626 then nsamp=combins;
  else if combins<52360 then nsamp=2138;
  else if combins<101270 then nsamp=1956;
  else if combins<521855 then nsamp=1934;
  else if combins<1028790 then nsamp=1912;
  else if combins<10009125 then nsamp=1908;
  else if combins<25637001 then nsamp=1904;
  else if combins<100290905 then nsamp=1903;
  else if combins<5031771045 then nsamp=1902;
  else if combins>=5031771045 then nsamp=1901;
END;

```

\*\*\* For npermsampT other than 1901, obtain nsamp with a convergence routine based on the first and second moments of the Collectors Problem distribution and using the nsamp calculated above as a basis for the starting values. Even for large npermsampT (e.g. 32,000) and conservatively defined Xstdev, convergence (based on false position) typically is achieved in less than five iterations;

```
ELSE DO;
```

\*\*\* Define X\*stdev (Xstdev) here conservatively, based on the size of npermsampT compared to 1901 (the base would be Xstdev = 2.875 since this is (approximately) true when npermsampT = 1901). Larger npermsampT allows for the use of smaller Xstdev, but smaller npermsampT requires larger Xstdev to maintain the same (approximate) probability of a redraw. Any functional relationship between Xstdev and npermsampT similar to the one below can be used (the exponent below (0.25) was chosen based on a wide range of values for npermsampT).;

```
Xstdev= (1901/&npermsampT)**0.25;
```

```

if combins<&npermsampT
  then startratio=-999;
else if combins<(&npermsampT*10626/1901)
  * Xstdev then startratio=-888;
else if combins<(&npermsampT*52360/1901)
  * Xstdev then startratio=2138/1901;
else if combins<(&npermsampT*101270/1901)
  * Xstdev then startratio=1956/1901;
else if combins<(&npermsampT*521855/1901)
  * Xstdev then startratio=1934/1901;
else if combins<(&npermsampT*1028790/1901)
  * Xstdev then startratio=1912/1901;
else if combins<&npermsampT*10009125/1901
  * Xstdev then startratio=1908/1901;
else if combins<&npermsampT*25637001/1901
  * Xstdev then startratio=1904/1901;

```

```

else if combins<&npermsampT*100290905/1901
  * Xstdev then startratio=1903/1901;
else if combins<&npermsampT*5031771045/1901
  * Xstdev then startratio=1902/1901;
else if combins>=&npermsampT*5031771045/1901
  * Xstdev then startratio=1.0;

IF startratio=-999 | startratio=1
  THEN nsamp=&npermsampT;
ELSE IF startratio=-888
  THEN nsamp=combins;
ELSE IF startratio>1 THEN DO;

*** Starting value for nsamp.;
  nsamp=ceil(startratio*&nresamp);
  nsampoldhigh=nsamp;
  nsampoldlow=(&nresamp*1);
  initgap=nsampoldhigh-nsampoldlow;

  colldist_avg = combins*(1-
    (1-1/combins)**nsampoldlow);

*** Numeric precision constraints prevent
calculation of the second moment for large
inputs, but a conservative (i.e. larger-than-
actual) approximation suffices in these
cases.;
  if (combins*(combins-1)*
    (1- 2/combins)**nsampoldlow) >
    100144465758007
  then colldist_stdev = 0.4;
  else
  colldist_stdev =
    sqrt(combins*(combins-1)*
      (1-2/combins)**nsampoldlow+
      combins*(1-1/combins)**nsampoldlow-
      combins**2*(1-1/combins)**
      (2*nsampoldlow));

  lowpoint =(colldist_avg - Xstdev *
    colldist_stdev - &nresamp*1);

  colldist_avg = combins*(1-
    (1-1/combins)**nsampoldhigh);

  if (combins*(combins-1)*
    (1- 2/combins)**nsampoldhigh) >
    100144465758007
  then colldist_stdev = 0.4;
  else
  colldist_stdev =
    sqrt(combins*(combins-1)*
      (1-2/combins)**nsampoldhigh+
      combins*(1-1/combins)**nsampoldhigh-
      combins**2*(1-1/combins)**
      (2*nsampoldhigh));

  highpoint = (colldist_avg - Xstdev *
    colldist_stdev-&nresamp*1);
  point=highpoint;

*** Use counter only to eliminate the
possibility of infinite loop.;

  DO z=1 to 1000;

*** Obtain nsamp only to within 4 of optimal
nsamp (when converging on nsamp from upper
bound) to prevent unnecessary looping.;

  TOPLOOPNSAMP:
  if point>4 then do;
    nsampoldhigh=nsamp;
    nsamp=ceil((nsampoldlow * highpoint -
      nsampoldhigh * lowpoint)
      /
      (highpoint-lowpoint));
  end;

*** If necessary, get upper bound above zero
on 1st loop (& increment lower bound
concurrently);

  else if z=1 & point<-1 then
  do y=1 to 1000;
    nsampoldlow = nsamp;
    nsamp = ceil(nsamp+initgap);
    colldist_avg = combins*(1-
      (1-1/combins)**nsamp);
    if (combins*(combins-1)*
      (1- 2/combins)**nsamp) >
      100144465758007
    then colldist_stdev = 0.4;
    else
    colldist_stdev =
      sqrt(combins*(combins-1)*
        (1-2/combins)**nsamp+
        combins*(1-1/combins)**nsamp-
        combins**2*(1-1/combins)**
        (2*nsamp));
    highpoint = (colldist_avg - Xstdev *
      colldist_stdev -
      &nresamp*1);
    point = highpoint;

    if point>4 then do;
      colldist_avg = combins*(1-
        (1-1/combins)**nsampoldlow);
      if (combins*(combins-1)*
        (1- 2/combins)**nsampoldlow) >
        100144465758007
      then colldist_stdev = 0.4;
      else
      colldist_stdev =
        sqrt(combins*(combins-1)*
          (1-2/combins)**nsampoldlow+
          combins*
          (1-1/combins)**nsampoldlow-
          combins**2*(1-1/combins)**
          (2*nsampoldlow));

      lowpoint = (colldist_avg -
        Xstdev*colldist_stdev -
        &nresamp*1);

      goto TOPLOOPNSAMP;
    end;

    else if -1<=point<=4
      then goto STOPCNVG;
  end;

*** Require a stricter convergence criterion
on optimal nsamp when converging from lower
bound;

```

```

else if point<-1 then do;
  nsampoldlow=nsamp;
  nsamp=ceil((nsampoldlow*highpoint -
             nsampoldhigh*lowpoint)
            /
            (highpoint-lowpoint));
end;

else if -1<=point<=4 then goto
  STOPCNVG;

if z = 1000 then do;
  noconverge = 1;
  goto STOPCNVG;
end;

*** For next iteration;
temp_avg = combins*
           (1-(1-1/combins)**nsamp);
if (combins*(combins-1)*
   (1- 2/combins)**nsamp) >
   100144465758007
then temp_stdev = 0.4;
else
temp_stdev = sqrt(combins*(combins-1)*
                 (1-2/combins)**nsamp
                 + combins*(1-
                 1/combins)**nsamp -
                 combins**2*
                 (1-1/combins)**(2*nsamp));

temp_point = (temp_avg - Xstdev *
             temp_stdev - &nresamp*1);
if temp_point >= 0 then do;
  highpoint = temp_point;
  point = highpoint;
end;
else do;
  lowpoint = temp_point;
  point = lowpoint;
end;
END;

STOPCNVG:
if noconverge = 1 then do;
  call symput('noconverge',
             compress(noconverge));
  stop;
end;
END;
END;

minrcomb=min(combins,nsamp);

minof3=min(combins,nsamp,&npermsampT);

if combins=minrcomb then bigcomb=0;
else if combins>minrcomb then bigcomb=1;

ncalls2pp=ceil(minrcomb*
              sum(studynobs,contrlnoobs)/2**31);
topdraws=floor(nsamp/ncalls2pp);
lastdraw=topdraws+mod(nsamp,ncalls2pp);

if studynobs<=contrlnoobs then
  smaller="stdy";
else smaller="cntl";

nobsmalr=min(studynobs,contrlnoobs);
sumofnobs=sum(studynobs,contrlnoobs);

run;

*** Although algorithm should always converge,
code should account for any contingency.;
%if &noconverge=1 %then %do;
  %put;
  %put WARNING: The permutation sample-size
algorithm did not converge.;
  %put Scrutinize the data and/or adjust the
functional relationship between Xstdev and
npermsampT.;
  %put;
  %goto EXITALL;
%end;

*** define outside of CREATSMP (which is
called in a loop) four macros used for
assigning By Variables and their values
(exactly as they exist on both the original
data (FULLDATA) and SUMDINPT) to the sampling
datasets generated by PROC PLAN in CREATSMP;

%MACRO GETVARLEN(varname=);
  %let dsetid=%sysfunc(open(fulldata));
  %let len=%sysfunc(varlen(&dsetid,
  %sysfunc(varnum(&dsetid,&varname))));
  %let dsetid=%sysfunc(close(&dsetid));
  &len
%MEND GETVARLEN;

%MACRO ASSIGNBYVRLENS;
  %do p=1 %to &num_byvars;
    &&byvar&p $%GETVARLEN(varname=&&byvar&p)
  %end;
%MEND ASSIGNBYVRLENS;

%MACRO ASSIGNBYVRVALS;
  %do q=1 %to &num_byvars;
    %let x=%scan(&byvars,&q,' ');
    %str(&x=resolve("&"||"&x"));
  %end;
%MEND ASSIGNBYVRVALS;

%MACRO GETBYVARVALUES;
  %do q=1 %to &num_byvars;
    %let x=%scan(&byvars,&q,' ');
    %str(byvarval=resolve("&"||"&x")); output;
  %end;
%MEND GETBYVARVALUES;

*** When multiple loops on PROC PLAN
required...;
*** ...use for combining datasets.;
%MACRO COMBBIGSAMPS;
  %do s=2 %to &ncalls2pp;
    ptemp&s.(in=in&s)
  %end;
%MEND COMBBIGSAMPS;

*** ..use for assigning DRAWNUM values.;
%MACRO ASSIGNDRAWNUM;

```

```

    %if &ncalls2pp>2 %then
        %do k=3 %to &ncalls2pp;
            %str(else if in&k then drawnum =
                drawnum+(&k-1)*&topdraws;
            %end;
%MEND ASSIGNDRAWNUM;

*** Obtains # of records in a dataset.;
%MACRO NOBS(dset);
    %if %sysfunc(exist(&dset)) %then %do;
        %let dsid=%sysfunc(open(&dset));
        %let nobss=%sysfunc(attrn(&dsid,nobs));
        %let dsid=%sysfunc(close(&dsid));
    %end;
    %else %let nobss=0;
    &nobs
%MEND NOBS;

%let seednum =-1;

%MACRO CREATSMP(recoutr = );

*** The automatic random seed for PROC PLAN,
based on the time of day, does not update as
fast as PROC PLAN is repeatedly called in the
loops below. Hence, ranuni() is used to
generate the seed, & its value is explicitly
checked to ensure the current random number is
different from the previous one. This ensures
random sampling is unrelated across tests.;

*** if combins <= r, choose all sample
combinations, then select npermsampT samples
from them.;

%if &bigcomb=0 %then %do;

    data _null_;
        x=1000000000*ranuni(-1);
        if compress(&seednum)=compress(" "||x)
            then x=x+1;
        call symput('seednum',compress(x));
    run;

    %if &nobsmalr=1 %then %do;
        proc plan seed=&seednum;
            factors drawnum = 1
                dataobsid = &minof3 of &combins
                    random / noprint;
            output out = psamp&recoutr;
        run;
    %end;

    %if &nobsmalr>1 %then %do;

*** cannot just select first npermsampT draws
because the comb option orders them, and the
data may be ordered in some way;

        proc plan seed=&seednum;
            factors drawnum = &combins
                dataobsid =&nobsmalr of &sumofnobs
                    comb / noprint;
            output out = psamp&recoutr;
        run;

    %if &combins>&npermsampT %then %do;
        data _null_;
            x=1000000000*ranuni(-1);
            if compress(&seednum)=
                compress(" "||x) then x=x+1;
            call symput('seednum',compress(x));
        run;

        proc plan seed=&seednum;
            factors drawnum = 1
                dataobsid=&npermsampT of &combins
                    random / noprint;
            output out = choosmp;
        run;

        data choosmp(keep=drawnum);
            set choosmp(drop=drawnum);
            drawnum=dataobsid;
        run;

        proc sort data=choosmp;
            by drawnum;
        run;

        proc sort data=psamp&recoutr;
            by drawnum;
        run;

        data psamp&recoutr;
            merge psamp&recoutr
                choosmp(in=inchoos);
            by drawnum;
            if inchoos then output psamp&recoutr;
        run;

        data psamp&recoutr(drop=drawnum2);
            set psamp&recoutr(drop=drawnum);
            retain drawnum2 0;
            if mod(_n_,&nobsmalr)=1
                then drawnum2 = drawnum2+1;
            drawnum=drawnum2;
        run;
    %end;
%end;
%end;

*** if combins > r, check whether PROC PLAN
needs to be looped multiple times -- if not,
simply select r samples, delete duplicates,
and keep npermsampT samples. If so, loop it
first to select r samples. In either case,
redraw samples if fewer than npermsampT unique
samples are drawn the first time around.;

%if &bigcomb=1 %then %do;

    %redraw1:
        data _null_;
            x=1000000000*ranuni(-1);
            if compress(&seednum)=
                compress(" "||x) then x=x+1;
            call symput('seednum',compress(x));
        run;

    %if &ncalls2pp=1 %then %do;

```

```

proc plan seed=&seednum;
factors drawnum = &minrcomb
      dataobsid= &nobsmalr of &sumofnobs
              random / noprint;
output out      = psamp&recontr;
run;

proc sort data=psamp&recontr;
  by drawnum;
run;

proc transpose data=psamp&recontr
      out=temp prefix=stdy;
  var dataobsid;
  by drawnum;
run;

proc sort data=temp out=temp nodupkey;
  by stdyl-stdy&nobsmalr;
run;

%let ndrawn=%nobs(temp);
%if &ndrawn < &npermsampT %then %do;
  %put;
  %put Fewer than &npermsampT unique
permutation samples (only &ndrawn) were drawn
in a &sumofnobs-choose-&nobsmalr draw;
  %put for the study - control group pair
and "by variable" values listed below;;
  %put
=====;
  %put Study Control &byvars;

  data holdvals;
    %GETBYVARVALUES
  run;

  proc sql noprint;
    select byvarval into
      :byvarvals separated by ' '
    from holdvals;
  quit;

  proc datasets library=work nolist;
    delete holdvals temp;
  run;

  %put &stdy &cntl &byvarvals;
  %put;
  %put A redraw has been performed.;
  %put;
  %goto redraw1;
%end;

%else %do;
  proc datasets library=work nolist;
    delete temp;
  run;
  %if &ndrawn>&npermsampT %then %do;
    data psamp&recontr;
      set psamp&recontr
        (where=(drawnum<=&npermsampT));
  run;
  %end;
%end;

%end;
%end;
%end;

%redraw2:
%if &ncalls2pp>1 %then
  %do q=1 %to &ncalls2pp;

  %if &q<&ncalls2pp %then %do;
    data _null_;
      x=1000000000*ranuni(-1);
      if compress(&seednum)=compress(" "||x)
        then x=x+1;
      call symput('seednum',compress(x));
    run;

    proc plan seed=&seednum;
      factors drawnum = &topdraws
            dataobsid = &nobsmalr of
                    &sumofnobs random / noprint;
      output out      = ptemp&q;
    run;
  %end;

  %if &q=&ncalls2pp %then %do;
    data _null_;
      x=1000000000*ranuni(-1);
      if compress(&seednum)=
        compress(" "||x) then x=x+1;
      call symput('seednum',compress(x));
    run;

    proc plan seed=&seednum;
      factors drawnum = &lastdraw
            dataobsid = &nobsmalr of
                    &sumofnobs random / noprint;
      output out      = ptemp&q;
    run;

    data psamp&recontr;
      set ptempl %COMBBIGSAMPs;
      if in2 then drawnum=drawnum+&topdraws;
      %ASSIGNDRAWNUM
    run;

    proc sort data=psamp&recontr;
      by drawnum;
    run;

    proc transpose data=psamp&recontr
          out=temp prefix=stdyn;
      var dataobsid;
      by drawnum;
    run;

    proc sort data=temp out=temp nodupkey;
      by stdyn1-stdyn&nobsmalr;
    run;

    %let ndrawn=%nobs(temp);
    %if &ndrawn < &npermsampT %then %do;
      %put;
      %put Fewer than &npermsampT unique
permutation samples (only &ndrawn) were drawn
in a &sumofnobs-choose-&nobsmalr draw;
    %end;
  %end;
%end;

```

```

    %put for the study - control group
pair and "by variable" values listed below:;
    %put
=====;
    %put Study Control &byvars;

    data holdvals;
        %GETBYVARVALUES
    run;

    proc sql noprint;
        select byvarval into
            :byvarvals separated by ' '
        from holdvals;
    quit;

    proc datasets library=work nolist;
        delete holdvals temp;
    run;

    %put &stdy &cntl &byvarvals;
    %put;
    %put A redraw has been performed.;
    %put;
    %goto redraw2;
%end;

%else %do;
    proc datasets library=work nolist;
        delete temp;
    run;
    %if &ndrawn>&npermsampT %then %do;
        data psamp&recountr;
            set psamp&recountr
                (where=(drawnum<=&npermsampT));
    run;
    %end;
%end;
%end;
%end;
%end;

*** assign By Variable values on the sampling
datasets generated by PROC PLAN in CREATSMP.;

data psamp&recountr;
    length %ASSIGNBYVRLENS;
    set psamp&recountr;
    %ASSIGNBYVRVALS
run;

%MEND CREATSMP;

*** In a loop, generate permutation samples
for each record of SUMDINPT.;

%let sumdsid=%sysfunc(open(sumdinpt));
%let topofloop=%sysfunc(attrn(&sumdsid,nobs));
%syscall set(sumdsid);
%do i=1 %to &topofloop;
    %let fo=%sysfunc(fetchobs(&sumdsid,&i));
    %CREATSMP(recountr=&i);
%end;
%let sumdsid=%sysfunc(close(&sumdsid));

*** After looping above, combine PROC PLAN
output datasets to merge with the original
unsummarized dataset (FULLDATA) by By
Variables & record id variable (dataobsid).
Use the variable "smaller" when calculating
the test statistic for every permutation
sample.;

%MACRO COMBSAMPS;
    %do i=1 %to &totsamps; psamp&i %end;
%MEND COMBSAMPS;

data samples; set %COMBSAMPS; run;

proc datasets library=work nolist;
    delete %COMBSAMPS;
run;

%EXITALL;
%MEND RUN_PRG;

%RUN_PRG;

```

## Analyzing Group by Time Effects in Longitudinal Two-Group Randomized Trial Designs With Missing Data

James Algina                      H. J. Keselman                      A. R. Othman  
University of Florida      University of Manitoba      Universiti Sains Malaysia

We investigated bias, sampling variability, Type I error and power of nine approaches for testing the group by time interaction in a repeated measures design under three types of missing data mechanisms. One procedure due to Overall, Ahn, Shivakumar, and Kalburgi (1999) performed reasonably well over a range of conditions.

Key words: Missing data, random coefficients model, pattern mixture model

### Introduction

Consider a design in which  $N$  participants are randomly assigned to  $K = 2$  treatments. The researcher plans to observe each participant  $J$  times on the dependent variable, with the first observation prior to initiating a treatment and the remaining  $J - 1$  observations following initiation of a treatment.

This design has been referred to as a longitudinal two-group randomized trial design (Delucchi & Bostrom, 1999), randomized parallel-groups design (Overall, Ghasser, Shobaki & Fiore, 1996), or split-plot repeated measures design (Littell, Milligan, Stroup, & Wolfinger, 1996; Maxwell & Delaney, 1990). The effect of primary interest, typically, is whether there are differential rates of change over time, that is, whether there is a group by time interaction.

James Algina (algina@ufl.edu) is a Professor of Educational Psychology, University of Florida. His research interests are in applied statistics and psychometrics. H. J. Keselman (kesel@ms.umanitoba.ca) is a Professor of Psychology, University of Manitoba. His research interests are in applied statistics. A. R. Othman (oarahman@usm.my) is Lecturer of Mathematics, Universiti Sains Malaysia. His research interests are in applied statistics and psychometrics. Work on this project was supported by a grant from the Social Sciences and Humanities Research Council of Canada.

Let  $Y_{ijk}$  denote a random variable underlying the score, in treatment  $k$  ( $k = 1, 2$ ), for participant  $i$  ( $i = 1, \dots, n_k$ ), on occasion  $j$  ( $j = 1, \dots, J$ ). A possible model for the subject-specific regression of the dependent variable on time of measurement is

$$y_{ik} = \mathbf{X} \mathbf{b}_{ik} + \mathbf{e}_{ik}$$

where  $\mathbf{y}'_{ik} = (Y_{1k}, \dots, Y_{Jk})$ ,  $\mathbf{b}_{ik}$  is an unobservable  $r$ -dimensional random vector,  $\mathbf{e}_{ik}$  is a  $J$ -dimensional random vector,

$$\mathbf{X} = \begin{bmatrix} 1 & t_1 & t_1^2 & \cdots & t_1^{r-1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & t_J & t_J^2 & \cdots & t_J^{r-1} \end{bmatrix},$$

and  $t_1, \dots, t_J$  indexes time of measurement. We assume  $\mathbf{e}_{ik} \sim N(0, \mathbf{s}^2 \mathbf{I}_J)$ .

In this paper we focus on situations in which it is reasonable to assume that the subject-specific regressions are well described by a linear trend. Therefore

$$\mathbf{X} = \begin{bmatrix} 1 & t_1 \\ \vdots & \vdots \\ 1 & t_J \end{bmatrix}$$

and  $\mathbf{b}'_{ik} = (\mathbf{b}_{0ik} \ \mathbf{b}_{1ik})$ . The between-subjects model for  $\mathbf{b}_{ik}$  is

$$\begin{bmatrix} \mathbf{b}_{0ik} \\ \mathbf{b}_{1ik} \end{bmatrix} = \begin{bmatrix} 1 & z & 0 & 0 \\ 0 & 0 & 1 & z \end{bmatrix} \begin{bmatrix} \mathbf{g}_{00} \\ \mathbf{g}_{01} \\ \mathbf{g}_{10} \\ \mathbf{g}_{11} \end{bmatrix} + \begin{bmatrix} u_o \\ u_1 \end{bmatrix} \quad (1)$$

where  $z = 0$  for the first treatment and 1 for the second treatment. More compactly  $\mathbf{b}_{ik} = \mathbf{W}\mathbf{g} + \mathbf{u}$ . We assume that  $\mathbf{u} \sim N(\mathbf{0}, \mathbf{D})$ .

In many studies, participants may not be observed on all occasions. In general, the correct method of analysis depends on the missing data mechanism. Using an incorrect method can result in inconsistent estimates of the parameters. Little (1995) reviewed two different classes of methods for use in longitudinal designs. The design considered in this paper is a special case of the longitudinal design considered by Little. Little presented his review in the context of monotone missing data patterns, a context we adopt here. That is, we assume that if a participant is not observed on a particular occasion, the participant is not observed on any subsequent occasion.

### Random Coefficient Models

Let  $J_{ik}$  denote the last occasion at which participant  $i$  in group  $k$  was observed and  $t_{J_{ik}}$  the value of  $t$  for this time point and  $\mathbf{y}_{ik}$  be partitioned as  $\mathbf{y}'_{ik} = (\mathbf{y}'_{obs, ik} \ \mathbf{y}'_{miss, ik})$ ,  $R_{ik} = J$  if the participant has complete data, and  $R_{ik} = J_{ik}$ , otherwise. The first class of methods is the random coefficient selection models. According to Little (1995), in this approach the joint distribution of  $\mathbf{y}_{ik}$ ,  $\mathbf{b}_{ik}$ , and  $R_{ik}$  is factored as

$$f(\mathbf{y}_{ik}, \mathbf{b}_{ik}, R_{ik} | \mathbf{X}, \mathbf{W}) = f(\mathbf{y}_{ik} | \mathbf{X}, \mathbf{W}, \mathbf{b}_{ik}) f(\mathbf{b}_{ik} | \mathbf{W}) f(R_{ik} | \mathbf{X}, \mathbf{W}, \mathbf{y}_{ik}, \mathbf{b}_{ik}).$$

In our context, the model for  $f(\mathbf{y}_{ik} | \mathbf{X}, \mathbf{W}, \mathbf{b}_{ik})$  is

$$(\mathbf{y}_{ik} | \mathbf{X}, \mathbf{W}, \mathbf{b}_{ik}) \sim N(\mathbf{W}\mathbf{g} + \mathbf{X}\mathbf{u}, \mathbf{s}^2 \mathbf{I}_J)$$

and

$$(\mathbf{b}_{ik} | \mathbf{W}) = \mathbf{u} \sim N(\mathbf{0}, \mathbf{D}).$$

The model for  $f(R_{ik} | \mathbf{X}, \mathbf{W}, \mathbf{y}_{ik}, \mathbf{b}_{ik})$  is the model for the missing data mechanism. The data are referred to as missing completely at random (MCAR) if

$$f(R_{ik} | \mathbf{X}, \mathbf{W}, \mathbf{y}_{ik}, \mathbf{b}_{ik}) = f(R_{ik})$$

(see Rubin, 1976; Little, 1995; Little & Rubin, 1987). That is, the data are MCAR if the probability of a particular data point being missing does not depend on either  $\mathbf{y}_{ik}$ ,  $\mathbf{b}_{ik}$ ,  $\mathbf{X}$  or  $\mathbf{W}$ . The missing data mechanism is called missing at random (MAR) if

$$f(R_{ik} | \mathbf{X}, \mathbf{W}, \mathbf{y}_{obs, ik}, \mathbf{y}_{miss, ik}, \mathbf{b}_{ik}) = f(R_{ik} | \mathbf{X}, \mathbf{W}, \mathbf{y}_{obs, ik})$$

that is, the probability of a particular data point being missing does not depend on either  $\mathbf{y}_{miss, ik}$  or  $\mathbf{b}_{ik}$ . Following Verbeke and Molenberghs (2000, p. 213), a missing data mechanism that does not meet either of these criteria can be referred to as missing not at random (MNAR). Consistent estimates for  $\mathbf{g}$  can be obtained from the likelihood for  $\mathbf{y}_{obs, ik}$  and  $R_{ik}$ . However if the data are MCAR or MAR (and if the parameters of the missing data mechanism are distinct from the parameters for the data), consistent estimates can be obtained by maximizing the likelihood for  $\mathbf{y}_{obs, ik}$ , a process that is called ignoring the missing data mechanism. Thus, for the purposes of estimating the fixed effects, the missing data mechanism is ignorable if the mechanism is MCAR or MAR, but the missing data mechanism is non-ignorable if the mechanism is MNAR.

As Hedeker and Gibbons (1997) noted “many instances of missing data are related to previous performance or other subject characteristics...” [See Little (1995, Section 2.2.2) and Schafer (1997, Ch. 2) for other examples of studies where MAR is a reasonable model of missingness]. Accordingly, MAR may very well be a reasonable process to presume for the missing data in one's study. Again, it should be noted for completeness, that in order to legitimately ignore the missing data mechanism for estimation



random but, as well, the parameters of the missing data mechanism must be independent of the parameters of the data model (Little, 1995; Little & Rubin, 1987; Schafer, 1997). This independence or distinctness of parameters is quite realistic in many contexts (See Schafer, 1997, pp. 11-15). When the missing data mechanism is ignorable, numerical results can easily be obtained with commercially available software, e.g., the SAS PROC (SAS, 1995) MIXED program (See Littell et al., 1996).

#### Pattern Mixture Models

The second class of models presented by Little (1995) is the class of random coefficient pattern-mixture models. As Little (1995, p. 1113) noted, "Pattern-mixture models stratify the population by the pattern of dropout, implying a model for the whole population that is a mixture over the patterns." An advantage of this procedure is that when drop-out depends on  $\mathbf{X}$ ,  $\mathbf{W}$  and  $\mathbf{b}_{ik}$  but not on  $\mathbf{y}_{ik}$ , the missing data mechanism does not have to be explicitly introduced into the likelihood function.

According to Little (1995), pattern-mixture models are based on the factorization

$$f(\mathbf{y}_{ik}, \mathbf{b}_{ik}, R_{ik} | \mathbf{X}, \mathbf{W}) = f(\mathbf{y}_{ik} | \mathbf{X}, \mathbf{W}, \mathbf{b}_{ik}, R_{ik}) f(\mathbf{b}_{ik} | \mathbf{W}, R_{ik}) f(R_{ik} | \mathbf{W}).$$

In this expression  $f(\mathbf{y}_{ik} | \mathbf{X}, \mathbf{W}, \mathbf{b}_{ik}, R_{ik})$  models the subject-specific regressions stratified by missing data pattern,  $f(\mathbf{b}_{ik} | \mathbf{W}, R_{ik})$  models the subject-specific regression coefficients as a function of the between-subjects variables and the missing-data pattern, and  $f(R_{ik} | \mathbf{W})$  models the proportions of each missing data pattern as functions of the between-subjects variables. The approach stratifies the sample by time and missing data pattern and models differences in the distributions of the dependent variables over these patterns.

Little (1995, p. 1118) presented a pattern-mixture model in which  $\mathbf{e}_{ik} \sim N(\mathbf{0}, \mathbf{s}^2 \mathbf{I}_j)$ , as in the model considered in this paper, and drop-out depends on  $\mathbf{W}$  and  $\mathbf{b}_{ik}$  but not on  $\mathbf{y}_{ik}$ . In this case

$$(\mathbf{y}_{ik} | \mathbf{X}, \mathbf{W}, \mathbf{b}_{ik}, R_{ik} = J_{ik}) \sim N(\mathbf{W}\mathbf{g}^{(j)} + \mathbf{X}\mathbf{u}, \mathbf{s}^2 \mathbf{I}_j) \quad (2)$$

and

$$(\mathbf{b}_{ik} | \mathbf{W}) = \mathbf{u} \sim N(\mathbf{0}, \mathbf{D}). \quad (3)$$

The notation  $\mathbf{g}^{(j)}$  indicates that the fixed effects introduced in equation (1) depend on drop-out time. Let  $\mathbf{p}_{jk}$  denote the probability that a participant in treatment  $k$  drops out after occasion  $j$ . The pattern-mixture model estimate of the treatment effect is

$$\sum_j \hat{\mathbf{p}}_{j2} (\hat{\mathbf{g}}_{10}^{(j)} + \hat{\mathbf{g}}_{11}^{(j)}) - \sum_j \hat{\mathbf{p}}_{j1} \hat{\mathbf{g}}_{10}^{(j)}. \quad (4)$$

Little pointed out that the  $\mathbf{g}^{(j)}$  can be estimated in PROC MIXED by introducing drop-out time as a categorical variable. The standard error can be computed using the delta method.

Another alternative is to use the unweighted least squares (UWLS) approach presented by Wang-Clow, Lange, Laird, and Ware (1995). As Little (1995, p. 1120) noted, UWLS is maximum likelihood for the pattern-mixture model described in equations (2) and (3). In the UWLS approach, the estimated treatment effect is

$$\frac{1}{n_1} \sum_{i=1}^{n_1} \hat{\mathbf{b}}_{1i1} - \frac{1}{n_2} \sum_{i=1}^{n_2} \hat{\mathbf{b}}_{1i2} \quad (5)$$

where  $\hat{\mathbf{b}}_{1ik}$  is the ordinary least squares (OLS) estimate of the subject-specific slope for the  $i$ th subject in the  $k$ th group. The standard error of the estimated treatment effect is the (2,2) element of

$$\sum_k \sum_i \frac{\hat{\mathbf{V}}_i}{n_k^2} \quad (6)$$

where  $\hat{\mathbf{V}}_i = \mathbf{s}^2 (\mathbf{X}_i' \mathbf{X}_i)^{-1} + \hat{\mathbf{D}}$  and

$$\mathbf{X}_i = \begin{bmatrix} 1 & t_1 \\ \vdots & \vdots \\ 1 & t_{j_i} \end{bmatrix}.$$

Wang-Clow et al. (1995) showed how to estimate  $\hat{\mathbf{s}}^2$  and  $\hat{\mathbf{D}}$  using the method of moments. These

quantities can also be estimated by using maximum likelihood.

Pattern-mixture modeling is potentially an important approach to analyzing longitudinal data collected in the design considered in this study. However, the method does have one drawback. The results of simulation studies reported by Wu and Carroll (1988), Wu and Bailey (1989), and Wang-Clow et al. (1995) indicated that when the pattern-mixture model in equations (2) and (3) is used the maximum likelihood estimate of the treatment effect may be highly inefficient. For example, Wang-Clow et al. compared various estimation procedures [e.g., un-weighted least squares, maximum likelihood, generalized least squares) under a number of missing data mechanisms (e.g., MAR and MNAR) in a two-group longitudinal design in which measurements were taken over 14 occasions. Wang-Clow et al. tabulated the sampling mean and standard deviation (sd) of the estimated treatment difference between mean slopes (see their Table II), and Type I error and power rates for the test of the treatment difference between mean slopes (see their Table III).

The treatment difference between mean slopes estimates the treatment effect. With regard to their Table II results, the sds for the UWLS method were frequently considerably larger than the other estimation procedures (e.g., under one of their MNAR cases, the UWLS sd was 41.62, while the values for the other estimators ranged from 16.97 to 18.05). The MSE for the UWLS estimator, again under one of the MNAR mechanisms, was 1730.80, a value much larger than those reported for the other estimators (range = 320.51-562.47).

Consequently, Wang-Clow et al. in their summary indicated that “the unweighted estimator is too inefficient to merit consideration.” (p. 294). (Of course, this conclusion may be limited to the conditions of their simulation.) They drew this conclusion despite the fact that the pattern-mixture model estimator of the treatment effect was unbiased in all conditions. Finally, Type I error rates were frequently conservative (range 3.2%-3.8%) and importantly, power to detect differences was considerably less than when other estimators were used (e.g., 15.3% vs. 10.5%-32%).

Hedeker and Gibbons (1997) presented an example illustrating application of the pattern-

mixture model approach to data collected in the design considered in this paper. Whereas Little’s (1995) presentation indicated stratifying participants into as many strata as there are missing data patterns, Hedeker and Gibbons argued that, when the number of participants in some of the strata is small, the strata containing these participants can be combined. In their example, Hedeker and Gibbons had two strata. One included all participants who had a measurement on the last measurement occasion; the other included all other participants. Both groups included participants with different missing data patterns.

The potential problem with this approach can be seen by contrasting it with the UWLS approach used by Wang-Clow et al. (1995). Recall that this approach is maximum likelihood for the pattern-mixture model described in equations (2) and (3). In UWLS, the OLS estimate of the subject-specific slope is calculated for each participant. The un-weighted average of these slopes is then computed for each treatment group and the estimated treatment effect is the difference between these averages. The same estimate would be obtained if participants were stratified into as many strata as there are missing data patterns and ML were applied. This follows because the ML estimate of the expected value of  $\mathbf{b}_{ik}$  within stratum  $j$  and treatment group  $k$  is

$$\hat{B}_{kj} = \frac{\sum_{i=1}^{n_{kj}} \hat{\mathbf{v}}_i^{-1} \hat{\mathbf{b}}_{ik}}{\sum_{i=1}^{n_{kj}} \hat{\mathbf{v}}_i^{-1}},$$

where  $\hat{\mathbf{b}}_{ik}$  is the OLS estimate of  $\mathbf{b}_{ik}$ . When there are as many strata as missing data patterns, within a stratum and treatment group  $\hat{\mathbf{v}}_i$  is a constant over  $i$  and  $\hat{B}_{kj}$  is the un-weighted average of the OLS estimates. Then, the estimated treatment effect is the second element of

$$\sum_j \hat{\mathbf{p}}_{j2} \hat{B}_{2j} - \sum_j \hat{\mathbf{p}}_{j1} \hat{B}_{1j},$$

which is equivalent to equations (4) and (5). On the other hand, when the strata are combined as

suggested by Hedeker and Gibbons, the  $\hat{V}_i$  are not constant over  $i$  and the ML estimate of the expected value of  $\mathbf{b}_{ik}$  within a stratum and treatment is a weighted average of the least squares estimates of the subject-specific slopes for that group. Then, if the expected values of the within-subject regression parameters vary over the missing data patterns that were combined into the missing-data groups, the Hedeker-Gibbons' approach, with two strata, to the pattern-mixture model is likely to yield inconsistent estimators even when the missing data conform to the missing data mechanism assumed by the model in equations (2) and (3).

The Hedeker and Gibbons (1997) model is

$$Y_{ijk} = \mathbf{b}_{0ik} + \mathbf{b}_{1ik}t_j + \mathbf{e}_{ijk} \quad (7)$$

$$\mathbf{b}_{0ik} = \mathbf{I}_{00} + \mathbf{I}_{01}z + \mathbf{I}_{02}z_2 + \mathbf{I}_{03}(z \times z_2) + u_{0i} \quad (8)$$

$$\mathbf{b}_{1ik} = \mathbf{I}_{10} + \mathbf{I}_{11}z + \mathbf{I}_{12}z_2 + \mathbf{I}_{13}(z \times z_2) + u_{1i} \quad (9)$$

where  $z_2$  is 0 for participants with complete data and 1 otherwise. Using the gamma coefficients defined in equation (1), this model can also be written explicitly as a pattern-mixture model

$$Y_{ijk} = \mathbf{g}_{00}^{(z_2)} + \mathbf{g}_{01}^{(z_2)}z + \mathbf{g}_{10}^{(z_2)}t_j + \mathbf{g}_{12}^{(z_2)}(z \times t_j) + u_{1j}t_j + u_{0j} + \mathbf{e}_{ijk} \quad (10)$$

where, as in equation (2), the superscript indicates the group (drop-out or completer) which the parameter describes. Using this notation  $\mathbf{g}_{12}^{(0)}$  is the treatment effect for the completers (i.e., the Time  $\times$  Treatment interaction for the completers) and  $\mathbf{g}_{12}^{(1)}$  is the treatment effect for the dropouts. Further,  $\hat{\mathbf{I}}_{11}$  estimates  $\mathbf{g}_{12}^{(0)}$  and  $\hat{\mathbf{I}}_{13}$  estimates  $\mathbf{g}_{12}^{(1)} - \mathbf{g}_{12}^{(0)}$  (the difference in the Time  $\times$  Treatment interaction for the drop-outs and completers). Therefore the estimated treatment effect is  $\hat{\mathbf{p}}_c \hat{\mathbf{I}}_{11} + \hat{\mathbf{p}}_d (\hat{\mathbf{I}}_{11} + \hat{\mathbf{I}}_{13})$  where  $\hat{\mathbf{p}}_c$  and  $\hat{\mathbf{p}}_d$  are the estimated proportion of participants who completed and dropped out, respectively. The estimated sampling variance is

$$\hat{\mathbf{p}}_c^2 V(\hat{\mathbf{I}}_{11}) + \hat{\mathbf{p}}_d^2 \times V(\hat{\mathbf{I}}_{11} + \hat{\mathbf{I}}_{13}) + \frac{\hat{\mathbf{p}}_c \hat{\mathbf{p}}_d \times \hat{\mathbf{I}}_{13}^2}{n_1 + n_2}$$

where

$$V(\hat{\mathbf{I}}_{11} + \hat{\mathbf{I}}_{13}) = V(\hat{\mathbf{I}}_{11}) + V(\hat{\mathbf{I}}_{13}) + 2C(\hat{\mathbf{I}}_{11}, \hat{\mathbf{I}}_{13}),$$

$V(\cdot)$  denotes a sampling variance and  $C(\cdot, \cdot)$  denotes a sampling covariance.

#### Alternative Methods

A number of other analytic methods, that use information about the pattern of missing data, have been suggested in the literature and one of our goals in this paper is to review alternative methods for analyzing effects in longitudinal designs in which data are missing; the second goal is to report the results of a simulation study which compares the methods.

Wu and Bailey (1989) presented an alternative method, which they called the linear minimum variance unbiased estimator. Later Wang-Clow et al. (1995) referred to the method as the ANCOVA method and we use the latter term in this paper. Provided participants are randomly assigned to groups and it is reasonable to assume that the subject-specific regressions of the dependent variable on time of measurement are well-described by the simple linear regression model, the test of the treatment effect focuses on the average slope (i.e., the population average) in each treatment. Specifically, to test for a treatment effect one tests whether the average slopes are equal for the treatment groups. Wu and Bailey proposed the following procedure:

1. Use OLS to estimate the slope for each participant in each treatment group.
2. Using the estimated slopes as the dependent variable, conduct an ANCOVA with treatment group as the between-subjects factor of interest. Wu and Bailey discussed including two types of covariates. The first is the time point after which the participant dropped out and the second comprises the pretreatment score on the variable of interest and other pretreatment measures that may be available. In this paper we investigate the model without the second type of covariate, as did Wu and Bailey and Wang-Clow et al. (1995). However, we also investigate a related procedure due to Overall, Ahn, Shivakumar, and Kalburgi (1999) that includes the pretest as the covariate.

Wu and Bailey showed that the error variance in this model will vary over dropout times and presented a weighted least squares procedure for estimation and hypothesis testing. The test for the treatment effect (i.e., the group  $\times$  time interaction) is the test of the treatment factor in the ANCOVA. In calculating the weights, Wu and Bailey assumed

$$\mathbf{b}_{ik} \sim N(\mathbf{B}_k, \mathbf{D}_k).$$

Wu and Bailey presented method of moment estimators for  $\mathbf{D}_k$  and  $\mathbf{s}^2$ . Alternatively, maximum likelihood estimates for  $\mathbf{D}_k$  and  $\mathbf{s}^2$  can be obtained by using PROC MIXED:

```
proc mixed method=ml;
class id group;
model score=time group
group*time/solution;
random intercept time/type=un
subject=id group=group;
```

The following are definitions of the variables used in the code:

- id—a categorical variable identifying the participant
- group—a categorical variable identifying the treatment group

In the random statement the code group=group specifies that the covariance matrix for the intercept and slope varies across treatment groups.

The procedure described by Wu and Bailey (1989) is fairly complicated to implement because of the necessity of estimating the weights and inserting them in a weighted least squares procedure. However, we show that a related procedure can be easily implemented in PROC MIXED. Wu and Bailey proposed using the following model to compare treatment groups:

$$\hat{\mathbf{b}}_{1ik} = \mathbf{I}_{10k} + \mathbf{I}_{11} t_{J_{ik}} + \mathbf{d}_{ik}.$$

They compare the groups by using

$$\hat{\mathbf{I}}_{10k} + \hat{\mathbf{I}}_{11} \bar{t}_k,$$

where  $\bar{t}_k$  is the average of  $t_{J_{ik}}$  for the  $k$ th group. If the model

$$\hat{\mathbf{b}}_{1ik} = \mathbf{I}_{10} + \mathbf{I}_{11} (t_{J_{ik}} - \bar{t}_k) + \mathbf{I}_{12} z + \mathbf{d}_{ik} \quad (11)$$

is estimated, then

$$\hat{\mathbf{I}}_{12} = (\hat{\mathbf{I}}_{102} - \hat{\mathbf{I}}_{101}) + \hat{\mathbf{I}}_{11} (\bar{t}_2 - \bar{t}_1).$$

An alternative to equation (11) is

$$\mathbf{b}_{1ik} = \mathbf{I}_{10} + \mathbf{I}_{11} (t_{J_{ik}} - \bar{t}_k) + \mathbf{I}_{12} z + u_{i1}. \quad (12)$$

Readers familiar with multilevel models will recognize this model as a level-2 model for the slope in the level-1 equation

$$Y_{ijk} = \mathbf{b}_{0ik} + \mathbf{b}_{1ik} t_j + \mathbf{e}_{ijk}. \quad (13)$$

We also formulate a level-2 model for the intercept:

$$\mathbf{b}_{0ik} = \mathbf{I}_{00} + \mathbf{I}_{01} (t_{J_{ik}} - \bar{t}_k) + \mathbf{I}_{02} z + u_{i0}. \quad (14)$$

The approach presented by Wu and Bailey (1989) does not include an equation for the intercept. Nevertheless, we include it because Bryk and Raudenbush (1992) have noted that omitting variables in one level-2 model can impact estimates in a second equation because of the correlated error terms for the level-2 models. By including  $(t_{J_{ik}} - \bar{t}_k)$  in equations (12) and (14), the model conditions on the missing data pattern and the model can be formulated as a pattern-mixture model.

PROC MIXED can estimate the model represented by equations (12) to (14). The PROC MIXED program we suggest using is:

```
proc mixed method=ml;
class id group;
model score=lobsc group time
time*lobsc time*group/solution;
random intercept time/type=un
subject=id group=group;
```

The variable  $\text{lobsc}$  is  $(t_{j_{ik}} - \bar{t}_k)$ . The inclusion of  $\text{lobsc}$  and  $\text{time} \times \text{lobsc}$  is intended to improve estimation and testing when drop-out depends on  $\mathbf{W}$  and  $\mathbf{b}_{ik}$  as in Little's (1995) pattern-mixture model presented in equations (2) and (3). If the data are MCAR or MAR valid estimates can be obtained with these terms excluded.

Overall et al. (1999) investigated an analysis similar to the pre-post score analysis advocated by Delucchi and Bostrom (1999), namely an endpoint analysis involving a simple change score from baseline to the last available measurement (p. 206). Their endpoint analysis is a two-stage procedure. At stage-one they obtained a simple change score from baseline to last available measurement and apply these change scores in an ANCOVA, again using pretest score on  $Y$  ( $Y_{i1k}$ ) and the number of available measurements for participant  $i$  ( $J_{ik}$ ) as covariates:

$$(Y_{ijk} - Y_{i1k}) = I_0 + I_1 J_{ik} + I_2 z + I_3 Y_{i1k} + d_{ik}.$$

Overall et al. (1999) employed pretest scores and number of available measurements as covariates because Overall et al., (1996) had shown that these covariates were necessary to control the Type I error rate in conditions where participants who drop out early show less change from the pretest than do later dropouts and completers.

Overall et al. (1999, pp. 205-209) also investigated an ANCOVA approach implemented by using PROC MIXED, though their approach differs from Wu and Bailey (1989). They included the pretest score on  $Y$  and the number of available measurements for participant  $i$  as covariates in order to have the same type of covariate control that they had in their change score analysis. Their model is

$$\begin{aligned} Y_{ijk} &= \mathbf{b}_{0ik} + \mathbf{b}_{1ik} t_j + \mathbf{e}_{ijk} \\ \mathbf{b}_{0ik} &= I_{00} + I_{01} J_{ik} + I_{02} z + I_{03} Y_{i1k} + u_{i0} \\ \mathbf{b}_{1ik} &= I_{10} + I_{12} z + u_{i1}. \end{aligned}$$

Substituting the right hand sides of the equations for the intercept and slope into the equation for the observed data

$$Y_{ijk} = I_{00} + I_{01} J_{ik} + I_{02} z + I_{03} Y_{i1k} + I_{10} t_j + I_{12} z \times t_j + u_{i0} + u_{i1} \times t_j + \mathbf{e}_{ijk},$$

we see that pretest scores appear in the model both as dependent variable scores and as independent variable scores. As Overall et al. (1999, pp. 213-214) and Ahn, Tonidandel, and Overall (2000, pp.278-279) pointed out, use of this model has not been without controversy. A less controversial alternative is to include the pretest as a covariate, but to exclude pretest score from the dependent variable. However, simulations conducted by Overall et al. indicated that the more controversial procedure worked adequately for testing the group  $\times$  time interaction.

Moreover, Ahn et al. compared the more controversial and less controversial procedure and showed that both had similar Type I error rates for testing the group  $\times$  time interaction, but the procedure developed by Overall and his colleagues had better power. PROC MIXED code for the Overall et al. model is

```
proc mixed method=ml;
class id group;
model score=nrm t1 group time
time*group/solution;
random intercept time/type=un
subject=id;
```

The variable  $\text{nrm}$  is the number of measurements available for a participant. The variable  $\text{t1}$  is the pretest score. There are three major differences between our code and theirs. First the time of last observation ( $\text{nrm}$ ) is not centered. Second  $\text{t1}$  is included in their model but not in ours. Third, the time by  $\text{nrm}$  interaction is excluded in their model.

Finally, Overall et al. (1999) investigated a two-stage ANCOVA procedure. They again used the pretest score on  $Y$  and the number of available measurements for participant  $i$  as covariates. Like the Wu and Bailey (1989) approach, Overall et al. used OLS in stage 1 to estimate the subject-specific regression coefficients. The slopes were multiplied by  $t_{j_{ik}}$  and then used in a second stage ANCOVA model:

$$t_{j_{ik}} \hat{\mathbf{b}}_{1ik} = I_{10} + I_{11} J_{ik} + I_{12} z + I_{13} Y_{i1k} + d_{ik}.$$

Thus, the previously described analyses can be used to analyze the important group by time interaction effect in longitudinal designs in which data are missing. In this report we compare these methods because prior research either had not compared all the procedures just enumerated in one study under a common set of manipulated conditions, or, the comparisons were not made on all of the measures we assess. These measures are rates of Type I error and power for the test of equality of average slopes, bias in the difference in the average slopes, and the variability in estimating this difference.

### Method

Nine methods of examining the group by time interaction effect in a between by within subjects repeated measures design were examined. Specifically, the methods (with their acronyms) were:

- (1) the PROC MIXED analysis that presumes the data are missing at random (PMMAR),
- (2) the un-weighted least squares (pattern-mixture) analysis (UWLS),
- (3) Hedeker and Gibbons' (1997) approach to estimating the pattern-mixture model (HGPM),
- (4) Overall et al.'s (1999) PROC MIXED analysis that uses  $t_1$  and  $nrm$  as covariates (OPMAOC),
- (5) Wu and Bailey's (1989) ANCOVA implemented in PROC MIXED (WBPMAOC),
- (6) the weighted least squares ANCOVA presented by Wang-Clow et al. (1995), where the weights for the weighted least squares part of the analysis are obtained from PROC MIXED (WLSAOC),
- (7) the weighted least squares ANCOVA presented by Wang-Clow et al. (1995), where the weights for the weighted least squares part of the analysis are obtained through the method of moments (See Wu & Bailey, 1998, p. 945) (WLSAOCMM),
- (8) Overall et al.'s (1999) two-stage ANCOVA (OTSAOC), and
- (9) Overall et al.'s (1999) two-stage endpoint ANCOVA (OEPAOC).

In the UWLS method standard errors were calculated by using the procedure presented in equation (6). However,  $\mathbf{s}^2$  and  $\mathbf{D}$  were estimated

by maximum likelihood rather than the method of moments.

We investigated two factors in our study: number of equally spaced levels of the repeated measures variable (5 and 9) and missing data mechanism (MCAR, MAR and MNAR). Overall and his colleagues (See Ahn, Tonidandel & Overall, 2000; Overall et al., 1999; Overall et al., 1996) examined the group by time interaction effect in a parallel-groups design containing a baseline score and eight additional repeated measurements; thus, for comparative purposes we had nine levels for one of our cases of number of repeated measurements. Overall and his colleagues designed their investigation to mirror design characteristics in clinical trials where a large number of repeated measurements would not be unusual. However, in behavioral science research, nine levels of the repeated measures variable may not be typical. Accordingly, we also included a smaller case, that is, five levels.

To compare the procedures, we simulated data for a situation in which participants are randomly assigned to treatments. We used the following equation to generate data for the  $i$ th participant, in group  $k$  on the  $j$ th occasion:

$$Y_{ijk} = \mathbf{b}_{0i} + \mathbf{b}_{1i}t_j + \mathbf{e}_{ijk}.$$

In each treatment group, data were simulated for 100 participants. The variable  $t_j$  was coded (0, 0.23077, 0.46154, 0.69231, 0.92308, 1.15385, 1.38462, 1.61538, 1.84615). To get the codes for conditions with five time points we eliminated the last four codes.

The mean for  $\mathbf{b}_{0i}$  was 50 in both groups, implying that both treatment groups had the same population pretest mean. For Type I error data, the mean for the slope was 4.5 in treatment 1 and treatment 2 [ $\mathbf{g}_{11} = 0$ , where  $\mathbf{g}_{11}$  is defined in equation (1)], indicating identical average rates of increase over time, hence, a null condition. For our power comparisons, the slope was 9.0 in treatment 2 and 4.5 in treatment 1 ( $\mathbf{g}_{11} = 4.5$ ) when there were nine occasions and 12.5 in treatment 2 and 4.5 in treatment 1 ( $\mathbf{g}_{11} = 8$ ) when there were five occasions. The slopes for treatment 2 were selected to provide similar power for both levels of the number of occasions factor. The errors ( $\mathbf{e}_{ijk}$ )

were assumed to be uncorrelated for different times of observation. This does not imply that the scores were uncorrelated over time. Allowing the slope and intercept to vary across participants implies that scores were correlated over time. The variance for the residuals, conditional on time, was 240. In all cases the covariance matrix ( $\mathbf{D}$ ) for the intercept and slope was

$$\mathbf{D} = \begin{bmatrix} 15.21 & -12.42 \\ -12.42 & 82.81 \end{bmatrix}.$$

The correlation between the slope and intercept was -.35, indicating that participants with higher pretest status increased less rapidly. We also replicated the entire study changing the covariance to 12.42 from -12.42 and retaining all other features of the design. Notable differences that emerged between the two sets of conditions will be highlighted in the Results section.

Without further complications to the method, the ANCOVA methods can only be applied to participants who have at least two observations and was formulated for the situation in which the missing data occur in a monotone pattern. That is, once a participant drops out, subsequent measurements are not available. Therefore in our simulated data, every participant had an observation at the pretest and the first two follow-up occasions.

Once the data were generated, data were eliminated according to a MCAR, a MAR, or one of two MNAR missing data mechanisms. As indicated in our introduction, when the missing data mechanism is MNAR, ignoring the mechanism can result in inconsistent estimates of the unknown parameters. Accordingly, unlike Delucchi and Bostrom (1999), we compared approaches under a MCAR, a MAR, and two MNAR mechanisms. To select missing observations we used the following model

$$Z_{ijk} = \mathbf{q}_{1j} + \mathbf{q}_2 \mathbf{b}_{0i} + \mathbf{q}_3 \mathbf{b}_{1i} + \mathbf{q}_4 Y_{i(j-1)k} + \mathbf{q}_5 Y_{ijk}.$$

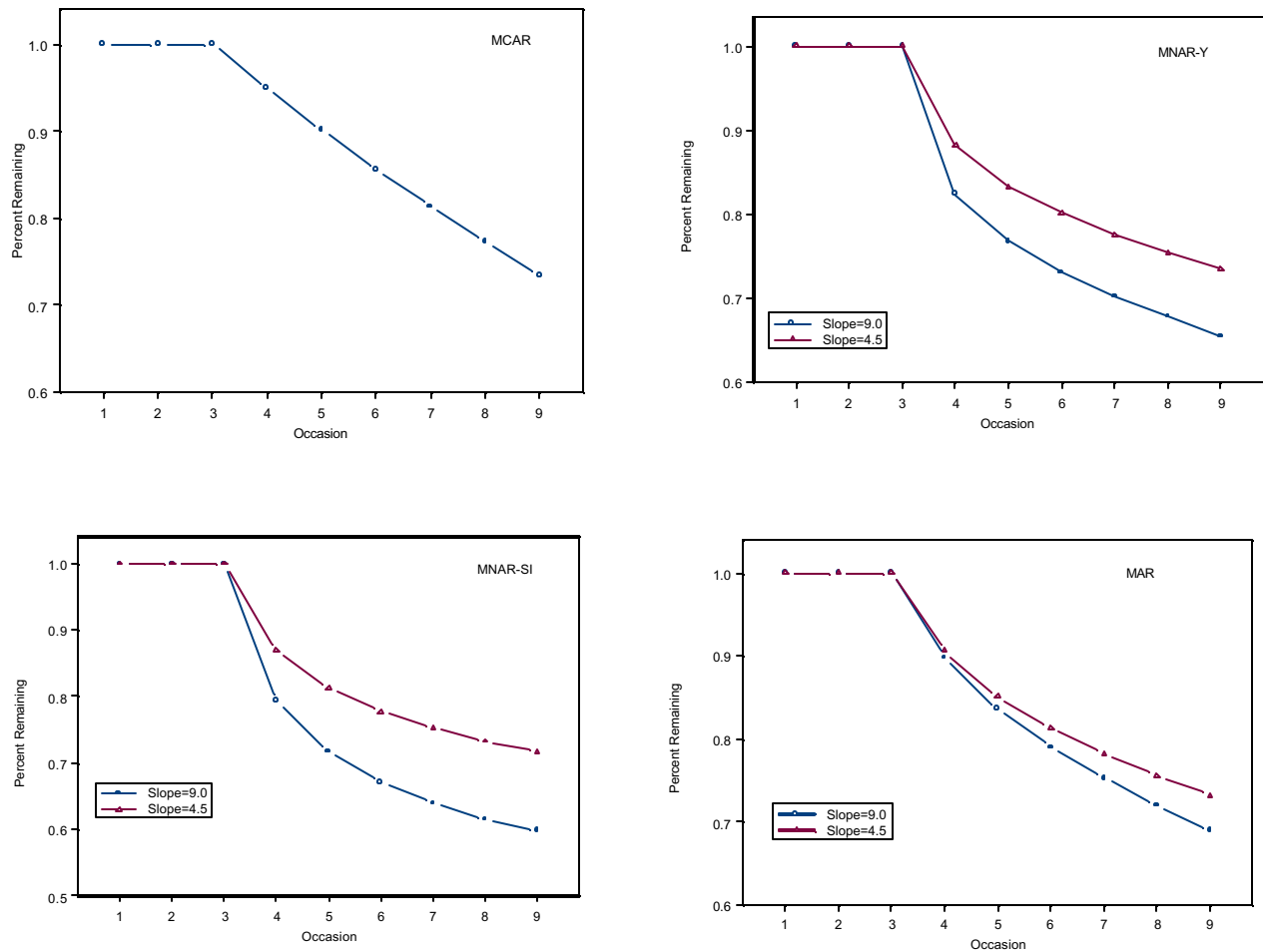
An observation was set as missing if  $U_{ijk} < \mathbf{f}(Z_{ijk})$  where  $U_{ijk}$  is a uniformly distributed random

variable and  $\mathbf{f}$  is the standard normal distribution. The missing data mechanism is MCAR if  $\mathbf{q}_2 = \mathbf{q}_3 = \mathbf{q}_4 = \mathbf{q}_5 = 0$ , MAR if  $\mathbf{q}_2 = \mathbf{q}_3 = \mathbf{q}_5 = 0$  and MNAR if  $\mathbf{q}_2$ ,  $\mathbf{q}_3$ , or  $\mathbf{q}_5$  is not equal to zero. In one MNAR mechanism only  $\mathbf{q}_2$  and  $\mathbf{q}_3$  were not equal to zero (MNAR-SI). This mechanism meets the assumption required for the pattern-mixture model in equations (2) and (3) to yield consistent estimates. In the other MNAR mechanism, only  $\mathbf{q}_5$  was not equal to zero (MNAR-Y). The values of  $\mathbf{q}_{1j}$  were selected to give cumulative missing data rates between 30% and 40% at the ninth occasion.

Figure 1 shows estimated proportions of participants remaining in the study at each occasion in the non-null condition with nine time points under the MCAR, MAR, MNAR-SI and MNAR-Y mechanisms. To obtain these estimates, 100,000 data points were generated for each treatment group. (For the MCAR mechanism, a total of 100,000 data points were generated since in our MCAR condition the dropout rate was the same in both treatments.) For our MAR condition the probability of dropping out at occasion  $j$  was positively related to the participant's score at occasion  $j-1$ . For our MNAR-SI condition the probability of dropping out at occasion  $j$  was positively related to the participant's intercept and slope. For our MNAR-Y condition the probability of dropping out at occasion  $j$  was positively related to the score the participant would have attained at occasion  $j$  if the participant had not dropped out. Thus in all panels of Figure 1, except the top right, drop-out rates are higher for the treatment with the average slope equal to 9 (treatment 2).

Drop out rates vary across type of missing data mechanism; however, because we will compare methods for a particular mechanism, and not the performance of a method across mechanisms, this variation in drop out rates across mechanisms is not problematic. Each condition was replicated 2,500 times. All hypotheses were conducted with a nominal alpha of .05.

Figure 1. Percent of Data that is Not Missing by Occasion and Missing Data Mechanism



## Results

Tabled results are for conditions in which the correlation between the slope and intercept was negative. Important differences that emerged when the correlation between the slope and intercept was positive will be noted in the text.

Type I error rates and power are reported in Table 1 for the MCAR and MAR conditions and in Table 2 for the MNAR conditions. All procedures exhibited adequate control of the Type I error rate. However, when the missing data mechanism was MAR and the correlation between the slope and intercept was positive WLSAOCMM, WLSAOC, and WBPMAOC had higher Type I error rates than those reported in Table 1. These error rates were .067, .068, and .069, respectively, when the number of time points

was five and .076, .112, and .115 for nine time points. Although in some conditions, UWLS, HGPMM, and/or OEPAOC were competitive with the other procedures in terms of power, they generally had lower power than the other procedures. Excluding HGPMM, UWLS, and OEPAOC from consideration, under the MCAR and MAR conditions, power differences were fairly small among the remaining methods. In the MCAR conditions, OTSAOC and PMMAR had the highest power estimates; in the MAR conditions WBPMAOC had the best power estimates. The slight advantage of WBPMAOC relative to PMMAR may reflect the fact that WBPMAOC resulted in treatment effect estimators with a positive bias (see Table 5) when the data were MAR, whereas, as expected



theoretically, PMMAR provided a consistent estimator of the treatment effect.

In the MNAR conditions the methods seem to separate into two groups; PMMAR, UWLS, OTSAOC, and OEPAOC tended to have lower power than the other procedures. Among OPMAOC, WBPMAOC, WLSAOC, and WLSAOCMM, WBPMAOC tended to have the highest power in MNAR-SI while WBPMAOC and OPMAOC tended to have the highest power in MNAR-Y.

The slope difference ( $g_{11}$ ) can be estimated by all procedures except OTSAOC and OEPAOC. For each condition in the study, the slope difference was estimated by using each of the remaining six methods. Table 3 contains means and standard deviations of these estimates for the MCAR and MAR conditions when  $g_{11} = 0$ . Table 4 contains the same information for the MNAR conditions. When  $g_{11} = 0$ , none of the procedures had an average estimate that was significantly different from zero. In Tables 3 and 4, UWLS and HGPMM tended to have larger standard deviations than the other procedures. The standard deviations for the remaining four procedures were similar in size.

Table 5 contains means and standard deviations of these estimates for the MCAR and MAR conditions when  $g_{11} \neq 0$ ; Table 6 contains the same information for the MNAR conditions. Bold entries are average estimated slope differences that were significantly different from the population slope difference. The results suggest that all of the procedures are unbiased when the data were MCAR. When the data were MAR, only PMMAR did not show any significant evidence of bias. For the condition with five time points OPMAOC and HGPMM were not significantly biased. This finding probably reflects the larger standard error for the condition with five time points: For each of HGPMM and OPMAOC, the amount of estimated bias was similar when there were five and nine time points. When the covariance between the slope and intercept was positive, HGPMM exhibited more bias (average  $\hat{g}_{11} = 7.680$  for five time points and  $\hat{g}_{11} = 3.967$  for nine time points).

In the MNAR-SI condition, missingness depends on the subject-specific intercepts and slopes and the pattern-mixture model presented in equations (2) and (3) is expected to result in a consistent estimator of the slope difference. As expected from theory, the UWLS procedure did not result in significant evidence of bias. HGPMM, which is also intended to be unbiased under MNAR-SI, was substantially biased. In fact HGPMM exhibited the second largest amount of bias, following PMMAR. WBPMAOC, WLSAOC, WLSAOCMM were also intended to be unbiased under MNAR-SI. WLSAOCMM was unbiased and WLSAOC exhibited a small but significant bias for nine time points. WBPMAOC was biased but its bias was much smaller than that for HGPMM.

In the MNAR-Y condition missingness depends on the participant's score at occasion  $j$ ; under MNAR-Y none of the procedures were expected to result in consistent estimators of the slope difference. PMMAR exhibited substantial bias for both five and nine time points. The other procedures had fairly large bias when there were five time points and less bias when there were nine time points. When the covariance between the slope and intercept was positive HGPMM was substantially biased when there were five measurement occasions; the average value of  $\hat{g}_{11}$  was 7.12.

The other procedures exhibited less evidence of bias in the positive covariance case than in the negative covariance case. Although OPMAOC did not exhibit significant evidence of bias when there were nine measurement occasions and a negative covariance, OPMAOC was substantially biased when the covariance between the slope and intercept was positive with an average value for  $\hat{g}_{11}$  of 4.04.

In both Tables 5 and 6 the standard deviations for UWLS and HGPMM are larger than for the other procedures which most likely accounts for their relatively poor power. The remaining procedures have similar standard deviations.

Table 1. Type I Error and Power Rates for MCAR and MAR Conditions.

Missing Data Mechanism	Method	5-levels		9-levels	
		Type I Error	Power	Type I Error	Power
MCAR	PMMAR	0.052	0.663	0.052	0.669
	UWLS	0.052	0.612	0.052	0.419
	HGPMM	0.053	0.631	0.054	0.577
	OPMAOC	0.052	0.658	0.055	0.662
	WPMAOC	0.053	0.650	0.051	0.662
	WLSAOC	0.052	0.647	0.050	0.654
	WLSAOCMM	0.052	0.645	0.049	0.620
	OTSAOC	0.052	0.711	0.050	0.669
	OEPAOC	0.050	0.625	0.050	0.554
MAR	PMMAR	0.056	0.638	0.054	0.630
	UWLS	0.054	0.564	0.051	0.371
	HGPMM	0.047	0.555	0.048	0.473
	OPMAOC	0.055	0.645	0.053	0.645
	WPMAOC	0.057	0.665	0.073	0.687
	WLSAOC	0.057	0.658	0.067	0.670
	WLSAOCMM	0.055	0.654	0.053	0.624
	OTSAOC	0.050	0.642	0.045	0.585
	OEPAOC	0.048	0.574	0.047	0.444

*Notes:* PMMAR-Proc Mixed MAR analysis; UWLS-Un-weighted least squares analysis which is ML for pattern-mixture models; HGPMM-Hedeker and Gibbons' (1997) approach to pattern-mixture models; OPMAOC-Overall et al.'s (1999) Proc Mixed ANCOVA; WPMAOC- Wu and Bailey's (1989) ANCOVA with PROC Mixed as defined in this paper; WLSAOC- Wang-Clow et al.'s (1995) ANCOVA analysis; WLSAOCMM-Wang-Clow et al.'s ANCOVA using the method of moments for estimation; OTSAOC- Overall et al.'s two-stage ANCOVA; OEPAOC- Overall et al.'s two-stage endpoint ANCOVA analysis.

Table 2. Type I Error and Power Rates for MNAR Conditions.

Missing Data		5-levels		9-levels	
Mechanism	Method	Type I Error	Power	Type I Error	Power
MNAR-SI	PMMAR	0.052	0.446	0.046	0.396
	UWLS	0.049	0.449	0.045	0.236
	HGPMM	0.053	0.364	0.044	0.273
	OPMAOC	0.056	0.531	0.048	0.505
	WBPMAOC	0.055	0.618	0.056	0.649
	WLSAOC	0.056	0.581	0.055	0.579
	WLSAOCMM	0.056	0.575	0.043	0.525
	OTSAOC	0.052	0.261	0.041	0.249
	OEPAOC	0.045	0.228	0.045	0.198
MNAR-Y	PMMAR	0.052	0.493	0.049	0.497
	UWLS	0.042	0.435	0.049	0.258
	HGPMM	0.046	0.488	0.053	0.430
	OPMAOC	0.048	0.556	0.051	0.607
	WBPMAOC	0.046	0.552	0.050	0.588
	WLSAOC	0.050	0.528	0.049	0.532
	WLSAOCMM	0.049	0.520	0.042	0.478
	OTSAOC	0.048	0.449	0.045	0.435
	OEPAOC	0.046	0.422	0.051	0.336

*Notes:* PMMAR-Proc Mixed MAR analysis; UWLS-Un-weighted least squares analysis which is ML for pattern-mixture models; HGPMM-Hedeker and Gibbons' (1997) approach to pattern-mixture models; OPMAOC-Overall et al.'s (1999) Proc Mixed ANCOVA; WBPMAOC- Wu and Bailey's (1989) ANCOVA with PROC Mixed as defined in this paper; WLSAOC- Wang-Clow et al.'s (1995) ANCOVA analysis; WLSAOCMM-Wang-Clow et al.'s ANCOVA using the method of moments for estimation; OTSAOC- Overall et al.'s two-stage ANCOVA; OEPAOC- Overall et al.'s two-stage endpoint ANCOVA analysis.

Table 3. Mean and Standard Deviation of the Difference between the Control and Treatment Group ( $g_{11} = 0$ ): MCAR and MAR Conditions.

Missing data		5-levels		9-levels	
Mechanism	Method	MEAN	SD	MEAN	SD
MCAR	PMMAR	0.008	3.402	-0.023	1.947
	UWLS	-0.028	3.625	-0.032	2.588
	HGPMM	-0.014	3.572	-0.029	2.150
	OPMAOC	0.005	3.408	-0.022	1.971
	WPMAOC	0.006	3.417	-0.023	1.961
	WLSAOC	0.004	3.416	-0.021	1.967
	WLSAOCMM	0.004	3.417	-0.021	1.972
MAR	PMMAR	0.019	3.449	0.051	1.959
	UWLS	0.006	3.875	0.084	3.000
	HGPMM	0.006	3.725	0.075	2.248
	OPMAOC	0.016	3.472	0.057	1.972
	WPMAOC	0.009	3.542	0.030	2.116
	WLSAOC	0.013	3.541	0.045	2.109
	WLSAOCMM	0.010	3.538	0.046	2.113

*Notes:* PMMAR-Proc Mixed MAR analysis; UWLS-Un-weighted least squares analysis which is ML for pattern-mixture models; HGPMM-Hedeker and Gibbons' (1997) approach to pattern-mixture models; OPMAOC-Overall et al.'s (1999) Proc Mixed ANCOVA; WPMAOC- Wu and Bailey's (1989) ANCOVA with PROC Mixed as defined in this paper; WLSAOC- Wang-Clow et al.'s (1995) ANCOVA analysis; WLSAOCMM-Wang-Clow et al.'s ANCOVA using the method of moments for estimation.

Table 4. Mean and Standard Deviation of the Difference between the Control and Treatment Group ( $g_{11} = 0$ ): MNAR Conditions.

Missing Data		5-levels		9-levels	
Mechanism	Method	MEAN	SD	MEAN	SD
MNAR-SI	PMMAR	0.000	3.523	0.012	1.950
	UWLS	0.086	4.008	-0.053	3.206
	HGPMM	0.063	3.903	0.016	2.376
	OPMAOC	0.028	3.545	-0.003	1.993
	WPMAOC	0.025	3.538	0.014	2.007
	WLSAOC	0.033	3.551	-0.013	2.037
	WLSAOCMM	0.035	3.554	-0.012	2.042
MNAR-Y	PMMAR	-0.043	3.520	-0.028	1.968
	UWLS	-0.008	3.860	-0.045	3.105
	HGPMM	-0.066	3.783	-0.024	2.351
	OPMAOC	-0.044	3.480	-0.022	1.956
	WPMAOC	-0.046	3.482	-0.021	1.936
	WLSAOC	-0.042	3.499	-0.023	1.970
	WLSAOCMM	-0.040	3.497	-0.020	1.978

*Notes:* PMMAR-Proc Mixed MAR analysis; UWLS-Un-weighted least squares analysis which is ML for pattern-mixture models; HGPMM-Hedeker and Gibbons' (1997) approach to pattern-mixture models; OPMAOC-Overall et al.'s (1999) Proc Mixed ANCOVA; WPMAOC- Wu and Bailey's (1989) ANCOVA with PROC Mixed as defined in this paper; WLSAOC- Wang-Clow et al.'s (1995) ANCOVA analysis; WLSAOCMM-Wang-Clow et al.'s ANCOVA using the method of moments for estimation.

Table 5. Mean and Standard Deviation of the Difference between the Control and Treatment Group ( $g_{11} \neq 0$ ): MCAR and MAR Conditions.

Missing Data		5-levels		9-levels	
		$g_{11} = 8.0$		$g_{11} = 4.5$	
Mechanism	Method	MEAN	SD	MEAN	SD
MCAR	PMMAR	8.036	3.357	4.501	1.895
	UWLS	8.094	3.597	4.542	2.560
	HGPMM	8.109	3.558	4.495	2.082
	OPMAOC	8.046	3.365	4.511	1.907
	WBPMAOC	8.026	3.381	4.503	1.899
	WLSAOC	8.032	3.381	4.513	1.901
	WLSAOCMM	8.033	3.382	4.514	1.902
MAR	PMMAR	8.006	3.544	4.489	1.969
	UWLS	8.253	3.993	4.805	3.031
	HGPMM	7.862	3.833	4.311	2.235
	OPMAOC	8.137	3.567	4.618	1.986
	WBPMAOC	8.374	3.645	4.888	2.124
	WLSAOC	8.338	3.644	4.865	2.113
	WLSAOCMM	8.334	3.644	4.863	2.117

*Notes:* PMMAR-Proc Mixed MAR analysis; UWLS-Un-weighted least squares analysis which is ML for pattern-mixture models; HGPMM-Hedeker and Gibbons' (1997) approach to pattern-mixture models; OPMAOC-Overall et al.'s (1999) Proc Mixed ANCOVA; WBPMAOC- Wu and Bailey's (1989) ANCOVA with PROC Mixed as defined in this paper; WLSAOC- Wang-Clow et al.'s (1995) ANCOVA analysis; WLSAOCMM-Wang-Clow et al.'s ANCOVA using the method of moments for estimation. Bold values indicate average estimates that are significantly different than the population slope difference.

Table 6. Mean and Standard Deviation of the Difference between the Control and Treatment Group ( $g_{11} \neq 0$ ): MNAR Conditions.

Missing Data		5-levels		9-levels	
		$g_{11} = 8.0$		$g_{11} = 4.5$	
Mechanism	Method	MEAN	SD	MEAN	SD
MNAR-SI	PMMAR	6.606	3.671	3.411	2.037
	UWLS	7.978	4.391	4.394	3.509
	HGPMM	6.992	4.344	3.660	2.541
	OPMAOC	7.489	3.676	4.057	2.052
	WPMAOC	8.318	3.733	4.809	2.082
	WLSAOC	8.069	3.737	4.588	2.127
	WLSAOCMM	8.066	3.739	4.582	2.136
MNAR-Y	PMMAR	6.893	3.437	3.964	1.997
	UWLS	7.395	3.978	4.301	3.320
	HGPMM	7.667	3.868	4.405	2.390
	OPMAOC	7.477	3.452	4.455	1.996
	WPMAOC	7.491	3.439	4.379	1.994
	WLSAOC	7.310	3.476	4.194	2.051
	WLSAOCMM	7.309	3.477	4.202	2.052

*Notes:* PMMAR-Proc Mixed MAR analysis; UWLS-Un-weighted least squares analysis which is ML for pattern-mixture models; HGPMM-Hedeker and Gibbons' (1997) approach to pattern-mixture models; OPMAOC-Overall et al.'s (1999) Proc Mixed ANCOVA; WPMAOC- Wu and Bailey's (1989) ANCOVA with PROC Mixed as defined in this paper; WLSAOC- Wang-Clow et al.'s (1995) ANCOVA analysis; WLSAOCMM-Wang-Clow et al.'s ANCOVA using the method of moments for estimation. Bold values indicate average estimates that are significantly different than the population slope difference.

### Additional Conditions and Results

Our results indicate that UWLS can be inefficient and have low power. As noted earlier the sampling variance of the UWLS estimator of the slope difference is the (2,2)

element of  $\sum_k \sum_i \frac{\widehat{V}_i}{n_k^2}$  where

$\widehat{V}_i = \mathcal{S}^2(\mathbf{X}'_i \mathbf{X}_i)^{-1} + \widehat{\mathbf{D}}$  and therefore depends on the relative sizes of the contributions of  $\mathcal{S}^2(\mathbf{X}'_i \mathbf{X}_i)^{-1}$  and  $\widehat{\mathbf{D}}$ . This being the case, in order to increase the generalizability of our results, we expanded our study by conducting additional simulations in which the  $\mathbf{X}$  matrix used to generate the data

$$\mathbf{X}' = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \end{bmatrix}$$

rather than

$$\mathbf{X}' = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & .23 & .46 & .69 & .92 & 1.15 & 1.38 & 1.61 & 1.85 \end{bmatrix}.$$

These simulations were limited to MCAR and MNAR-SI missing data mechanisms. For the MAR and MNAR-Y missing data mechanisms in our study, it is not possible to change the initial  $\mathbf{X}$  matrix without either increasing the rate of missing data or reducing the dependence of the missing data indicator on the variables in the missing data model to maintain the rates of missing data that occurred with the original  $\mathbf{X}$  matrix. In either case, the change in the  $\mathbf{X}$  matrix would be confounded with another feature of the data. For these simulations we used 1000 replications. All other features of the simulation were unchanged. Given that we only changed was the  $\mathbf{X}$  matrix, the change simulates conducting a study over a longer time period.

In the MCAR and MNAR-SI conditions with the  $\mathbf{X}$  matrix, all procedures controlled the Type I error rate well. The same result was found with the revised  $\mathbf{X}$  matrix except when the covariance between the slope and intercept was positive and the data were MNAR-SI. Then WLSAOCMM, WLSAOC, and WBMAOC had higher

Type I error rates than with the original  $\mathbf{X}$  matrix. The error rates were .072, .072, and .076, respectively, when the number of time points was five and .078, .083, and .084 for nine time points.

In general, with the new  $\mathbf{X}$  matrix the UWLS procedure was more competitive in terms of sampling variability (see Tables 7 and 8, which contain results for the condition with a negative correlation between the slope and intercept) and thus in power. Thus, contrary to the results in Wang-Clow et al. (1995), UWLS can be reasonably efficient in some situations. Apparently, the efficiency improves as the sampling variance of the OLS estimators of the within-subjects regression model improves, as might happen when data are collected over a longer time span.

With the initial  $\mathbf{X}$  matrix, UWLS was unbiased, as expected, in the MNAR-SI condition but HGPMM exhibited substantial bias when  $g_{11} \neq 0$  and therefore had less power. This result also occurred with the revised  $\mathbf{X}$  matrix (see Table 8).

PMMAR performed well in the MCAR condition in terms of bias and power. As expected from theory, PMMAR performed less well in the MNAR-SI condition. In particular, when  $g_{11} \neq 0$ , PMMAR exhibited evidence of bias and was not among the more powerful procedures. Similar results occurred with the revised  $\mathbf{X}$  matrix (see Table 8).

With the initial  $\mathbf{X}$  matrix,  $g_{11} \neq 0$ , and MNAR-SI missing data mechanisms, OPMAOC, tended to show evidence of bias, with bias ranging from 6% to 17% of the population slope difference. The bias was reduced with the revised  $\mathbf{X}$  matrix, ranging from 3% to 5%. Similarly WBMAOC tended to show evidence of bias with the original  $\mathbf{X}$  matrix, with bias ranging from 2% to 7%. Bias was reduced with the revised  $\mathbf{X}$  matrix. In the MNAR-SI condition WLSAOC, and WLSAOCMM tended to exhibit very little bias and this was true with the revised  $\mathbf{X}$  matrix also (see Table 8).



Table 7. Mean and Standard Deviation of the Difference between the Control and Treatment Group for the revised  $\mathbf{X}$  matrix and  $\mathbf{g}_{11} = 0$ : MCAR and MNAR-SI Conditions.

Missing Data		5-levels		9-levels	
Mechanism	Method	MEAN	SD	MEAN	SD
MCAR	PMMAR	0.017	1.486	0.075	1.386
	UWLS	0.019	1.501	0.060	1.399
	HGPMM	0.017	1.509	0.070	1.390
	OPMAOC	0.023	1.488	0.069	1.387
	WBPMAOC	0.016	1.487	0.078	1.388
	WLSAOC	0.019	1.488	0.076	1.387
	WLSAOCMM	0.019	1.488	0.076	1.387
MNAR-SI	PMMAR	0.011	1.453	-0.002	1.389
	UWLS	0.001	1.527	-0.017	1.485
	HGPMM	-0.002	1.468	0.007	1.385
	OPMAOC	0.010	1.476	-0.001	1.406
	WBPMAOC	0.008	1.494	-0.011	1.418
	WLSAOC	0.009	1.492	-0.009	1.420
	WLSAOCMM	0.009	1.492	-0.009	1.421

*Notes:* PMMAR-Proc Mixed MAR analysis; UWLS-Un-weighted least squares analysis which is ML for pattern-mixture models; HGPMM-Hedeker and Gibbons' (1997) approach to pattern-mixture models; OPMAOC-Overall et al.'s (1999) Proc Mixed ANCOVA; WBPMAOC- Wu and Bailey's (1989) ANCOVA with PROC Mixed as defined in this paper; WLSAOC- Wang-Clow et al.'s (1995) ANCOVA analysis; WLSAOCMM-Wang-Clow et al.'s ANCOVA using the method of moments for estimation. Bold values indicate average estimates that are significantly different than the population slope difference.

Table 8. Mean and Standard Deviation of the Difference between the Control and Treatment Group for the revised  $\mathbf{X}$  matrix and  $\mathbf{g}_{11} \neq 0$ : MCAR and MNAR-SI Conditions.

Missing Data		5-levels		9-levels	
		$\mathbf{g}_{11} = 8.0$		$\mathbf{g}_{11} = 4.5$	
Mechanism	Method	MEAN	SD	MEAN	SD
MCAR	PMMAR	8.024	1.438	4.462	1.307
	UWLS	8.017	1.457	4.468	1.342
	HGPMM	8.013	1.468	4.464	1.320
	OPMAOC	8.022	1.442	4.462	1.313
	WPMAOC	8.024	1.437	4.461	1.309
	WLSAOC	8.024	1.439	4.461	1.309
	WLSAOCMM	8.024	1.439	4.461	1.309
MNAR-SI	PMMAR	7.545	1.515	4.218	1.366
	UWLS	7.964	1.600	4.497	1.476
	HGPMM	6.999	1.621	3.867	1.413
	OPMAOC	7.751	1.533	4.304	1.378
	WPMAOC	8.106	1.534	4.561	1.380
	WLSAOC	8.030	1.538	4.520	1.388
	WLSAOCMM	8.025	1.538	4.518	1.387

*Notes:* PMMAR-Proc Mixed MAR analysis; UWLS-Un-weighted least squares analysis which is ML for pattern-mixture models; HGPMM-Hedeker and Gibbons' (1997) approach to pattern-mixture models; OPMAOC-Overall et al.'s (1999) Proc Mixed ANCOVA; WPMAOC- Wu and Bailey's (1989) ANCOVA with PROC Mixed as defined in this paper; WLSAOC- Wang-Clow et al.'s (1995) ANCOVA analysis; WLSAOCMM-Wang-Clow et al.'s ANCOVA using the method of moments for estimation. Bold values indicate average estimates that are significantly different than the population slope difference.

### Conclusion

The purpose of our article was to introduce and examine a number of methods of analysis for longitudinal designs in which data may be missing. Random coefficients selection models may be used to obtain estimates of parameters when data are not completely observed, that is when data are missing. As Little (1995) and others have noted, when random coefficients selection models are used, biased estimates can result if the data are MNAR and the missing data mechanism is not accounted for in the estimation procedure. An alternative method is random coefficients pattern-mixture modeling due to Little.

Little has presented a random coefficients pattern-mixture model that yields consistent estimators of the fixed effects when the missing data mechanism is MNAR-SI (i.e., the pattern of missingness is predictable from the random coefficients). Because recent evidence suggests that this pattern-mixture model can result in inefficient estimates, we presented and examined other methods of analysis that, also according to the literature, may result in better estimation of unknown parameters and which take MNAR-SI missingness into account in their analyses. In particular, we investigated methods due to Wu and Bailey (1988, 1989) and Wang-Clow et al. (1995). We also investigated several methods due to Overall et al. (1999) and we included the random coefficients selection model that ignores the missing data mechanism and an implementation of Little's pattern-mixture model that is due to Hedeker and Gibbons (1997).

All procedures except WBMAOC, WLSAOC, and WLSAOCMM controlled the Type I error rates well in all conditions. The latter three procedures had elevated Type I error rates in several conditions, although the elevation was severe only when there were nine time points. Even with nine time points, WLSAOCMM performed reasonably well, with a maximum Type I error rate of .076 for a nominal .05 test.

WBMAOC and WLSAOC performed reasonably well when there were five time points with maximum estimated Type I error rates of .076 and .072 respectively.

Although no single procedure dominated the other in terms of power, WBMAOC tended to be among the more powerful procedures in all conditions. This occurred in conditions in which WBMAOC controlled the Type I error rate well in addition to the conditions in which it did not. Procedures that tended to be competitive with WBMAOC over a range of conditions were OPMAOC, WLSAOC, and WLSAOCMM.

All procedures produced estimators that were unbiased when the population treatment effect was null. Thus in the following all references to bias refer to conditions in which the treatment effect was non-null. UWLS was unbiased in MCAR and MNAR-SI conditions and had reasonably small biases in the other conditions. Consistent with evidence reported by Wu and Bailey (1989) and Wang-Clow et al. (1995), our results indicate that UWLS can be inefficient and have low power in some conditions. However, our results also indicate that UWLS can be competitive with the other procedures in terms of efficiency and power. The improved performance for UWLS occurred when the design permitted more accurate OLS estimates of the within-subject slopes. In these conditions, the standard errors produced by UWLS were fairly similar to those produced by PMMAR. Therefore a comparison of standard errors may be a useful diagnostic for determining when UWLS should be used.

HGPPM can be inefficient and have low power in some conditions though it tends to be as or more efficient than UWLS. And like UWLS, efficiency and power for HGPPM improved when the design permitted more accurate OLS estimates of the within-subject slopes. Unlike UWLS, HGPPM produced a substantially biased estimate of the treatment effect in the MNAR-SI condition.

This is a serious weakness because the pattern-mixture model is designed to be unbiased in the MNAR-SI condition. It should be noted, however, that the bias of the Hedeker and Gibbons' approach might improve if participants with different missing data patterns were combined into several missing data groups based on the similarity of the time points at which the data were missing. In addition if, within each treatment group, the expected value of the slope is the same for all participants with incomplete data, then the Hedeker and Gibbons' approach should result in an unbiased estimator of the treatment effect.

WBMAOC tended to have levels of bias similar to UWLS except with the original  $\mathbf{X}$  matrix in the MNAR-SI condition. Then WBMAOC was slightly more biased. Similarly, OPMAOC also tended to have levels of bias similar to those of UWLS except in the MNAR-SI condition with the original  $\mathbf{X}$  matrix. Then it tended to exhibit more bias than WBMAOC. WLSAOC and WLSAOCMM tended to have levels of bias similar to UWLS except with the original  $\mathbf{X}$  matrix, nine measurement occasions, and the MNAR-Y missing data mechanism. Then WLSAOC and WLSAOCMM were more biased than UWLS, WBMAOC, and OPMAOC. PMMAR was unbiased in MCAR and MAR conditions, but exhibited fairly substantial bias in the MNAR conditions.

Our analyses of bias, sampling variability, Type I error and power indicated that no one procedure performed best for all missing data mechanisms. Clearly if one were to have valid information about the type of missing data, the information should be taken into account in selecting a procedure. Nevertheless, in our view, the Overall et al. (1999) ANCOVA (OPMAOC) performed better than the others over the range of conditions considered in the research, even though in any particular condition it may have been outperformed by one of the remaining procedures. The main drawback in OPMAOC was its negative bias in the MNAR-SI conditions; this bias made it less competitive in terms of power with

other procedures, in particular with the Wu and Bailey (1989) procedure (WLSAOCMM), the Wu and Bailey procedure implemented with our PROC MIXED program (WBPMAOC), and the Wang-Clow et al. (1995) ANCOVA procedure with weights estimated using results from PROC MIXED (WLSAOC).

WLSAOCMM also tended to perform well in terms of bias, sampling variability, Type I error and power over a range of conditions. Its main weakness was a somewhat elevated Type I error rate in some conditions. However, its maximum estimated Type I error rate was .078. WBPMAOC and WLSAOC performed well when there were five time points, but showed elevated Type I error rates in some conditions with nine time points. Because these procedures tended to be among the most powerful in conditions in which they controlled the Type I error rate, they may be attractive when there are relatively few time points.

Of course, as is true of all empirical studies, the generalizability of our results is limited by the design of the study. The procedures may perform differently if different models for dropping out are adopted. Of particular interest are conditions in which the parameters for the missing data model vary across treatment groups.

#### References

- Ahn, C., Tonidandel, S., & Overall, J. E. (2000). Issues in use of SAS Proc.Mixed to test the significance of treatment effects in controlled clinical trials. *Journal of Biopharmaceutical Statistics*, 10(2), 265-286.
- Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models: Applications and data analysis methods*. Newbury Park, CA: Sage.
- Delucchi, K., & Bostrom, A. (1999). Small sample longitudinal clinical trials with missing data: A comparison of analytic methods. *Psychological Methods*, 4, 158-172.

- Hedeker, D., & Gibbons, R. D. (1997). Application of random-effects pattern-mixture models for missing data in longitudinal studies. *Psychological Methods*, 2, 64-78.
- Littell, R. C., Milliken, G. A., Stroup, W. W., & Wolfinger, R. D. (1996). *SAS system for mixed models*. Cary, NC: SAS Inc.
- Little, R. J. A. (1993). Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association*, 88, 125-134.
- Little, R. J. A. (1994). A class of pattern-mixture models for normal incomplete data. *Biometrics*, 81, 471-483.
- Little, R. J. A. (1995). Modeling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association*, 90, 1112-1121.
- Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data*. New York: Wiley.
- Overall, J. E., Ahn, C., Shivakumar, C., & Kalburgi, Y. (1999). Problematic formulations of SAS Proc. Mixed models for repeated measurements. *Journal of Biopharmaceutical Statistics*, 9, 189-216.
- Overall, J. E., Ghasser, S., & Fiore, J. (1996). Random regression with imputed values for dropouts. *Psychopharmacology Bulletin*, 1, 377-388.
- Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63, 581-592.
- SAS Institute. (1995). *Introduction to the MIXED procedure: Course Notes*. Cary, NC: Author.
- Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. New York: Chapman & Hall/CRC.
- Verbeke, G., & Molenberghs. (2000). *Linear mixed models for longitudinal data*. New York: Springer.
- Wang-Clow, F., Lange, M., Laird, N. M., & Ware, J. H. (1995). A simulation study of estimators for rates of change in longitudinal studies with attrition. *Statistics in Medicine*, 14, 283-297.
- Wu, M. C., & Carroll, R. J. (1988). Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics*, 44, 175-188.
- Wu, M. C., & Bailey, K. R. (1989). Estimation and comparison of changes in the presence of informative right censoring: Conditional linear model. *Biometrics*, 45, 939-955.

## A Parametric Bootstrap Version of Hedges' Homogeneity Test

Wim Van den Noortgate      Patrick Onghena  
Katholieke Universiteit Leuven, Belgium

---

Hedges'  $Q$ -test is frequently used in meta-analyses to evaluate the homogeneity of effect sizes, but for several kinds of effect size measures it does not always appropriately control the Type 1 error probability. Therefore we propose a parametric bootstrap version, which shows Type 1 error control under a broad set of circumstances. This is confirmed in a small simulation study.

Key words:  $Q$ -test, homogeneity, effect sizes, parametric bootstrap, Type 1 error

---

### Introduction

A meta-analysis cumulates the findings of previous research. Often fixed effects techniques are used to summarize the findings of several studies into one single result. The individual effect size estimates are averaged (usually with each effect size weighted by the size of the study or by the inverse of its sampling variance), to obtain an estimate of the overall effect size. These techniques of course are only appropriate if studies can be assumed to be sharing a common population effect size or if in the meta-analysis no inference to a broader population of effect sizes is aimed at (Hedges & Vevea, 1998).

The suitability of the fixed effects techniques therefore is usually statistically tested by means of a homogeneity test. If effect sizes are found heterogeneous, study characteristics are included in the model as covariates to investigate their moderating effect, resulting in a fixed effects regression model. Alternatively, or in addition to the inclusion of moderator variables, the heterogeneity may be explicitly modeled, by defining random study effects. This results in a random effects model or a random effects regression model (see Raudenbush, 1994, for more details). The homogeneity test thus often plays a crucial role in a meta-analysis, since its results are often used to decide if the simple fixed effects model is to be extended with moderator variables and/or random effects, and fixed effects and random effects meta-analytic models often give dissimilar results (Van den Noortgate & Onghena, in press).

Probably the most frequently used statistical test of the homogeneity of a set of effect sizes is the  $Q$ -test, which was described by Hedges (1982) and by DerSimonian and Laird (1986) and therefore is often referred to as the Hedges' or the DerSimonian and Laird's homogeneity test, although it was proposed before by Cochran (1954).

The test statistic for this test is calculated as

$$Q = \sum_{i=1}^k \frac{(t_i - \bar{t})^2}{\hat{\sigma}_{(t_i)}^2} \quad (1),$$

---

Wim Van den Noortgate is a postdoctoral researcher at the Department of Education at the Katholieke Universiteit Leuven (Belgium). His research interests include multilevel analysis, meta-analysis, resampling methods, single-case designs and item response theory. Email: Wim.VandenNoortgate@ped.kuleuven.ac.be.  
Patrick Onghena is professor of Educational Methodology and Statistics at the Katholieke Universiteit Leuven (Belgium). His research interests include resampling inference, exact nonparametric inference, multilevel analysis, meta-analysis, and single-case designs. Email: Patrick.Onghena@ped.kuleuven.ac.be.

with  $k$  the number of studies,  $t_i$  the observed effect size in study  $i$ ,  $\bar{t}$  the precision weighted mean of the observed effect sizes, with the (estimated) precision of study  $i$  defined as  $1/\hat{\sigma}_{(t_i)}^2$ , and  $\hat{\sigma}_{(t_i)}^2$  the estimate of  $\sigma_{(t_i)}^2$ , the sampling variance of the observed effect size given the ‘true’ effect size in study  $i$ .

Under the null hypothesis of homogeneous effect sizes,  $Q$  follows a  $\chi^2$  distribution with  $k-1$  degrees of freedom, given relatively large study sizes, and given that  $\sigma^2(t_i)$  is independent of  $t_i$  (DerSimonian & Laird, 1986; Takkouche, Cadarso-Suárez, & Spiegelman, 1999).

Although several simulation studies showed the advantages of the  $Q$ -test compared to other kinds of homogeneity tests (e.g., Baydoun, 1995; Sanchez-Meca & Marin-Martinez, 1997; Takkouche, et al., 1999), using the  $Q$ -test is not without problems. Besides the problem that the  $Q$ -test, like other homogeneity tests, suffers from a lack of power (Harwell, 1997; Sanchez-Meca & Marin-Martinez, 1997; Takkouche, et al., 1999), the Type 1 error rate of the  $Q$ -test is not always under control, since the underlying assumptions are usually only approximately met. The degree of the violation of the assumptions, and therefore the behavior of the homogeneity test, depends on the kind of effect size measure that is used and on the conditions under which it is applied.

The proportion of Type 1 errors for instance was found inflated if the  $Q$ -test is used for evaluating the homogeneity of correlation coefficients, but close to the nominal level if the correlation coefficients are first transformed to Fisher’s  $z$ -values (Alexander, Scozzaro, & Borodkin, 1989; Sagie & Koslowsky, 1993; Spector & Levine, 1987). Gavaghan, Moore and McQuay (2000) found a slightly inflated number of Type 1 errors when using the risk difference as a measure of effect size. The results of the  $Q$ -test for Hedges’  $d$  are found highly liberal if used to test the homogeneity of a sample of Hedges’ standardized mean differences ( $d$ ), in case within studies the group sizes and population variances are unequal and the smaller group size is associated with the largest population variance (Harwell, 1997). If under both conditions scores are normally distributed with a common variance,

the  $Q$ -test has been shown slightly conservative, especially if the study sizes are relatively small compared to the number of studies (Hedges & Olkin, 1985; Harwell, 1997).

In the following, we present a parametric bootstrap version of the  $Q$ -test, intended to estimate more closely the reference null distribution of  $Q$  in case the  $\chi^2$ -distribution is inappropriate due to a violation of the underlying assumptions. In a small simulation study, we evaluate the performance of the bootstrap  $Q$ -test for different conditions and different effect size measures.

## Methodology

### A Parametric Bootstrap Version of the $Q$ -test

In the bootstrap, the empirical data are used to estimate the population distribution(s), and samples are simulated from the estimated distribution(s) in order to approximate the sampling distribution of a certain quantity. For the application of the bootstrap procedure to the  $Q$ -test we propose the following procedure:

1. Perform a meta-analysis using techniques for fixed effects models (Hedges & Olkin, 1985), calculate and store the  $Q$ -statistic.
2. Simulate new raw data that could have been observed under the null hypothesis of homogeneity (see below).
3. Calculate for the simulated data of each study the measure of effect size that was used in the initial meta-analytic data set.
4. Perform a meta-analysis on those new effect sizes, calculate and store the  $Q$ -statistic.
5. Repeat step 2-4 a large number of times  $B$ , for instance 1000.
6. Compare the initial  $Q$ -value with the empirical distribution of  $Q$ -values from the  $B$  bootstrap samples. The bootstrap  $p$ -value is the proportion of the  $Q$ -values that is larger than or equal to the initial  $Q$ -value.

In step 2, new raw data are sampled from the estimated population distributions, holding constant the study sizes and the number of studies. A general principle for estimating the population distributions is that for each study the population distributions must show the same effect size (fulfilling the null hypothesis of homogeneity). Furthermore, the population distributions are estimated based on the initial data and additional assumptions. The estimation of the distributions can easily be adapted according to the measure of effect size that is used and to the assumptions one is willing to make.

We give some examples. First, suppose the correlation coefficient is used as the measure of effect size, and data can be assumed bivariate normal. In this case, we can draw new raw data for each study from a bivariate normal distribution. Since the data are used only to calculate the correlation coefficient, means and variances of the distributions can be chosen freely. The population correlation for each bivariate normal distribution is set equal to the overall estimated correlation coefficient. One could for instance draw new data from bivariate normal distributions with zero mean, variances equal to 1 and a covariance equal to the estimated overall correlation coefficient.

As another example, suppose the risk difference or the difference between proportions is used as the effect size. If for each study the proportions for both groups can be retrieved (as is often the case), we can estimate the population proportions under both conditions by means of a precision weighted mean of the observed proportions, assuming equal population proportions in each study. For the bootstrap samples, new data are sampled for each study from two Bernoulli distributions, defined by the estimated population proportions.

Third, if the standardized mean difference is used as a measure of effect size, and raw data under both conditions can be assumed normally distributed with a common variance, for each study data are drawn from two normal distributions with the same variance, and with standardized mean difference that is the same for each study. This standardized mean difference is estimated by the precision-weighted average of the observed effect sizes. One could for instance draw data from  $N(\bar{d}, 1)$  and  $N(0,1)$  for both groups

respectively. Note that drawing data from normal distributions with other variances and means will not alter the results, as long as the variances are equal and the effect size is unchanged, since the raw data are used only to calculate the standardized mean difference.

The situation is somewhat more complicated if the population variances under both conditions cannot be assumed equal. If in the studies the observed within group variance estimates are reported, for study  $i$  these are  $s_{Ai}^2$  and  $s_{Bi}^2$ , one can calculate the pooled within group variance estimate for each study (Hedges, 1981). Multiplying the square root of this pooled variance with the estimated mean standardized mean difference estimate, results for study  $i$  in the estimated study-specific unstandardized mean difference,  $Est(\mu_{Ai} - \mu_{Bi})$ . Raw data can subsequently be drawn from  $N(Est(\mu_{Ai} - \mu_{Bi}), s_{Ai}^2)$  and  $N(0, s_{Bi}^2)$ .

#### A Simulation Study

In order to evaluate the parametric bootstrap version of the  $Q$ -test, we compared its results with the results of the ordinary  $Q$ -test, by means of a small simulation study. Here we show the results of both homogeneity tests for relatively extreme situations, in which (as described above) the ordinary  $Q$ -test has been shown in previous research failing to keep the proportion of Type 1 errors under control. More specifically, we simulated:

- sets of correlation coefficients,
- sets of risks differences,
- sets of standardized mean differences with small group sizes paired with large population variances (called negative paired variances and group sizes by Harwell, 1997),
- large sets of standardized mean differences stemming from small studies, and
- sets of values ("effect sizes") sampled from a normal distribution, with sampling variances independent of the effect sizes, intended as a control condition (see below).



The characteristics of the simulated data sets are summarized in Table 1. The values are chosen such that the situations are comparable with those discussed in previous research. For each of the five situations, we simulated 1000 homogeneous as well as 1000 heterogeneous data sets, 10 000 in total, making possible the assessment of both the proportion of Type 1 and Type 2 errors. The bootstrap as well as the ordinary  $Q$ -test was used for each set to evaluate its heterogeneity. For each data set, we drew 1000 bootstrap samples and calculated  $Q$  for each sample in order to approximate its null distribution. Bootstrap samples were drawn as described above. (Table 1 appears on following page.)

Based on the results of previous research described above, we expect that the proportion of Type 1 errors when using the ordinary  $Q$ -test will be too high in the first three situations, while it will be lower than the nominal level in the fourth situation. When sampling effect sizes from a normal distribution (with a variance that is independent of the effect size), we expect that the proportion of Type 1 errors will be close to the nominal level.

In Figure 1 (following page), histograms present the distributions of the  $p$ -values resulting from the ordinary  $Q$ -test and the bootstrap  $Q$ -test in case of homogeneous data. If the reference distribution is close to the true null distribution, we expect an approximately uniform distribution of the  $p$ -values. This means that under the null hypothesis, we expect that 1% of the  $p$ -values will be smaller than .01, 5% smaller than .05, 10 % smaller than .10 and so on, or otherwise stated, that regardless of the nominal  $\alpha$ -level chosen, the nominal and the actual  $\alpha$ -level correspond.

As expected, the distribution of the  $p$ -values for the ordinary  $Q$ -test is skewed in the first four situations. The ordinary  $Q$ -test gives too much relatively small  $p$ -values when using  $r$ , when using risk differences, or when using  $d$  in case  $n$  and the within group variance are negatively paired, while it yields too much relatively large  $p$ -values when using  $d$  with a small  $N/k$  ratio. This means that for a homogeneous set of effect sizes, the null hypothesis of homogeneity is too often rejected in the first three situations, but less than optimal in the fourth situation. As an example, in Table 2 the proportion of Type 1 errors is presented for a nominal level of .05. Note that in case the

sampling variance of the effect sizes is independent of the effect sizes, the distribution of the  $p$ -values is approximately uniform, and the proportion of Type 1 errors is near to the nominal level.

Figure 1 and Table 2 (following page) furthermore reveal that the  $p$ -values of the bootstrap procedure are approximately uniformly distributed in all situations, yielding a relatively accurate proportion of Type 1 errors, although there seems to be a slightly liberal tendency.

In Table 3, we see that both procedures are equally powerful when testing a set of normally distributed effect sizes with sampling variances that are independent of the effect sizes. In other situations, it is difficult to compare the power of both procedures, because for the ordinary  $Q$ -test the rejection rates are biased since the proportion of Type 1 errors is not under control. Anyway, we see that using the bootstrap procedure instead of the ordinary procedure affects the proportion of rejections in the same way in the homogeneous and the heterogeneous case. In case the  $Q$ -test is used for testing the homogeneity of a set of correlation coefficients, of a set of risk differences, or of a set of standardized mean differences with small group sizes paired with large variances, the proportion of rejections is lower if the bootstrap version is used. In contrast, the bootstrap version of the  $Q$ -test rejects the null hypothesis more often if the homogeneity of a large set of standardized mean differences stemming from small studies is tested.

### Conclusion

Although the  $Q$ -test is very often used in meta-analysis to test the homogeneity of effect sizes, it has been shown in previous research that in several situations the test fails to keep the proportion of Type 1 errors under control. In this article, we therefore present a parametric bootstrap version of the test, which allows freeing one or more assumptions underlying the  $Q$ -test or the calculation of the effect size measures and their sampling distribution. The results of a small simulation study suggest that even in situations where the ordinary  $Q$ -test does not succeed controlling the proportion of Type 1 errors, the Type 1 error rate for the bootstrap version is still close to the nominal level.

Table 1. Characteristics of the simulated data sets.

	Population distribution				
	K	N	Homogeneous case		Heterogeneous case
			80 %		20 %
Correlation coefficient	50	$N=20$	Raw data $\approx N\left(\begin{matrix} 0 \\ 0 \end{matrix}, \begin{bmatrix} 1 & \\ & .50 \end{bmatrix}\right)$	Raw data $\approx N\left(\begin{matrix} 0 \\ 0 \end{matrix}, \begin{bmatrix} 1 & \\ & .45 \end{bmatrix}\right)$	Raw data $\approx N\left(\begin{matrix} 0 \\ 0 \end{matrix}, \begin{bmatrix} 1 & \\ & .55 \end{bmatrix}\right)$
Risk difference	50	$n_A = n_B = n = 50$	Data group A $\approx \text{Bin}(.2, 1)$ Data group B $\approx \text{Bin}(.5, 1)$	Data group A $\approx \text{Bin}(.2, 1)$ Data group B $\approx \text{Bin}(.45, 1)$	Data group A $\approx \text{Bin}(.2, 1)$ Data group B $\approx \text{Bin}(.55, 1)$
Hedges' $d$ , negative pairing	50	$n_A = 10$ $n_B = 20$	Data group A $\approx N(0.6, 2)$ Data group B $\approx N(0, 1)$	Data group A $\approx N(0.3, 2)$ Data group B $\approx N(0, 1)$	Data group A $\approx N(1, 2)$ Data group B $\approx N(0, 1)$
Hedges' $d$ , small N/k	100	$n_A = n_B = n = 5$	Data group A $\approx N(0.5, 1)$ Data group B $\approx N(0, 1)$	Data group A $\approx N(0.1, 1)$ Data group B $\approx N(0, 1)$	Data group A $\approx N(0.8, 1)$ Data group B $\approx N(0, 1)$
Control condition	50	$n_A = n_B = n = 10$	Effect size $\approx N(0.5, 2/n)$	Effect size $\approx N(0.3, 2/n)$	Effect size $\approx N(0.8, 2/n)$

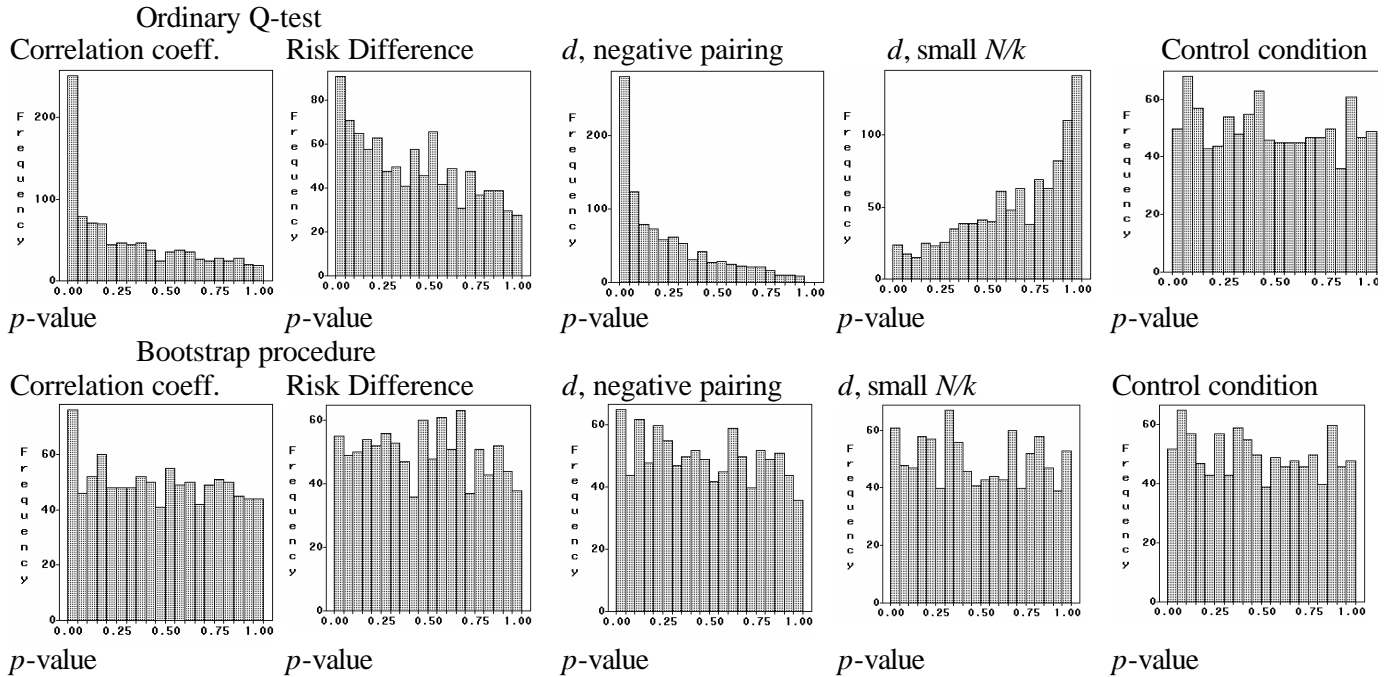


Figure 1. Distribution of the  $p$ -values in case of true homogeneity

Table 2. Rejection rates of the null hypothesis (with a nominal  $\alpha$  of .05) in the homogeneous case (proportion Type 1 errors).

	Correlation coefficient	Risk Difference	$d$ , negative pairing	$d$ , small $N/k$	Control condition
Ordinary	.251	.091	.280	.024	.050
Bootstrap	.076	.055	.065	.061	.052

Table 3. Rejection rates of the null hypothesis (with a nominal  $\alpha$  of .05) in the heterogeneous case (power).

	Correlation coefficient	Risk Difference	$d$ , negative pairing	$d$ , small $N/k$	Control condition
Ordinary	.720	.349	.731	.116	.247
Bootstrap	.347	.258	.367	.302	.252

Moreover, in case the assumptions of the ordinary  $Q$ -test are met, and the test yields appropriate Type 1 error rates, the bootstrap version seems to be equally powerful. A disadvantage of the bootstrap version of the test is that for some situations additional data are required, that may not always be available. E.g., for testing the homogeneity of a set of risk differences, the proportions for each of the groups must be available.

Based on the encouraging results of our simulation study, we suggest comparing the  $Q$ -statistic to the approximate null distribution based on the bootstrap, rather than to a  $\chi^2$ -distribution, whenever possible. Meanwhile however, we note that the power of both versions of the homogeneity test is low and recommend a prudent use of the tests in both modeling and evaluating the heterogeneity.

#### References

- Alexander, R.A., Scozzaro, M.J., & Borodkin, L.J. (1989). Statistical and empirical examination of the chi-square test for homogeneity of correlations in meta-analysis. *Psychological Bulletin*, 106, 329-331.
- Baydoun, R.B. (1995). A Monte Carlo investigation of the Type 1 error rate and power of the Hedges and Olkin moderator search method. *Dissertation Abstracts International: Section B: The Sciences and Engineering*, 55, 4152.
- Cochran, W. G. (1954). The combination of estimates from different experiments. *Biometrics*, 10, 101-129.
- DerSimonian, R., & Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials*, 7, 177-188.
- Gavaghan, D.J., Moore, R.A., & McQuay, H.J. (2000). An evaluation of homogeneity tests in meta-analyses in pain using simulations of individual patient data. *Pain*, 85, 415-424.
- Harwell, M. (1997). An empirical study of Hedges's homogeneity test. *Psychological Methods*, 2, 219-231.
- Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, 6, 107-128.
- Hedges, L. V. (1982). Fitting categorical models to effect sizes from a series of experiments. *Journal of Educational Statistics*, 7, 119-137.
- Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Orlando : Academic Press.
- Hedges, L. V., & Vevea, J. L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods*, 3, 486-504.
- Raudenbush, S. W. (1994). Random effects models. In H. Cooper, & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 301-321). New York: Russell Sage Foundation.

Sagie, A., & Koslowsky, M. (1993). Detecting moderators with meta-analysis: an evaluation and comparison of techniques. *Personnel Psychology, 46*, 629-640.

Sanchez-Meca, J., & Marin-Martinez, F. (1997). Homogeneity tests in meta-analysis: A Monte Carlo comparison of statistical power and Type 1 error. *Quality and Quantity, 31*, 385-399.

Spector, P. E., & Levine, E. L. (1987). Meta-analysis for integrating study outcomes: a Monte Carlo study of its susceptibility to Type I and Type II errors. *Journal of Applied Psychology, 72*, 3-9.

Takkouche, B., Cadarso-Suarez, C., & Spiegelman, D. (1999). Evaluation of old and new tests of heterogeneity in epidemiologic meta-analysis. *American Journal of Epidemiology, 150*, 206-215.

Van den Noortgate, W., & Onghena, P. (In press). Multilevel meta-analysis: A comparison with traditional meta-analytical procedures. *Educational and Psychological Measurement*.

## Randomization Technique, Allocation Concealment, Masking, And Susceptibility Of Trials To Selection Bias

Vance W. Berger  
National Cancer Institute  
& University of Maryland-  
Baltimore County

Costas A. Christophi  
George Washington University

---

It is widely believed that baseline imbalances in randomized clinical trials must necessarily be random. Yet even among masked randomized trials conducted with allocation concealment, there are mechanisms by which patients with specific covariates may be selected for inclusion into a particular treatment group. This selection bias would force imbalance in those covariates, measured or unmeasured, that are used for the patient selection. Unfortunately, few trials provide adequate information to determine even if there was allocation concealment, how the randomization was conducted, and how successful the masking may have been, let alone if selection bias was adequately controlled. In this article we reinforce the message that allocation details should be presented in full. We also facilitate such reporting by identifying and clarifying the role of specific reportable design features. Because the designs that eliminate all selection bias are rarely feasible in practice, our development has important implications for not only the implementation, but also the reporting and interpretation, of randomized clinical trials.

Key words: Baseline imbalance, confounding, masking, randomized clinical trials, validity

---

### Introduction

When lecturing on selection bias, we have addressed audience questions about how selection bias can occur in randomized clinical trials (RCTs). After all, it may be argued, if any subversion occurred, then the trial was not truly randomized. This statement implies that randomization confers absolute protection against any subversion, so that any covariate imbalances must be random. Similar abilities are often ascribed to allocation concealment or masking. Yet the effect of an action may differ from its objective; washed dishes, e.g., may remain dirty; cooked food may remain cold; and treated patients may remain sick.

It is in this light that we critically evaluate the ability of masking, allocation concealment, and randomization *as actually implemented* to produce treatment groups that differ only randomly. If they cannot do so, then observed covariate imbalances may be systematic, and may reflect selection bias. Observed treatment effects could then be attributable to biases, and not to the treatments themselves.

Selection bias can compromise the credibility of standard between-group comparisons, especially when the trial is conducted by a sponsor with a vested interest in the outcome (Hogel & Gaus, 1999). Yet details sufficient to assess the success of randomization, allocation concealment, and masking are rarely reported (Kyriakidi & Ioannidis, 2002).

This draws into question the reliability of the results of many RCTs that have been otherwise well conducted. In fact, if randomization is defined so as to eliminate the possibility of any subversion, then we question whether there has ever been a truly randomized trial. The irony is that until sufficient design details are routinely reported, it will be impossible to quantify the

---

Correspondence should be sent to Vance W. Berger, Ph.D., Biometry Research Group, National Cancer Institute, Executive Plaza North, Suite 3131, 6130 Executive Boulevard, MSC 7354, Bethesda, MD 20892-7354. Phone: (301) 435-5303. Fax: (301) 402-0816. E-mail: vb78c@nih.gov

extent to which selection bias actually occurs in RCTs, yet this lack of reporting is likely due to failure to appreciate the extent to which selection bias occurs in RCTs. Our development clarifies those details that should be presented in RCT reports. It is our hope that more RCT reports will provide these details, and test for selection bias explicitly (Berger & Exner, 1999).

#### What Are Randomization, Allocation Concealment, and Masking?

In a discussion of the distinction between a claim of masking and true masking, Oxtoby et al. (1989) pointed out that “the presumption that a plan to which one has aspired has come to fruition by virtue of aspiration alone is not science, and is particularly inapposite for a profession which should have a reputation for making clear distinctions between fantasy and reality”. This profound remark highlights the distinction between an action and its effect. Masking may be defined as either the process (researchers not revealing treatment codes until the database is locked) or the result (complete ignorance of all trial participants as to which patients received which treatments). A masking claim indicates only the former; this may help to ensure the ignorance of some parties, but is unlikely to ensure the desired state of complete ignorance.

As the legal term “inevitable discovery” suggests, knowledge transfers by various mechanisms. It may be possible to fool all of the people some of the time, or some of the people all of the time, but it is not possible to fool all of the people all of the time. Just as a speed limit is a statement not about how fast drivers drive but rather about how fast they are *encouraged* to drive, so too is a policy of masking a statement not about who knew what (and when) but rather about a process.

Masking is often said to be possible only some of the time, while allocation concealment (Schulz, 1995a,b; 1996), which is essentially the masking of each allocation just until it is executed, is always possible. This confusion of the two definitions is a double-standard. If masking is possible only some of the time, then clearly reference is being made to the result, and not the process.

To be fair, then, one would have to ask if the *result* of allocation concealment is always

possible. Sealed envelopes have been held to lights, phantom patients have been enrolled, and locked files have been raided to determine upcoming treatment allocations in successful subversions of allocation concealment (Schulz, 1995a). Also, it may be clear what a given patient would receive, if enrolled, if cluster randomization (Jordhoy et al., 2002) or minimization (Pocock & Simon, 1975) is used. Drug bottle numbers can also lead to prediction (Kuznetsova, 2002). So only the *process* of allocation concealment, but not its result, can be ensured. Without the result of allocation concealment, selection bias remains a concern.

#### Mechanisms for Selection Bias, and Specific Countermeasures

To focus ideas, we confine our attention to selection bias that interferes with internal validity (a fair comparison, Mark, 1997); we do not consider external validity. Groups of patients to be compared may differ in important ways even before any intervention is applied (Prorok, Hankes, & Bundy, 1981). These baseline imbalances cannot be attributed to the interventions, but they can interfere with and overwhelm the comparison of the interventions (Green & Byar, 1984).

If treatments are independent of patient characteristics, then any baseline imbalances (even if statistically significant) are due to chance variation only. This is one reason often cited for using randomization.

On the other hand, a systematic explanation for the imbalances, known or unknown, would constitute selection bias, even if the imbalances are not statistically significant, or even readily observed (Berger & Exner, 1999). We present a sequence of mechanisms by which selection bias may occur, starting with observational studies in Section A, and such countermeasures as randomization, allocation concealment, and masking (see Table 1).

Table 1: What to Report in Randomized Clinical Trials To Control Selection Bias

<u>Concern</u>	<u>Report</u>
Differential Allocation Discretion	Planned allocation proportions Number of screened and randomized patients by the group to which they were or would have been randomized had they been randomized
Deferred Enrollment	List patients who were screened twice or more, or that there were none
Allocation Concealment	Specific means of concealing the future allocations
Predicted Allocations	Specific restrictions on the randomization (including block sizes) Specific methods of concealing the past allocations (masking) Evidence of unmasking (including differential rates of observable adverse events, any emergencies requiring intentional unmasking, and rates of correct treatment group guesses at de-briefing)
Baseline Imbalances	Compare baseline covariates across treatment groups
Selection Bias	Graph key covariates against P{active}, as in Berger and Exner (1999) Graph response against P{active} within each treatment group, per Berger and Exner (1999). List stratification errors (if any), or that there were none

#### A. Selection Bias in Observational Studies or with Consumer Randomization

Investigators may assign treatments based on patient characteristics (Green & Byar, 1984; Rubin, 1977). Patients may select either their treatment or, with consumer randomization (Bird, 2001), their randomization probability, at least from among a given set of choices. Allocation discretion may be available to the patient, the investigator, both, or neither (dictated allocation). Those patients selecting one treatment or probability may differ systematically from those selecting another (Green & Byar, 1984), so dictated allocation (no freedom of choice) is a countermeasure to prevent patient characteristics from influencing the allocation sequence through either overt treatment assignment based on patient characteristics or self-selection.

#### B. Selection Bias with Dictated Allocation

If allocation is alternated, then either patients with even accession numbers or patients

with odd accession numbers receive the active treatment. The others receive the control. This dictated allocation would prevent the type of selection bias considered in Section A. But with sequential accrual, knowledge of the upcoming treatment, and enrollment discretion (Chalmers, 1990), an investigator could deny enrollment to patients lacking the characteristics that would make them “suitable” to receive the upcoming treatment (Schulz, 1995a; Schulz & Grimes, 2002a).

The selection bias enabled by the predictable allocation sequence (Schulz & Grimes, 2002b) can be controlled by creating instead an unpredictable allocation sequence, or randomizing (Rosenberger & Lachin, 2002). The second countermeasure is the use of actual (not virtual, quasi-, or pseudo-) randomization (Berger & Bears, 2003) to prepare the allocation sequence.

### C. Selection Bias with Dictated Allocation and Randomization

Urn randomization (Wei & Lachin, 1988) is conducted by tossing a (possibly biased) coin each time a patient is to be allocated. Heads indicates active treatment, and tails indicates control. There is no actual allocation discretion, yet having screened and evaluated a given patient, the investigator might exercise *de facto* allocation discretion to reject the toss and repeat until the preferred allocation is observed.

Another mechanism for selection bias with dictated allocation and randomization would be possible if minimization, or dynamic randomization (Pocock & Simon, 1975), were used to force balance with respect to certain covariates. The allocation is determined by minimizing an imbalance function, and randomization may be used to break the ties. So there is both dictated allocation and randomization. Yet because most allocations will be deterministic, it would be possible to determine the allocation to be made once a patient has been identified. A patient enrollment decision may be based on a combination of the treatment to be assigned and values of observed covariates that were not used to define the imbalance function. Randomization is *conventional* if the allocation sequence is generated in advance of screening any patients, and *unconventional* otherwise. Conventional randomization prevents the types of selection bias discussed in this section, and is our third countermeasure.

### D. Selection Bias with Dictated Allocation and Conventional Randomization

As in Section B, selection bias may result from enrollment discretion and advance knowledge of the allocation sequence; the latter may be facilitated by conventional randomization, as the allocation sequence may be posted publicly before patients are screened (Schulz & Grimes, 2002a). A countermeasure to eliminate this advance knowledge is that each allocation be determined only after the patient to be enrolled is identified (Clarke, 2002), as occurs with minimization (Pocock & Simon, 1975). Either the allocation to be made or the patient to be enrolled has to be selected first; whichever it is may influence the other, and the biases possible with unconventional randomization (Section C) are at

least as serious as the biases possible with conventional randomization.

Unconventional randomization may not be able to eliminate advance knowledge of patient characteristics, but one might hope to eliminate advance knowledge of the allocation sequence with conventional randomization and the fourth countermeasure, allocation concealment, which is often taken to mean precisely this lack of advance knowledge. But recall that allocation concealment signifies only that the allocation codes are not intentionally revealed. Even with steps to ensure that these codes cannot be observed, e.g. by holding an envelope to a light (Schulz, 1995a,b), it is not possible to enumerate, and rule out, all mechanisms by which allocations can be observed. We are not prepared to take the success of allocation concealment on faith in an actual trial; we do so for the purpose of this article to demonstrate that even in this unrealistically optimistic case, subversion is still possible.

### E. Selection Bias with (D) and Allocation Concealment

In a randomized depression study of nurse telehealth care (Hunkeler et al., 2000), the initial 40:60 randomization to two groups later became 40:20 to those same two groups, with the remaining 40% allocated to a new third group. If the change in allocation proportions was planned (which need not be the case; see Lippman et al., 2001), then even with allocation concealment it may still be possible to *predict* (but not observe) future allocations. Knowing that more late patients than early patients would be allocated to the third group constitutes advance knowledge of the allocations which, though imperfect, allows for deferred enrollment (Schulz, 1996) of those subjects most “suitable” for the third group until after the new proportions took effect. The fifth countermeasure, then, is the fixed allocation proportions that prevent this.

### F. Selection Bias with (E) and Fixed Allocation Proportions

Randomization is *unrestricted* (Schulz & Grimes, 2002b) if a patient’s likelihood of receiving either treatment is independent of all previous allocations, and is *restricted* (ter Riet & Kessels, 1995) otherwise. The random allocation rule (Schulz & Grimes, 2002b), in which both



treatment groups must be assigned equally often, is one form of restricted randomization, as the final allocation would be determined by the prior ones. Even with allocation concealment and fixed allocation proportions, patterns created by restrictions on the randomization allow prediction of the allocation sequence. Berger and Exner (1999) quantified this extent of advance knowledge with the probability,  $P\{\text{active}\}$ , of a given patient being allocated to the active group given the previous allocations.

With 1:1 allocation,  $P\{\text{active}\}=0.5$  for the first patient; with alternation (Section B),  $P\{\text{active}\}$  is always either 0 or 1. Note that  $P\{\text{active}\}$  reflects the restrictions on the allocation sequences, and becomes a patient characteristic only after that patient is randomized. With enrollment discretion,  $P\{\text{active}\}$  may be used, in conjunction with the estimated potential outcomes of each patient to each treatment, say  $\mathbf{Y}=\{Y(A),Y(C)\}$  for the active and control treatments, respectively, as a basis for enrollment decisions.

Gender, age, race, pre-existing medical conditions, or other baseline characteristics may be considered in deriving the value of  $\mathbf{Y}$  for a given patient. Based on  $\mathbf{Y}$ , the investigator might select a range of  $P\{\text{active}\}$  values for which the patient would be enrolled. If the  $P\{\text{active}\}$  value at the time this patient is screened happens to fall outside of this patient's  $P\{\text{active}\}$  range, then the patient will be denied enrollment, and another patient will be screened. Only when a patient is found with a  $P\{\text{active}\}$  range to match the actual  $P\{\text{active}\}$  value will the patient be enrolled.

Selection bias occurs if the  $P\{\text{active}\}$  range is restricted based on  $\mathbf{Y}$ . It would be possible, e.g., to enroll patients only if  $P\{\text{active}\}$  and  $\mathbf{Y}$  are both large (suppose that larger  $\mathbf{Y}$  values indicate better responses) or both small, but not if they are discordant (Schulz, 1995a). This possibility is depicted in Table 2, using randomized blocks of size four to calculate  $P\{\text{active}\}$  (Berger & Exner, 1999). Notice that not only does treatment assignment for randomized patients depend upon the allocation sequence, but in fact Patients #S5, #S7, #S9, and #S10 may or may not be randomized depending on the allocation sequence, and Patient #S3 cannot get the control.

## Discussion

Few RCT reports make any effort to address the potential for selection bias. Presumably, this is due to unrealistically optimistic definitions of randomization, allocation concealment, and masking. Unfortunately, even in combination, these design features *as implemented* cannot eliminate selection bias. One may argue that while selection bias is possible in theory, its mechanisms are implausible, especially when the main analyses have low p-values.

Unfortunately, history has demonstrated the fallibility of the plausibility test; at best low p-values rule out (probabilistically) chance events, but they do not rule out biases (Berger, 2000; Berger et al., 2000; Grimes and Schulz, 2002). Because of the one-sponsor problem (Hogel & Gaus, 1999) and the vested interest the one sponsor usually has in the outcome of the trial, the best way to offer a convincing argument that a trial was free of a certain bias is to eliminate the possibility of its occurrence. Hence, the burden needs to be on the researchers to demonstrate the reliability of their results. In this article we have presented a number of countermeasures, few combinations of which would eliminate the potential for selection bias. In most cases, then, it is unrealistically optimistic to believe that RCTs are insulated from severe bias (Schulz, 1996).

We are hopeful that the information presented in Table 1 will accompany reports of future trials, preferably in the text of the article, but possibly in an accompanying web site. Such transparency would enable readers to determine the extent to which various mechanisms for selection bias were possible in a given trial, and the extent to which it appears as though there actually was selection bias. The refined measures of trial quality could be used in determining the extent to which specific trials influence policy and meta-analyses. This would exert pressure on those who design trials to design better trials. We are hopeful that journal editors, regulators, and granting institutions will rely, in part, on this information to make their important decisions.

Table 2: Selection Bias with Randomization and Allocation Concealment.

S	P{active} Range*	{(A C C A); (C C A A)}		{(A C A C); (C A A C)}	
		P{active}	Randomized	P{active}	Randomized
S1	[0.50,1.00]	0.50	Active	0.50	Active
S2	[0.00,0.33]	0.33	Control	0.33	Control
S3	[1.00,1.00]	0.50	-	0.50	-
S4	[0.00,0.50]	0.50	Control	0.50	Active
S5	[0.50,1.00]	1.00	Active	0.00	-
S6	[0.00,0.50]	0.50	Control	0.00	Control
S7	[0.00,0.50]	0.67	-	0.50	Control
S8	[0.67,1.00]	0.67	Control	0.67	Active
S9	[0.67,1.00]	1.00	Active	0.50	-
S10	[0.00,0.50]	1.00	-	0.50	Active
S11	[0.33,0.67]	1.00	-	0.00	-
S12	[0.00,1.00]	1.00	Active	0.00	Control

\*The range of P{active} values for which the patient gets randomized. P{active} computed according to the formula of Berger and Exner [3] using the randomized block procedure with a fixed block size of four. Not only does treatment assignment for randomized patients depend upon the allocation sequence, but in fact Patients #S5, #S7, #S9, and #S10 may or may not be randomized depending on the allocation sequence, and Patient #S3 cannot get the control.

### References

Berger, V. W. (2000). Pros and cons of permutation tests in clinical trials. *Statistics in Medicine*, 19, 1319-1328.

Berger, V. W., & Bears, J. (2003). When can a clinical trial be called 'randomized'? *Vaccine*, 21, 468-472.

Berger, V. W., & Exner, D. V. (1999). Detecting selection bias in randomized clinical trials. *Controlled Clinical Trials*, 20, 319-327.

Berger, V. W., Lunneborg, C., Ernst, M. D., & Levine, J.G. (2002). Parametric analyses in randomized clinical trials, *Journal of Modern Applied Statistical Methods*, 1(1), 74-82.

Bird, S. M. (2001). Dissemination of decisions on interim analyses needs wider debate. *BMJ*, 323, 1424.

Chalmers, T. C. (1990). Discussion of biostatistical collaboration in medical research by Jonas H. Ellenberg. *Biometrics*, 46, 20-22.

Clarke, M. (2002). Last moment randomization and concealment. *British Medical Journal*, 323, <http://bmj.com/cgi/eletters/323/7310/446/>

Day, S. (1998). Blinding or masking in the encyclopedia of biostatistics. P. Armitage & T. Colton (Eds.), Vol. 1. Chichester: John Wiley and Sons, 410-417.

Green, S. B., & Byar, D. B. (1984). Using observational data from registries to compare treatments: The fallacy of omnimetrics. *Statistics in Medicine*, 3, 361-370.

Grimes, D. A., & Schulz, K. F. (2001). Randomized controlled trials in *contraception*: The need for "consort" guidelines. *Contraception*, 64, 139-142.

Grimes, D. A., & Schulz, K. F. (2002). Bias and causal associations in observational research. *Lancet*, 359, 248-252.

- Hogel, J., & Gaus, W. (1999). The procedure of new drug application and the philosophy of critical rationalism or the limits of quality assurance with good clinical practice. *Controlled Clinical Trials*, 20, 511-518.
- Hunkeler, E. M., Meresman, J. F., & Hargreaves, W. A. (2000). Efficacy of nurse telehealth care and peer support in augmenting treatment of depression in primary care. *Archives of Family Medicine*, 9(8), 700-708.
- Jordhoy, M. S., Fayers, P. M., Ahlner-Elmqvist, M., & Kaasa, S. (2002) Lack of concealment may lead to selection bias in cluster randomized trials of palliative care. *Palliative Medicine*, 16, 43-49.
- Kuznetsova, O. M. (2002). Why permutation is even more important in ivrs drug codes schedule generation than in patient randomization schedule generation. *Controlled Clinical Trials*, 22, 69-71.
- Kyriakidi, M., & Ioannidis, J. P. A. (2002). Design and quality considerations for randomized controlled trials in systematic sclerosis. *Arthritis Care and Research*, 47, 73-81.
- Lippman, S. M., Lee, J. J., & Kurp, D. D. (2001). Randomized phase iii intergroup trial of isotretinoin to prevent second primary tumors in stage i non-small-cell lung cancer. *Journal of the National Cancer Institute*, 93, 605-618.
- Mark, D. H. (1997). Interpreting the term selection bias in medical research. *Family Medicine*, 29(2), 132-136.
- Oxtoby, A., Jones, A., & Robinson, M. (1989). Is your 'double-blind' design truly double-blind? *British Journal of Psychiatry*, 155, 700-701.
- Pocock, S. J., & Simon, R. (1975). Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial. *Biometrics*, 31, 103-115.
- Prorok, P. C., Hankes, B. F., & Bundy, B. N. (1981). Concepts and problems in the evaluation of screening programs. *Journal of Chronic Diseases*, 34, 159-171.
- Rosenberger, W., & Lachin, J. M. (2002). Randomization in clinical trials: Theory and practice. NY: John Wiley and Sons.
- Rubin, D. B. (1977). Assignment to treatment group on the basis of a covariate. *Journal of Educational Statistics*, 2(1), 1-26.
- Schulz, K. F. (1995a). Subverting randomization in controlled trials. *JAMA*, 274, 1456-1458.
- Schulz, K. F. (1995b). Unbiased research and the human spirit: The challenges of randomized controlled trials. *Canadian Medical Association Journal*, 153, 783-786.
- Schulz, K. F. (1996). Randomised Trials, Human Nature, and Reporting Guidelines. *Lancet*, 348, 596-598.
- Schulz, K. F., & Grimes, D. A. (2002a). Allocation concealment in randomized trials: defending against deciphring. *Lancet*, 359, 614-618.
- Schulz, K. F., & Grimes, D. A. (2002b). Generation of allocation sequences in randomized trials: chance, not choice. *Lancet*, 359, 515-519.
- ter Riet, G., & Kessels, A. G. H. (1995). Restricted randomization in randomized controlled trials. *JAMA*, 274, 1835.
- Wei, L. J., & Lachin, J. M. (1988). Properties of the urn randomization in clinical trials. *Controlled Clinical Trials*, 9, 345-364.

## Screening Properties And Design Selection Of Certain Two-Level Designs

H. Evangelaras      C. Koukouvinos  
Department of Mathematics  
National Technical University of Athens

---

Screening designs are useful for situations where a large number of factors ( $q$ ) is examined but only few ( $k$ ) of these are expected to be important. It is of practical interest for a given  $k$  to know all the inequivalent projections of the design into the  $k$  dimensions. In this paper we give all the inequivalent projections of inequivalent Hadamard matrices of order 28 into  $k=3$  and 4 dimensions and furthermore, we give partial results for  $k=5$ . Then, we sort these projections according to their generalized resolution and their generalized aberration.

Key words: Hadamard matrices, inequivalent projections, screening designs, factorial designs, generalized resolution, generalized aberration, generalized wordlength pattern.

---

### Introduction

In the early stages of an experimental situation, a large number of factors is likely to have been identified as possibly having an influence on the response. However, it is believed that only a few of these actually have a substantial effect, a situation known as factor sparsity. The small number of active factors can be identified through a screening experiment. Screening designs are frequently used by experimenters to help understand the impact of a large number of factors in relatively few trials. Traditionally Hadamard matrices have been used for this purpose. A lot of work has been done in this area (see [7, 10, 11, 16]).

A design suitable for screening out the  $k$  relevant factors from the total factors is called a screening design, see [2, 7, 11]. An  $n$ -dimensional Hadamard matrix is an  $n$  by  $n$  matrix of 1's and -1's with  $H^T H = H H^T = nI_n$ .

A Hadamard matrix is said to be *normalized* if it has its first row and column all 1's. If not we can normalize the Hadamard matrix by multiplying rows and columns by -1 where is needed. In these matrices,  $n$  is necessarily 2 or a multiple of 4. Two Hadamard matrices  $H_1$  and  $H_2$  are called equivalent (or H-equivalent) if one can be obtained from the other by a sequence of row negations, row permutations, column negations and column permutations.

Their usefulness in statistical analysis is as follows. There are two general questions to be answered. (i) If  $q$  factors are to be studied, which  $q$  columns should be assigned to the  $q$  factors? Since any set of  $q$  columns are orthogonal, we must compare them in terms of their ability in entertaining  $m$  two-factor interactions in addition to the  $q$  main effects. (ii) For each assignment, main effect analysis may reveal that only  $k$  factors (i.e.  $k$  columns),  $k \leq q$  are significant.

We can then raise the question (i) for these  $k$  factors. Since the projection onto  $k$  columns varies with the outcome of the analysis, it will be desirable to study this problem for all (or most) projections. The information obtained will be useful for experimenters in contemplating the choice of designs. The choice of  $k$  factors is equivalent to the choice of a  $n \times k$  submatrix of a Hadamard matrix of order  $n$ . Two such matrices are said to be (combinatorially) equivalent if one

---

Correspondence concerning this article should be addressed to H. Evangelaras, Department of Mathematics, National Technical University of Athens, Zografou 15773, Athens, Greece. Email: harris11@central.ntua.gr.

can be obtained from the other by permutation of rows, columns and sign changes in columns. In the context of design theory we refer to this equivalence as (combinatorial) equivalence of two factor assignments.

#### Classification Criteria

Orthogonal factorial designs can be classified into two categories: the regular fractional factorials, that have simple aliasing structure in which any two effects are either orthogonal or fully aliased and the non-regular fractional factorials, that have complex aliasing structure in which effects are neither orthogonal nor fully aliased.

Fractional factorial designs are the most popular experimental designs used in various fields. There are many useful criteria for comparing and ranking fractional factorial designs, such as resolution [2], minimum aberration [6], estimation capacity [3] and uniformity [5]. Among them, the minimum aberration is the most used criterion, but it can be applied only to regular factorials.

It is of practical use to rank and compare non-regular factorial designs in a systematic manner. Deng and Tang [4] proposed *generalized resolution* as a criterion to rank such designs in a similar way as the resolution criterion is used for regular designs. According to this criterion, an orthogonal design is regarded as a set of  $m$  columns  $D = \{d_1, \dots, d_m\}$ . Then, for  $1 \leq k \leq m$  and any  $k$ -subset  $s = \{d_{j_1}, \dots, d_{j_k}\}$  define

$$J_k(s) = |\sum d_{ij_1} \dots d_{ij_k}|.$$

If  $r$  is the smallest integer such that  $\max_{|s|=r} J_r(s) > 0$  and the maximization is over all the subsets of  $r$  distinct columns of  $D$ , then the generalized resolution of  $D$  is defined to be:

$$R(D) = r + [1 - \max_{|s|=r} J_r(s)/n].$$

Then, using simple calculations, we are able to calculate the generalized resolution of any fractional factorial design and therefore we can rank and compare any set of inequivalent projections of Hadamard matrices in any order  $n \equiv 0 \pmod{4}$  and especially when  $n$  is not a power

of 2. Designs with greater generalized resolution from the others are preferred.

The previously stated criterion of generalized resolution is not strong enough to rank such designs since there are cases where two or more fractional factorial designs have the same generalized resolution (see Table 4, where there are 3 such designs with the same generalized resolution). Ma and Fang [12] proposed a stronger criterion that can be applied to all regular and non-regular factorials. Let  $D$  be a fractional factorial design with  $n$  runs and  $s$  factors, each factor in  $q$  levels. The new criterion appends to the design  $D$  its *generalized wordlength pattern*, which is defined by:  $W^g(D) = \{A_1^g(D), \dots, A_s^g(D)\}$  where

$$A_i^g(D) = \frac{1}{n(q-1)} \sum_{j=0}^s P_i(j; s) E_j(D), \quad i=1, \dots, s$$

$P_i(j; s)$  are the Krawtchouk polynomials and  $E_j(D)$ ,  $j=0, \dots, s$  is the distance distribution of  $D$ , defined - in a similar way with Hamming distance- as:

$$E_i(D) = \frac{\#\{(c, d) \mid c, d \in D, d_H(c, d) = i\}}{n}$$

where  $d_H(\mathbf{c}, \mathbf{d})$  is the Hamming distance between two runs  $\mathbf{c}$  and  $\mathbf{d}$  of  $D$ . For the undefined terms in coding theory, we refer the interested reader to [13] and [15].

Let now  $D_1$  and  $D_2$  be two inequivalent designs. Let  $t$  be the smallest integer for which  $A_t^g(D_1) \neq A_t^g(D_2)$  in their generalized wordlength patterns. Then, if  $A_t^g(D_1) < A_t^g(D_2)$  we say that  $D_1$  has less generalized aberration from  $D_2$  and hence it is preferred. A design  $D$  has minimum generalized aberration if no other design has less generalized aberration than it.

By an algorithm which relies on the definition, we have found all the inequivalent projections for  $n=28$ ,  $k=3, 4$  and  $5$  as well as their frequencies. Then by simple computations, we sort these projections according to their generalized resolution and aberration in order to present the best classification.

Inequivalent Hadamard Matrices Of Order 28 And Their Projections

We know that by adding a column of 1's to a Plackett and Burman design [14], we obtain a Hadamard matrix H which satisfies  $H^T H = nI$ . For  $n=12$ , H is unique, but for higher n this is not true. Inequivalent Hadamard matrices have different projection properties.

For  $n=28$  there are 487 inequivalent Hadamard matrices [8, 9] but only one of them corresponds to a Plackett and Burman design designated as H28.487 here, that is, only one provides a 28-run design of the type whose projections are widely known and studied [1], [11]. We will now discuss the projection patterns of all the types, which we designate as  $H28.1, H28.2 \dots H28.487$  as found in <http://www.research.att.com/~njas/hadamard/>.

From now on, in this paper we will denote each projection with  $(k.\#)$  where k are the factors included in the projection and # is the number of the projection. We present each projection as a set of k vectors to save space. In each such vector we have used the letters from A to Z to denote the position of the +1 in each column but since these letters are 26, we need two more characters for the positions 27 and 28. So, we used # for position 27 and \* for position 28. For example, the vector ABEGIJLORUVXZ\* applies to the +++-+-----+ column.

For  $k=3$  there are three different possible projections listed in Table 1. All of them contain a  $2^3$  full factorial design.

Table 1: Inequivalent projections of all 28-run inequivalent Hadamard matrices into k=3 dimensions.

No.	Projection
(3.1)	ABEGIJLORUVXZ*, ACDFIKLORTWYZ*, AHIJKLMNOPQRS*
(3.2)	ABEFGKLMOPSTUW, ADEFGHIJNOPWXY, AHIJKLMNOPQRS*
(3.3)	ACEFGKLMNQRVXY, ADEFGHIJNOPWXY, AHIJKLMNOPQRS*

Table 2 shows the generalized resolution and the generalized wordlength pattern of the three inequivalent projections of Hadamard matrices of order 28 in 3 factors. Projection (3.2) has the best properties than the other two and hence it is preferred from the others.

The frequencies of appearance of each projection in every Hadamard matrix are available on request. It is worth mentioning that the Plackett and Burman design does not provide us the projection (3.1).

Table 2: Sorting of the inequivalent projections of Hadamard matrices of order 28 in 3 dimensions according to their generalized resolution and their generalized wordlength pattern.

Projection number	Generalized Resolution	Generalized Wordlength Pattern
(3.2)	3.856	(0, 0, 0.2)
(3.3)	3.571	(0, 0, 0.18)
(3.1)	3.286	(0, 0, 0.51)

For  $k=4$  there are seven different possible projections listed in Table 3. Projection (4.6) contains a full  $2^4$  factorial design while projections (4.2) and (4.5) contain a half fraction of the full  $2^4$  factorial design with defining relation  $I=ABCD$  contrary to the projections (4.3) and (4.7) that contain a half fraction with defining relation  $I=-ABCD$ . Finally, projections (4.1) and (4.4) do not have any geometrical property.

The frequencies of appearance of each projection in every Hadamard matrix are available on request. The Plackett and Burman design does not provide us the projections (4.1) and (4.2). It is also worth to mentioning that over the 90% of the projections in each Hadamard matrix contain a half fraction of the full  $2^4$  factorial design and furthermore, projection (4.6), which is the best under geometric approach as it contains a full  $2^4$  factorial design, can be recognized in more than 50% out of the whole 17550 possible projections of the 27 columns of each Hadamard matrix of order 28 in 4 factors.

Table 3: Inequivalent projections of all 28-run inequivalent Hadamard matrices into  $k=4$  dimensions.

Number	Projection
(4.1)	ABEGIJLORUVXZ*, ACDFIKLORTWYZ*, ADEFHGHIJNOPWXY, AHIJKLMNOPQRS*
(4.2)	ADEFIMNOQTUV#*, ADEGJKPRSTVY#*, ABCDHFJLNPRUV, ABCDGHIKOQSUVY
(4.3)	ADEGJKPRSTVY#*, ABCDEIJMQRSTWX, ABCDGHIKOQSUVY, ACEFGKLMNQRVXY
(4.4)	ABCDHFJLNPRUV, ACEFGKLMNQRVXY, ADEFHGHIJNOPWXY, AHIJKLMNOPQRS*
(4.5)	ABCDGHIKOQSUVY, ABEFGKLMOPSTUW, ACEFGKLMNQRVXY, ADEFHGHIJNOPWXY
(4.6)	ABCDGHIKOQSUVY, ABEFGKLMOPSTUW, ADEFHGHIJNOPWXY, AHIJKLMNOPQRS*
(4.7)	ABEFGKLMOPSTUW, ACEFGKLMNQRVXY, ADEFHGHIJNOPWXY, AHIJKLMNOPQRS*

Table 4 shows the generalized resolution and the generalized wordlength pattern of the seven inequivalent projections of Hadamard matrices of order 28 in 4 factors. The classification has been made firstly by their generalized resolution and then by their generalized wordlength pattern. So, there are three projections with generalized resolution equal to 3.857 but projection (4.6) is the best since it has better generalized wordlength pattern. On the

other hand, projection (4.1) is the worst since it has the least generalized resolution among all.

Table 4: Sorting of the inequivalent projections of Hadamard matrices of order 28 in 4 dimensions according to their generalized resolution and their generalized wordlength pattern.

Projection number	Generalized Resolution	Generalized Wordlength Pattern
(4.6)	3.857	(0, 0, 0.08, 0.02)
(4.5)	3.857	(0, 0, 0.08, 0.18)
(4.2)	3.857	(0, 0, 0.08, 0.51)
(4.7)	3.571	(0, 0, 0.24, 0.02)
(4.3)	3.571	(0, 0, 0.24, 0.18)
(4.4)	3.571	(0, 0, 0.41, 0.02)
(4.1)	3.286	(0, 0, 0.57, 0.02)

For  $k=5$ , we give partial results since the combinatorial equivalence algorithm we applied requires vast computational time which increases rapidly as the number of factors enlarges. In particular, we have studied the problem for only the first thirty matrices listed in <http://www.research.att.com/~njas/hadamard/>. From these Hadamard matrices, 126 inequivalent projections arise and they are listed in Table 5. It is worth mentioning that projections (5.91) and (5.101) contain a  $2^{5-1}_V$  fraction with defining relations  $I=-ABCDE$  and  $I=ABCDE$  respectively.

The classification of these 126 projections under the generalized resolution and aberration criteria is presented in Table 6. From this table one can notice that projection (5.124) is the best under the classification criteria concerned and on the other hand, projections (5.2) and (5.29) are the worst ones under the same criteria. It is worth mentioning that several inequivalent projections have the same generalized resolution and wordlength pattern.

Table 5: Inequivalent projections of all 28-run inequivalent Hadamard matrices into k=5 dimensions.

Number	Projection
(5.1)	ABEFHIKNQSTWZ#, ABDGILMNSTXYZ*, ABCDGHIKNQRUVY, ADEFHGHIJNOPWXY, AHIJKLMNOPQRS*
(5.2)	AHIJKLMTUVWXY#, ABEFGKLMNOPTUV, ACEFGKLMQRSWXY, ADEFHGHIJNOPWXY, AHIJKLMNOPQRS*
(5.4)	ABEGHJMOQSUYZ#, ABCDFHJLOQSTVX, ACEFGKLMQRSWXY, ADEFHGHIJNOPWXY, AHIJKLMNOPQRS*
(5.5)	ACDEHLMNORVWZ#, ACDGJKMNPQTXZ#, ABDFJKMNSVWYZ*, ABCFIMNOQUWX##*, ABCGHLNPSTWY##*
(5.6)	ACDGJKMNPQTXZ#, ABDEHKLQPQUWXZ*, ABCGHLNPSTWY##*, ABCDEIJMPRSTUW, AHIJKLMNOPQRS*
(5.7)	ACDGJKMNPQTXZ#, ABDFJKMNSVWYZ*, ABCGHLNPSTWY##*, ABCDEIJMPRSTUW, ABCDFHJLOQSTVX
(5.8)	ABDEHKLQPQUWXZ*, ACEFHIMPQTVYZ*, ABCGHLNPSTWY##*, ADFGHMPRSUVX##*, AHIJKLMNOPQRS*
(5.9)	ABDEHKLQPQUWXZ*, ACFGHJKORTUWZ*, ADFGHMPRSUVX##*, ABCDFHJLOQSTVX, ADEFHGHIJNOPWXY
(5.10)	ABDEHKLQPQUWXZ*, ADFGHMPRSUVX##*, ABCDEIJMPRSTUW, ACEFGKLMQRSWXY, AHIJKLMNOPQRS*



(5.11)	ABDFJKMNSVWYZ*, ACEGIJLNSUVXZ*, ABCGHLPSTWY##*, ADEGIKOQSTVW##*, ADFGHMPRSUVX##*
(5.12)	ABDFJKMNSVWYZ*, ABCEJKOPRVXY##*, ADEFJLNQRTUY##*, ABCDGHIKNQRUVY, ADEFGHIJNOPWXY
(5.13)	ABDGILMORTXYZ*, ACEFHIMPQTVYZ*, ABCFIMNOQUWX##*, ADEGIKOQSTVW##*, AHIJKLMNOPSRS*
(5.14)	ABDGILMORTXYZ*, ACFGHJKORTUWZ*, ABCEJKOPRVXY##*, ADEGIKOQSTVW##*, AHIJKLMNOPSRS*
(5.15)	ABDGILMORTXYZ*, ABCEJKOPRVXY##*, ABCFIMNOQUWX##*, ADEFJLNQRTUY##*, ABCDFHJLOQSTVX
(5.16)	ABDGILMORTXYZ*, ABCFIMNOQUWX##*, ABCGHLPSTWY##*, ADEFJLNQRTUY##*, ADFGHMPRSUVX##*
(5.17)	ACEFHIMPQTVYZ*, ACEGIJLNSUVXZ*, ABCEJKOPRVXY##*, ADEGIKOQSTVW##*, ABCDEIJMPRSTUW
(5.18)	ACEFHIMPQTVYZ*, ACEGIJLNSUVXZ*, ADEFJLNQRTUY##*, ADFGHMPRSUVX##*, AHIJKLMNOPSRS*
(5.19)	ACEFHIMPQTVYZ*, ACFGHJKORTUWZ*, ABCEJKOPRVXY##*, ABCFIMNOQUWX##*, ABCGHLPSTWY##*
(5.20)	ACEFHIMPQTVYZ*, ACFGHJKORTUWZ*, ABCGHLPSTWY##*, ABCDFHJLOQSTVX, ADEFGHIJNOPWXY

(5.21)	ACEFHIMPQTVYZ*, ADFGHMPRSUVX#*, ABCDEIJMPRSTUW, ABEFGKLMNOPTUV, AHIJKLMNOPQRS*
(5.22)	ACEGIJLNSUVXZ*, ACFGHJKORTUWZ*, ABCEJKOPRVXY#*, ABCGHLNPSTWY#*, AHIJKLMNOPQRS*
(5.23)	ACEGIJLNSUVXZ*, ABCEJKOPRVXY#*, ABCFIMNOQUWX#*, ABCFHJLOQSTVX, ACEFGKLMQRSWXY
(5.24)	ACEGIJLNSUVXZ*, ABCEJKOPRVXY#*, ADEFJLNQRTUY#*, ABCFHJLOQSTVX, ABEFGKLMNOPTUV
(5.26)	ACEGIJLNSUVXZ*, ABCGHLNPSTWY#*, ACEFGKLMQRSWXY, ADEFGHIJNOPWXY, AHIJKLMNOPQRS*
(5.27)	ACFGHJKORTUWZ*, ABCEJKOPRVXY#*, ABCFIMNOQUWX#*, ADFGHMPRSUVX#*, AHIJKLMNOPQRS*
(5.28)	ACFGHJKORTUWZ*, ABCEJKOPRVXY#*, ABCGHLNPSTWY#*, ADEGIKOQSTVW#*, AHIJKLMNOPQRS*
(5.29)	ACFGHJKORTUWZ*, ABCEJKOPRVXY#*, ACEFGKLMQRSWXY, ADEFGHIJNOPWXY, AHIJKLMNOPQRS*
(5.30)	ACFGHJKORTUWZ*, ABCFIMNOQUWX#*, ADEFJLNQRTUY#*, ADFGHMPRSUVX#*, ABCDGHIKNQRUVY
(5.31)	ACFGHJKORTUWZ*, ABCGHLNPSTWY#*, ADEGIKOQSTVW#*, ADFGHMPRSUVX#*, ABCDGHIKNQRUVY

(5.32)	ACFGHJKORTUWZ*, ABCGHLNPSTWY#*, ADEGIKOQSTVW#*, ADFGHMPRSUVX#*, ABEFGKLMNOPTUV
(5.33)	ACFGHJKORTUWZ*, ABCGHLNPSTWY#*, ADFGHMPRSUVX#*, ABCDFHJLOQSTVX, ADEFHGHIJNOPWXY
(5.34)	ABCEJKOPRVXY#*, ABCFIMNOQUWX#*, ADEFJLNQRTUY#*, ABCDEIJMPRSTUW, ADEFHGHIJNOPWXY
(5.35)	ABCEJKOPRVXY#*, ABCFIMNOQUWX#*, ADEGIKOQSTVW#*, ABCDEIJMPRSTUW, ABCDFHJLOQSTVX
(5.37)	ABCEJKOPRVXY#*, ABCFIMNOQUWX#*, ABCDFHJLOQSTVX, ABCDGHIKNQRUVY, ACEFGKLMQRSWXY
(5.38)	ABCEJKOPRVXY#*, ABCGHLNPSTWY#*, ADEGIKOQSTVW#*, ABCDGHIKNQRUVY, ADEFHGHIJNOPWXY
(5.39)	ABCEJKOPRVXY#*, ADEFJLNQRTUY#*, ADEGIKOQSTVW#*, ABCDEIJMPRSTUW, ABEFGKLMNOPTUV
(5.40)	ABCEJKOPRVXY#*, ADEFJLNQRTUY#*, ADEGIKOQSTVW#*, ABCDFHJLOQSTVX, ACEFGKLMQRSWXY

(5.41)	ABCEJKOPRVXY#*, ADEFJLNQRTUY#*, ADFGHMPRSUVX#*, ABCDEIJMPRSTUW, AHIJKLMNOPQRS*
(5.42)	ABCEJKOPRVXY#*, ABCDGHIKNQRUVY, ACEFGKLMQRSWXY, ADEFGHIJNOPWXY, AHIJKLMNOPQRS*
(5.43)	ABCFIMNOQUWX#*, ABCGHLPSTWY#*, ADEFJLNQRTUY#*, ADEGIKOQSTVW#*, ADFGHMPRSUVX#*
(5.44)	ABCFIMNOQUWX#*, ABCGHLPSTWY#*, ADEFJLNQRTUY#*, ABCDEIJMPRSTUW, AHIJKLMNOPQRS*
(5.45)	ABCFIMNOQUWX#*, ABCGHLPSTWY#*, ADEGIKOQSTVW#*, ADFGHMPRSUVX#*, ACEFGKLMQRSWXY
(5.47)	ABCFIMNOQUWX#*, ABCGHLPSTWY#*, ABCDFHJLOQSTVX, ACEFGKLMQRSWXY, ADEFGHIJNOPWXY
(5.48)	ABCFIMNOQUWX#*, ADEFJLNQRTUY#*, ADFGHMPRSUVX#*, ABCDEIJMPRSTUW, ABCDGHIKNQRUVY
(5.49)	ABCFIMNOQUWX#*, ADEGIKOQSTVW#*, ADFGHMPRSUVX#*, ABCDEIJMPRSTUW, AHIJKLMNOPQRS*
(5.50)	ABCFIMNOQUWX#*, ADEGIKOQSTVW#*, ADFGHMPRSUVX#*, ABCDGHIKNQRUVY, ABEFGKLMNOPTUV

(5.51)	ABCFIMNOQUWX#*, ADEGIKOQSTVW#*, ABCDEIJMPRSTUW, ABCDGHIKNQRUVY, ADEFGHIJNOPWXY
(5.52)	ABCFIMNOQUWX#*, ADEGIKOQSTVW#*, ABCDEIJMPRSTUW, ADEFGHIJNOPWXY, AHJKLMNOPS*
(5.53)	ABCFIMNOQUWX#*, ADEGIKOQSTVW#*, ABCDFHJLOQSTVX, ABIEFGKLMNOPTUV, ACEFGKLMQRSWXY
(5.54)	ABCFIMNOQUWX#*, ADEGIKOQSTVW#*, ABCDFHJLOQSTVX, ABIEFGKLMNOPTUV, AHJKLMNOPS*
(5.55)	ABCFIMNOQUWX#*, ADFGHMPRSUVX#*, ABCDFHJLOQSTVX, ABCDGHIKNQRUVY, ACEFGKLMQRSWXY
(5.56)	ABCFIMNOQUWX#*, ADFGHMPRSUVX#*, ABCDFHJLOQSTVX, ACEFGKLMQRSWXY, ADEFGHIJNOPWXY
(5.58)	ABCGHLNPSTWY#*, ADEFJLNQRTUY#*, ADFGHMPRSUVX#*, ABCDEIJMPRSTUW, ADEFGHIJNOPWXY
(5.59)	ABCGHLNPSTWY#*, ADEFJLNQRTUY#*, ADFGHMPRSUVX#*, ABCDEIJMPRSTUW, AHJKLMNOPS*
(5.60)	ABCGHLNPSTWY#*, ADEFJLNQRTUY#*, ABCDEIJMPRSTUW, ABCDGHIKNQRUVY, ADEFGHIJNOPWXY
(5.61)	ABCGHLNPSTWY#*, ADEFJLNQRTUY#*, ABCDFHJLOQSTVX, ACEFGKLMQRSWXY, ADEFGHIJNOPWXY

(5.62)	ABCGHLNPSTWY#*, ADEGIKOQSTVW#*, ADFGHMPRSUVX#*, ABCDEIJMPRSTUW, ABCDFHJLOQSTVX
(5.63)	ABCGHLNPSTWY#*, ADEGIKOQSTVW#*, ADFGHMPRSUVX#*, ABCDEIJMPRSTUW, ADEFGHIJNOPWXY
(5.64)	ABCGHLNPSTWY#*, ADEGIKOQSTVW#*, ADFGHMPRSUVX#*, ABCDGHIKNQRUVY, ACEFGKLMQRSWXY
(5.65)	ABCGHLNPSTWY#*, ADEGIKOQSTVW#*, ADFGHMPRSUVX#*, ABCDGHIKNQRUVY, ADEFGHIJNOPWXY
(5.66)	ABCGHLNPSTWY#*, ADEGIKOQSTVW#*, ADFGHMPRSUVX#*, ABCDGHIKNQRUVY, AHJKLMNOPS*
(5.67)	ABCGHLNPSTWY#*, ADEGIKOQSTVW#*, ABCDEIJMPRSTUW, ABCDFHJLOQSTVX, ADEFGHIJNOPWXY
(5.69)	ABCGHLNPSTWY#*, ADEGIKOQSTVW#*, ABCDFHJLOQSTVX, ABCDGHIKNQRUVY, ADEFGHIJNOPWXY
(5.70)	ABCGHLNPSTWY#*, ADEGIKOQSTVW#*, ABCDGHIKNQRUVY, ABEFGKLMNOPTUV, ADEFGHIJNOPWXY
(5.71)	ABCGHLNPSTWY#*, ADFGHMPRSUVX#*, ABCDEIJMPRSTUW, ABCDFHJLOQSTVX, ADEFGHIJNOPWXY

(5.72)	ABCGHLNPSTWY#*, ADFGHMPRSUVX#*, ABCDEIJMPRSTUW, ABCDFHJLOQSTVX, AHJKLMNOPSRS*
(5.73)	ABCGHLNPSTWY#*, ADFGHMPRSUVX#*, ABCDEIJMPRSTUW, ABCDGHKIQURVY, AHJKLMNOPSRS*
(5.74)	ABCGHLNPSTWY#*, ADFGHMPRSUVX#*, ABCDEIJMPRSTUW, ABIEFGKLMNOPTUV, AHJKLMNOPSRS*
(5.75)	ABCGHLNPSTWY#*, ADFGHMPRSUVX#*, ABCDEIJMPRSTUW, ACEFGKLMQRSWXY, AHJKLMNOPSRS*
(5.76)	ABCGHLNPSTWY#*, ADFGHMPRSUVX#*, ABCDFHJLOQSTVX, ABCDGHKIQURVY, ADEFGHIJNOPWXY
(5.77)	ABCGHLNPSTWY#*, ADFGHMPRSUVX#*, ABCDFHJLOQSTVX, ABIEFGKLMNOPTUV, ADEFGHIJNOPWXY
(5.78)	ABCGHLNPSTWY#*, ADFGHMPRSUVX#*, ABCDFHJLOQSTVX, ACEFGKLMQRSWXY, AHJKLMNOPSRS*
(5.79)	ABCGHLNPSTWY#*, ADFGHMPRSUVX#*, ABCDFHJLOQSTVX, ADEFGHIJNOPWXY, AHJKLMNOPSRS*
(5.80)	ABCGHLNPSTWY#*, ADFGHMPRSUVX#*, ABCDGHKIQURVY, ADEFGHIJNOPWXY, AHJKLMNOPSRS*

(5.81)	ABCGHLNPSTWY#*, ABCDEIJMPRSTUW, ABCDFHJLOQSTVX, ABCDGHKIQNRUVY, ADEFGHIJNOPWXY
(5.82)	ABCGHLNPSTWY#*, ABCDEIJMPRSTUW, ABCDFHJLOQSTVX, ACEFGKLMQRSWXY, ADEFGHIJNOPWXY
(5.83)	ABCGHLNPSTWY#*, ABCDEIJMPRSTUW, ABCDFHJLOQSTVX, ADEFGHIJNOPWXY, AHJKLMNQPORS*
(5.84)	ABCGHLNPSTWY#*, ABCDFHJLOQSTVX, ABCDGHKIQNRUVY, ACEFGKLMQRSWXY, ADEFGHIJNOPWXY
(5.85)	ABCGHLNPSTWY#*, ABCDFHJLOQSTVX, ABCDGHKIQNRUVY, ADEFGHIJNOPWXY, AHJKLMNQPORS*
(5.86)	ABCGHLNPSTWY#*, ABCDGHKIQNRUVY, ABEFGKLMNOPTUV, ACEFGKLMQRSWXY, ADEFGHIJNOPWXY
(5.87)	ABCGHLNPSTWY#*, ABCDGHKIQNRUVY, ACEFGKLMQRSWXY, ADEFGHIJNOPWXY, AHJKLMNQPORS*
(5.88)	ADEFJLNQRTUY#*, ADFGHMPRSUVX#*, ABCDEIJMPRSTUW, ABCDGHKIQNRUVY, ACEFGKLMQRSWXY
(5.89)	ADEFJLNQRTUY#*, ADFGHMPRSUVX#*, ABCDEIJMPRSTUW, ABCDGHKIQNRUVY, AHJKLMNQPORS*
(5.90)	ADEFJLNQRTUY#*, ADFGHMPRSUVX#*, ABCDFHJLOQSTVX, ACEFGKLMQRSWXY, ADEFGHIJNOPWXY



(5.91)	ADEFJLNQRTUY#*, ABCDFHJLOQSTVX, ABCDGHIKNQRUVY, ADEFHGHIJNOPWXY, AHIJKLMNOPQRS*
(5.92)	ADEGIKOQSTVW#*, ADFGHMPRSUVX#*, ABCDEIJMPRSTUW, ABCDFHJLOQSTVX, ABEFGKLMNOPTUV
(5.93)	ADEGIKOQSTVW#*, ADFGHMPRSUVX#*, ABCDEIJMPRSTUW, ABCDFHJLOQSTVX, ACEFGKLMQRSWXY
(5.94)	ADEGIKOQSTVW#*, ADFGHMPRSUVX#*, ABCDEIJMPRSTUW, ABCDFHJLOQSTVX, ADEFHGHIJNOPWXY
(5.95)	ADEGIKOQSTVW#*, ADFGHMPRSUVX#*, ABCDEIJMPRSTUW, ABCDFHJLOQSTVX, AHIJKLMNOPQRS*
(5.96)	ADEGIKOQSTVW#*, ADFGHMPRSUVX#*, ABCDEIJMPRSTUW, ABCDGHIKNQRUVY, AHIJKLMNOPQRS*
(5.97)	ADEGIKOQSTVW#*, ADFGHMPRSUVX#*, ABCDEIJMPRSTUW, ACEFGKLMQRSWXY, ADEFHGHIJNOPWXY
(5.98)	ADEGIKOQSTVW#*, ADFGHMPRSUVX#*, ABCDEIJMPRSTUW, ADEFHGHIJNOPWXY, AHIJKLMNOPQRS*
(5.99)	ADEGIKOQSTVW#*, ABCDEIJMPRSTUW, ABCDGHIKNQRUVY, ACEFGKLMQRSWXY, ADEFHGHIJNOPWXY

(5.100)	ADEGIKOQSTVW#*, ABCDEIJMPRSTUW, ABCDGHIKNQRUVY, ADEFHGHIJNOPWXY, AHIJKLMNOPQRS*
(5.101)	ADEGIKOQSTVW#*, ABCFHJLOQSTVX, ABEFGKLMNOPTUV, ADEFHGHIJNOPWXY, AHIJKLMNOPQRS*
(5.102)	ADFGHMPRSUVX#*, ABCDEIJMPRSTUW, ABCFHJLOQSTVX, ABCDGHIKNQRUVY, ABEFGKLMNOPTUV
(5.103)	ADFGHMPRSUVX#*, ABCDEIJMPRSTUW, ABCFHJLOQSTVX, ABCDGHIKNQRUVY, ADEFHGHIJNOPWXY
(5.104)	ADFGHMPRSUVX#*, ABCDEIJMPRSTUW, ABCFHJLOQSTVX, ABEFGKLMNOPTUV, ACEFGKLMQRSWXY
(5.105)	ADFGHMPRSUVX#*, ABCDEIJMPRSTUW, ABCFHJLOQSTVX, ADEFHGHIJNOPWXY, AHIJKLMNOPQRS*
(5.106)	ADFGHMPRSUVX#*, ABCDEIJMPRSTUW, ABCDGHIKNQRUVY, ABEFGKLMNOPTUV, AHIJKLMNOPQRS*
(5.107)	ADFGHMPRSUVX#*, ABCDEIJMPRSTUW, ABCDGHIKNQRUVY, ACEFGKLMQRSWXY, AHIJKLMNOPQRS*
(5.108)	ADFGHMPRSUVX#*, ABCDEIJMPRSTUW, ABCDGHIKNQRUVY, ADEFHGHIJNOPWXY, AHIJKLMNOPQRS*
(5.109)	ADFGHMPRSUVX#*, ABCDEIJMPRSTUW, ABEFGKLMNOPTUV, ACEFGKLMQRSWXY, ADEFHGHIJNOPWXY

(5.110)	ADFGHMPRSUVX#*, ABCDEIJMPRSTUW, ABEFGKLMNOPTUV, ACEFGKLMQRSWXY, AHIJKLMNOPQRS*
(5.111)	ADFGHMPRSUVX#*, ABCDEIJMPRSTUW, ABEFGKLMNOPTUV, ADEFHGHIJNOPWXY, AHIJKLMNOPQRS*
(5.112)	ADFGHMPRSUVX#*, ABCDEIJMPRSTUW, ACEFGKLMQRSWXY, ADEFHGHIJNOPWXY, AHIJKLMNOPQRS*
(5.113)	ADFGHMPRSUVX#*, ABCFHJLOQSTVX, ABCDGHIKNQRUVY, ABEFGKLMNOPTUV, ADEFHGHIJNOPWXY
(5.114)	ADFGHMPRSUVX#*, ABCFHJLOQSTVX, ABCDGHIKNQRUVY, ACEFGKLMQRSWXY, AHIJKLMNOPQRS*
(5.115)	ADFGHMPRSUVX#*, ABCFHJLOQSTVX, ABCDGHIKNQRUVY, ADEFHGHIJNOPWXY, AHIJKLMNOPQRS*
(5.116)	ADFGHMPRSUVX#*, ABCFHJLOQSTVX, ABEFGKLMNOPTUV, ACEFGKLMQRSWXY, ADEFHGHIJNOPWXY
(5.117)	ADFGHMPRSUVX#*, ABCFHJLOQSTVX, ABEFGKLMNOPTUV, ADEFHGHIJNOPWXY, AHIJKLMNOPQRS*
(5.118)	ADFGHMPRSUVX#*, ABCFHJLOQSTVX, ACEFGKLMQRSWXY, ADEFHGHIJNOPWXY, AHIJKLMNOPQRS*
(5.119)	ADFGHMPRSUVX#*, ABCDGHIKNQRUVY, ABEFGKLMNOPTUV, ACEFGKLMQRSWXY, ADEFHGHIJNOPWXY

(5.120)	ADFGHMPRSUVX#*, ABCEFGKLMNOPTUV, ACEFGKLMQRSWXY, ADEFGHIJNOPWXY, AHIJKLMNOPQRS*
(5.121)	ABCDEIJMPRSTUW, ABCDFHJLOQSTVX, ABCDGHIKNQRUVY, ABCEFGKLMNOPTUV, AHIJKLMNOPQRS*
(5.122)	ABCDEIJMPRSTUW, ABCDFHJLOQSTVX, ABCDGHIKNQRUVY, ADEFGHIJNOPWXY, AHIJKLMNOPQRS*
(5.123)	ABCDFHJLOQSTVX, ABCDGHIKNQRUVY, ABCEFGKLMNOPTUV, ACEFGKLMQRSWXY, ADEFGHIJNOPWXY
(5.124)	ABCDFHJLOQSTVX, ABCDGHIKNQRUVY, ABCEFGKLMNOPTUV, ADEFGHIJNOPWXY, AHIJKLMNOPQRS*
(5.125)	ABCDFHJLOQSTVX ABCDGHIKNQRUVY ACEFGKLMQRSWXY ADEFGHIJNOPWXY AHIJKLMNOPQRS*
(5.126)	ABCDGHIKNQRUVY ABCEFGKLMNOPTUV ACEFGKLMQRSWXY ADEFGHIJNOPWXY AHIJKLMNOPQRS*

Table 6: Sorting of the inequivalent projections of Hadamard matrices of order 28 in 5 dimensions according to their generalized resolution and their generalized wordlength pattern.

Projection number	Generalized Resolution	Generalized Wordlength Pattern
(5.124)	3.857	(0, 0, 0.204, 0.102, 0)
(5.91)	3.857	(0, 0, 0.204, 0.102, 0.082)
(5.104)	3.857	(0, 0, 0.204, 0.102, 0.082)
(5.101)	3.857	(0, 0, 0.204, 0.102, 0.327)
(5.102)	3.857	(0, 0, 0.204, 0.265, 0)
(5.119)	3.857	(0, 0, 0.204, 0.265, 0.082)
(5.122)	3.857	(0, 0, 0.204, 0.265, 0.082)
(5.121)	3.857	(0, 0, 0.204, 0.265, 0.327)
(5.114)	3.857	(0, 0, 0.204, 0.429, 0)
(5.123)	3.857	(0, 0, 0.204, 0.429, 0)
(5.109)	3.857	(0, 0, 0.204, 0.429, 0.082)
(5.93)	3.857	(0, 0, 0.204, 0.592, 0)
(5.88)	3.857	(0, 0, 0.204, 0.592, 0.082)
(5.92)	3.857	(0, 0, 0.204, 0.592, 0.082)
(5.117)	3.571	(0, 0, 0.367, 0.102, 0)
(5.61)	3.571	(0, 0, 0.367, 0.102, 0.082)
(5.108)	3.571	(0, 0, 0.367, 0.102, 0.082)
(5.113)	3.571	(0, 0, 0.367, 0.102, 0.082)
(5.34)	3.571	(0, 0, 0.367, 0.102, 0.327)
(5.40)	3.571	(0, 0, 0.367, 0.102, 0.327)
(5.86)	3.571	(0, 0, 0.367, 0.265, 0)
(5.90)	3.571	(0, 0, 0.367, 0.265, 0)
(5.106)	3.571	(0, 0, 0.367, 0.265, 0)
(5.107)	3.571	(0, 0, 0.367, 0.265, 0)
(5.116)	3.571	(0, 0, 0.367, 0.265, 0)
(5.58)	3.571	(0, 0, 0.367, 0.265, 0.082)
(5.97)	3.571	(0, 0, 0.367, 0.265, 0.082)
(5.103)	3.571	(0, 0, 0.367, 0.265, 0.082)
(5.111)	3.571	(0, 0, 0.367, 0.265, 0.082)
(5.115)	3.571	(0, 0, 0.367, 0.265, 0.082)
(5.78)	3.571	(0, 0, 0.367, 0.429, 0)
(5.96)	3.571	(0, 0, 0.367, 0.429, 0)
(5.41)	3.571	(0, 0, 0.367, 0.429, 0.082)
(5.99)	3.571	(0, 0, 0.367, 0.429, 0.082)
(5.100)	3.571	(0, 0, 0.367, 0.429, 0.082)
(5.110)	3.571	(0, 0, 0.367, 0.429, 0.082)
(5.45)	3.571	(0, 0, 0.367, 0.429, 0.327)
(5.89)	3.571	(0, 0, 0.367, 0.592, 0)
(5.95)	3.571	(0, 0, 0.367, 0.592, 0)
(5.47)	3.571	(0, 0, 0.531, 0.102, 0)
(5.73)	3.571	(0, 0, 0.531, 0.102, 0)
(5.83)	3.571	(0, 0, 0.531, 0.102, 0)
(5.84)	3.571	(0, 0, 0.531, 0.102, 0)
(5.98)	3.571	(0, 0, 0.531, 0.102, 0)

## 105 SCREENING PROPERTIES AND DESIGN SELECTION OF TWO-LEVEL DESIGNS

(5.49)	3.571	(0, 0, 0.531, 0.102, 0.082)
(5.80)	3.571	(0, 0, 0.531, 0.102, 0.082)
(5.82)	3.571	(0, 0, 0.531, 0.102, 0.082)
(5.85)	3.571	(0, 0, 0.531, 0.102, 0.082)
(5.105)	3.571	(0, 0, 0.531, 0.102, 0.082)
(5.50)	3.571	(0, 0, 0.531, 0.265, 0)
(5.52)	3.571	(0, 0, 0.531, 0.265, 0)
(5.74)	3.571	(0, 0, 0.531, 0.265, 0)
(5.77)	3.571	(0, 0, 0.531, 0.265, 0)
(5.35)	3.571	(0, 0, 0.531, 0.265, 0.082)
(5.44)	3.571	(0, 0, 0.531, 0.265, 0.082)
(5.63)	3.571	(0, 0, 0.531, 0.265, 0.082)
(5.70)	3.571	(0, 0, 0.531, 0.265, 0.082)
(5.75)	3.571	(0, 0, 0.531, 0.265, 0.082)
(5.43)	3.571	(0, 0, 0.531, 0.429, 0)
(5.46)	3.571	(0, 0, 0.531, 0.429, 0)
(5.54)	3.571	(0, 0, 0.531, 0.429, 0)
(5.55)	3.571	(0, 0, 0.531, 0.429, 0)
(5.56)	3.571	(0, 0, 0.531, 0.429, 0)
(5.57)	3.571	(0, 0, 0.531, 0.429, 0)
(5.64)	3.571	(0, 0, 0.531, 0.429, 0)
(5.51)	3.571	(0, 0, 0.531, 0.429, 0.082)
(5.53)	3.571	(0, 0, 0.531, 0.429, 0.082)
(5.60)	3.571	(0, 0, 0.531, 0.429, 0.082)
(5.62)	3.571	(0, 0, 0.531, 0.592, 0)
(5.18)	3.571	(0, 0, 0.531, 0.592, 0.082)
(5.94)	3.571	(0, 0, 0.531, 0.592, 0.082)
(5.59)	3.571	(0, 0, 0.694, 0.102, 0)
(5.76)	3.571	(0, 0, 0.694, 0.102, 0)
(5.17)	3.571	(0, 0, 0.694, 0.102, 0.082)
(5.32)	3.571	(0, 0, 0.694, 0.102, 0.082)
(5.33)	3.571	(0, 0, 0.694, 0.102, 0.082)
(5.38)	3.571	(0, 0, 0.694, 0.102, 0.082)
(5.13)	3.571	(0, 0, 0.694, 0.102, 0.327)
(5.36)	3.571	(0, 0, 0.694, 0.265, 0)
(5.65)	3.571	(0, 0, 0.694, 0.265, 0)
(5.66)	3.571	(0, 0, 0.694, 0.265, 0)
(5.67)	3.571	(0, 0, 0.694, 0.265, 0)
(5.69)	3.571	(0, 0, 0.694, 0.265, 0)
(5.72)	3.571	(0, 0, 0.694, 0.265, 0)
(5.79)	3.571	(0, 0, 0.694, 0.265, 0)
(5.11)	3.571	(0, 0, 0.694, 0.265, 0.082)
(5.37)	3.571	(0, 0, 0.694, 0.265, 0.082)
(5.71)	3.571	(0, 0, 0.694, 0.265, 0.082)
(5.68)	3.571	(0, 0, 0.694, 0.429, 0)
(5.81)	3.571	(0, 0, 0.694, 0.429, 0)
(5.6)	3.571	(0, 0, 0.694, 0.429, 0.082)
(5.10)	3.571	(0, 0, 0.694, 0.429, 0.082)
(5.16)	3.571	(0, 0, 0.694, 0.429, 0.082)

(5.19)	3.571	(0, 0, 0.694, 0.592, 0)
(5.39)	3.571	(0, 0, 0.694, 0.592, 0)
(5.24)	3.571	(0, 0, 0.857, 0.102, 0)
(5.30)	3.571	(0, 0, 0.857, 0.102, 0)
(5.31)	3.571	(0, 0, 0.857, 0.102, 0.082)
(5.25)	3.571	(0, 0, 0.857, 0.265, 0)
(5.15)	3.571	(0, 0, 0.857, 0.265, 0.082)
(5.9)	3.571	(0, 0, 0.857, 0.429, 0)
(5.20)	3.571	(0, 0, 0.857, 0.429, 0)
(5.7)	3.571	(0, 0, 0.857, 0.429, 0.082)
(5.21)	3.571	(0, 0, 0.857, 0.429, 0.082)
(5.48)	3.571	(0, 0, 0.857, 0.592, 0)
(5.8)	3.571	(0, 0, 0.857, 0.592, 0.082)
(5.23)	3.571	(0, 0, 0.857, 0.592, 0.082)
(5.5)	3.571	(0, 0, 1.02, 0.102, 0.082)
(5.120)	3.286	(0, 0, 0.694, 0.102, 0)
(5.27)	3.286	(0, 0, 0.694, 0.102, 0.082)
(5.125)	3.286	(0, 0, 0.694, 0.102, 0.082)
(5.126)	3.286	(0, 0, 0.694, 0.265, 0)
(5.118)	3.286	(0, 0, 0.857, 0.102, 0)
(5.87)	3.286	(0, 0, 0.857, 0.265, 0)
(5.14)	3.286	(0, 0, 0.857, 0.265, 0.082)
(5.28)	3.286	(0, 0, 0.857, 0.265, 0.082)
(5.112)	3.286	(0, 0, 0.857, 0.429, 0)
(5.22)	3.286	(0, 0, 0.857, 0.592, 0)
(5.42)	3.286	(0, 0, 1.02, 0.265, 0)
(5.12)	3.286	(0, 0, 1.02, 0.429, 0)
(5.3)	3.286	(0, 0, 1.02, 0.429, 0.082)
(5.26)	3.286	(0, 0, 1.02, 0.429, 0.082)
(5.4)	3.286	(0, 0, 1.184, 0.265, 0)
(5.1)	3.286	(0, 0, 1.184, 0.592, 0)
(5.2)	3.286	(0, 0, 1.184, 0.592, 0.082)
(5.29)	3.286	(0, 0, 1.184, 0.592, 0.082)

## References

- [1] Box, G. E. P., & Hunter, J. S. (1961). The  $2^{k-p}$  fractional factorial designs. *Technometrics*, 27 (1961), 173–180.
- [2] Box, G. E. P., Hunter, G. E. P., & Hunter, J. S. (1978). *Statistics for experimenters*. NY: Wiley.
- [3] Cheng, C. S., & Mukerjee, R. (1998). Regular fractional factorial designs with minimum aberration and maximum estimation capacity. *Ann. Statist.*, 26, 2289-2300.
- [4] Deng, L.-Y., & Tang, B. (1999). Generalized resolution and minimum aberration criteria for Plackett-Burman and other nonregular factorial designs, *Statistica Sinica*, 9, 1071–1082.
- [5] Fang, K. T., & Mukerjee, R. (2000). Connection between uniformity and aberration in regular fractions of two-level factorials, *Biometrika*, 87, 193-198.
- [6] Fries, A., & Hunter, W. G. (1980). Minimum aberration  $2^{k-p}$  designs, *Technometrics*, 22, 601-608.

- [7] Goh, D., & Street, D. J. (1998). Projective properties of small Hadamard matrices and fractional factorial designs, *J. Combin. Math. Combin. Comput.*, 28, 141-148.
- [8] Kimura, H. (1994). Classification of Hadamard matrices of order 28 with Hall sets, *Discrete Math*, 128, 257-268.
- [9] Kimura, H. (1994). Classification of Hadamard matrices of order 28, *Discrete Math.*, 133, 171-180.
- [10] Lin, D. K. J., & Draper, N. R. (1992). Projection properties of Plackett and Burman designs, *Technometrics*, 34, 423-428.
- [11] Lin, D. K. J., & Draper, N. R. (1995). Screening properties of certain two-level designs, *Metrika*, 42, 99-118.
- [12] Ma, C.-X., & Fang, K.-T. (2001). A note on generalized aberration in factorial designs, *Metrika*, 53, 85-93.
- [13] MacWilliams, F. J., & Sloane, N. J. A. (1977). *The theory of error-correcting codes*. Amsterdam: North Holland Pub. Co.
- [14] Plackett, R. L., & Burman, J. P. (1946). The design of optimum multifactorial experiments, *Biometrika*, 33, 305-325.
- [15] Roman, S. (1992). *Coding and information theory*. NY: Springer-Verlag.
- [16] Wang, J. C., & Wu, C. F. J. (1995). A hidden projection property of Plackett-Burman and related designs, *Statistica Sinica*, 235-250.



## Incorporating Sampling Weights Into The Generalizability Theory For Large-Scale Analyses

Christopher W.T. Chiu  
Law School Admission Council

Ronald S. Fecso  
National Science Foundation

---

Large scale studies frequently use complex sampling procedures, disproportionate sampling weights, and adjustment techniques to account for potential bias due to nonresponses and to ensure that results from the sample can be generalized to a larger population. Survey researchers are concerned about measurement error and the use of weights in developing models. Consequently, multiple weighting factors are used and these weighting factors are manifested as a final survey (composite) weight available for analysis. We developed a method to incorporate an external weighting factor like this for analyses of measurement errors in the theory of generalizability to provide researchers with a tool to evaluate the measurement error components of survey quality and undesirable error components of large-scale assessment programs such as national and state assessments.

Key words: Generalizability theory, large-scale performance assessment, rater reliability, sampling, Survey of Doctorate Recipients (SDR), variance component, weighting

---

### Introduction

The focus of this research is to illustrate how to incorporate weights in the framework of generalizability theory (Brennan, 1992a; Cronbach, Gleser, Nanda, and Rajaratnam, 1972; and Shavelson and Webb, 1991) when it is applied to large-scale studies such as national surveys and educational assessments.

This research is important because educational researchers need to determine variance components and reliability coefficients to accurately reflect measurement errors in statewide or nationwide assessment programs, which often test only a sample of students for accountability purposes. Generalizability theory is a well-known method in educational and psychological research, but today, no one has examined the effect of sample survey data on the method. In addition, survey researchers can use such knowledge to understand, monitor, and improve survey quality. If a weighting scheme was used but researchers ignored the weights in generalizability studies (G studies), as is often the case with such a model, the estimated errors will be biased (Rosenbaum, 1987). In addition, the standard error of the variance component estimates will be inappropriate.

---

Chris W. T. Chiu is a Research Scientist, Psychometrics Group, Law School Admission Council (LSAC), 661 Penn Street, Newtown, PA 18940. Email: cchiu@lsac.org, Ronald S. Fecso is Chief Statistician, National Science Foundation (NSF). This work was partially supported by the American Statistical Association (ASA) through a grant from NSF, Division of Science Resources Statistics (grant number: SRS-0004192). A portion of the research was conducted while Chris Chiu was a professor at the University of Pittsburgh. The authors thank Robert Brennan, Neil Timm, and Loan Tran for their suggestions and comments. Information in this article represents the opinions of the authors and is not NSF, the ASA, the University of Pittsburgh, and the LSAC.

A very popular model in generalizability theory is the two-facet crossed model, which is frequently used in monitoring measurement errors (e.g., Brennan et al., 1995, Brennan, 2000b; Chiu and Wolfe, 2002; Lane et al., 1996) when human judgments are involved. The model can partition error variances into specific sources so that researchers can determine which error source(s) is/are most in need for reduction. For example, one

can determine the score consistency in high-stake examinations where test-takers respond to a set of test questions scored by a group of raters (i.e., a *person x item x rater* two-facet model). Alternatively, one can use a two-facet crossed model (i.e., *respondent x item x coding method*) to determine the coding consistency in survey analysis where survey responses are coded using different schemes (e.g., self-report versus objectively coded responses).

Despite the common applications of the generalizability theory in survey studies (Adam and Ujwal, 1999; Johnson and Bell, 1985; Shipper, et al., 1986), we did not find references discussing how one could incorporate weights into G studies — we searched monographs on G theory (Brennan, 1992a; Brennan, 2001b; Chiu, 2001; Cronbach, et. al., 1972; Fyans, 1983; Shavelson and Webb, 1991) and on variance estimations (Rao, 1997; and Wolter, 1985) using the five major modes of searching: footnote chasing, consultation, searches in subject indices, browsing, and citation searchers (White, 1994). Also, we contacted experts in G theory (Brennan, 2001b; Cronbach, 2000) and searched journal articles and electronic databases (PSYINFO, 1887–2001; ERIC, 1966-2001; MEDLINE, 1966-2001; JSTOR, 1887-1996; Sociological Abstracts, 1963-2001).

In the current study, we first reviewed the purposes and importance of survey weights followed by a summary of the traditional variance component estimation procedures. Second, we discussed the concepts and essential steps of a new weighting method in G studies (i.e., the Chiu-Fecoso G-method, denoted CFG hereafter). Specifically, we used two examples to illustrate the method. The first example was a hypothetical dataset with a context in educational assessment and the other was an operational dataset from a large-scale survey used for research on science and engineering education. (The Survey of Doctorate Recipients is a longitudinal survey administered by the Division of Science Resources Statistics (SRS) at the National Science Foundation (NSF). Details of the survey can be found in the homepage of SRS: <http://www.nsf.gov/sbe/srs>). We intentionally used a simple case in the first example to demonstrate the computational procedures of the new method. The example was simple enough for

hand calculation. The second example, based on an operational dataset from a national study, was used to show the capacity of the method for a real data set. Given the wide applications of the two-facet crossed model, we focus our discussions on the two-facet model throughout the manuscript.

#### Basic Concepts of G Theory and Weighting

An extension of the Classical Test Theory (Crocker and Algina, 1986) and the Analysis of Variance (ANOVA) methods, G theory has been applied to examine the reliability and validity of measurement procedures in educational assessments, psychological measurement, program evaluations, and survey analysis. As Shavelson and Webb (1991) stated:

“The strength of G theory is that multiple sources of error in a measurement can be estimated separately in a single analysis. Consequently, in a manner similar to the way the Spearman-Brown ‘prophecy formula’ is used to forecast reliability as a function of test length in classical test theory, G theory enables the decision maker to determine how many occasions, test forms, and administrators are needed to obtain dependable scores. In the process, G theory provides a summary coefficient reflecting the level of dependability, a generalizability coefficient that is analogous to classical test theory’s reliability coefficient.” (p. 2)

Brennan (1992a, 1992b, and 2000a) and Shavelson and Webb (1991) provided a succinct treatment of the essential features of G theory. Chiu (1999a, 2001) developed a subdividing method to estimate variance components in large-scale performance assessments with missing observations. Brennan (2000a) discussed the misconceptions about the theory. Brennan and Johnson (1995) and Cronbach, Linn, Brennan, and Haertel (1997) covered basic concepts in G theory. Brennan (1997) and Shavelson and Webb (1981) summarized the history of the G theory. Despite the popularity of G theory, all of the

aforementioned studies assumed that simple random sampling was used.

Traditionally, G theory assumes less than or equal to simple random sampling (Bell, 1985; Brennan, 1992a; Cronbach et al., 1972), only that every person has the same probability of being sampled from a population or, that every element is assigned a unit weight. Such an assumption is not viable in national studies where complex sampling procedures (e.g., disproportionate sampling of smaller demographic groups) are used. To create representative estimates in such cases, variable probabilities of selection or variable weights are needed.

Another purpose of weighting is to adjust for the effects of non-respondents (Kish, 1995; Lee, Forthofer, and Lorimer, 1989; and Sarndal, 1980). Bailer, Bailey, and Corby (1978) summarized the purposes and compared some adjustment and weighting procedures (e.g., reweighting, substitution, regression) that were actually used at the US Bureau of the Census, for survey data. The National Science Foundation provided a concise summary of using survey weights, for the Survey of Doctorate Recipients (SDR) — a longitudinal panel survey of individuals who have received their doctorates mainly in the sciences or engineering fields (the data of this survey is used as an example in subsequent sections):

Sampling weights were defined as the reciprocal of the probability of selection for each sampled units, and the weights were adjusted by using weighting class or poststratification adjustment procedures. The final adjusted sampling weights become the analysis weights [also called Final Survey Weights], which have been added to each individual's record in the survey database. (Author, 2002)

Instead of making available multiple weights to researchers, survey developers create a single composite weight also called the final survey weight (e.g., in the Survey of Doctorate Recipients) for analysis. Designed as a proxy for

all the weighting factors in the survey, the Final Survey Weights may be the only weighting information available in the survey data. In this paper, we first derived the methodological adjustments to incorporate such a composite weight on G theory estimation. We then applied the methodology in the context of a large-scale survey to examine the impact of the methodological change and substantively the occupational stability in the engineering profession of the United States. The methodology developed here can be used directly in any crossed design with two facets. The three principles of the weighting method discussed in this paper, however, can be used for other designs with any number of facets. However, our intention is to focus on a two facet crossed design, which has a variety of applications in measurement.

### Methodology

#### Detecting Measurement Errors and Estimating Variance Components

Many have contributed to the methods in monitoring measurement errors and in estimating variance components. In the survey research context, Biemer and Fecso (1995), Rao and Sitter (1997), and Reiser, Fecso, and Chua (1992) discussed methods to characterize measurement errors. In the statistics and educational assessment context, Brennan (1992a), Chiu (1999a, 1999b), Chiu and Wolfe (1997), Corbeil and Searle (1976), Millman and Glass (1967), and Searle, Casella, and McCulloch (1992) among others, provided in-depth discussions on variance component estimation methods. Brennan (1992a) offered an extensive treatment on the topic geared toward generalizability theory. Also, he used synthetic datasets to illustrate the computational steps for variance component estimations. Instead of repeating the details, we summarized the general procedures below and used the summary as building blocks to develop a weighted variance component method based on G theory discussed in the subsequent sections.

In G theory, variance component estimates can be obtained by solving a set of Expected Mean Square (EMS) equations (Brennan, 1992a, chapter 2 and 3; appendices A through B) relating the variance components and mean squares. In the sections that follow, we used a fully crossed two-

faceted design (Brennan, 1992a) as an example. Unless stated otherwise, the universe of admissible observations contains person (p), item (i), and rater (r). The EMS equations can be expressed in the following matrix formula,

$$\hat{\mathbf{s}}^2 = \mathbf{C} \hat{\mathbf{a}}^2 \quad (1)$$

where  $\mathbf{C}$  is an  $f \times f$  upper-triangular matrix of coefficients for the variance components estimated, and  $f = 1, 2, \dots, 7$  represent the seven

variance component estimates in a two faceted design. The column vector  $\hat{\mathbf{a}}^2$  is a set of mean squares for the effects observed in the data (Brennan, 1992a). One can also explicitly write out the elements in  $\mathbf{C}$  and  $\hat{\mathbf{a}}^2$  as follows.

$$\begin{bmatrix} \hat{\mathbf{S}}_p^2 \\ \hat{\mathbf{S}}_i^2 \\ \hat{\mathbf{S}}_r^2 \\ \hat{\mathbf{S}}_{pi}^2 \\ \hat{\mathbf{S}}_{pr}^2 \\ \hat{\mathbf{S}}_{ir}^2 \\ \hat{\mathbf{S}}_{pir}^2 \end{bmatrix} = \begin{bmatrix} (n_i n_r)^{-1} & 0 & 0 & (n_r)^{-1} & (n_i)^{-1} & 0 & 1 \\ 0 & (n_p n_r)^{-1} & 0 & (n_r)^{-1} & 0 & (n_p)^{-1} & 1 \\ 0 & 0 & (n_p n_i)^{-1} & 0 & (n_i)^{-1} & (n_p)^{-1} & 1 \\ 0 & 0 & 0 & (n_r)^{-1} & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & (n_i)^{-1} & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & (n_p)^{-1} & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} MS_p \\ MS_i \\ MS_r \\ MS_{pi} \\ MS_{pr} \\ MS_{ir} \\ MS_{pir} \end{bmatrix} \quad (2)$$

The mean squares vector  $\hat{\mathbf{a}}^2$ , in the above, can be estimated by dividing the set of “sum of squared means” by their corresponding degrees of freedom (Brennan, 1992a, p. 36). We represented such computations using Equation (3), whose elements are explicitly shown in Equation (4).

$$\hat{\mathbf{a}}^2 = \mathbf{D} \mathbf{t} \quad (3)$$

$$\hat{\mathbf{a}}^2 = \begin{bmatrix} MS_p \\ MS_i \\ MS_r \\ MS_{pi} \\ MS_{pr} \\ MS_{ir} \\ MS_{pir} \end{bmatrix} = \begin{bmatrix} (n_p - 1)^{-1} & 0 & 0 & 0 & 0 & 0 & 0 & -(n_p - 1)^{-1} \\ 0 & (n_i - 1)^{-1} & 0 & 0 & 0 & 0 & 0 & -(n_i - 1)^{-1} \\ 0 & 0 & (n_r - 1)^{-1} & 0 & 0 & 0 & 0 & -(n_r - 1)^{-1} \\ \bullet (n_p - 1)^{-1} & \bullet (n_i - 1)^{-1} & 0 & \bullet (n_i - 1)^{-1} & 0 & 0 & 0 & \bullet (n_p - 1)^{-1} \\ \bullet (n_p - 1)^{-1} & 0 & \bullet (n_r - 1)^{-1} & 0 & \bullet (n_p - 1)^{-1} & 0 & 0 & \bullet (n_r - 1)^{-1} \\ 0 & \bullet (n_i - 1)^{-1} & \bullet (n_r - 1)^{-1} & 0 & 0 & \bullet (n_i - 1)^{-1} & 0 & \bullet (n_r - 1)^{-1} \\ \bullet (n_p - 1)^{-1} & \bullet (n_p - 1)^{-1} & \bullet (n_p - 1)^{-1} & \bullet (n_p - 1)^{-1} & \bullet (n_p - 1)^{-1} & \bullet (n_p - 1)^{-1} & \bullet (n_p - 1)^{-1} & \bullet (n_p - 1)^{-1} \\ \bullet (n_i - 1)^{-1} & \bullet (n_i - 1)^{-1} & \bullet (n_i - 1)^{-1} & \bullet (n_i - 1)^{-1} & \bullet (n_i - 1)^{-1} & \bullet (n_i - 1)^{-1} & \bullet (n_i - 1)^{-1} & \bullet (n_i - 1)^{-1} \\ \bullet (n_r - 1)^{-1} & \bullet (n_r - 1)^{-1} & \bullet (n_r - 1)^{-1} & \bullet (n_r - 1)^{-1} & \bullet (n_r - 1)^{-1} & \bullet (n_r - 1)^{-1} & \bullet (n_r - 1)^{-1} & \bullet (n_r - 1)^{-1} \end{bmatrix} \begin{bmatrix} T_p \\ T_i \\ T_r \\ T_{pi} \\ T_{pr} \\ T_{ir} \\ T_{pire} \\ T_m \end{bmatrix} \quad (4)$$

The elements of the **D** matrix in equations (3) and (4) are the sample sizes ( $n_p, n_i, n_r$ ) involved in the seven variance components of the two faceted crossed design. The “sum of squared mean” denoted  $T_f$  is computed for each facet and for the grand mean, such that  $\mathbf{t}=[T_1, \dots, T_f]'$ . The rightmost side of equations (3) and (4),  $\mathbf{t}$ , can be computed by summing individual scores, taking the average, squaring the mean, and multiplying the squared mean by the number of levels in the facet(s) other than the facet for which the sum of squared mean is computed. See equation (5).

$$\mathbf{t} = \begin{bmatrix} T_p \\ T_i \\ T_r \\ T_{pi} \\ T_{pr} \\ T_{ir} \\ T_{p i r e} \\ T_m \end{bmatrix} = \begin{bmatrix} n_i n_r \sum_p \bar{x}_{pr}^2 \\ n_p n_i \sum_i \bar{x}_{ir}^2 \\ n_p n_i \sum_r \bar{x}_{pr}^2 \\ n_r \sum_p \sum_i \bar{x}_{pir}^2 \\ n_i \sum_p \sum_r \bar{x}_{pir}^2 \\ n_p \sum_r \sum_i \bar{x}_{pir}^2 \\ \sum_p \sum_i \sum_r x_{pir}^2 \\ n_p n_i n_r \bar{x}^2 \end{bmatrix} = \begin{bmatrix} n_i n_r \sum_p \left( \frac{1}{n_i n_r} \sum_i \sum_r x_{pir} \right)^2 \\ n_p n_r \sum_i \left( \sum_p \left( \frac{1}{n_r} \sum_r x_{pir} \right) \right)^2 \\ n_p n_i \sum_r \left( \sum_p \left( \frac{1}{n_i} \sum_i x_{pir} \right) \right)^2 \\ n_r \sum_p \sum_i \left( \frac{1}{n_r} \sum_r x_{pir} \right)^2 \\ n_i \sum_p \sum_r \left( \frac{1}{n_i} \sum_i x_{pir} \right)^2 \\ n_p \sum_i \sum_r \sum_p (x_{pir})^2 \\ \sum_p \sum_i \sum_r x_{pir}^2 \\ n_p n_i n_r \left( \frac{1}{n_p n_r} \sum_p \sum_i \sum_r x_{pir} \right)^2 \end{bmatrix} \quad (5)$$

Conceptual Framework of the Chiu-Fecso G-Method

One limitation of the traditional method is that it assumes that every person carries the same weight in an analysis. This assumption is often violated in sample surveys where persons typically receives a different weight as a result of complex sampling and valid response adjustments discussed earlier (See Basic Concepts of G Theory and Weighting). The Chiu-Fecso method enables such a weight (a composite weight supplied to analysts by survey developers and statisticians) to be incorporated in generalizability studies. See Equation (5) for the “sum of squared mean” shown in the  $\mathbf{t}$  vector. Prior to a thorough treatment in computing the weighed sum of squared means, we introduced three fundamental principles used in the Chiu-Fecso G-method.

Multiplication Principle

The summations in Equation (5) simply add up individual scores, assuming that each score occurs once in the data. For example, the total of a set of scores {2, 1, 3, 4} is obtained by  $1 \bullet 2 + 1 \bullet 1 + 1 \bullet 3 + 1 \bullet 4 = 10$ . This approach, assuming that each score received a unit weight, is used in the traditional framework of G theory (Brennan, 1992a, 1992b), discussed in the previous section. The Chiu-Fecso approach relaxed such assumption by allowing each score to have a different weight. This difference is critical when incorporating survey weights for computing the “sum of squared means” because the idea of using survey weights is equivalent to replicating an observed value by the number of times specified in the weights. Rosenbaum (1987) called such weighting approach “direct adjustment.” He pointed out that direct adjustment has two attractive properties: (a) it does not require explicit modeling of the stratification in the sampling design and (b) it produces parallel adjustments in the original statistical procedures so that only little modifications are needed in adapting the original procedures. Consistent with Rosenbaum (1987), Lee, Forthofer, and Lorimor (1989) advocated the use of weights, which they called the weights “expansion weights,” to compute unbiased estimates for means and sums. However, they did not develop a method for variance components. This limitation motivates the current study. To begin, we review the expansion weights. First, assume that the first two scores {2, 1} in the previous example came from a minority group, and each received a composite weight of 49. Further assume that the last two scores came from a majority group and thus received a unit composite weight. The total became  $49 \bullet 2 + 49 \bullet 1 + 1 \bullet 3 + 1 \bullet 4 = 154$ . In the following two sections, we modified the “expansion weight” to obtain the adjusted degrees of freedom (using the Adjustment Principle) and the weighted mean (using the Relative Weighting Principle). These two quantities serve as the building blocks for the weighted variance components discussed in the subsequent section (Computational Equation of the Chiu-Fecso Method).

### Adjustment Principle

The goal of inferential statistics is to determine the extent to which we can infer the results from a sample to a target population. A critical factor in making correct inferences is to determine the correct degrees of freedom reflecting the sample size. In the previous example, a sample size of 4 was collected and each person received a weight assigned by survey developers, statisticians, or policy makers. As shown earlier, if we were to apply the multiplication principle directly, we would obtain a total of 154 ( $49 \bullet 2 + 49 \bullet 1 + 1 \bullet 3 + 1 \bullet 4 = 154$ ). However, this approach is problematic because it assumes that a sample of 100 was collected ( $49+49+1+1$ ). Put differently, this approach erroneously expanded the degrees of freedom. To correct for this problem, we use an adjustment principle so that the weights reflect the actual sample size ( $n = 4$ ) and also the correct degrees of freedom. Such adjustment is accomplished through dividing each weight in the vector of weight  $\mathbf{w} = [49 \ 49 \ 1 \ 1]$  by the mean of the weights ( $\sum w_p/n$ ). After the adjustment, the "adjusted expansion weights" became  $\mathbf{w} / (\sum w_p / n) = [49 \ 49 \ 1 \ 1] / 25 = [1.96 \ 1.96 \ 0.04 \ 0.04]$ . Note that the total of the adjusted expansion weights matches the sample size ( $n = 4$ ) and the ratio between the first and third cases remains 49 to 1. In general, the ratios among all the cases remain unchanged.

### Relative Weighting Principle

One way to obtain the weighted mean for a set of values is to add up all the weighted scores in a set and then divided the total by the total weight or the number of scores in the set, ( $\sum wx/\sum w$ ). An alternative is to multiply each unique value of a set of scores by its relative frequency and then add up the products (i.e.,  $\sum f(x) \bullet x$ ). For instance, the weighted average of the previous example is  $0.49 \bullet 2 + 0.49 \bullet 1 + 0.01 \bullet 3 + 0.01 \bullet 4 = 1.54$ , where 0.49 was obtained by dividing the sampling weight for the first case by the total weight of the four cases (i.e.,  $49 / 100$ ). Hereafter we referred to  $f(x)$  as the relative frequency.

With the multiplication principle, the adjustment principle, and the relative weighting principle, we have computed the adjusted total, adjusted degrees of freedom, and adjusted means

in the above sections. Next we introduce the CFG method to analytically compute the weighted variance component estimates.

### Computational Equation of the Chiu-Fecoso Method

An assumption and three steps are involved in our modification of the G theory. We assume that a set of composite weights is given and stored in a row vector  $\mathbf{w}$ . With this set of weights, we first compute the adjusted expansion weights (using the adjustment principle). Second, we compute the relative weights based on the adjusted expansion weights (using the relative weighting principle). Third, we apply two decision rules to determine when and how to use the two sets of weights obtained in steps 1 and 2.

#### Step 1: Compute Adjusted Expansion Weights

In general, a row vector of the adjusted expansion weights ( $\mathbf{w}_p$ ) is obtained by dividing each of the weights in  $\mathbf{w}$  by the mean of all the weights. That is,  $\mathbf{w}_p = [w_1 \ w_2 \ w_3 \ \dots \ w_p] / (\sum w/n)$ .

#### Step 2: Compute Relative Weights

The relative weights, denoted  $\mathbf{w}_{f(p)}$ , are obtained by dividing each of the adjusted expansion weights above by the sum of these weights. That is,  $\mathbf{w}_{f(p)} = [w_{p_1} \ w_{p_2} \ w_{p_3} \ \dots \ w_{p_p}] / (\sum w_p)$ . Since the sum of all the adjusted expansion weight equals to the sample size, an alternative is:  $\mathbf{w}_{f(p)} = [w_{p_1} \ w_{p_2} \ w_{p_3} \ \dots \ w_{p_p}] / n$ .

#### Step 3: Apply Decision Rules

*Rule #1:* When finding the weighted sum in a facet of interest, we pre-multiply the adjusted expansion weighting vector ( $\mathbf{w}_p$ , a row vector) to a set

of scores ( $\mathbf{s}$ , a column vector), resulting in  $\mathbf{w}_p \bullet \mathbf{s}$ .

*Rule #2:* When finding the weighted average score in the facet of interest, we pre-multiply the vector of relative weights to the column vector of scores (i.e.,  $\mathbf{w}_{f(p)} \bullet \mathbf{s}$ ).

How do we apply the two decision rules to the theory of generalizability? We replace all  $\sum_p$  in Equation (5) with  $\sum_p w_p$  when the facet of interest involves the weighting facet (in this case, the Object of Measurement, person); otherwise, we replace  $\sum_p$  in Equation (5) with  $\sum_p w_{f(p)}$ . For

example, the first entry in  $\mathbf{t}$  of Equation (5) is the Object of Measurement (p), which is also the weighting facet, so we insert  $w_p$  to  $\sum_p$ , resulting

$\sum_p w_p$ . In the second entry of  $\mathbf{t}$  of Equation (5), the facet of interest involves item (i) and does not involve the weighting facet (p), so we replace  $\sum_p$

with  $\sum_p w_{f(p)}$ . By the same token, we apply the same rule to the remaining entries in  $\mathbf{t}$  of Equation (5). Consequently, we have Equation (6). We highlighted  $w_p$  in circle and  $w_{f(p)}$  in square to show where to insert the weights.

$$\mathbf{t}^{(w)} = \begin{bmatrix} T_p \\ T_i \\ T_r \\ T_{pi} \\ T_{pr} \\ T_{ir} \\ T_{pir,e} \\ T_m \end{bmatrix} = \begin{bmatrix} n_i n_r \sum_p \bar{x}_{p.}^2 \\ n_p n_r \sum_i \bar{x}_{i.}^2 \\ n_p n_i \sum_r \bar{x}_{r.}^2 \\ n_r \sum_p \sum_i \bar{x}_{pi.}^2 \\ n_i \sum_p \sum_r \bar{x}_{pr.}^2 \\ n_p \sum_i \sum_r \bar{x}_{ir.}^2 \\ \sum_p \sum_i \sum_r x_{pir}^2 \\ n_p n_i n_r \bar{x}^2 \end{bmatrix} = \begin{bmatrix} n_i n_r \sum_p w_p \left( \frac{1}{n_i n_r} \sum_i \sum_r x_{pir} \right)^2 \\ n_p n_r \sum_i \left( \sum_p w_{f(p)} \left( \frac{1}{n_r} \sum_r x_{pir} \right) \right)^2 \\ n_p n_i \sum_r \left( \sum_p w_{f(p)} \left( \frac{1}{n_i} \sum_i x_{pir} \right) \right)^2 \\ n_r \sum_p w_p \left( \sum_i \left( \frac{1}{n_r} \sum_r x_{pir} \right) \right)^2 \\ n_i \sum_p w_p \left( \sum_r \left( \frac{1}{n_i} \sum_i x_{pir} \right) \right)^2 \\ n_p \sum_i \sum_r \sum_p \left( w_{f(p)} x_{pir} \right)^2 \\ \sum_p w_p \sum_i \sum_r x_{pir}^2 \\ n_p n_i n_r \left( \frac{1}{n_p n_i n_r} \sum_p w_p \sum_i \sum_r x_{pir} \right)^2 \end{bmatrix} \quad (6)$$

where  $w_p$  is the adjusted expansion weight for person  $p$  and  $w_{f(p)}$  is the relative weight for person  $p$ .

With the updated “sum of mean scores” in Equation (6), we obtained the weighted variance component estimates using the following steps. First, compute the weighted “sum of mean scores” vector ( $\mathbf{t}^{(w)}$ ) as shown in Equation (6). Second,

substitute  $\mathbf{t}^{(w)}$  back to Equation (4) to obtain the updated Mean Squares  $[\hat{\mathbf{a}}^2]^{(w)}$ , which in turn is substituted back to equation (2) to obtain weighted variance component estimates  $[\hat{\mathbf{s}}^2]^{(w)}$ . In summary, we estimate the weighted variance component estimates using:

$$\begin{bmatrix} \hat{\mathbf{s}} \\ \mathbf{2} \end{bmatrix}_{7 \times 1}^{(w)} = \mathbf{C} \mathbf{D} \mathbf{t}^{(w)} \quad (7)$$

$\begin{matrix} 7 \times 7 & 7 \times 8 & 8 \times 1 \end{matrix}$

The standard error of the weighted variance components can be obtained by substituting the weighted means squares  $MS_j^{(w)}$ , their coefficients  $c_j$ , and degrees of freedom  $df_j$  into Equation (8). Brennan (1992a) and Chiu (1999a) provided an in-depth discussion for the unweighted standard error equations. Chiu (2001, p. 127, Equations 34 through 40) expressed the standard errors in terms of variance components and the number of levels in each facet. Brennan (1992a, p. 101, equation 6.2.1) provided the general form of the equation. We modified the general equation to incorporate the composite weights as follows:

$$[SE(\hat{\mathbf{s}}_j^2)]^{(w)} = \sqrt{\sum_j \frac{2(c_j MS_j^{(w)})^2}{df_j + 2}} \quad (8)$$

One cautious note to Equation (8) is the distinction between the subscripts  $f$  and  $j$ . The former denotes the  $f^{\text{th}}$  variance component and the latter denotes the  $j^{\text{th}}$  Mean Square term for the  $f^{\text{th}}$  variance component. As shown in Equation (2), each variance component estimate involves a different number of Mean Square terms and for this reason,  $J$ , the total number of mean square terms varies for each variance component estimate. For simplicity and consistency with the G theory literature, we use a single subscript notation  $j$  as opposed to the double subscript notation  $j_f$ , although they are interchangeable in this context.

## Results

### Validation of the Weighted Method

Being able to incorporate weights in generalizability studies are particularly important when the weights differ greatly among the samples. We used a published data set with 10 hypothetical cases and purposely assigned highly disproportionate weights to the data set (one case received a weight of 10 while the rest received a unit weight). As a result, the ratio of the weighted

and unweighted variance component estimates was between 0.3459 and 2.9865, for the seven components, indicating that the weighted estimates could be almost three times larger or three times lower than the unweighted estimates (See Appendix B). Such a result reminds researchers that weighted estimates could be different from their unweighted counterparts when extreme values appear in the weights. The extent to which the two types of estimates would become drastically different depends on the weighting scheme provided in the survey.

We purposely chose an extreme example to contrast the weighted and unweighted results. Such an example is realistic because when applying a two-facet model where test items or tasks are involved, researchers may desire to explore the effect of assigning a much larger weight to one important item — a 300 word essay requiring 45 minutes of testing time may be weighted as much as 10 times more than a multiple-choice question requiring lower than two minutes of testing time.

The aforementioned example (discussed fully in Appendix B) also served as a benchmark comparison between the Chiu-Fecso method and the traditional unweighted method (Brennan, 1992a). Appendix B shows that the unweighted method was a special case of the weighted method because when the weights were set to unity, the CFG method yielded identical variance component estimates to the traditional method.

### Example 1: Performance Assessment

Performance assessment has been popular in the recent decades (Bejar and Braun, 1999; Bennett and Sebrechts, 1996; Braun, Bennett, Frye, and Soloway, 1990; Brennan, 2000b; Chiu, 2001; Clauser, 2000). Many educational and professional testing programs employ constructed-response items to assess performance (e.g. the National Assessment of Educational Progress, the Texas Assessments of Academic Skills, and the United States Medical Licensing Examination). Generalizability analysis is one of the popular techniques to examine the quality of test scores and it can provide guidance regarding the potential to reduce measurement error (Brennan, 2000b; Clauser, 2000). Of the many models in G theory, the two-facet crossed model (Brennan, 2000; Chiu, 2001) is frequently used. Utilizing a two-faceted



model, the following hypothetical data set (3 items  $\times$  2 raters) demonstrates the computational procedures of the Chiu-Fecso method. As shown in the data matrix  $\mathbf{X}$ , each of the four persons has six scores arranged in a row. Columns one through three represent the scores on the three items judged by the first rater; Columns four through six represent the scores on the same three items judged by the second rater. The gap between the third and fourth columns is intended to visually separate the scores for the two raters.

$$\mathbf{X} = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \quad (9)$$

Assume that a final survey weight is derived by survey developers and it is the only weighting information available in the data given to the analyst. Further assume that the weights for the four persons are stored in a row vector [2 3 4 1] which is given to the analyst. We then obtained the adjusted expansion weights and relative weights as follows.

$$\mathbf{w}_p = [0.8 \ 1.2 \ 1.6 \ 0.4] = [2 \ 3 \ 4 \ 1] / ((2 + 3 + 4 + 1) / 4) \text{ and}$$

$$\mathbf{w}_{f(p)} = [0.2 \ 0.3 \ 0.4 \ 0.1] = [0.8 \ 1.2 \ 1.6 \ 0.4] / ((0.8 + 1.2 + 1.6 + 0.4)).$$

With the  $\mathbf{w}_p$  and  $\mathbf{w}_{f(p)}$  computed, we used Equation (6) to obtain  $\mathbf{t}^{(w)}$  as shown below (see Appendix A for the step-by-step illustrations).

$$\mathbf{t}^{(w)} = \begin{bmatrix} T_p \\ T_i \\ T_r \\ T_{pi} \\ T_{pr} \\ T_{ir} \\ T_{p i r e} \\ T_m \end{bmatrix}^{(w)} = \begin{bmatrix} 6.8667 \\ 6.6200 \\ 6.4133 \\ 7.4000 \\ 8.4000 \\ 7.0000 \\ 12.4000 \\ 6.4067 \end{bmatrix} = \begin{bmatrix} 3 \times 2 \times 1.1445 \\ 4 \times 2 \times 0.8275 \\ 4 \times 3 \times 0.5344 \\ 2 \times 3.7000 \\ 3 \times 2.8000 \\ 4 \times 1.7500 \\ 12.4000 \\ 4 \times 3 \times 2 \times 0.2669 \end{bmatrix} \quad (10)$$

By using  $n_p = 4$ ,  $n_j = 3$ , and  $n_r = 2$ , and equation (4), we post-multiplied  $\mathbf{t}^{(w)}$  to  $\mathbf{D}$ . The product became the weighted mean square vector  $[\mathbf{a}^2]^{(w)}$ . See equation (11)

$$[\mathbf{a}^2]^{(w)} = \begin{bmatrix} MS_p \\ MS_i \\ MS_r \\ MS_{pi} \\ MS_{pr} \\ MS_{ir} \\ MS_{pire} \end{bmatrix}^{(w)} = \begin{bmatrix} 0.1533 \\ 0.1067 \\ 0.0067 \\ 0.0533 \\ 0.5089 \\ 0.1867 \\ 0.5156 \end{bmatrix}$$

$$= \begin{bmatrix} (3)^{-1} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -(3)^{-1} \\ 0 & (2)^{-1} & 0 & 0 & 0 & 0 & 0 & 0 & -(2)^{-1} \\ 0 & 0 & (1)^{-1} & 0 & 0 & 0 & 0 & 0 & -(1)^{-1} \\ -(3)^{-1} & -(3)^{-1} & 0 & (3)^{-1} & 0 & 0 & 0 & 0 & (3)^{-1} \\ (2)^{-1} & (2)^{-1} & 0 & (2)^{-1} & 0 & 0 & 0 & 0 & (2)^{-1} \\ -(3)^{-1} & 0 & -(3)^{-1} & 0 & (3)^{-1} & 0 & 0 & 0 & (3)^{-1} \\ (1)^{-1} & 0 & (1)^{-1} & 0 & (1)^{-1} & 0 & 0 & 0 & (1)^{-1} \\ 0 & -(3)^{-1} & -(2)^{-1} & 0 & 0 & (2)^{-1} & 0 & 0 & (2)^{-1} \\ (1)^{-1} & (1)^{-1} & (1)^{-1} & 0 & 0 & (1)^{-1} & 0 & 0 & (1)^{-1} \\ (3)^{-1} & (3)^{-1} & (3)^{-1} & -(3)^{-1} & -(3)^{-1} & -(3)^{-1} & (3)^{-1} & -(3)^{-1} \\ (2)^{-1} & (2)^{-1} & (2)^{-1} & (2)^{-1} & (2)^{-1} & (2)^{-1} & (2)^{-1} & (2)^{-1} \\ (1)^{-1} & (1)^{-1} & (1)^{-1} & (1)^{-1} & (1)^{-1} & (1)^{-1} & (1)^{-1} & (1)^{-1} \end{bmatrix} \begin{bmatrix} 6.8667 \\ 6.6200 \\ 6.4133 \\ 7.4000 \\ 8.4000 \\ 7.0000 \\ 12.4000 \\ 6.4067 \end{bmatrix} \quad (11)$$

Next, we post-multiplied the mean square vector  $[\mathbf{a}^2]^{(w)}$  to the  $\mathbf{C}$  matrix to obtain the variance component estimates. See equation (12). Note that negative variance component estimates occurred in the hypothetical example because we used a randomly generated hypothetical data set, which had only a small sample ( $n_p = 4$ ). Also, for simplicity, no distribution assumptions were specified in generating the data. In practice, one may not obtain negative estimates. Cronbach et. al. (1972) and Brennan (1992a) discussed the causes of negative variance components and developed methods to avoid negative variance component estimates. Those methods include Algorithm 2 (Brennan, 1992a) and Bayesian procedures (see Box and Tiao, 1973; Searle, et al., 1992).

$$\begin{bmatrix} \hat{S}_{p}^2 \\ \hat{S}_{i}^2 \\ \hat{S}_{r}^2 \\ \hat{S}_{pi}^2 \\ \hat{S}_{pr}^2 \\ \hat{S}_{ir}^2 \\ \hat{S}_{pir}^2 \end{bmatrix}^{(w)} = \begin{bmatrix} 0.0178 \\ 0.0478 \\ (-0.0144) \\ (-0.2311) \\ (-0.0022) \\ (-0.0822) \\ 0.5156 \end{bmatrix}$$

$$= \begin{bmatrix} (3 \cdot 2)^{-1} & 0 & 0 & (2)^{-1} & (3)^{-1} & 0 & 1 \\ 0 & (4 \cdot 1)^{-1} & 0 & (2)^{-1} & 0 & (4)^{-1} & 1 \\ 0 & 0 & (4 \cdot 3)^{-1} & 0 & (3)^{-1} & (4)^{-1} & 1 \\ 0 & 0 & 0 & (2)^{-1} & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & (3)^{-1} & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & (4)^{-1} & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 0.1533 \\ 0.1067 \\ 0.0067 \\ 0.0533 \\ 0.5089 \\ 0.1867 \\ 0.5156 \end{bmatrix} \quad (12)$$

Example 2: Large-Scale Survey Analysis

A panel sample of 2388 Engineers was obtained from a longitudinal survey for doctorate recipients. The survey was administered biennially. All survey respondents in the selected sample (a) were under the age 76, in 1999; (b) received at least one research doctorate in Science or Engineering from a U.S. institution in or prior to 1990; (c) were residing in the States on April 15 in four survey years analyzed in the current study (1993, 95, 97, and 99); and (d) were employed in the Engineering profession for at least one of the four aforementioned survey years. The panel of 2388 Engineers represented a population of approximately 50832 Engineers in the U.S. Engineers were broadly defined as those employed in professions such as Aerospace Engineering, Chemical Engineering, Civil and Architectural Engineering, Electrical, Electronic, Computer and Communications Engineering, Industrial Engineering, Mechanical Engineering, Postsecondary Engineering Teaching, and other Engineering fields. Using their age in 1999, the

2388 Engineers with Ph.D degrees can be divided into the following age groups.

Age Groups	Below 30	35-39	40-44	45-49	50-54
Sample Size	3	202	439	400	440
Age Groups	55-59	60-64	65-69	Above 70	
Sample Size	392	256	140	116	

Respondents were given a list of 126 job codes and were asked to choose the most appropriate title for their principal jobs (i.e., self-reported job codes). In addition, the respondents also reported their employment history and background information (e.g., sector of employment, work activities, number of people supervised directly). Such information was used to derive a second measure of occupational title, which was called the “best codes” of occupational titles. The best codes were derived using employment history, job activities, and such. Comprehensive discussions of the best coding process can be found in Hardy and Eisenhower (1994), McGuinness (1997), Rak, Chen, and Gray (1997).

Due to complex sampling and adjustment of nonresponse rate, respondents were selected with a different probability and thus a weighting scheme was used to ensure the representativeness of the sample. The average weight for Engineers was 21.29 (SD = 9.71; median = 22.98; minimum = 1.05; maximum = 46.72).

We conducted a generalizability study with a crossed design (G study, Brennan, 1992a; 1992b) to measure occupational changes. Specifically, we employed the  $p \times y \times m$  design (person  $\times$  year  $\times$  method) in which all survey respondents ( $p$ ) provided their occupational title in all four survey years ( $y$ ). Whether or not one was classified as an Engineer was determined by two methods ( $m$ ), namely the best and self coded methods. The universe of admissible observations (UAO, Brennan, 1992a), therefore, contains 50,832 doctorate recipients who were ever employed in the Engineering profession between 1993 and 1999. For any particular survey year, an

Engineer received a value 1 if s/he was employed in Engineering and a 0 otherwise. The generalizability analysis allowed one to determine the extent to which (1) the professionals were employed the same number of years in Engineering; (2) the Engineering occupation employed a similar number of Ph.D.s across the survey years; (3) survey respondents reported their occupations as consistently as the objectively derived occupation; and (4) the interactions of these three factors.

Similar to Example 1, we estimated seven variance components ( $p, y, m, py, pm, ym, pym,e$ ). Table 1 shows the estimates for the seven variance components and their corresponding standard errors. Both the weighted and unweighted methods yielded very similar results in the point estimate and the standard error of the variance components. For example, the ratio between the unweighted and weighted standard errors of the person effects was close to one because  $0.00299 / 0.00296 = 1.0102$  (i.e.,  $SE[\hat{\mathbf{S}}_p]^2 / SE[\hat{\mathbf{S}}_p^{(w)}]^2$ ).

Table 1: Comparisons of Variance Component Estimates (Weighted VS Unweighted)

	$\hat{\mathbf{S}}_p^2$	$\hat{\mathbf{S}}_y^2$	$\hat{\mathbf{S}}_m^2$	$\hat{\mathbf{S}}_{py}^2$	$\hat{\mathbf{S}}_{pm}^2$	$\hat{\mathbf{S}}_{ym}^2$	$\hat{\mathbf{S}}_{pyme}^2$
	<i>person</i>	<i>year</i>	<i>method</i>	<i>person by year</i>	<i>person by method</i>	<i>year by method</i>	<i>person by year by method, other errors</i>
Weighted	0.0675	0.0002	0.0008	0.0980	0.0047	0.0009	0.0477
Unweighted	0.0690	0.0002	0.0007	0.0969	0.0047	0.0008	0.0471
Ratio	1.0217	0.8984	0.9077	0.9888	0.9970	0.8485	0.9868
Weighted SE	0.0030	0.0006	0.0009	0.0021	0.0005	0.0006	0.0008
Unweighted SE	0.0030	0.0005	0.0008	0.0021	0.0005	0.0005	0.0008
Ratio	1.0102	0.8711	0.8940	0.9884	0.9893	0.8514	0.9868

Note: "Ratio" is the ratio of the unweighted estimates to the weighted estimates. The ratios were computed before the estimates were rounded to four decimal places.

Table 2 shows the percent contribution for each of the variance component estimates. The largest component was  $\hat{\mathbf{S}}_{py}^2$  (0.098), which contributed to approximately 44.6% of the total variance in measuring occupational changes. Such results suggested that one can differentiate those who worked in the Engineering occupations for the same number of year by their job-switching patterns, where a job-switching pattern is characterized by the survey years in which a Ph.D. was employed in the Engineering profession as well as the years the doctorate was employed in other non-Engineering occupations (we summarize job switching patterns below and Chiu and Fecso,

under review, offer an in-depth discussion). For example, two Ph.Ds. can be considered to have a different job-switching pattern even though they were both employed in an Engineering occupation for only one of the four survey years — hypothetically speaking, person A could work in an Engineering profession in 1993 but in a non-engineering profession in the subsequent years (the occupation pattern for person A would be [0 0 0 1], where the first, second, third, and fourth entries are binary variables for an Engineering employment in 1999, 1997, 1995, and 1993, respectively); person B could work in an non-engineering profession prior to becoming an

Engineer in 1999 (person B would have an occupation pattern [1 0 0 0]). Indeed, among the 487 doctorate recipients employed in Engineering for only one of the four survey years, 212 were employed in an Engineering occupation in only 1993; 90 were in only 1995; 61 were in only 1997; and 124 were in only 1999. The aforementioned differential job-switching pattern explained the relatively large  $\hat{S}_{py}^2$ .

Table 2: Comparisons of Variance Component Estimates Weighted VS Unweighted (Percent Contribution)

	$\hat{S}_p^2$	$\hat{S}_y^2$	$\hat{S}_m^2$	$\hat{S}_{py}^2$
Weighted	30.7%	0.1%	0.4%	44.6%
Unweighted	31.4%	0.1%	0.3%	44.2%
	$\hat{S}_{pm}^2$	$\hat{S}_{ym}^2$	$\hat{S}_{pyme}^2$	
Weighted	2.1%	0.4%	21.7%	
Unweighted	2.1%	0.4%	21.5%	

The second large variance component estimate was  $\hat{S}_p^2$ , which indicated that, on average across all survey years and measurement methods, some Engineers had been employed in the profession for a longer duration than the others and the difference in duration accounted for approximately one third (30%) of the total job change variation.

Comparing the number of professionals employed in Engineering in different years can shed light in the stability of the occupation — having a similar number of Engineers across different years can provide some evidence of stability whereas having a drastically different number of Engineers can provide some evidence of instability. The result that  $\hat{S}_y^2$  accounted for only 0.1% of variation of the total job change suggested that the profession employed a similar number of Engineers in the survey years.

Like  $\hat{S}_y^2$ , the  $\hat{S}_m^2$  accounted for only a small portion of total job change variation (0.4%) suggesting the objectively derived (best coding practice) and self-reported methods were relatively consistent in coding the Engineering profession. Resembling the  $\hat{S}_y^2$  and the  $\hat{S}_m^2$ , the  $\hat{S}_{ym}^2$  was

relatively small suggesting that the two measurement methods were implemented consistently across the survey years.

The variance component estimate  $\hat{S}_{pm}^2$ , however, contributed to a larger share (2.1%) of the total variation than  $\hat{S}_y^2$  and  $\hat{S}_m^2$ . One can interpret  $\hat{S}_{pm}^2$  as an interaction between the variations due to person and method. It showed that the two occupational-determining methods were slightly more consistent for some survey respondents than the others but such differential consistency was relatively small comparing to the other sources of variation.

The person-by-year-by-method with any systematic and unsystematic variability  $\hat{S}_{pyme}^2$  accounted for 21.7% of the total variation, suggesting that about one fifth of the job change variability in Engineering was due to: (a) the observation that Engineers changed jobs differentially in different survey years and the extent to which such a differential change occurred depends on which method was used to measure occupational titles; (b) any systematic variability such as the possibility that Engineers in some geographical regions were more mobile; and/ or (c) any unsystematic variability that was not measured.

## Conclusion

The goal of incorporating sampling or survey weights into the framework of generalizability is to ensure that variance component are correctly estimated. The Chiu-Fecso method is designed for this purpose. In practice, the CFG method can be applied to educational assessment, psychological measurement, professional testing, and survey research where generalizability studies are called for to examine desirable variations and undesirable variations (measurement errors). Regardless of its dependence on sampling, the traditional G Theory framework assumes that simple random sampling is used. Indeed, national surveys and large-scale assessment programs use a variety of disproportional sampling techniques to ensure sample representations and account for non-responses. To this end a composite weight (final survey weight) is provided to analysts. Given that

the composite weight is frequently the only weighting information available to analysts, the current study extended the capacity of the G theory so that it can allow weights to be used.

In this article, we first introduced three principles in deriving the weighting method by showing how to estimate means and sums correctly. We then used the same principles to illustrate how to estimate variance components. Rules and step-by-step procedures were discussed. We validated the method using a published data set. The validation study suggested that weighted and unweighted variance component estimates can differ drastically if some cases receive a weight differ drastically from the others. Also, we showed that the traditional generalizability analysis is a special case of the weighted generalizability analysis. Two examples were provided to illustrate the applications of the weighting method in performance assessment and survey analysis. The weighted and unweighted variance component estimates of a large-scale operational data set yielded very similar conclusions.

Although the object of measurement, person, was the weighting facet in the two examples, this is not necessary to be the case. In practice, the weighting facet can be any facet in a crossed-two-faceted design (the main effect facets or the interaction effect facets). For instance, in standardized psychological or educational testing programs, researchers may desire to designate the item facet to be the weighting facet. This can be useful in examining the reliability of test scores when examinees do not respond to all items within the standard time. In the event that speededness happens, researchers can assign a lower weight to “not reached” items (those presented in the end of the test) than items presented in the beginning. Reese (1999) found that the true ability of low performing examinees is overestimated and that of high performing examinees is underestimated, when items are “locally dependent” or not reached by examinees (e.g., due to fatigue). The CFG method discussed in the current paper can be used to assign lower weights to not reached or locally dependent items. Future research can further investigate the extent to which different weights will change the reliability of test scores. Due to the page limits, it is not our intention to examine this topic in the current study.

Sometimes researchers are interested in assigning weights to multiple facets. For example, in educational assessment, one might be interested in oversampling minority students from the target population (i.e., weighting is used to adjust for the design effect). The weights to oversample minority students can be incorporated into a G study by assigning them to the facet related to persons (i.e., the object of measurement, Brennan, 1992a). In addition to assigning weights to the object of measurement, one can also weight the person-by-item facet. This can allow items to be weighted differently for individual students. Such an adaptive weighting mechanism can enable psychometricians to take into consideration the “opportunity to learn” when deciding the importance of an item on the test score. For example, one might assign a lower weight to an item when it is responded by a student whose school does not emphasize the learning objective of the item than when it is responded by another student who came from a school with a strong emphasis on the same item.

Similarly, in survey analysis, statisticians may desire to assign one set of weights to the sample of respondents and a completely different set of weights to the measurement methods. By doing so, survey statisticians could put a stronger emphasis on one measurement method (e.g., objective method) than the other (e.g., self-reported method) in evaluating quality of survey data. The aforementioned goal can be accomplished by developing a method to incorporate weighting schemes into multiple facets of a generalizability study (e.g., person and person-by-item). Future pursuit in developing a multifacet weighting scheme can apply the three principles discussed in the current study.

## References

- Adam, F., & Ujwal, K. (1999). Unmasking a phantom: A psychometric assessment of mystery shopping. *Journal of Retailing*, 75(2), 195-217.
- Author. (2002). *Weighting Strategy*. [On-Line]: <http://srsstats.sbe.nsf.gov/techinfo.html> (Accessed Jan 25, 2002)
- Bailar, B. A., Bailey, L., & Corby, C. (1978). A comparison of some adjustment and weighting procedures for survey data. In N. K. Namboodiri (Ed.), *Survey sampling and measurement*. New York: Academic Press.
- Bejar, I. I., & Braun, H. I. (1999). *Architectural simulations: From research to implementation* (99-2). Princeton, NJ: Educational Testing Service.
- Bell, J. F. (1985). Generalizability theory: The software problem. *Journal of Educational Statistics*, 10 (1), 19-29.
- Bennett, R., E. & Sebrechts, M. M. (1996). The accuracy of expert-system diagnoses of mathematical problem solutions. *Applied Measurement in Education*, 9, 133-150.
- Biemer, P. P., & Fecso, R. S. (1995). Evaluating and controlling measurement error in business surveys. In B. Cox, Chinnappa, Christianson, Colledge, Kott. (Ed.), *Business survey methods* (257-281): Wiley & Sons, Inc.
- Box, G. E.P., & Tiao, G. C. (1973). *Bayesian inference in statistical analysis*. Reading, MA: Addison-Wesley.
- Braun, H. I., Bennett, R. E., Frye, D., & Soloway, E. (1990). Scoring constructed responses using expert systems. *Journal of Educational Measurement*, 27, 93-108.
- Brennan, R. L. (1992a). *Elements of generalizability theory*. Iowa City, IA: American College Testing.
- Brennan, R. L. (1992b). NCME instructional module: Generalizability theory. *Educational measurement issues and practice*, 11(4), 27-34.
- Brennan, R. L. (1997). A perspective on the history of generalizability theory. *Educational measurement: issues and practice*, 16(4), 14-20.
- Brennan, R. L. (2000a). (Mis)Conceptions at about generalizability theory. *Educational Measurement: Issues and Practice*, 19 (1), 5-10.
- Brennan, R. L., Gao, S., & Colton, D. (1995). Generalizability analyses of work keys listening and writing tests. *Educational and Psychological Measurement*, 55(2), 157-176.
- Brennan, R. L. (2000b). Performance assessments from the perspective of generalizability theory. *Applied Psychological Measurement*, 24 (4), 339-353.
- Brennan, R. L. (2001a). *Weights in generalizability theory*. Personal Communication in summer, 2001.
- Brennan, R. L. (2001b). *Generalizability theory*. New York: Springer.
- Brennan, R. L., & Johnson, E. G. (1995). Generalizability of performance assessments. *Educational Measurement : Issues and Practice*, 14 (4), 9-12,27.
- Chiu, C. W. T., & Fecso, S. R. (in press review). *SEER: A graphical tool for multidimensional and categorical data*. *Journal of Data Science*.
- Chiu, C. W. T., & Wolfe, E. W. (1997, April). *Generalizability theory: A new approach to analyze non-crossed performance assessment data*. Paper presented at the American Educational Research Association annual meeting, Chicago, IL.
- Chiu, C. W. T., & Wolfe, E. W. (2002). A Method for Analyzing Sparse Data Matrices in the Generalizability Theory Framework. *Applied Psychological Measurement*.26(3), 319-336.
- Chiu, C. W. T. (1999a). *Scoring performance assessments based on judgments: Utilizing meta-analysis to estimate variance components in generalizability theory for unbalanced situations*. Unpublished Dissertation. Michigan State University, Lansing, MI.
- Chiu, C.W.T. (1999b, April). *Scoring performance assessments*. Poster for the Graduate Student Session at the 1999 Annual Meeting of the National Council on Measurement in Education, Montreal, Canada.
- Chiu, C. W. T. (2001). *Scoring performance assessments based on human judgments: Generalizability theory*: Boston, MA: Kluwer Academic Publisher.
- Clauser, B. E. (2000). Recurrent issues and recent advances in scoring performance assessments. *Applied Psychological Measurement*, 24 (4), 310-324.

- Corbeil, R. R., & Searle, S. R. (1976). Restricted maximum likelihood (REML) estimation of variance components in the mixed model. *Technometrics*, 18 (1), 31-38.
- Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York : Holt, Rinehart, and Winston.
- Cronbach, L. J. (2000). *Weights in generalizability theory*. Personal Communication in summer, 2000.
- Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York, NY: Wiley.
- Cronbach, L. J., Linn, R. L., Brennan, R. L., & Haertel, E. H. (1997). Generalizability analysis for performance assessments of student achievement or school effectiveness. *Educational and Psychological Measurement*, 57 (3), 373-399.
- Fyans, L. J. J. (Ed.). (1983). *Generalizability theory: Inferences and practical applications*. (Vol. 18): Jossey-Bass.
- Hardy, L. P., & Eisenhower, D. L. (1994). *Developing methods for collecting and coding the occupation of persons with college degrees*. Paper presented at the Proceedings of the Section on Survey Research Methods. American Statistical Association, Alexandria, VA.
- Holt, D. E., D. (1991). Methods of weighting for unit non-response. *The statistician*, 40, 333-342.
- Johnson, S., & Bell, J. F. (1985). Evaluating and predicting survey efficiency using generalizability theory. *Journal of Educational Measurement*, 22 (2), 107-119.
- Kish (1995). *Survey sampling*. NY, New York: John Wiley & Son, Inc.
- Lane, S., Liu, M., Ankenmann, R. D., & Stone, C. A. (1996). Generalizability and validity of a mathematics performance assessment. *Journal of Educational Measurement*, 33(1), 71-92.
- Lee, E. S., Forthofer, R. N., & Lorimor, R. J. (1989). *Analyzing complex survey data*. London: Sage.
- McGuinness, R. (1997). *1995 NSCG Coding Quality Evaluation*. Washington, DC: U.S. Department of Commerce, Bureau of the Census.
- Millman, J., & Glass, G. V. (1967). Rules of thumb for writing the ANOVA table. *Journal of Educational Measurement*, 4(2), 41-51.
- Rak, R., Chen, S., & Gray, L. (1997). *Occupation Coding: Best Coding and CATI Coding Methods*. (Research Compendium ). Rockville, MD: Westat, Inc.
- Rao, C. R. (1988). *Estimation of Variance Components and Applications*. New York: Elsevier Science.
- Rao, J. N. K., & Sitter, R. R. (1997). Variance estimation under stratified two-phase sampling with applications to measurement bias. In B. Lyberg, Collins, de Leeuw, Dippo, Schwarz, Trewin (Ed.), *Survey measurement and process quality* (pp. 753-768): John Wiley & Sons, Inc.
- Rao, P. S. R. S. (1997). *Variance components estimation: Mixed models methodologies and applications*. New York: Chapman & Hall.
- Reese, L. M. (1999). *Impact of local item dependence on item response theory scoring in CAT*. Computerized Testing Report 98-08. The Law School Admission Council. Newtown, PA.
- Reiser, M., Fecso, R., & Chua, M. K. (1992). Some aspects of measurement error in the United States objective yield survey. *Journal of Official Statistics*, 8 (3), 351-375.
- Rosenbaum, P. R. (1987). Model-based direct adjustment. *Journal of the American Statistical Association*, 82 (398), 387-394.
- Sarndal, C. (1980). On pie-inverse weighting versus best linear unbiased weighting in probability sampling. *Biometrika*, 67 (3), 639-50.
- Schott, J. R. (1997). *Matrix analysis for statistics*. New York: John Wiley & Sons, Inc.
- Searle, S. R., Casella, G., & McCulloch, C. E. (1992). *Variance components*. New York: Wiley.
- Shavelson, R. J., & Webb, N. M. (1981). Generalizability theory 1973-1980. *British Journal of Mathematical and Statistical Psychology*, 34, 133-166.
- Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Newbury Park, CA: Sage.
- Shavelson, R. J., & Ruiz-Primo, M. A. (2000). On the psychometrics of assessing science understanding. In J. J. Mintzes & J. H. Wandersee (Eds.), *Assessing science understanding: A human*

*constructivist view* (303-341). San Diego: Academic Press, Inc.

Shipper, F. (1986). A study of four psychometric properties of the Jenkins Activity Survey Type. A scale with suggested modifications and validation. *Educational and Psychological Measurement*, 46 (3), 551-64.

Suter, N., Harter, R., & Selfa, L. (1999). *1997 Survey of doctorate recipients methodology*

*report*. Chicago, IL: National Opinion Research Center.

White, H. D. (1994). Scientific communication and literature retrieval. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (41-53). New York: Russell Sage Foundation.

Wolter, K. M. (1985). *Introduction to variance estimation*. New York: Springer-Verlag.

## Appendix A: Derivations and Computational Examples for the Sum of Mean Scores used in Example One

Matrix notations were adopted from Scott (1997).

- **diag** is the operator to create a diagonal matrix.
- $\otimes$  is the Kronecker product operator, which multiplies the entire matrix in the right side of the operator to every element in the matrix to the left of the Kronecker operator. If  $\mathbf{A}$  is an  $m \times n$  matrix and  $\mathbf{B}$  is a  $p \times q$  matrix, then the Kronecker product of A and B, denoted  $\mathbf{A} \otimes \mathbf{B}$ , is the  $mp \times nq$  matrix.

$$\begin{bmatrix} a_{11}\mathbf{B} & a_{12}\mathbf{B} & \dots & a_{1n}\mathbf{B} \\ a_{21}\mathbf{B} & a_{22}\mathbf{B} & \dots & a_{2n}\mathbf{B} \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ a_{m1}\mathbf{B} & a_{m2}\mathbf{B} & \dots & a_{mn}\mathbf{B} \end{bmatrix}$$

- We defined  $\mathbf{w}$  and  $\mathbf{w}_{f(p)}$  as row vectors. They are equivalent to the traditional matrix notation (Scott, 1997), which would define the two row vectors as transposes (i.e.,  $\mathbf{w}^T$  and  $\mathbf{w}_{f(p)}^T$ ).
- $\odot$ , the Hadamard operator, is the elementwise multiplication operator for two matrices. The traditional Hadamard operator  $\odot$  requires that two quantities to be expressed separately in the left and in the right sides of the operator. This becomes cumbersome when the two quantities are identical, because one would have to repeat a quantity twice. For example, to perform an elementwise multiplication of  $(\mathbf{w}_{f(p)} \bullet \mathbf{X} \bullet \mathbf{I}_{n_r \times n_r} \otimes \mathbf{1}_{n_i \times 1} \bullet (1/n_i))$  to itself, one would write:  $(\mathbf{w}_{f(p)} \bullet \mathbf{X} \bullet \mathbf{I}_{n_r \times n_r} \otimes \mathbf{1}_{n_i \times 1} \bullet (1/n_i)) \odot (\mathbf{w}_{f(p)} \bullet \mathbf{X} \bullet \mathbf{I}_{n_r \times n_r} \otimes \mathbf{1}_{n_i \times 1} \bullet (1/n_i))$ . To save space, we defined a parsimonious version of the Hadamard operator, to represent an elementwise power multiplication. For example,  $\mathbf{X} \odot^2$  indicates that the elements in  $\mathbf{X}$  were raised to the second



power. Using the new operator, the aforementioned cumbersome notation can be simplified as follows.  $(\mathbf{w}_{f(p)} \bullet \mathbf{X} \bullet \mathbf{I} \otimes \mathbf{1} \bullet (1/n_i)) \odot^2$ . In summary,  $\mathbf{X} \odot^2 = \mathbf{X} \odot \mathbf{X}$ .

- In each of the following equations, the first line shows the summation notation of the sums of squared means and the second line shows the matrix notation of the same quantity.

$$\begin{aligned} \sum_p \bar{X}_p^2 &= \sum_p w_p \left( \frac{1}{n_i n_r} \sum_i \sum_r x_{pir} \right)^2 \\ &= \mathbf{w} \bullet \text{diag} \left( (\mathbf{X} \bullet \mathbf{1}) \bullet (1/n_r) \right) \bullet ((\mathbf{X} \bullet \mathbf{1}) \bullet (1/n_r)) \end{aligned} \quad (13)$$

$$\begin{aligned} \sum_i \bar{X}_i^2 &= \sum_i \left( \sum_p w_{f(p)} \left( \frac{1}{n_r} \sum_r x_{pir} \right) \right)^2 \\ &= (\mathbf{w}_{f(p)} \bullet \mathbf{X} \bullet \begin{bmatrix} \mathbf{I} \\ \mathbf{I} \end{bmatrix}) \bullet (1/n_r) \bullet \text{diag} (w_{f(p)} \bullet \mathbf{X} \bullet \begin{bmatrix} \mathbf{I} \\ \mathbf{I} \end{bmatrix}) \bullet (1/n_r) \bullet \mathbf{1} \end{aligned} \quad (14)$$

$$\begin{aligned} \sum_r \bar{X}_r^2 &= \sum_r \left( \sum_p w_{f(p)} \left( \frac{1}{n_i} \sum_i x_{pir} \right) \right)^2 \\ &= (\mathbf{w}_{f(p)} \bullet \mathbf{X} \bullet \mathbf{I} \otimes \mathbf{1} \bullet (1/n_i)) \odot^2 \bullet \mathbf{1} \end{aligned} \quad (15)$$

$$\begin{aligned} \sum_p \sum_i \bar{X}_{pi}^2 &= \sum_p w_p \left( \sum_i \left( \frac{1}{n_r} \sum_r x_{pir} \right) \right)^2 \\ &= \mathbf{w} \bullet ((\mathbf{X} \bullet (\mathbf{1} \otimes \mathbf{I})) \bullet (1/n_r)) \odot^2 \bullet \mathbf{1} \end{aligned} \quad (16)$$

$$\begin{aligned} \sum_p \sum_r \bar{X}_{pr}^2 &= \sum_p w_p \left( \sum_r \left( \frac{1}{n_i} \sum_i x_{pir} \right) \right)^2 \\ &= \mathbf{w} \bullet (\mathbf{X} \bullet (\mathbf{I} \otimes \mathbf{1}) \bullet (1/n_i)) \odot^2 \bullet \mathbf{1} \end{aligned} \quad (17)$$

$$\begin{aligned} \sum_i \sum_r \bar{X}_{ir}^2 &= \sum_i \sum_r \sum_p (w_{f(p)} x_{pir})^2 \\ &= (\mathbf{w}_{f(p)} \bullet \mathbf{X}) \odot^2 \bullet \mathbf{1} \end{aligned} \quad (18)$$

$$\begin{aligned} \sum_p \sum_i \sum_r X_{pir}^2 &= \sum_p w_p \sum_i \sum_r x_{pir}^2 \\ &= \mathbf{w} \bullet \mathbf{X} \odot^2 \bullet \mathbf{1} \end{aligned} \quad (19)$$

$$\begin{aligned} \bar{X}^2 &= \left( \frac{1}{n_p n_i n_r} \sum_p w_p \sum_i \sum_r x_{pir} \right)^2 \\ &= ((1/n_{pir}) \bullet \mathbf{w} \bullet \mathbf{X} \bullet \mathbf{1}) \odot^2 \end{aligned} \quad (20)$$

$$\mathbf{X}_{n_p \times n_{ir}} = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \quad (21)$$

$$\mathbf{w}_p_{1 \times n_p} = [0.8 \quad 1.2 \quad 1.6 \quad 0.4] \quad (22)$$

$$\mathbf{w}_{f(p)}_{1 \times n_p} = [0.2 \quad 0.3 \quad 0.4 \quad 0.1] \quad (23)$$

$$1/n_i = 1/3 \quad (24)$$

$$1/n_r = 1/2 \quad (25)$$

$$1/n_{ir} = 1/6 \quad (26)$$

$$1/n_{pir} = 1/24 \quad (27)$$

$$\mathbf{1}_{n_i \times 1} = [1 \quad 1 \quad 1]^T \quad (28)$$

$$\mathbf{1}_{n_r \times 1} = [1 \quad 1]^T \quad (29)$$

$$\mathbf{1}_{n_{ir} \times 1} = [1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1]^T \quad (30)$$

$$\mathbf{I}_{n_i \times n_i} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad (31)$$

$$\mathbf{I}_{n_r \times n_r} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \quad (32)$$

$$\mathbf{X}^{\odot 2} = \begin{bmatrix} 1^2 & 1^2 & 1^2 & 0^2 & 0^2 & 0^2 \\ 0^2 & 1^2 & 1^2 & 1^2 & 0^2 & 1^2 \\ 1^2 & 0^2 & 0^2 & 0^2 & 1^2 & 1^2 \\ 0^2 & 0^2 & 0^2 & 0^2 & 0^2 & 1^2 \end{bmatrix} = \mathbf{X} \odot \mathbf{X} \quad (33)$$

$$\mathbf{I}_{n_r \times n_r} \otimes \mathbf{1}_{n_i \times 1} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \otimes \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} \quad (34)$$

By substituting (21) through (34) into the corresponding elements in (13) through (20), following results are obtained and used to compute the weighted sum of squared mean shown in (10).

$$\sum_p \bar{X}_{p..}^2 = 1.444 \quad (35)$$

$$\sum_i \bar{X}_{i..}^2 = 0.8275 \quad (36)$$

$$\sum_r \bar{X}_{r..}^2 = 0.5344 \quad (37)$$

$$\sum_p \sum_i \bar{X}_{pi.}^2 = 3.7000 \quad (38)$$

$$\sum_p \sum_r \bar{X}_{pr.}^2 = 2.8000 \quad (39)$$

$$\sum_i \sum_r \bar{X}_{ir.}^2 = 1.7500 \quad (40)$$

$$\sum_p \sum_i \sum_r X_{pir}^2 = 12.400 \quad (41)$$

$$\bar{X}^2 = 0.2669 \quad (42)$$

## Appendix B: A Comparison between the Traditional Unweighted and the Chiu-Fecso Weighted Methods.

	Unweighted VC	Unweighted VC	Weighted VC	Ratio:
	Brennan, 1992 (p.38)	Chiu-Fecso Unit Weights	Chiu-Fecso Disproportionate Weights	Weighted VC / Unweighted VC
<i>P</i>	0.5528	0.5528	0.9634	1.7428
<i>I</i>	0.4417	0.4417	0.5656	1.2805
<i>R</i>	0.0074	0.0074	0.0221	2.9865
<i>Pi</i>	0.575	0.575	0.4432	0.7708
<i>Pr</i>	0.1009	0.1009	0.0349	0.3459
<i>Ir</i>	0.1565	0.1565	0.0562	0.3591
<i>pir,e</i>	0.9352	0.9352	0.5776	0.6176

Notes: Unit weights:  $w = [1 \ 1 \ 1 \ 1 \ 1 \ 1 \ 1 \ 1 \ 1]$ , Disproportionate weights:  
 $w = [1 \ 1 \ 1 \ 1 \ 1 \ 1 \ 1 \ 1 \ 10]$ . Data source: Brennan (1992a, p.38).

## Appendix C: Weighted Variance Component Estimates by Age Group for Example Two.

## Variance Component Estimates

Age Group	Below 30	35-39	40-44	45-49	50-54	55-59	60-64	65-69	Above 70
<i>p</i>	0.1667	0.055	0.0658	0.0753	0.0732	0.0661	0.0677	0.0561	0.0398
<i>y</i>	0	0.001	0.0002	0	0	0	0	0.0157	0.0203
<i>m</i>	0	0.0003	0.0003	0.0007	0.0008	0.0013	0.0015	0.0002	0.0002
<i>py</i>	0.0833	0.0848	0.0805	0.0943	0.0865	0.0973	0.0972	0.1314	0.1528
<i>pm</i>	0	0.0022	0.0029	0.0029	0.0064	0.0076	0.0067	0.001	0.0056
<i>ym</i>	0	0.0002	0.0001	0.0006	0.0011	0.0013	0.0015	0.001	0.0025
<i>pym,e</i>	0	0.0332	0.044	0.0509	0.0525	0.0523	0.0475	0.0425	0.0348

*p*: person, *y*: year, *m*: method; *py* = person by year; *pm*: person by method,  
*ym*: year by method, *pym,e*: person by year by method and other errors.

## Percent Contribution

Age Group	Below 30	35-39	40-44	45-49	50-54	55-59	60-64	65-69	Above 70
<i>p</i>	66.7%	31.1%	34.0%	33.5%	33.2%	29.3%	30.5%	22.6%	15.5%
<i>y</i>	0.0%	0.6%	0.1%	0.0%	0.0%	0.0%	0.0%	6.3%	7.9%
<i>m</i>	0.0%	0.1%	0.2%	0.3%	0.4%	0.6%	0.7%	0.1%	0.1%
<i>py</i>	33.3%	48.0%	41.5%	42.0%	39.2%	43.1%	43.8%	53.0%	59.7%
<i>pm</i>	0.0%	1.3%	1.5%	1.3%	2.9%	3.4%	3.0%	0.4%	2.2%
<i>ym</i>	0.0%	0.1%	0.0%	0.3%	0.5%	0.6%	0.7%	0.4%	1.0%
<i>pym,e</i>	0.0%	18.8%	22.7%	22.6%	23.8%	23.2%	21.4%	17.1%	13.6%
Sample Size	3	202	439	400	440	392	256	140	116

## A Different Future For Social And Behavioral Science Research

Shlomo S. Sawilowsky  
Educational Evaluation and Research  
College of Education  
Wayne State University

---

The dissemination of intervention and treatment outcomes as effect sizes bounded by confidence intervals in order to think meta-analytically was promoted in a recent article in *Educational Researcher*. I raise concerns with unfettered reporting of effect sizes, point out the con in *confidence interval*, and caution against thinking meta-analytically. Instead, cataloging effect sizes is recommended for sample size estimation and power analysis to improve social and behavioral science research.

Key words: Effect size encyclopedia, bracketed interval, confidence interval, sample size, power

---

### Introduction

Recently, an article appeared in *Educational Researcher* describing a possible future of social science research. It was one in which research results were reported in terms of effect sizes bounded by so-called confidence intervals. The notion of thinking meta-analytically was touted, and to that end, the publication of effect sizes was promoted (Thompson, 2002).

#### Bracketed Intervals (BI)

I prefer the phrase “bracketed interval” (BI) instead of confidence interval, for reasons discussed below. The Frequentist perspective of the BI was described by Thompson (2002) as a 95% degree of confidence that the interval contains the parameter in question. According to this view it would be inappropriate to say there is a 95% probability that  $\mu$ , the population mean, is within the interval, but it would not be inappropriate to say there is a 95% level of confidence that  $\mu$  is in the interval.

The first intervals of a statistical nature were developed by de Moivre between 1733 - 1742, but they were not positioned for interval estimation. That feat was first accomplished by Lagrange in 1776.

De Moivre stated that the interval refers to “the probability that the value of [a parameter] is enclosed between the [upper and lower] limits” (cited by Hald, 1998, p. 23). Thus, in modern classification schemes, the original expression of bracketed intervals was from a Frequentist perspective.

Now, return to the term *confidence*. The general idea originated with Pytkowski (1932), but the first use of the phrase *confidence interval* and its theoretical development was by Neyman (1934, 1937, 1939). He referred to

determining certain intervals, which I propose to call the confidence intervals (see Note 1), in which we may assume are contained the values of the estimated characters of the population, the probability of an error in a statement of this sort being equal to or less than  $1 - \epsilon$ , where  $\epsilon$  is any number  $0 < \epsilon < 1$ , chosen in advance. The number  $\epsilon$  I call the confidence coefficient. (1934, p. 562)

---

Shlomo S. Sawilowsky is Professor of Educational Evaluation and Research, and Wayne State University Distinguished Faculty Fellow. He is the editor of *Journal of Modern Applied Statistical Methods*. Email him at [shlomo@wayne.edu](mailto:shlomo@wayne.edu).

He opined that “the solution of the problem which I described as the problem of confidence intervals has been sought by the greatest minds since the work of Bayes 150 years ago” (Neyman, 1934, p. 563). However, because Jerzy Neyman, along with Egon Sharpe Pearson, originated the Frequentist version of modern statistics (Neyman & Pearson, 1928a, 1928b), his definition was purposefully not “Bayesian”, and instead followed the Frequentist paradigm.

The student of Bayes would demur, claiming it doesn’t make sense to ascribe the 95% moniker to  $\mu$  being found within the interval. The  $1-\alpha\%$  probability only pertains prior to the collection of data, whereas afterwards either the parameter falls within the interval or it doesn’t.

Instead, the Bayesian perspective is that the judicious usage of specific prior information regarding the estimate is the only meaningful way to obtain such a probability. Thompson (2002) characterized this as “a better definition” (p. 26).

The weakness of the Bayesian approach, (which Fisher, Neyman, Wald, and others rejected) is the reliance on subjective prior information. I cannot resolve the philosophical debate between the Frequentist and the Bayesian, but it is inappropriate to call either perspective “better”, as did (Thompson, 2002, p. 26).

Furthermore, the philosophical controversy Thompson (2002) alluded to is not relevant in practical application. What is of importance is the role of interval estimation vs hypothesis tests. There has been a flurry of activity since the early 1990s where the usage of hypothesis tests was taken to task, particularly within the American Educational Research Association (AERA) and other professional organizations. For example, Carver (1992) presented a paper to the AERA attempting to make a case against statistical significance testing, and recommended banning its usage altogether.

Amazingly and inexplicably, proponents of the case against hypothesis testing are also proponents of the usage of interval estimation. The root of their misconception is the misnomer *confidence*, as if bracketed intervals have a certain amount of confidence to them that hypothesis tests do not. There is no more confidence associated with an interval based on  $(1-\alpha)100\%$  than in a point null hypothesis based on  $\alpha$ .

Thompson (2002) incorrectly construed

my position in *Educational Researcher*, claiming I “erroneously equate CIs and statistical significance tests” (p. 29). In an article with Thomas Knapp, I pointed out that the statistical criteria regarding the probabilities associated with bracketed intervals are the same as those for point null hypothesis tests, but certainly the two procedures cannot be equated. Regarding the equivalency of probabilities: (1) Is zero really not in the interval? (Type I error), (2) is zero really in the interval? (Type II error), and (3) is the width of the interval at a minimum (comparative statistical power)? The probabilities associated with these criteria are exactly the same (Knapp & Sawilowsky, 2001).

These three points are congruent with a careful examination of Neyman (1934). He equated the boundaries of the interval with the probabilities of classical Fisherian “fiducial” limits of  $\theta_1(x)$  and  $\theta_2(x)$ , which represent the lower and upper bound of the bracketed interval. With a passing reference to the famous debate in the literature on what Sir Ronald Fisher meant by fiducial, Neyman (1934) did not dissociate the so-called confidence of the bracketed interval from the probabilities used in its construction:

Since the word “fiducial” has... caused misunderstandings I have already referred to, and which in reality cannot be distinguished from the ordinary concept of probability, I prefer to avoid the term and call the intervals  $[\theta_1(x), \theta_2(x)]$  the confidence intervals. (p. 590)

Although Wald (1950) subsumed both hypothesis tests and interval estimation in a single model, and expressed them as specific cases of the general theory of statistical decision functions, that does not mean the two procedures are equivalent in every respect. After pointing out the probabilities associated with BIs and hypothesis tests are the same, I noted there is an advantage of BIs over point null hypothesis tests. It results in a range of possible values wherein the parameter might fall, whereas hypothesis tests do not.

This doesn’t appear to be the tremendous advantage that many proponents claim it to be. What added benefit is there in knowing, for

example, that the BI for a student's *Wechsler IQ* was 97-103 from an educator's perspective? Furthermore, in Knapp and Sawilowsky (2001), we mentioned specific data analysis situations where the BI would be preferred over the hypothesis test, as well as the reverse.

I also pointed out there are areas of concern in unbridled promotion of BIs (Knapp & Sawilowsky, 2001): (1) Some statistics are not amenable to the determination of standard errors, relying instead on theoretically interesting but practically questionable asymptotic variances (which are mathematical inventions pertaining to the world of infinite sample sizes). This may make the BI yield poorer statistical properties than point hypothesis testing. (2) There is the question of whether or not the interval should be symmetric about the sample statistic (Low, 1997).

(3) There is the problem of the effects of measurement error in constructing the interval (Nunnally, 1978). (4) Here, I add yet another concern: Bienaymé's complaint in 1852 against using BIs based on a single parameter expressed as a continuum on a line. Instead, he proposed the concept of Bracketed Ellipsoids, where simultaneous regions are constructed taking into account multiple parameters. For example, two parameters result in an ellipsoid continuum on a Cartesian plane.

#### Meta-Analysis

These issues regarding BIs apply to all statistics, including effect sizes. Thompson (2002) focused on effect sizes to provide fodder for meta-analyses. This became necessary following Gene Glass' presidential address on meta-analysis to the AERA in April of 1976, because modern meta-analysis depends on the proliferation of effect sizes.

Thompson (2002) viewed effect sizes as the enabler in thinking meta-analytically. His exuberance with meta-analysis led him to recommend that effect sizes "can and should be reported and interpreted in all studies, regardless of whether or not statistical tests are reported" (Thompson, 1996, p. 29), and "even [for] non-statistically significant effects" (Thompson, 1999, p. 67). The same argument had previously been made by Carver (1979, 1993).

However, Sawilowsky and Yoon (2001, 2002) reported a brief Monte Carlo simulation

demonstrating the trouble with reporting research findings via effect size in the absence of statistical significance. The practice will wreak havoc in the literature, as the Monte Carlo simulation demonstrated that an intervention of random numbers will produce typical effect sizes that are not near zero, but rather, are at a magnitude Cohen (1988) calls a small treatment effect.

Roberts and Henson (2002) purported to rebut these results. However, their study was not a Monte Carlo simulation of typical effect sizes produced under the truth of the null hypothesis. Instead, it was a Monte Carlo study of the bias in  $d$ , a topic irrelevant to the point being made. See the ensuing *Invited Debate* in this issue of the *Journal of Modern Applied Statistical Methods*.

There have been many articles published here and there by a variety of authors, including myself, that addressed specific methodological and substantive issues with meta-analyses. In addition, I have raised questions about thinking meta-analytically (e.g., Knapp & Sawilowsky, 2001). Rather than reviewing that literature here, I find it more instructive to recite an excerpt from Glass' (2000) most recent vision of research synthesis:

In the twenty-five years between the first appearance of the word "meta-analysis" in print and today, there have been several attempts to modify the approach, or advance alternatives to it, or extend the method to reach auxiliary issues. If I may be so cruel, few of efforts have added much... If our efforts to research and improve education are to prosper, meta-analysis will have to be replaced by more useful and more accurate ways of synthesizing research findings.

#### Sample Size Estimation and Power Analysis

The role of effect sizes in sample size determination and power analysis is an entirely different matter from that of meta-analysis. The first part of my professorial career could be summarized by the many consultations I had with students, teachers, faculty, and researchers outside of academe on the "how large should my sample be?" question. The bottleneck was obtaining an

estimate of the effect size, which is necessary to enter Cohen's (1988) sample size and power tables. I was not alone; every colleague I discussed this matter with in the past twenty years has reported the same difficulty.

I wrestled with this problem for a decade. During that time I had a series of written and telephone conversations with, and initiated by, Jacob Cohen. He recognized the weaknesses in educated guessing (Cohen, 1988, p. 12) or using his rules of thumb for small, medium, and large effect sizes (p. 532). I suggested cataloging and cross-referencing effect size information for sample size estimation and power analysis as a more deliberate alternative.

Cohen expressed keen interest in this project. His support led to me to delivering a paper at the annual meeting of the AERA on the topic of a possible encyclopedia of effect sizes for education and psychology (Sawilowsky, 1996). The idea was to create something like the "physician's desk reference", but instead of medicines, the publication would be based on effect sizes. (I presented papers every year at AERA from 1985 - 2000, but this session had a higher attendance than most of them put together.) I doubt any of those listening to the presentation envisioned a future for quantitative social and behavioral science research with sample size estimation and power analysis forever relegated to prestidigitation.

Encouraged by colleagues, in 1999 and again in 2000, I submitted proposals to the U. S. Department of Education to fund a print and electronic encyclopedia project. Thirty-five experts on effect sizes and meta-analysis wrote supportive letters (Table 1). A summit would be held with these experts, the most recent ten years of ninety journals in education and psychology would be culled for effect sizes and cataloged, and an internet-based data-base would be created in which authors/journal editors could submit additions or updates. Alas, the proposals were not judged to be a funding priority. Subsequently, I had a series of e-mail and telephone conversations with Herbert Walberg on creating the encyclopedia sans funding, but the enormity of the project was prohibitive.

Table 1. Supporters of the Encyclopedia of Effect Sizes Project:

---

William Asher, Purdue University
Betsy Becker, Michigan State University
John Behrens, Arizona State University
Patricia Busk, University of San Francisco
C. Mitchel Dayton, University of Maryland
Robert Donmoyer, Ohio State University
Susan Embretson, University of Kansas
Gene Glass, Arizona State University
Robert Grissom, San Francisco State University
John Hunter*, Michigan State University
Carl Huberty, University of Georgia
Harvey Keselman, University of Manitoba
John Kim, San Francisco State University
Roger Kirk, Baylor University
Thomas Knapp, Ohio State University
Dennis Leitner, Southern Illinois University
Joel Levin, University of Wisconsin-Madison
Lisa Lix, Private Scholar
Jorge Mendoza, University of Oklahoma
Theodore Micceri, University of South Florida
Isadore Newman, University of Akron
Steve Olejnik, University of Georgia
Liora Pedhazur-Schmelkin, Hofstra University
Bob Rosenthal, University of California-Riverside
Donald Rubin, Harvard University
Frank Schmidt, University of Iowa
Michael Seaman, University of South Carolina
Ronald Serlin, University of Wisconsin-Madison
Juliet Shaffer, University of California-Berkeley
Bruce Thompson, Texas A&M University
Howard Wainer, ETS
Herbert Walberg, University of Illinois-Chicago
Rand Wilcox, University of Southern California
Joe Wisenbaker, University of Georgia
Bruno Zumbo, University of N. British Columbia

---

*Notes:* \*Deceased. Affiliations were accurate in 1999-2000.

### Conclusion

Sample size estimation and power analysis in every grant funded by the U. S. Department of Education and every article published in AERA journals are based on guessing or Cohen's (1988) rules of thumb. Those practices could be discontinued in a different future of social and behavioral science research. Along with a re-commitment to true experimental design



(Sawilowsky, 1999), a compendium of effect sizes could improve research design in education and psychology, and propel disciplined inquiry forward in a scientific fashion.

The encyclopedia could be a globally cooperative effort among professional organizations and learned societies, their journal editors, and authors. It could be internet-based and updated in real-time, cross-referenced by discipline/sub-discipline and independent variable, have effect size entries categorized by statistically significant studies at various  $\alpha$  levels, and classified according to whether the journal was peer reviewed. Finally, entries should be categorized based on whether the effect size arose from a true experimental design vs. quasi-experimental, post hoc, survey, and other non-experimental designs.

#### References

- Carver, R. P. (1979). The case against statistical significance testing. *Harvard Educational Review*, 48, 378-399.
- Carver, R. P. (April, 1992). *The case against statistical significance hypothesis testing, revisited*. Annual Meeting of the American Educational Research Association, San Francisco, CA.
- Carver, R. P. (1993). The case against statistical significance testing, revisited. *Journal of Experimental Education*, 61, 287-292.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. (2<sup>nd</sup> ed.) Hillsdale, NJ: Erlbaum.
- Glass, G. (2000). *Meta-analysis at 25*. <http://glass.ed.asu.edu/gene/papers/meta25.html>.
- Hald, A. (1998). *A history of mathematical statistics from 1750 to 1930*. NY: Wiley & Sons.
- Knapp, T., & Sawilowsky, S. (2001). Constructive criticisms of methodological and editorial practices. *Journal of Experimental Education*, 70, 65-79.
- Low, M. G. (1997). On nonparametric confidence intervals. *The Annals of Statistics*, 25, 2547-2554.
- Neyman, J. (1934). On the Two Different Aspects of the Representative Method, *Journal of the Royal Statistical Society*, 97, 558-625.
- Neyman, J. (1937). Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of the Royal Society of London*, A(236), 333-380.
- Neyman, J. (1939). L'estimation statistique traitée comme un problème classique de probabilité. *Actualités Scientifiques et Industrielles*, 739, 25-57.
- Neyman, J., & Pearson, E. S. (1928a). On the use and interpretation of certain test criteria for purposes of statistical inference: Part 1. *Biometrika*, 20, 175-240.
- Neyman, J., & Pearson, E. S. (1928b). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London*, A(231), 289-337.
- Nunnally, J. C. (1978). *Psychometric theory* (2<sup>nd</sup> ed). NY: McGraw-Hill.
- Pytkowski, W. (1932). Outline of the income in small farms upon their area, the outlay and the capital invested in cows. Monograph #31, *Biblioteka Pulawska*, Poland: Agricultural Research Institute of Pulaway.
- Roberts, J. K., & Henson, R. K. (2002). Correction for bias in estimating effect sizes. *Educational and Psychological Measurement*, 62, 241-253.
- Sawilowsky, S. S. (April, 1996). *Encyclopedia of educational and psychological effect sizes*. Annual Meeting of the American Educational Research Association, Division D, Measurement and Research Methodology, NY, NY.
- Sawilowsky, S. S. (1999). Quasi-experimental design: The legacy of Campbell and Stanley. In (Bruno D. Zumbo, Ed.) *Social indicators/quality of life research methods: Methodological developments and issues, Yearbook 1999*. Norwell, MA: Kluwer Academic Publishers.
- Sawilowsky, S. S., & Yoon, J. (2001). *The trouble with trivials (p > .05)*. 53<sup>rd</sup> Session of the International Statistical Institute, Seoul, South Korea.
- Sawilowsky, S. S., & Yoon, J. (2002). The trouble with trivials (p > .05). *Journal of Modern Applied Statistical Methods*, 1, 143-144.
- Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. *Educational Researcher*, 25, 26-30.
- Thompson, B. (1999). Five methodology errors in educational research: A pantheon of statistical significance and other faux pas. In B. Thompson (Ed.), *Advances in social science methodology*, 5, 23-86.
- Thompson, B. (April, 2002). What future quantitative social science research could look like: Confidence intervals for effect sizes. *Educational Researcher*, 31, p. 25-32.
- Wald, A. (1950). *Statistical decision functions*. NY: Wiley.

## Without Supporting Statistical Evidence, Where Would Reported Measures of Substantive Importance Lead? To No Good Effect

Anthony J. Onwuegbuzie  
University of South Florida

Joel R. Levin  
University of Arizona

---

Although estimating substantive importance (in the form of reporting effect sizes) has recently received widespread endorsement, its use has not been subjected to the same degree of scrutiny as has statistical hypothesis testing. As such, many researchers do not seem to be aware that certain of the same criticisms launched against the latter can also be aimed at the former. Our purpose here is to highlight major concerns about effect sizes and their estimation. In so doing, we argue that effect size measures *per se* are not the hoped-for panaceas for interpreting empirical research findings. Further, we contend that if effect sizes were the only basis for interpreting statistical data, social-science research would not be in any better position than it would if statistical hypothesis testing were the only basis. We recommend that hypothesis testing and effect-size estimation be used in tandem to establish a reported outcome's believability and magnitude, respectively, with hypothesis testing (or some other inferential statistical procedure) retained as a "gatekeeper" for determining whether or not effect sizes should be interpreted. Other methods for addressing statistical and substantive significance are advocated, particularly confidence intervals and independent replications.

Key words: Effect-size concerns, statistical inference, substantive importance

---

### Introduction

Statistical hypothesis testing has been implemented to assess the believability, or non-"chanceness" (Levin, 1998b; Levin & Robinson, 1999), of research findings for more than 75 years, stemming from the seminal works of Fisher (1925/1941) and Neyman and Pearson (1928). Despite the widespread use of hypothesis testing during most of the last century through today, its practice has been controversial. Indeed, over the past few decades testing for statistical significance has come under close scrutiny.

---

Anthony J. Onwuegbuzie is Associate Professor, Department of Educational Measurement and Research, College of Education, University of South Florida, 4202 East Fowler Avenue, EDU 162, Tampa, Florida 33620-7750. Email him at [tonyonwuegbuzie@aol.com](mailto:tonyonwuegbuzie@aol.com)). Joel R. Levin is Professor, Department of Educational Psychology, College of Education, University of Arizona. E-Mail: [jrlevin@u.arizona.edu](mailto:jrlevin@u.arizona.edu)

Since 1950, for example, the number of articles published in the fields of education, psychology, ecology, and medicine criticizing hypothesis testing has been increasing at an exponential rate (Anderson, Burnham, & Thompson, 2000). Additionally:

(a) professional journals (e.g., *The Journal of Experimental Education* and *Research in the Schools*) have devoted special theme issues to statistical hypothesis testing; and

(b) symposia have been held at national annual meetings, such as the American Educational Research Association, the American Psychological Association, and the American Psychological Society. Even an edited book, *What if there were no significance tests?* (Harlow, Mulaik, & Steiger, 1997), has been devoted exclusively to the topic.

### The Case Against Statistical Hypothesis Testing

Some of the staunchest critics of statistical hypothesis testing contend that this practice has been extremely harmful to scientific progress in the social sciences. For example, Meehl (1978, p.

817) stated that it “is a terrible mistake, a basically unsound, poor scientific strategy, and one of the worst things that ever happened in the history of psychology.” Rozeboom (1997) continued:

Null-hypothesis significance testing is surely the most bone-headedly misguided procedure ever institutionalized in the rote training of science students... [I]t is a sociology-of-science wonderment that this statistical practice has remained so unresponsive to criticism. (p. 335)

Similarly, Tryon (1998) complained:

[T]he fact that statistical experts and investigators publishing in the best journals cannot consistently interpret the results of these analyses is extremely disturbing. Seventy-two years of education have resulted in minuscule, if any, progress toward correcting this situation. It is difficult to estimate the handicap that widespread, incorrect, and intractable use of a primary data analytic method has on a scientific discipline, but the deleterious effects are undoubtedly substantial. (p. 796)

Schmidt and Hunter (1997, p. 37) claimed that “[s]tatistical significance testing retards the growth of scientific knowledge; it never makes a positive contribution,” and Thompson (1992b, p. 436) added: “[Statistical significance testing] has created considerable damage as regards the cumulation of knowledge.”

As a result of the purported flaws that statistical hypothesis testing has been accused of, several researchers have recommended that it be banned completely (e.g., Bakan, 1966; Cahan, 2000; Carver, 1978, 1993; Cohen, 1994; Guttman, 1985; Loftus, 1996; Meehl, 1967, 1978; Nix & Barnette, 1998; Rozeboom, 1960; Schmidt, 1992; 1996; Schmidt & Hunter, 1997). Although we: (a) agree that statistical hypothesis testing has been misused, and (b) concur with many of the criticisms of it that have been offered, it is quite a

leap to charge that hypothesis testing by itself has stunted “the cumulation of knowledge” (Thompson, 1992b, p. 436), is “one of the worst things that ever happened in the history of psychology” (Meehl, 1978, p. 817), or “retards the growth of scientific knowledge... [and]... never makes a positive contribution” (Schmidt & Hunter, 1997, p. 37).

Furthermore, some of the assertions made in an attempt to invalidate the hypothesis-testing practice either have been accompanied by unsubstantiated claims or represent flawed logic. As noted by Krantz (1999):

It is one thing to accuse scientists of showing their ignorance of statistical reasoning in the course of their science, but this does not imply that their ultimate conclusions will be incorrect, nor even that their efficiency in reaching correct conclusions will be impaired. A causal attribution of this sort needs to be supported by careful empirical arguments. (p. 1378)

The foregoing concerns aside, valid criticisms of statistical hypothesis testing have nonetheless been made. Fan (2001) provided a summary of some of these criticisms:

Thompson (1993) discussed three relevant criticisms for (*sic.*) statistical significance testing: (a) overdependency on sample size, (b) some nonsensical comparisons, and (c) some inescapable dilemmas created by statistical significance testing (e.g., testing for assumption vs. testing for the research hypothesis). In a similar vein, Kirk (1996) discussed three major criticisms of statistical significance testing: (a) Significance testing does not tell researchers what they want to know, but rather, it creates the illusion of probabilistic proof by contradiction (Falk & Greenbaum,

1995). (b) Statistical significance testing is often a trivial exercise because it simply indicates the power of the design (which primarily depends on the sample size) to reject the false null hypothesis. (c) Significance testing “turns a continuum of uncertainty into a dichotomous reject-do-not-reject decision,” and this dichotomous decision process may “lead to the anomalous situation in which two researchers obtain identical treatment effects but draw different conclusions” (Kirk, p. 748) because of the slight differences in their design (e.g., sample sizes). (p. 276)

Because of these and other concerns, many researchers have called for the reporting of measures of practical significance (or substantive importance, as reflected by effect size or strength of relationship indices), either in addition to or instead of testing for statistical significance. Indeed, the most recent edition of the influential *Publication Manual of the American Psychological Association* (2001) states:

The general principle to be followed...is to provide the reader not only with information about statistical significance but also with enough information to assess the magnitude of the observed effect or relationship. (p. 26)

Certain anti-hypothesis-testers (e.g., Carver, 1993) even go so far as to endorse effect-size estimates as *replacements* for statistical significance testing – that is, they contend that effect sizes are all that are needed to make inferences about empirical research outcomes. As is argued throughout the remainder of this manuscript, however, we believe that such practice would only lead to no good effect!

Debates about the value and warrants of statistical hypothesis testing can be traced back to Boring (1919) and Berkson (1938, 1942). Over the last decade, many researchers have seemingly jumped on the effect-size bandwagon without

scrutinizing its use to the same degree as has occurred for hypothesis testing. Moreover, what appears to have been lost in all this fervor for effect-size provision – and as we illustrate later – is that many of the same criticisms launched against statistical hypothesis testing can also be aimed at effect sizes. As one salient illustration, cautions concerning hypothesis testing and its interpretation can be found in such sources as the aforementioned *APA Publication Manual* (2001) – namely, that *p*-values (statistical significance probabilities) do not directly reflect “the magnitude of an effect or the strength of a relationship” (p. 25). Yet, no such cautions about effect-size measures are found in that pivotal reference source.

#### Concerns and Cautions About Effect Sizes

In what follows we highlight several major concerns about effect sizes and their estimation, in what might be called *nine effect-size nuisances and no-no's*. In doing so, we consider several rarely acknowledged limitations of effect-size measures. We (as others before us) argue that effect-size measures are influenced by, and therefore must be interpreted with respect to, a number of critical factors. As a preliminary comment, we regard certain of these considerations as being especially relevant when effect sizes are reported as *sole* indicators of an empirical study's significance (i. e., as reflected in Carver's, 1993, “effect-size only” recommendation). We return to this fundamental issue in a later section.

According to Wilkinson and the Task Force on Statistical Inference (1999, p. 599) “[R]eporting and interpreting effect sizes...is essential to good research.” Unfortunately, this statement might suggest to some that the provision of effect sizes necessarily improves the quality of empirical studies. Yet, the uncritical acceptance of effect size measures is problematic because, as is now discussed, such measures are sensitive to a number of factors, such as: the research objective; sampling design (including the levels of the independent variable, choice of treatment alternatives, and statistical analysis employed); sample size and variability; type and range of the measures used; and score reliability (see, for example, Fern & Monroe, 1996; Frick, 1995;

O'Grady, 1982; Olejnik & Algina, 2000; and Sechrest & Yeaton, 1982).

1. *The research objective.* According to Fern and Monroe (1996), one's interpretation of an effect size should vary, depending on whether the objective of the study is what they call theory application or effects application. In theory-application research (or explanatory studies) the goal is to identify theories that increase our understanding of phenomena. Studies involving theory application, which consist primarily of theory generation and theory testing, typically focus on generalizing theories beyond the underlying sample and/or context. More specifically, in explanatory studies, the goal is to determine the "shape or functional nature of a relationship" (O'Grady, 1982, p. 770).

In such investigations, a large effect size is not necessarily of interest. Indeed, a large effect may be viewed as a negative outcome if it was not predicted by theory. That is, in theory-application research, a small effect may be more informative and useful than a large effect (Calder, Phillips, & Tybout, 1981). In fact, using "large" effect-size guidelines (e.g., Cohen, 1988) as the criterion for choosing among several independent variables in explanatory studies may culminate in misleading final theoretical models being selected. Conversely, in effects-application research (or predictive studies), researchers usually are not interested in generalizing the results beyond the levels of the variables selected. That is, in effects-application studies, the interest is more on the size of the effect than on determining the generalizability of a particular theory. This suggests that effect sizes should not be interpreted without taking into account whether one's research objective is essentially explanatory or predictive in nature.

2. *Choice of a specific research design and experimental conditions.* The selected research design also affects interpretation of effect sizes. Specifically, because within-subject sampling designs typically are more efficient than are between-subject sampling designs – inasmuch as they tend to minimize error variance (Maxwell & Delaney, 1990) – they tend to yield larger effect sizes (Keppel, 1991; O'Grady, 1982). Therefore, in interpreting effect sizes, consideration should be given to the sampling design used.

Although experimental studies allow the strongest causal inferences to be made and typically result in relatively smaller error variance in comparison to correlational studies, experimental designs also tend to yield smaller effect sizes than do correlational designs. This is because in experimental research the independent variable is artificially created specifically for the study and thus is weaker than it is in the population (Kerlinger, 1973). As such, comparing effect sizes stemming from experimental studies and those generated from correlational studies easily can be the equivalent of comparing apples and oranges. Moreover, in fixed-effects models, the magnitude of the omnibus effect size depends on the specific levels of the variables of interest. If different levels of the independent variable are studied, the effect sizes are not comparable (Olejnik & Algina, 2000).

Further, the number of experimental conditions (or levels of the independent variable) used in a study can either increase or decrease the effect size. O'Grady (1982, p. 773) provides a striking example of a two-conditions study (yielding  $M_1 = 10$  and  $M_2 = 18$ , with common  $SDs$  of 2 and  $ns$  of 10) in which the proportion of variance accounted for by the treatment factor (sample  $\eta^2$ ) is .82. Yet, had the same two conditions been part of a study that also included three additional experimental conditions, whose resulting means ranged in equal increments between the two original means (i.e.,  $M_3 = 12$ ,  $M_4 = 14$ , and  $M_5 = 16$ ), with the same  $SDs$  and  $ns$  as before, the proportion of variance accounted for by the treatment factor is reduced to .69. Of course, had the proportion of variance associated with *just* the two focal conditions been calculated and reported (i.e., the sample  $\eta^2$  associated with the Treatment 1 vs. Treatment 2 *contrast*), it would be equal to the original .82.

Interpretive problems resulting from omnibus, as opposed to contrast, strength-of-relationship reporting were pointed out by Levin (1967). Such problems can be further illustrated by another hypothetical example, which represents the "flip side" of the one just presented. Suppose that a researcher compares two different experimental treatments and finds that  $M_1 = 16$  and  $M_2 = 17$ , with common  $SDs$  of 2.5 and  $ns$  of 8. Here, the sample  $\eta^2$  can be found to be a fairly "small" .04. However, had these two treatments

been part of a study that included a low-scoring “control” group ( $M_3 = 6$ ) with the same  $SD$  and  $n$  as in the other two conditions, now the sample  $\eta^2$  would be found to leap to an “impressive” .81. As long as the researcher focused on the Treatment 1 vs. Treatment 2 contrast (for which  $\eta^2 = .04$ ), the same conclusion about a “small” treatment difference would have been reached as before. Unfortunately, however, many researchers routinely report and interpret the omnibus measure (here,  $\eta^2 = .81$ ), to the detriment of the unquestioning consumer. In multifactor designs a similar opportunity arises for misleading the consumer – namely, by not recognizing Kirk’s (1995, p. 261) distinction between omnibus and partial strength-of-relationship measures.

The design of an experimental study also refers to the manner in which participants are assigned to experimental conditions and treatments administered (generally characterized as between-subjects designs, within-subjects designs, mixed designs, blocking designs, and hierarchical designs), whether or not concomitant variables (covariates) are included, and the statistical analyses employed. Effect-size measures are affected by all such factors in a design, compromising comparisons of effect sizes across studies that differ in their specifics (Oljenik & Algina, 2000).

In particular, when one or more factors in a comparison-of-means analysis represents an individual difference factor (e.g., a covariate or blocking variable), problems arise with respect to what to use as the standardizer in an effect-size index. For example, in a two-factor design in which one factor is a manipulated factor and the other an individual difference factor, it is often a matter of debate whether the standardizer should be computed by ignoring or controlling for the individual difference factor (Oljenik & Algina, 2000). Whichever approach is taken leads to a different effect size being computed and, therefore, effect sizes using these two different standardizers are not comparable. In fact, as noted by Oljenik and Algina (2000): “depending on the sample size and effect sizes associated with the individual difference and interaction factors in a two-factor design, the effect size estimated for the manipulated factor can vary from trivial to quite large” (p. 250).

The difference in effect sizes is even greater if the individual difference factors vary across studies. Because varying standardizers for computing effect sizes are used in different studies, researchers should compare effect sizes only if they are completely aware of the standardizer that was used in each study of interest. Unfortunately, most researchers do not specify which standardizer was used in their effect-size computation. This discussion should make it clear that a researcher can make an effect size look larger or smaller by defining an effect size in terms of the specific design and control-variable characteristics just mentioned – basically, by incorporating (or not) any design features that serve to affect the error variance – and which may have ethical implications as well.

*3. Selection of an effect-size measure.* We now turn our attention to another potentially ethically sensitive effect-size issue. Although there is general agreement that the provision of effect-size information is valuable, recommendations concerning the specific measure that should be reported for a particular study are typically absent. In our view, such recommendations are critical, for as one of us noted previously:

Which of, say, half a dozen different effect-size measures that could be summoned up for a given problem should a researcher report? The one that is most informative, the one that is most conservative, or the one that enhances the researcher’s case and misleads the unsuspecting reader? For example, researchers might report percent agreement measures or percentages of variance accounted for that have not been corrected for chance, or researchers might seek out a goodness-of-fit measure that places their data in the most favorable light. For dependent measures where a frame of reference is needed or helpful, providing scale-free (relative) effect sizes (e.g., Cohen’s  $d$  or percentages of variance accounted for) is not nearly as substantively

interpretable as is providing the scale-dependent (absolute) measures in addition or instead...In many domains, not even knowledgeable statisticians agree on what the “best” or “most informative” effect-size measure actually is. (Levin & Robinson, 1999, p. 151)

Levin (1998b, pp. 45-46) similarly provided the following hypothetical example of the perplexing situation that effect sizes can create for researchers, readers, and other interpreters of the importance of an empirical finding:

Suppose that an investigator wants to help older adults remember an ordered set of ten important daily tasks that must be performed (insert and turn on a hearing aid, take certain pills, make a telephone call to a caregiver, etc.). In a sample of six elderly adults, three are randomly assigned to each of two experimental conditions. In one condition (A), no special task instruction is given; and in the other (B<sub>1</sub>), participants are instructed in the use of self-monitoring strategies. Following training, the participants are observed with respect to their success in performing the ten tasks...[T]he average number of tasks the participants correctly remembered to perform was 1.33 [*SD* = .577, raw scores = 1, 1, and 2] and 3.33 [*SD* = .577, raw scores = 3, 3, and 4] for the no-instruction (A) and self-monitoring (B<sub>1</sub>) conditions, respectively. For [these data], it can be determined that the “conditions” factor accounts for a hefty 82% of the total variation in task performance (i.e., the squared point-biserial correlation is .82, which for the two-sample case, is equivalent to the sample  $\eta^2$ ). Alternatively, the self-monitoring

mean is 3-1/2 within-group standard deviations higher than the no-instruction mean (i.e., Cohen’s *d* is 3.5). From either effect-size perspective ( $\eta^2$  or *d*), certainly this represents an impressive treatment effect, doesn’t it? Or does it?

Suppose that instead of self-monitoring training, participants were taught how to employ “mnemonic” (systematic memory-enhancing) techniques (B<sub>2</sub>) ...with the results [yielding a mean number correct of 7.67 (*SD* = 2.517, raw scores = 5, 8, and 10)]...[A] comparison with no-instruction Condition A surprisingly reveals that once again, the conditions factor accounts for 82% of the total variation in task performance (equivalently, *d* again equals 3.5). Thus, when expressed in standardized/relative terms (either  $\eta^2$  or *d*), the effect sizes associated with the two instructional conditions (B<sub>1</sub> and B<sub>2</sub>) are exactly the same, and substantial in magnitude. Yet, when expressed in absolute terms and with respect to the task’s maximum, there are important differences in the “effects” of B<sub>1</sub> and B<sub>2</sub>: Increasing participants’ average performance from 1.33 to 3.33 tasks remembered seems much less impressive than does increasing it from 1.33 to 7.67. Helping these adults remember an average of only 3 of their 10 critical tasks might be regarded as a dismal failure, whereas helping them remember an average of almost 8 out of 10 tasks would be a stunning accomplishment. Yet, the conventional effect-size measures are the same in each case.

To help shed light on this seeming paradoxical situation, Levin (1998b) pointed out:

The major problem in this example arises from the conditions' differing variabilities. That problem could be accounted for by defining alternative  $d$ -like effect-size measures based on just the control condition's (Condition A's) standard deviation... Interpreting effect sizes, in the absence of raw data, remains a problem for  $\eta^2$  and Cohen's  $d$ , however. (p. 53)

Insofar as different effect-size measures are suitable for different types of data (e.g., Hogarty & Kromrey, 2001), it is surprising that some researchers do not even indicate the index to which they are referring when reporting effect sizes (Kirk, 1996). Neither do researchers appear to indicate whether the effect-size measure interpreted represents an adjusted or unadjusted index. The lack of information provided is disturbing because meta-analyses involve aggregating and comparing effect sizes across studies. How can effect sizes be aggregated if it is not clear whether they are based on the same type of index? Unfortunately, the practice of some meta-analysts to omit unlabeled effect sizes from the aggregate index introduces bias.

4. *Varying, and generally arbitrary, guidelines for interpreting effect-size magnitudes.* As was noted earlier, a way in which statistical hypothesis testing is abused occurs when a dichotomous decision (i.e., reject vs. do not reject) comprises the sole determinant of the significance (read importance) of an observed outcome. This is done by comparing the outcome's significance probability ( $p$ -value) to some predetermined standard significance level ( $\alpha$  level), such as .05. Yet, many researchers who interpret effect sizes appear to use equally rigid categorical criteria such as those provided by Cohen (1988), who popularized the use of effect-size reporting. This occurs even though recommendations vary with respect to how effect sizes should be interpreted (McLean, O'Neal, & Barnette, 2000) and despite Cohen's (1988) admonishment that effect-size

values are dependent on the specific content and methods that prevail in a given research context.

For example, in interpreting effect sizes associated with differences between two groups (i.e., Cohen's  $d$ ), Cohen (1988) recommended demarcations of .20 for small effects, .50 for medium effects, and .80 for large effects. In stark contrast, McLean (1995) suggested the following criteria: .50 for small effects, between .50 and 1.00 for moderate effects, and above 1.00 for large effects. Regardless of which criteria are used, it is clear that adherence to such cutpoints has the effect of trichotomizing interpretations in much the same way as  $p$ -values dichotomize statistical decision making. As noted by Shaver (1993): "There already is a tendency to use criteria, such as Cohen's (1988) standards for small, medium, and large effect sizes, as mindlessly as has been the practice with the .05 criterion in statistical significance testing" (p. 311). Similarly, Thompson (2001) stated: "If people interpreted effect sizes [using fixed benchmarks] with the same rigidity that  $\alpha = .05$  has been used in statistical testing, we would merely be being stupid in another metric" (p. 82-83).

In addition, blending the previous concern (different effect-size measures may lead to different conclusions) with the present one (effect-size descriptors are arbitrary and vary by context) we consider the following confusing/conflicting medical-study conclusion presented by Rosenthal and DiMatteo (2001). The results of a study designed to examine the effect of taking aspirin on heart-attack prevention (Steering Committee of the Physicians' Health Study Research Group, 1988) yielded what is typically regarded as a tiny Pearson  $r$  of .034. Yet, when the same outcome is interpreted from the perspective of Rosenthal and Rubin's (1982) binomial effect size display (BESD), the "finding is, in fact, very important and translates into substantial reductions in morbidity and mortality" (Rosenthal & DiMatteo, 2001, p. 78). For related discussion on the potential importance of conventionally small effect sizes, see Prentice and Miller (1992).

5. *Sample size and sampling variability.* The interpretation of effect sizes also varies as a function of sample size. Studies with smaller sample sizes often result in effect sizes being overestimated, whereas investigations with large sample sizes tend to lead to effect sizes being



underestimated (Bakan, 1966; Fern & Monroe, 1996; Hedges & Olkin, 1985). Empirically, Barnette and McLean (1999) demonstrated that standardized effect-size variation is systematic rather than random. In their Monte Carlo investigation, these authors found that the number of groups and sample sizes were almost perfectly predictive (i.e.,  $R^2 = .999$ ) of standardized effect sizes. Thus, comparing effect sizes across studies with very different sample sizes can be misleading.

One of the most repeated criticisms of statistical hypothesis testing is its over-reliance on sample size (Cohen, 1994; Fan, 2001; Kirk, 1996; Onwuegbuzie & Daniel, 2003, in press; Schmidt & Hunter, 1997; Thompson, 1993). Yet, as was noted recently by Fan (2001): “effect size can also be misleading because sample size influences the sampling variability of an effect-size measure” (p. 275). Using Monte Carlo methods, Fan demonstrated that an observed finding that appears to have practical significance (i.e., a large effect size) actually could be the result of sampling error, thereby making any resultant conclusions unreliable and potentially misleading – which lends empirical support to a major facet of the argument promoted by Levin and Robinson (2000; see also Sawilowsky & Yoon, 2002), summarized later. Fan (2001) recommended that information about both statistical significance and effect sizes be reported for observed findings:

Statistical significance testing and effect size are two related sides that together make a coin; they complement each other but do not substitute for one another. Good research practice requires that, for making sound quantitative decisions in educational research, both sides should be considered. (p. 275)

It should come as no surprise that effect sizes are affected by sample size in much the same way as are  $p$ -values. Indeed, effect-size statistics represent random variables. Consequently, effect-size measures are affected by sampling variability, as dictated by its underlying sampling distribution. In turn, the amount of sampling variability of an effect-size estimate is influenced by the underlying

sample size, in much the same way that  $p$ -values are affected by the number of cases utilized in the study. When the sample size is small, the discrepancy between the sample effect size and population effect size is larger (i.e., large bias) than when the sample size is large. Also, effect sizes are affected by nonrandom sampling, a condition that applies to the vast majority of empirical studies in education and psychology. Thus, solutions to compensate for the problems stemming from the role of sample size in statistical hypothesis testing (e.g., use of confidence intervals) should also apply to effect sizes.

A valid criticism of hypothesis testing that is supported by data pertains to the low statistical power that prevails in many studies. Indeed, the average power of null hypothesis significance tests typically ranges from .40 to .60 in empirical studies (Cohen, 1962, 1965, 1988, 1994; Schmidt, 1996; Sedlmeier & Gigerenzer, 1989). With an estimated mean across-study power of .50 (Cohen, 1962, 1997), Schmidt and Hunter (1997) decry that “[t]his level of accuracy is so low that it could be achieved just by flipping a (unbiased) coin!” (p. 40). Yet, the finding that power is unacceptably low in most studies indicates to us that researchers’ application of statistical hypothesis testing, rather than its logic, is to blame. Indeed, it can be argued that low statistical power represents more of a research design issue than a statistical issue, since acceptable power can be rectified by incorporating a larger sample.

Unfortunately, as was discussed earlier, effect sizes also can fall victim to poor research designs, in general, and to small sample sizes, in particular. In fact, an obsession with effect sizes without considering the associated sample sizes can have the effect of promoting weak research designs. As such, in making decisions about which articles should be published, journal editors should focus less on  $p$ -values and effect sizes and more on the quality of the underlying research design (for related discussion and references, see Levin, 1998b, p. 45).

6. *Distribution nonnormality.* Although this may surprise or disturb some readers, many of the commonly used effect-size measures rely heavily on the parametric hypothesis-testing assumptions of normality and homogeneity of variance (see, for example, Fan, 2001, Barnette & McLean, 1999, and Hogarty & Kromrey, 2001).

The numerator of common effect-size measures involves means and mean differences, which are sensitive to extreme observations, especially when sample sizes are small (Huck, 2000). In the small-sample case, an extreme observation in one of the conditions (e.g., the experimental group) can seriously distort the true mean difference, thereby unduly influencing the effect-size estimate. Just as outlying observations affect the  $t$ -statistic and associated  $p$ -values (statistical significance), in the independent-samples test of means they also influence the effect size (practical significance). For this reason, nonparametric effect-size measures have been developed and considered.

Applying Monte Carlo methods, Hogarty and Kromrey (2001) demonstrated that the most frequently used effect-size estimates (e.g., Cohen's  $d$  and Hedges & Olkin's  $g$ ) are sensitive to departures from normality and variance homogeneity (discussed next). Even trimmed effect-size measures (Hedges & Olkin, 1985; Yuen, 1974) exhibit bias when sample sizes are small, as do several nonparametric effect-size indices, including  $Y_i$  (Kraemer & Andrews, 1982) and the Common Language (CL) effect-size statistic (McGraw & Wong, 1992).

7. *Score variability (both between and within samples)*. Other characteristics of the sample also affect interpretation of effect sizes. In particular, the more heterogeneous the sample is with respect to the variable of interest, the greater the effect size typically tends to be. This is the case for both explanatory and predictive studies (O'Grady, 1982). Moreover, homogeneous samples, which more often arise from convenience sampling, can result in range restriction and, subsequently, attenuate effect sizes (Pedhazur & Schmelkin, 1991). Recognition of this complicating situation can be seen in a recent critique of a report challenging the effectiveness of teacher education programs by Darling-Hammond and Youngs (2002):

The effect size also depends on other context factors, such as the range of variability in the measure used, which can change in different locations and time periods. For example, in some eras and in some locations virtually all teachers held content

degrees or were fully certified, so these variables do not strongly predict variations in outcomes. When much more variability is present, these variables are strongly predictive of outcomes. Thus, several studies have found strong measured influences of certification status on student achievement in states like California and Texas during the 1990s when there were wide differences in teachers' qualifications. (p. 15)

It is also possible for variance heterogeneity to reduce the effect size. This can be the case when the sample is *too* diverse and the heterogeneity increases error variance, thereby attenuating the effect size (Lesser, 1959).

Regardless of whether the effect size is increased or decreased by heterogeneous samples, interpreting effect sizes that arise from samples with different degrees of heterogeneity is inadvisable. In particular, researchers should exercise caution in comparing effect sizes across convenience samples. In fact, Daniel and Onwuegbuzie (2000) refer to sampling bias error that results in inconsistency of results across studies as a Type IX error. According to these authors, this type of error relates to "disparities in results generated from numerous convenience samples across a multiplicity of similar studies" (p. 23).

Further, because the denominator of common effect-size measures incorporates the pooled within-conditions variance, heterogeneity of variance affects effect-size estimation similarly to the way that it affects statistical hypothesis testing (and confidence-interval building - as was seen in Levin, 1998b, p. 53). Moreover, the problems caused by departures from normality and heterogeneity of variance when statistical significance testing is involved are very much an issue for effect-size measures associated with more complex family members of the general linear model. For example, the standard effect-size indices (e.g.,  $\eta^2$ ,  $\epsilon^2$ , and  $\omega^2$ ) that are often calculated for OVA-type analyses (e.g., ANOVA, ANCOVA, MANOVA) assume equal variances -

an assumption that is not always met (Onwuegbuzie & Daniel, 2003).

However, these weaknesses do not imply that effect sizes should be banned or replaced by some other sort of index, echoing what some researchers (e.g., Carver, 1993) recommend should be the fate of statistical significance testing. Indeed, in cases where such violations come to the fore, nonparametric effect sizes (e.g.,  $Y_I$  and CL) may be more appropriate, in much the same way that nonparametric inferential statistics often are more appropriate when the parametric assumptions are violated. The above limitations pertaining to effect sizes identified above suggest that: (a) assumptions underlying the selected effect-size method should be subjected to the same stringent scrutiny as are statistical significance tests; (b) combining statistical significance testing and effect-size indices, after checking all pertinent assumptions, provides an additional safety net from false or misleading conclusions, compared to using either technique alone; and (c) researchers should pay much more attention to maximizing the quality of their research designs (e.g., by selecting an appropriate or optimal sample size) in order to minimize threats to the model assumptions that pertain to both the statistical test and the accompanying effect-size measure of interest.

8. *Reliability of the outcome measure (measurement error)*. Reliability is a concept that receives disproportionately scant attention in the interpretation of an observed finding (Onwuegbuzie & Daniel, 2000, 2001, 2003, in press; Onwuegbuzie, Daniel, & Roberts, in press; Roberts & Onwuegbuzie, 2003; Roberts, Onwuegbuzie, & Eby, 2001; Onwuegbuzie & Weems, in press; Weems & Onwuegbuzie, 2001). Reliability (or more precisely, unreliability) can adversely affect the internal validity of findings via “instrumentation” problems (e.g., Campbell & Stanley, 1963; Onwuegbuzie, 2003), through a reduction in statistical power. Specifically, Onwuegbuzie and Daniel (in press) demonstrated that subgroups with scores that generate markedly different reliability estimates can seriously reduce statistical power, even when the full-sample (i.e., across-groups) reliability coefficient is adequate.

Importantly, however, low reliability indices adversely affect not just statistical hypothesis testing; they also negatively impact effect-size measures. After all, low reliability

coefficients stem from scores that do not behave in a consistent manner (Onwuegbuzie & Daniel, 2000, 2001) and it is these scores that are used to calculate both inferential test statistics and effect-size measures. Thus, effect-size measures are subject to the same limitations stemming from inadequate reliability as are  $p$ -values. Indeed, effect sizes should always be interpreted with respect to the reliability of the outcome measure, just as has been recommended for statistical hypothesis testing.

Specifically, there is an inverse relationship between the reliability of any of the variables of interest (whether the independent or dependent variable) and the corresponding effect size. In fact, such reliability provides an upper bound for the effect size (Lord & Novick, 1968; Nunnally & Bernstein, 1994). Because a study’s reliability is a function of the study’s obtained scores rather than *a priori* test norms (Onwuegbuzie & Daniel, 2000, 2002a, 2002b; Thompson & Vacha-Haase, 2000; Vacha-Haase, Kogan, & Thompson, 2000; Wilkinson & Task Force on Statistical Inference, 1999), effect sizes should not be compared across studies without taking into account the individual studies’ outcome-measure reliabilities. For further discussion of reliability and effect size in both correlational and experimental study contexts, see O’Grady (1982, pp. 767-770).

9. *Scale of measurement*. The type and range of measure used can affect the size of the effect. It is not unusual for researchers studying a phenomenon to use different measures. In particular, in a study of an affective variable, whereas one researcher might use a Likert-type scale, another researcher might employ a rating scale. Still another researcher might employ a semantic differential scale or a Thurstone or Guttman scale. Similarly, in an investigation of a cognitive outcome, whereas one researcher might administer a multiple-choice test, another researcher might administer some other type of closed-ended instrument (e.g., true-false, matching), and still another researcher might administer an open-ended measure such as an essay.

Although all of these measures yield scores that can be analyzed statistically, each type of scale might not be measuring exactly the same construct. For instance, multiple-choice and essay

examinations often target different levels of learning in Bloom's taxonomy of cognitive objectives (Bloom, 1956). As such, the effect size likely would vary as a function of the type of measure used. Although this apples-and-oranges situation is typically offered as the primary rationale for meta-analytic effect-size combinations (e.g., Hunt, 1997; Rosenthal & DiMatteo, 2001), it rarely is recognized as a study-comparison concern.

Even if scales with the same item format (e.g., a Likert-scale format) are used across studies, both the number/type of items and the number/type of response options employed can affect the size of the effect. With respect to the former, compared to their counterparts with more items, scales with a smaller number of items lead to restriction of range, thereby attenuating effect sizes. Similarly, the proportion of negatively worded and positively worded items can influence the effect size (Onwuegbuzie & Weems, in press; Weems & Onwuegbuzie, 2001). With regard to the latter, the number of response options also can influence the effect size. Specifically, a reduction in the number of response options attenuates the range of scores, which, in turn, may reduce the magnitude of the effect.

Similarly, and as was mentioned earlier, a restriction in the variability of one or more variables typically decreases the effect size. This holds for a study's independent variables, as well as its outcome measures. As noted by Onwuegbuzie and Daniel (2003), lacking the realization that nearly all parametric analyses represent the general linear model, many analysts inappropriately categorize independent variables in nonexperimental research designs in order to perform analyses such as analysis of variance. Disturbingly, findings from such analyses are then used to make causal inferences, when all that has occurred is a discarding of relevant variance – see, for example Cliff (1987); Pedhazur (1982); Prosser (1990); and Thompson (1986, 1988, 1992a).

Yet, categorizing a continuous variable has been found repeatedly to reduce the effect size. For instance, a median split of a continuous variable can reduce the observed correlation by 20% (Cohen, 1983; Hunter & Schmidt, 1990) – see also Vargha, Rudas, Delaney, & Maxwell (1996). If the cutpoint used for splitting the

continuous variable differs from the median, then the reduction in the relationship between the variables can be expected to be even larger (Fern & Monroe, 1996).

Moreover, as the number of categorized groups decreases, less variance in the dependent variable is accounted for by the categorical variable, compared to the continuous variable, and thus the effect size is attenuated (Peet, 1999). With regard to type of response options, the use of midpoint categories (e.g., neutral response options) has been found to affect both score reliability and effect size (Weems & Onwuegbuzie, 2001). Therefore, comparing effect sizes across studies using different types and formats of scales is questionable.

In addition, it does not appear to be obvious to some researchers that effect sizes are a function of the scale of measurement used. Evidence of this is provided by McLean et al. (2000), who demonstrated that “gain” effect sizes were different for the raw scores, scaled scores, and Normal Curve Equivalent (NCE) scores for students in Grades 4, 6, and 8 on a national norm-referenced test. Specifically, as McLean et al. expected, the effect sizes for NCE scores were lower than those for raw and scaled scores. The researchers appropriately concluded that when effect sizes are computed, researchers should take into account the scale of measurement on which they are based.

### Summary

We have highlighted nine general concerns about effect-size indices. When researchers design their studies, they must make numerous decisions. Each of these decisions can affect the magnitude of the effect-size estimate. Unfortunately, the extent to which the effect-size index is influenced by the decisions is almost always unknown. This suggests that researchers are not justified in reflexively applying Cohen's (1988) effect-size magnitude and adjectival guidelines across studies in different domains or across studies that have different research design and analytical factors. Even more importantly, because effect sizes vary as a function of research-related factors, effect sizes should be compared only when all of these factors are comparable. Assessing the substantive significance of an observed finding based solely on the effect size

may be misleading and no more diagnostic than is a test of a statistical hypothesis (Fern & Monroe, 1996).

This does not mean that effect sizes are useless. As noted by Fern and Monroe (1996), if the goal of the researcher is to determine the size of an effect given the unique combination of factors that underlie the data, then a computed measure of effect size is informative. On the other hand, effect sizes cannot be used as a meaningful basis for comparison across studies “unless the researcher understands what, if any, unique factors contributed to the effect-size estimate” (Fern & Monroe, 1996, p. 102). In any case, when reporting effect sizes, researchers should always specify as many design, analysis, and psychometric characteristics as possible to help subsequent researchers decide the extent to which they can compare their effect sizes with previous estimates. In other words, researchers should contextualize their effect sizes (i.e., they should interpret their effect sizes within study’s specific parameters).

Many researchers who criticize statistical hypothesis testing, in general, and those who advocate replacing  $p$ -values with effect size measures, in particular, fail to mention any of the limitations associated with effect-size reporting. Thus, methodologists who criticize hypothesis testing without also discussing the limitations of effect sizes are not providing a balanced analysis but are focusing on the bad practices that have traditionally been linked to the former approach. Unfortunately, the just-mentioned concerns about effect sizes typically are not mentioned by their advocates. In discussing the limitations, we argue that effect sizes are not the hoped-for panacea for empirical research in the social sciences.

Further, we contend that if *only* effect sizes were used to interpret statistical data, social-science research would not be in any better position than it would if only statistical hypothesis testing were used in quantitative studies. In fact, in an effect-size-only world, we submit that social-science research would be in a *worse* position, in that progress would be retarded (Thompson, 1992b) to an even greater extent than that imagined by hypothesis-testing critics, in that statistically “chance” findings would unjustifiably be promoted by researchers as “real.” We

reconsider that unfortunate situation in the following concluding section.

#### Toward a Détente

The effect-size flaws that we have reviewed support the assertion that statistical hypothesis testing and effect-size reporting should be used in combination. A logical, internally consistent, way of combining these two procedures is through Robinson and Levin’s (1997) two-step suggestion for analyzing empirical data – namely, that effect sizes are reported if and only if the observed finding is statistically significant.

That is, statistical hypothesis testing should serve as a gatekeeper, guarding against spurious effect-size estimation. As noted by Robinson and Levin (1997), the goal of these two complementary approaches is to prevent the over-interpretation of seemingly impressive effect sizes “in the absence of formal assessments of their likelihood” (p. 23). We therefore recommend that statistical hypothesis testing and effect-size estimation be used in tandem to establish a reported outcome’s believability and magnitude, respectively. As such, tests of significance serve a valuable purpose in determining whether effect-size measures should be ignored or reported, a position endorsed by Fan (2001), Levin (1993), Robinson and Levin (1997), Knapp and Sawilowsky (2001), and even – we think – Gliner, Leech, and Morgan (2002).

Let us take a moment to consider the last part of the foregoing sentence. We say “even” because Gliner et al.’s recommendation appeared in a journal whose editorial policy specifically calls for effect-size inclusions even in the absence of statistical confirmation: “Furthermore, authors are required to report and interpret magnitude-of-effect measures in conjunction with every  $p$  value that is reported” (Journal of Experimental Education, 2002, p. 94). We say “we think” because Gliner et al. are internally inconsistent in their position about *always* reporting and interpreting effect sizes in their position.

For example, they agree with Levin and Robinson’s (2000) distinction between single-study investigations and multiple-study syntheses: “Our opinion is that effect sizes should accompany all reported  $p$  values for possible future meta-analytic use, but they should not be presented as

findings in a single study in the absence of statistical significance” (Gliner et al., 2000, p. 86). Yet, in the penultimate sentence of their article they write: “We also recommend reporting effect size for nonsignificant outcomes” (p. 91). Addressing this blanket effect-size reporting recommendation, one of us has pointed out previously:

This practice is absurdly pseudoscientific and opens the door to encouraging researchers to make something of an outcome that may be nothing more than a “fluke,” a chance occurrence. Without an operationally replicable screening device such as statistical hypothesis testing, there is no way of separating the wheat (statistically “real” relationships or effects) from the chaff (statistically “chance” ones), where “real” and “chance” are anchored in reference to either conventional or researcher-established risks or “confidence levels.”...In its extreme form, effect-size-only reporting degenerates to strong conclusions about differential treatment efficacy that are based on comparing a single score of one participant in one treatment condition with that of another participant in a different condition. (Levin, 1998b, p. 45)

Moreover, in a recent survey of the editorial board members of four educational-research journals (Capraro & Capraro, 2003), the 97 respondents (estimated from the data provided) greeted the recommendation that their journals *require* effect-size reporting with overwhelming indifference: On a 7-point Likert scale ranging from “very strongly disagree” to “very strongly agree” the mean rating was 4.26,  $t(96) = 1.33$ ,  $p = .19$ , for testing the hypothesis that respondents’ mean ratings do not differ from the scale midpoint of 4. Given the study’s relatively large sample size, this nonrejection of the indifference hypothesis should be taken with more than a grain of salt.

One additional internal-inconsistency irony – or at least an example of journal non-policing – is worth mentioning. In an article published by one of the present authors (Hwang & Levin, 2002) in the same issue of the *Journal of Experimental Education* that proclaims the above effect-size policy, effect sizes were *not* reported for every  $p$ -value included; nor were they reported for statistically nonsignificant outcomes. Yet, somehow, some way, the article was published anyway! And this is not an isolated event.

A colleague, Dan Robinson, has experienced effect-size nonenforcement with two of his articles that were published in the same journal (Katayama & Robinson, 2000; Robinson, Katayama, Dubois, & Devaney, 1998), a journal that has promoted its effect-size policy since 1997 (D. H. Robinson, personal communication, January 13, 2003). As with Thompson’s (e.g., 1996) argument in other contexts, perhaps *JEE* should be *encouraged* to take a closer look at its own editorial policy, for in that journal effect-size endorsement clearly does not translate into effect-size enforcement. As an informative aside, the *Journal of Experimental Education* is apparently not alone in its effect-size non-enforcement practices for D. H. Robinson (personal communication, January 22, 2003) indicates a similar phenomenon with another effect-size mandated journal, *Contemporary Educational Psychology*. Out of 11 intervention experiments that he tallied for that journal in 2001, only two were accompanied by effect-size estimates.

Even those who contend that effect sizes should *replace* statistical significance testing (e.g., Carver, 1993; Schmidt, 1996) recommend the use of confidence intervals alongside effect sizes. A two-sided confidence interval, characterized by lower and upper bounds, identifies a probable range of magnitudes for the effect size (Abelson, 1997). As such, confidence intervals can be used to estimate the range of the effect’s practical significance – for related discussion, see Onwuegbuzie (2001) and Thompson (2002).

Moreover, insofar as confidence intervals include all the information provided by statistical hypothesis tests, and more (Cohen, 1994; Levin, 1998b; Serlin, 1993), constructing them allows researchers to conduct the corresponding hypothesis tests, if desired (Krantz, 1999). In that sense, then, the provision of an inferential

confidence interval (instead of a hypothesis test) has logical appeal because that approach kills two birds (statistical and practical significance) with one stone. So as not to confuse the issue, it should be made clear that the kind of confidence-interval approach we are endorsing is the single-interval procedure based on a pre-experimentally established Type I error probability, which is inferentially equivalent to applying a Neyman-Pearson statistical test of hypothesis. This approach is fundamentally and logically different from that espoused by certain hypothesis-testing critics, which would have researchers simultaneously provide multiple confidence intervals (for either raw or standardized effects) based on different confidence levels, such as 99%, 95%, 90%, 80%, etc. – see, for example, Schmidt & Hunter (1997) and Thompson (2002).

Alternatively, hypothesis testing *per se* can be substantially improved (strengthened) by applying it in forms that are more intelligent than the one that is currently practiced. Such more intelligent forms call for researchers to formulate/test more theoretically driven and precise hypotheses, to determine (through power calculations) optimal sample sizes to test those hypotheses, and to incorporate equivalence-testing procedures (e.g., Seaman & Serlin, 1998) for better establishing the truth of the null hypothesis (see, for example, Levin, 1998a, pp. 329-330).

At the same time, we contend that hypothesis tests, confidence intervals, and effect sizes do not go far enough in the way of maximizing a domain's knowledge base. This can be accomplished only through independent replications of results (i.e., two or more independent studies yielding similar findings that produce statistically and substantively compatible outcomes). We believe that “a replication is worth a thousandth  $p$  value” (Levin, 1995), as well as its being worth more than a large effect size based on a single study. In contrast to Carver (1978), however, we do not believe that “replicated results should automatically make statistical significance unnecessary” (p. 393). Such independent replications not only will make “invaluable contributions to the cumulative knowledge in a given domain” (Robinson & Levin, 1997, p. 25) but will also help empirical researchers achieve a common goal.

## Conclusion

As was noted by Onwuegbuzie (2003), a primary objective of empirical research – especially research designed to posit causal relationships – is to collect and analyze data that help a researcher make inferences from the sample(s) to the underlying population, leading to meaningful conclusions in which as many rival explanations as possible are eliminated. This is the goal that drives both statistical hypothesis testing and effect-size reporting. The extant literature has documented the limitations of hypothesis testing, whereas in this paper we have illustrated that effect-size interpretation is not without its flaws. No single index by itself is the magic bullet for analyzing and interpreting data. Rather, using both methods in combination, or combining confidence intervals and effect sizes, helps to rule out more rival threats to statistical-conclusion validity (Cook & Campbell, 1979; Shadish, Cook & Campbell, 2002) than would occur if either method were used alone to interpret observed findings. At the same time, however, to minimize both statistical-conclusion validity and external validity threats there is no substitute for independent replications.

## References

- Abelson, R. P. (1997). A retrospective on the significance test ban of 1999 (If there were no significance tests, they would be invented). In L.L. Harlow, S.A. Mulaik, & J.H. Steiger (Eds.), *What if there were no significance tests?* Mahwah, NJ: Erlbaum, p. 117-141.
- American Psychological Association. (2001). *Publication manual of the American Psychological Association* (5th ed.). Washington, DC: Author.
- Anderson, D. R., Burnham, K. P., & Thompson, W. L. (2000). Null hypothesis testing: Problems, prevalence, and an alternative. *Journal of Wildlife Management*, *64*, 912-923.
- Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin*, *66*, 423-437.

- Barnette, J. J., & McLean, J. E. (1999, November). *Empirically based criteria for determining meaningful effect size*. Paper presented at the annual meeting of the Mid-South Educational Research Association, Point Clear, AL.
- Berkson, J. (1938). Some difficulties of interpretation encountered in the application of the chi-square test. *Journal of the American Statistical Association*, *33*, 526-536.
- Berkson, J. (1942). Tests of significance considered as evidence. *Journal of the American Statistical Association*, *37*, 325-335.
- Bloom, B. S. (Ed.). (1956). *Taxonomy of educational objectives: The classification of educational goals, Handbook I: Cognitive domain*. New York: Longman, Green.
- Boring, E. G. (1919). Mathematical vs. scientific importance. *Psychological Bulletin*, *16*, 335-338.
- Cahan S. (2000). Statistical significance is not a "Kosher Certificate" for observed effects: A critical analysis of the two-step approach to the evaluation of empirical results. *Educational Researcher*, *29*(1), 31-34.
- Calder, B. J., Phillips, L. W., & Tybout, A. M. (1981). Designing research for application. *Journal of Consumer Research*, *8*, 197-207.
- Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally.
- Capraro, R. M., & Capraro, M. M. (April, 2003). *Exploring the APA fifth edition Publication Manual's impact on the preferences of journal editorial board members*. Paper presented at the annual meeting of the American Educational Research Association, Chicago.
- Carver, R. P. (1978). The case against statistical significance testing. *Harvard Educational Review*, *48*, 378-399.
- Carver, R. P. (1993). The case against statistical significance testing, revisited. *Journal of Experimental Education*, *61*, 287-292.
- Cliff, N. (1987). *Analyzing multivariate data*. San Diego: Harcourt Brace Jovanovich.
- Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal Psychology*, *65*, 145-153.
- Cohen, J. (1965). Some statistical issues in psychological research. In B.B. Wolman (Ed.), *Handbook of clinical psychology*. NY: McGraw-Hill, p. 95-121.
- Cohen, J. (1983). The cost of dichotomization. *Applied Psychological Measurement*, *7*, 249-253.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. New York: John Wiley.
- Cohen, J. (1994). The earth is round ( $p < .05$ ). *American Psychologist*, *49*, 997-1003.
- Cohen, J. (1997). The earth is round ( $p < .05$ ). In L.L. Harlow, S.A. Mulaik, & J.H. Steiger (Eds.), *What if there were no significance tests?* Mahwah, NJ: Erlbaum, p. 117-141.
- Cook, T. D & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Chicago: Rand-McNally.
- Daniel, L. G., & Onwuegbuzie, A. J. (2000, November). *Toward an extended typology of research errors*. Paper presented at the annual meeting of the Mid-South Educational Research Association, Bowling Green, KY.
- Darling-Hammond, L., & Youngs, P. (2002). Defining "highly qualified teachers": What does "scientifically-based research" actually tell us? *Educational Researcher*, *31*(9), 13-25.
- Falk, R., & Greenbaum, C. W. (1995). Significance tests die hard: The amazing persistence of a probabilistic misconception. *Theory & Psychology*, *5*, 75-98.
- Fan, X. (2001). Statistical significance and effect size in education research: Two sides of a coin. *Journal of Educational Research*, *94*, 275-282.
- Fern, E. F., & Monroe, K. B. (1996). Effect-size estimates: Issues and problems in interpretation. *Journal of Consumer Research*, *23*, 89-105.
- Fisher, R. A. (1925/1941). *Statistical methods for research workers* (84th ed.) Edinburgh, Scotland: Oliver & Boyd. (Original work published in 1925).
- Frick, R. W. (1995). *Using statistics: Prescription versus practice*. Unpublished manuscript, Department of Psychology, State University of New York at Stony Brook.



- Gliner, J. A., Leech, N. L., & Morgan, G. A. (2002). Problems with null hypothesis significance testing (NHST): What do the textbooks say? *Journal of Experimental Education, 71*, 83-92.
- Guttman, L. B. (1985). The illogic of statistical inference for cumulative science. *Applied Stochastic Models and Data Analysis, 1*, 3-10.
- Harlow, L. L., Mulaik, S. A., & Steiger, J. H. (1997, Eds.). *What if there were no significance tests?* Mahwah, NJ: Erlbaum.
- Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Orlando, FL: Academic Press.
- Hogarty, K. Y., & Kromrey, J. D. (2001, April). *We've been reporting some effect sizes: Can you guess what they mean?* Paper presented at the annual meeting of the American Educational Research Association, Seattle, WA.
- Huck, S. W. (2000). *Reading statistics and research* (3rd ed.). New York: Addison Wesley Longman.
- Hunt, M. (1997). *How science takes stock: The story of meta-analysis*. New York: Russell Sage.
- Hunter, J. E., & Schmidt, F. L. (1990). *Methods of meta-analysis: Correcting error and bias in research findings*. Newbury Park, CA: Sage.
- Hwang, Y., & Levin, J. R. (2002). Examination of middle-school students' independent use of a complex mnemonic system. *Journal of Experimental Education, 71*, 25-38.
- Keppel, G. (1991). *Design and analysis: A researcher's handbook*. Englewood Cliffs, NJ: Prentice-Hall.
- Katayama, A. D., & Robinson, D. H. (2000). Getting students "partially" involved in note-taking using graphic organizers. *Journal of Experimental Education, 68*, 119-133.
- Kerlinger, F. N. (1973). *Foundations of behavioral research*. New York: Holt, Rinehart, & Winston.
- Kirk, R. E. (1995). *Experimental design: Procedures for the behavioral sciences* (3rd ed.). Pacific Grove, CA: Brooks/Cole.
- Kirk, R. E. (1996). Practical significance. A concept whose time as come. *Education and Psychological Measurement, 56*, 746-759.
- Knapp, T. R., & Sawilowsky, S. S. (2001). Constructive criticisms of methodological and editorial practices. *Journal of Experimental Education, 70*, 65-79.
- Kraemer, H. C., & Andrews, G. A. (1982). A nonparametric technique for meta analysis effect size calculation. *Psychological Bulletin, 91*, 404-412.
- Krantz, D. H. (1999). The null hypothesis testing controversy in psychology. *Journal of the American Statistical Association, 94*, 1372-1381.
- Lesser, G. S. (1959). Population difference in construct validity. *Journal of Consulting Psychology, 23*, 60-65.
- Levin, J. R. (1967). Misinterpreting the significance of "explained variation." *American Psychologist, 22*, 675-676.
- Levin, J. R. (1993). Statistical significance testing from three perspectives. *Journal of Experimental Education, 61*, 378-382.
- Levin, J. R. (1995, April). *The consultant's manual of researchers' common statistical disorders*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.
- Levin, J. R. (1998a). To test or not to test  $H_0$ ? *Educational and Psychological Measurement, 58*, 313-333.
- Levin, J. R. (1998b). What if there were no more bickering about statistical significance tests? *Research in the Schools, 5*, 43-53.
- Levin, J. R., & Robinson, D. H. (1999). Further reflections on hypothesis testing and editorial policy for primary research journals. *Educational Psychological Review, 11*, 143-155.
- Levin, J. R., & Robinson, D. H. (2000). Statistical hypothesis testing, effect-size estimation, and the conclusion coherence of primary research studies. *Educational Researcher, 29*(1), 34-36.
- Loftus, G. R. (1996). Psychology will be a much better science when we change the way we analyze data. *Current Directions in Psychology, 5*, 161-171.
- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, Massachusetts: Addison-Wesley.
- Maxwell, S. E., & Delaney, H. D. (1990). *Designing experiments and analyzing data: A model comparison perspective*. Belmont, CA: Wadsworth.

- McGraw, K. O., & Wong, S. P. (1992). A common language effect size statistic. *Psychological Bulletin*, *111*, 361-365.
- McLean, J. E. (1995). *Improving education through action research: A guide for administrators and teachers*. Thousand Oaks, CA: Corwin Press.
- McLean, J. E., O'Neal, M. R., & Barnette, J. J. (November, 2000). *Are all effect sizes created equal?* Paper presented at the annual meeting of the Mid-South Educational Research Association, Bowling Green, KY.
- Meehl, P. E. (1967). Theory testing in psychology and physics: A methodological paradox. *Philosophy of Science*, *34*, 103-115.
- Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, *46*, 806-834.
- Neyman, J., & Pearson, E. S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference. *Biometrika*, *29A*, Part I: 175-240; part II 263-294.
- Nix, T. W., & Barnette, J. J. (1998). The data analysis dilemma: Ban or abandon. A review of null hypothesis significance testing. *Research in the schools*, *5*, 3-14.
- Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.
- O'Grady, K. E. (1982). Measures of explained variance: Cautions and limitations. *Psychological Bulletin*, *92*, 766-777.
- Olejnik, S., & Algina, J. (2000). Measures of effects size for comparative studies: Applications, interpretations, and limitations. *Contemporary Educational Psychology*, *25*, 241-286.
- Onwuegbuzie, A. J. (2001). *Towards a framework for comprehensive reporting of empirical findings: The role of statistical significance, theoretical significance, practical significance, and clinical significance*. Unpublished manuscript, Howard University, Washington, DC.
- Onwuegbuzie, A. J. (2003). Expanding the framework of internal and external validity in quantitative research. *Research in the Schools*, *10*, 71-90.
- Onwuegbuzie, A. J., & Daniel, L. G. (2000, November). *Reliability generalization: The importance of considering sample specificity, confidence intervals, and subgroup differences*. Paper presented at the annual meeting of the Mid-South Educational Research Association, Bowling Green, KY.
- Onwuegbuzie, A. J., & Daniel, L. G. (2001, April). *Indices of score reliability and their applications*. Paper presented at the annual meeting of the American Educational Research Association, Seattle, WA.
- Onwuegbuzie, A. J., & Daniel, L. G. (2002a). A framework for reporting and interpreting internal consistency reliability estimates. *Measurement and Evaluation in Counseling and Development*, *35*, 89-103.
- Onwuegbuzie, A. J., & Daniel, L. G. (2002b). Uses and misuses of the correlation coefficient. *Research in the Schools*, *9*, 73-90.
- Onwuegbuzie, A.J., & Daniel, L.G. (2003, February 12). Typology of analytical and interpretational errors in quantitative and qualitative educational research. *Current Issues in Education* [On-line], *6*(2). Available at <http://cie.ed.asu.edu/volume6/number2/>
- Onwuegbuzie, A. J., & Daniel, L. G. (in press). Reliability generalization: The importance of considering sample specificity, confidence intervals, and subgroup differences. *Research in the Schools*.
- Onwuegbuzie, A. J., Daniel, L. G., & Roberts, J. K. (in press). A proposed new "what if" reliability analysis for assessing the statistical significance of bivariate relationships. *Measurement and Evaluation in Counseling and Development*
- Onwuegbuzie, A. J., & Weems, G. H. (in press). Characteristics of item respondents who frequently utilize midpoint response categories on rating scales. *Research in the Schools*.
- Pedhazur, E. J. (1982). *Multiple regression in behavioral research: Explanation and prediction* (2nd ed.). New York: Holt, Rinehart and Winston.
- Pedhazur, E. J., & Schmelkin, L. P. (1991). *Measurement, design, and analysis: An integrated approach*. Hillsdale, NJ: Erlbaum.

Peet, M. W. (1999, November). *The importance of variance in statistical analysis: Don't throw the baby out of the bathwater*. Paper presented at the annual meeting of the Mid-South Educational Research Association, Point Clear, Alabama.

Prentice, D. A., & Miller, D. T. (1992). When small effects are impressive. *Psychological Bulletin*, *112*, 160-164.

Prosser, B. (1990, January). *Beware the dangers of discarding variance*. Paper presented at the annual meeting of the Southwest Educational Research Association, Austin, TX. (ERIC Reproduction Service No. ED 314 496)

Roberts, J. K., & Onwuegbuzie, A. J. (2003). Alternative approaches for interpreting alpha with homogeneous subsamples. *Research in the School*, *10*, 63-69.

Roberts, J. K., Onwuegbuzie, A. J., & Eby, R. (2001, April). *Alternative approaches for interpreting alpha with homogeneous subsamples: The introduction of a new measure of homogeneous alpha*. Paper presented at the annual meeting of the American Educational Research Association, Seattle, WA.

Robinson, D. H., Katayama, A. D., Dubois, N. F., & Devaney, T. (1998). Interactive effects of graphic organizers and delayed review on concept acquisition. *Journal of Experimental Education*, *67*, 17-31.

Robinson, D. H., & Levin, J. R. (1997). Reflections on statistical and substantive significance, with a slice of replication. *Educational Researcher*, *26*(5), 21-26.

Rosenthal, R., & DiMatteo, M. R. (2001). Meta-analysis: Recent developments in quantitative methods for literature reviews. *Annual Review of Psychology*, *52*, 59-82.

Rosenthal, R., & Rubin, D. B. (1982). A simple, general purpose display of magnitude of experimental effect. *Journal of Educational Psychology*, *74*, 166-169.

Rozeboom, W. W. (1960). The fallacy of the null hypothesis significance test. *Psychological Bulletin*, *57*, 416-428.

Rozeboom, W. W. (1997). Good science is abductive, not hypothetic-deductive. In L.L. Harlow, S.A. Mulaik, & J.H. Steiger (Eds.), *What if there were no significance tests?* Mahwah, NJ: Erlbaum, p. 335-392.

Sawilowsky, S. S., & Yoon, J. (2002). The trouble with trivials ( $p > .05$ ). *Journal of Modern Applied Statistical Methods*, *1*, 143-144.

Schmidt, F. L. (1992). What do data really mean? *American Psychologist*, *47*, 1173-1181.

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for the training of researchers. *Psychological Methods*, *1*, 115-129.

Schmidt, F. L., & Hunter, J. E. (1997). Eight common but false objections to the discontinuation of significance testing in the analysis of research data. In L.L. Harlow, S.A. Mulaik, & J.H. Steiger (Eds.), *What if there were no significance tests?* Mahwah, NJ: Erlbaum, 37-64.

Seaman, M. A., & Serlin, R. C. (1998). Equivalence confidence intervals for two-group comparisons of means. *Psychological Methods*, *3*, 403-411.

Sechrest, L., & Yeaton, W. H. (1982). Magnitudes of experimental effects in social science research. *Evaluation Review*, *6*, 579-600.

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin*, *105*, 309-316.

Serlin, R. C. (1993). Confidence intervals and the scientific method: A case for Holm on the range. *Journal of Experimental Education*, *61*(4), 350-360.

Shadish, W. R., Cook, T. D., & Campbell, D.T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.

Shaver, J. P. (1993). What statistical significance testing is, and what it is not. *Journal of Experimental Education*, *61*, 293-316.

Steering Committee of the Physicians' Health Study Research Group (1988). Findings from the aspirin component of the ongoing Physicians' Health Study. *New England Journal of Medicine*, *318*, 162-264.

Thompson, B. (1986). ANOVA versus regression analysis of ATI designs: An empirical investigation. *Educational and Psychological Measurement*, *46*, 917-928.

Thompson, B. (1988). Discard variance: A cardinal sin in research. *Measurement and Evaluation in Counseling and Development*, *21*, 3-4.

Thompson, B. (1992a, April). *Interpreting regression results: Beta weights and structure coefficients are both important*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Thompson, B. (1992b). Two and one-half decades of leadership in measurement and evaluation. *Journal of Counseling and Development, 70*, 434-438.

Thompson, B. (1993). The use of statistical significance research: Bootstrap and other alternatives. *The Journal of Experimental Education, 61*, 361-377.

Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. *Educational Researcher, 25*(2), 26-30.

Thompson, B. (2001). Significance, effect sizes, stepwise methods, and other issues: Strong arguments move the field. *Journal of Experimental Education, 70*, 80-93.

Thompson, B. (2002). What future quantitative social science research could look like: Confidence intervals for effect sizes. *Educational Researcher, 31*(3), 25-32.

Thompson, B., & Vacha-Haase, T. (2000). Psychometrics is datametrics: The test is not reliable. *Educational and Psychological Measurement, 60*, 174-195.

Tryon, W. W. (1998). The inscrutable null hypothesis. *American Psychologist, 53*, 796.

Vacha-Haase, T., Kogan, L. R., & Thompson, B. (2000). Sample compositions and variabilities in published studies versus those in test manuals: Validity of score reliability inductions. *Educational and Psychological Measurement, 60*, 509-522.

Vargha, A., Rudas, T., Delaney, H. D., & Maxwell, S. E. (1996). Dichotomization, partial correlation, and conditional independence. *Journal of Educational and Behavioral Statistics, 21*, 264-282.

Weems, G. H., & Onwuegbuzie, A. J. (2001). The impact of midpoint responses and reverse coding on survey data. *Measurement and Evaluation in Counseling and Development, 34*, 166-176.

Wilkinson, L., & the Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*, 594-604.

Yuen, K. K. (1974). The two-sample trimmed *t* for unequal population variances. *Biometrika, 61*, 165-170.

## Performing Two-Way Analysis of Variance Under Variance Heterogeneity

Scott J. Richter  
Department of Mathematical Sciences  
University of North Carolina at Greensboro

Mark E. Payton  
Department of Statistics  
Oklahoma State University

---

Small sample properties of the method proposed by Brunner et al. (1997) for performing two-way analysis of variance are compared to those of the normal based ANOVA method for factorial arrangements. Different effect sizes, sample sizes, and error structures are utilized in a simulation study to compare type I error rates and power of the two methods. An SAS program is also presented to assist those wishing to implement the Brunner method to real data.

Key words: Factorial arrangement of treatments, heterogeneity of variance

---

### Introduction

Normal theory methods for analysis of variance depend on the assumption of homogeneity of the variance of the error distribution. For a one-way treatment structure, modifications are available when the homogeneity of variance assumption is violated. Milliken and Johnson (1992) suggest a method due to Box (1954) when sample sizes are equal. When sample sizes are unequal, they suggest Welch's (1951) test.

For multifactor layouts, however, there are few options available for testing effects of interaction and main effects. A parametric approach to this problem was presented by Weerahandi (1995), but it requires complex and intensive computing and isn't yet practical for use on real data. Papers by Akritas (1990), Thompson (1991) and Akritas and Arnold (1994) present nonparametric rank test statistics in a multi-way ANOVA setting. One should see Brunner, et al. (1997) for a survey of references relating to this topic.

One method that does not require the equal variance assumption is based on a Wald statistic, which has an asymptotic chi-square distribution. This method tends to reject too frequently under the null hypothesis for small samples. In fact, simulations of Brunner, et al. (1997) show the test to be liberal (by as much as 0.05) for small to moderate sample sizes, and they suggest a small sample improvement over the Wald statistic.

Their approach is to use a generalization of chi-square approximations dating back to Patnaik (1949) and Box (1954). Simulation results indicate that this adjustment greatly improves the performance of the Wald statistic, and is effective for sample sizes as small as  $n=7$  per factor combination. They also point out that for equal sample sizes, their statistic is identical to the classical ANOVA F-statistic, and thus their method can be regarded as a robust extension of the classical ANOVA to heteroscedastic designs. They recommend that their method should always be preferred (even in the homoscedastic case) to the classical ANOVA. However, they do not investigate how the performance of their statistic compares to the ANOVA F-statistic.

In this paper, we present results of a simulation study comparing the performance of the Brunner statistic to the ANOVA F-statistic, make a recommendation for the Brunner statistic for moderate sample sizes ( $n \geq 7$ ), and also present a SAS program (SAS Institute, Cary, N.C.) for implementing the method.

---

Scott Richter is an assistant professor in the Mathematical Sciences Department at the University of North Carolina at Greensboro. His email address is [sjricht2@uncg.edu](mailto:sjricht2@uncg.edu). Mark Payton is a professor in the Department of Statistics at Oklahoma State University. His email address is [mpayton@okstate.edu](mailto:mpayton@okstate.edu).

Brunner Method

The method of Brunner et al. (1997) is a small sample adjustment to the well-known Wald statistic, which permits heterogeneous variance but is known to have inflated Type I error rates for small sample sizes. Consider a two-way layout  $a$  levels of factor  $A$  and  $b$  levels of factor  $B$ . Assume a set of independent random variables  $X_{ij} \sim N(\mathbf{m}_i, \mathbf{s}_i^2)$ ,  $i = 1, \dots, ab$ .

Let  $\boldsymbol{\mu} = (\mathbf{m}_1, \mathbf{m}_2, \dots, \mathbf{m}_{ab})'$  denote the vector containing the  $a \cdot b$  population means. Then the hypotheses of no main effects and interaction can be written as

$H_0(A) : \mathbf{M}_A \boldsymbol{\mu} = 0$
$H_0(B) : \mathbf{M}_B \boldsymbol{\mu} = 0$
$H_0(AB) : \mathbf{M}_{AB} \boldsymbol{\mu} = 0$

where

$\mathbf{M}_A = \mathbf{P}_a \otimes \frac{1}{b} \mathbf{J}_b$
$\mathbf{M}_B = \frac{1}{a} \mathbf{J}_a \otimes \mathbf{P}_b$
$\mathbf{M}_{AB} = \mathbf{P}_a \otimes \mathbf{P}_b$

Here,  $\mathbf{P}_c = \mathbf{I}_c - \frac{1}{c} \mathbf{J}_c$ , where  $\mathbf{I}_c$  is a  $c \times c$  identity matrix,  $\mathbf{J}_c$  a  $c \times c$  matrix of 1's, and the symbol  $\otimes$  represents the Kronecker product of the matrices. The vector of observed cell means is denoted by  $\bar{\mathbf{X}} = (\bar{X}_1, \dots, \bar{X}_{ab})'$  and the estimated covariance matrix is given by

$$\hat{\mathbf{S}}_N = N \bullet \text{diag} \left\{ \frac{S_1^2}{n_1}, \dots, \frac{S_{ab}^2}{n_{ab}} \right\}, \text{ where } S_i^2 \text{ is the } i^{\text{th}} \text{ sample variance and } N = \sum_{i=1}^{ab} n_i .$$

For a complete cross-classification, the test statistic is  $FB = \frac{N \bullet \bar{\mathbf{X}}' \mathbf{M} \bar{\mathbf{X}}}{\frac{1}{(n-1)} \text{tr}(\hat{\mathbf{S}}_N)}$ , which has an approximate  $F$  distribution with

$$f_{num} = \frac{1}{(n-1)^2} \bullet \left[ \text{tr}(\hat{\mathbf{S}}_N) \right]^2 \text{ numerator and}$$

$$f_{den} = \frac{\left[ \text{tr}(\hat{\mathbf{S}}_N) \right]^2}{\text{tr}(\hat{\mathbf{S}}_N^2)} \text{ denominator degrees of}$$

freedom, where  $\Lambda = \text{diag} \left\{ \frac{1}{n_1 - 1}, \dots, \frac{1}{n_{ab} - 1} \right\}$  (Brunner, 1997).

Results

A simulation study was performed using SAS version 8.02 for a two-way layout with  $a = 4$  and  $b = 3$ , for various sample sizes. The model used for all simulations was

$$Y_{ijk} = a_i + b_j + ab_{ij} + \mathbf{e}_{ijk},$$

$$i = 1, 2, 3, 4, j = 1, 2, 3,$$

$$k = 1, \dots, n_{ij}, \mathbf{e}_{ijk} \sim N(0, \mathbf{s}_{ij}^2)$$

The classical  $F$  test from ANOVA (denoted by  $F$ ), assuming normality and equal variances, and the adjusted  $F$ -test (denoted by  $FB$ ) of Brunner, et al. (1997) were calculated for 5000 samples and the probabilities of rejection estimated using an  $\alpha = 0.05$ . Differences in Type I error rates and powers are investigated for different sample sizes, effect sizes, and variance structures.

Case 1: Homogeneous errors, equal sample sizes. For this case, we let  $k = 1, \dots, n, \mathbf{e}_{ijk} \sim N(0, \mathbf{s}_i^2)$ . Table 1 shows nominal Type I error rate for both methods, for various sample sizes. Note that the  $FB$  statistic

underestimates the nominal level when  $n$  is small, but for sample size as small as  $n = 7$ , the nominal rates are comparable to the classical ANOVA test. As sample size increases beyond  $n = 7$ , the nominal rate remains stable near the target  $\alpha = 0.05$ .

Tables 2 and 3 give proportion of rejections when factor A effect is present, and when both main effects are present, respectively, for  $n = 3$  and  $n = 7$ . When  $n = 3$ , the test based on the FB statistic has less power than the F statistic,

and underestimates the nominal rate, especially for the test of interaction and when the effect size is small. When  $n = 7$ , power and nominal rate are very similar, with the exception that the nominal rate for interaction is still a bit too low.

Table 4 shows that when interaction only is present, the FB statistic again has less power for the small sample size case. When the sample size is  $n = 7$ , power is comparable for both tests, especially when effect sizes are not very small.

Table 1. Proportion of rejections at  $\alpha = 0.05$ , normally distributed errors, equal variance, based on 5000 samples, no effects present, equal cell sample sizes.

Test for:	Method	$n$	2	3	5	7	10	20
Main Effect A	F		.0492	.0496	.0478	.0482	.0494	.052
	FB		.0130	.0284	.0412	.0448	.0462	.0512
Main Effect B	F		.0466	.0522	.0526	.0530	.052	.0466
	FB		.0142	.0360	.0448	.0502	.0502	.0466
Interaction	F		.0458	.0470	.0474	.0512	.053	.0488
	FB		.0086	.0222	.0326	.0402	.0456	.0462

Table 2. Proportion of rejections at  $\alpha = 0.05$ , normally distributed errors, equal variance, based on 5000 samples, factor A effect present ( $a_1=c, a_3=-c$ ), equal cell sample sizes.

Test for:	Method	$n = 3$			$n = 7$		
		$c$			$c$		
		.5	1.0	1.5	.5	1.0	1.5
Main Effect A	F	.3446	.9302	1.000	.7530	.9998	1.000
	FB	.2642	.8876	.9992	.7370	.9998	1.000
Main Effect B	F	.0522	.0522	.0522	.0530	.0530	.0530
	FB	.0360	.0360	.0360	.0502	.0502	.0502
Interaction	F	.0470	.0470	.0470	.0512	.0512	.0512
	FB	.0222	.0222	.0222	.0402	.0402	.0402

Table 3. Proportion of rejections at  $\alpha = 0.05$ , normally distributed errors, equal variance, based on 5000 samples, factor A and B effects present ( $a_2=b_1=c$ ,  $a_3=b_2=-c$ ), equal cell sample sizes.

		$n = 3$			$n = 7$		
		$c$			$c$		
Test for:	Method	.5	1.0	1.5	.5	1.0	1.5
Main Effect A	F	.3440	.9214	.9998	.7422	1.000	1.000
	FB	.2604	.8780	.9986	.7276	1.000	1.000
Main Effect B	F	.5268	.9902	1.000	.9140	1.000	1.000
	FB	.4576	.9830	1.000	.9100	1.000	1.000
Interaction	F	.0470	.0470	.0470	.0512	.0512	.0512
	FB	.0222	.0222	.0222	.0402	.0402	.0402

Table 4. Proportion of rejections at  $\alpha = 0.05$ , normally distributed errors, equal variance, based on 5000 samples, interaction effect present ( $ab_{11}=ab_{33}=c$ ,  $ab_{13}=ab_{31}=-c$ ), equal cell sample sizes.

		$n = 3$			$n = 7$		
		$c$			$c$		
Test for:	Method	.5	1.0	1.5	.5	1.0	1.5
Main Effect A	F	.0496	.0496	.0496	.0482	.0482	.0482
	FB	.0284	.0284	.0284	.0448	.0448	.0448
Main Effect B	F	.0522	.0522	.0522	.0530	.0530	.0530
	FB	.0360	.0360	.0360	.0502	.0502	.0502
Interaction	F	.1584	.5976	.9460	.4276	.9828	1.000
	FB	.0842	.4368	.8734	.3864	.9762	1.000

Case 2: Heterogeneous errors, equal sample sizes.  
Here we consider:

$$k = 1, \dots, n, \mathbf{e}_{ijk} \sim N(0, \mathbf{s}_{ij}^2 = (1 + i * j / 2)^2),$$

(errors increasing with the levels of A). Tables 5, 6 and 7 are heterogeneous analogs to Tables 2, 3 and 4, respectively. They compare the tests under variance heterogeneity. Note that the classical F

test shows inflated nominal rates for all effects, with the test for interaction the most inflated. The inflation becomes more severe as the ratio between smallest and largest variances becomes larger. The test using the Box-type adjustment, however, maintains the correct nominal rate in all conditions considered.



Table 5. Proportion of rejections at  $\alpha = 0.05$ , normally distributed errors with unequal variance (variance increasing with factor A levels, ratio of largest to smallest variance  $\approx 10$  to 1), based on 5000 samples, factor A effect present ( $a_1=c, a_3=-c$ ), equal cell sample size:  $n_i=7$ .

Test for:	Method	$c$	0	.5	1.5	2.5
Main Effect A	F		.0592	.1684	.9518	.9998
	FB		.0490	.1384	.9266	.9998
Main Effect B	F		.0564	.0564	.0564	.0564
	FB		.0482	.0482	.0482	.0482
Interaction	F		.0728	.0728	.0728	.0728
	FB		.0486	.0486	.0486	.0496

Table 6. Proportion of rejections at  $\alpha = 0.05$ , normally distributed errors with unequal variance (variance increasing with factor A levels, ratio of largest to smallest variance  $\approx 22$  to 1), based on 5000 samples, factor A effect present ( $a_1=c, a_3=-c$ ), equal cell sample size:  $n_i=7$ .

Test for:	Method	$c$	0	.5	1.5	2.5
Main Effect A	F		.0652	.1008	.5324	.9672
	FB		.0488	.0750	.4408	.9392
Main Effect B	F		.0612	.0612	.0612	.0612
	FB		.0488	.0488	.0488	.0488
Interaction	F		.0824	.0824	.0824	.0824
	FB		.0494	.0494	.0494	.0494

Table 7. Proportion of rejections at  $\alpha = 0.05$ , normally distributed errors with unequal variance (variance increasing with factor A levels, ratio of largest to smallest variance  $\approx 22$  to 1), based on 5000 samples, factor A and B effects present ( $a_2=b_1=c, a_3=b_2=-c$ ), equal cell sample size:  $n_i=7$ .

Test for:	Method		.5	1.5	2.5
Main Effect A	F		.1030	.5234	.9518
	FB		.0784	.4422	.9220
Main Effect B	F		.1228	.7868	.9980
	FB		.1014	.7298	.9962
Interaction	F		.0824	.0824	.0824
	FB		.0494	.0494	.0494

Case 3: Homogeneous errors, unequal sample sizes.

In this case we consider:

$$k = 1, \dots, n_{ij}, \mathbf{e}_{ijk} \sim N(0, 1),$$

where  $n_{1j} = 7, n_{2j} = 8, n_{3j} = 9, n_{4j} = 10$ . Here there was little difference in the performance of the two tests (See Tables 8 and 9). The Box-adjusted test showed slightly higher power in some cases.

Case 4: Heterogeneous errors, unequal sample sizes.

Here we consider:

$$k = 1, \dots, n_{ij}, \mathbf{e}_{ijk} \sim N(0, \mathbf{s}_i^2),$$

with  $n_{1j} = 7, n_{2j} = 8, n_{3j} = 9, n_{4j} = 10$ . When the largest variance was associated with the smallest sample the classical F-test always had inflated nominal Type I error rates (often more than twice the nominal rate) for any effects not present, while the Box-adjusted test maintained expected nominal Type I error rates (See Tables 10, 11 and 12). The classical F-test had greater power for small effect sizes, but the power advantage became negligible as the effect size increased.

Although not shown here, when the largest variance was associated with the largest sample the power of the two tests was essentially equivalent, with the Box-adjusted test often having a slight power advantage. The classical F-test tended to underestimate the Type I error rate for effects not present.

Table 8. Proportion of rejections at  $\alpha = 0.05$ , normally distributed errors with unequal sample sizes ( $n_{1j} = 7, n_{2j} = 8, n_{3j} = 9, n_{4j} = 10$ ) and equal variances, based on 5000 samples, factor A effect present ( $a_1 = c, a_3 = -c$ ).

		C		
Test for:	Method	0	.5	1.5
Main Effect A	F	.0482	.7962	1.000
	FB	.0500	.8258	1.000
Main Effect B	F	.0518	.0552	.0598
	FB	.0514	.0514	.0514
Interaction	F	.0500	.0502	.0462
	FB	.0414	.0414	.0414

Table 9. Proportion of rejections at  $\alpha = 0.05$ , normally distributed errors with unequal sample sizes ( $n_{1j} = 7, n_{2j} = 8, n_{3j} = 9, n_{4j} = 10$ ) and equal variances, based on 5000 samples, factors A and B effects present ( $a_2 = b_1 = c, a_3 = b_2 = -c$ ).

		C	
Test for:	Method	.5	1.5
Main Effect A	F	.8002	1.000
	FB	.8302	1.000
Main Effect B	F	.9596	1.000
	FB	.9564	1.000
Interaction	F	.0498	.0496
	FB	.0420	.0420

Table 10. Proportion of rejections at  $\alpha = 0.05$ , normally distributed errors with unequal sample sizes ( $n_{1j} = 7, n_{2j} = 8, n_{3j} = 9, n_{4j} = 10$ ) and unequal variances ( $\mathbf{s}_{1j}^2 = 10, \mathbf{s}_{2j}^2 = 5, \mathbf{s}_{3j}^2 = 2, \mathbf{s}_{4j}^2 = 1$ ), based on 5000 samples, factor A effect present ( $a_1 = c, a_3 = -c$ ).

		<i>c</i>		
Test for:	Method	0	.5	1.5
Main Effect A	F	.1056	.2902	.9850
	FB	.0476	.1666	.9422
Main Effect B	F	.1000	.1024	.1034
	FB	.0418	.0418	.0418
Interaction	F	.1244	.1246	.1230
	FB	.0494	.0494	.0494

Table 11. Proportion of rejections at  $\alpha = 0.05$ , normally distributed errors with unequal sample sizes ( $n_{1j} = 7, n_{2j} = 8, n_{3j} = 9, n_{4j} = 10$ ) and unequal variances ( $\mathbf{s}_{1j}^2 = 10, \mathbf{s}_{2j}^2 = 5, \mathbf{s}_{3j}^2 = 2, \mathbf{s}_{4j}^2 = 1$ ), based on 5000 samples, factor A and B effects present ( $a_2 = b_1 = c, a_3 = b_2 = -c$ ).

		<i>C</i>		
Test for:	Method	.5	1.0	1.5
Main Effect A	F	.3070	.8176	.9944
	FB	.1634	.6660	.9788
Main Effect B	F	.4522	.9450	.9992
	FB	.3174	.8852	.9980
Interaction	F	.1242	.1224	.1208
	FB	.0494	.0494	.0494

Table 12. Proportion of rejections at  $\alpha = 0.05$ , normally distributed errors with unequal sample sizes ( $n_{1j} = 7, n_{2j} = 8, n_{3j} = 9, n_{4j} = 10$ ) and unequal variances ( $\mathbf{s}_{1j}^2 = 10, \mathbf{s}_{2j}^2 = 5, \mathbf{s}_{3j}^2 = 2, \mathbf{s}_{4j}^2 = 1$ ), based on 5000 samples, interaction effect present ( $ab_{11} = ab_{33} = c, ab_{13} = ab_{31} = -c$ ).

		<i>C</i>		
Test for:	Method	.5	1.5	2.5
Main Effect A	F	.1060	.1046	.1016
	FB	.0476	.0476	.0476
Main Effect B	F	.1032	.1018	.1026
	FB	.0418	.0418	.0418
Interaction	F	.2128	.8278	.9996
	FB	.0938	.6324	.9898

### Conclusion

Based on our results and the results of Brunner, et al. (1997), we agree with those authors that there is no reason to use the classical ANOVA F-test, as long as cell sample size is at least 7. For smaller samples, when the normal theory assumptions hold, we prefer the classical ANOVA F-test, since the FB statistic becomes very conservative in this case. When samples are very small and variances are not equal, the ANOVA test suffers from inflated nominal levels and thus should be used with caution. The FB test, on the other hand, is always conservative in these situations, and thus is a good choice for those concerned mostly with avoiding making Type I errors. The obvious trade-off for small sample sizes, however, is that the FB test is virtually powerless to detect small to moderate effects.

#### Example 1.

We illustrate the method using an example given in Sokal and Rohlf (1995). The data are from an experiment to examine differences in food consumption when rancid lard was substituted for fresh lard in the diet of rats. The data are classified by fat (fresh, rancid) and gender (male, female). The amount of food eaten (in grams) is given in the following table:

	Fats	
	Fresh	Rancid
Gender		
Male	709	592
	679	538
	699	476
Female	657	508
	594	505
	677	539

A SAS program (available from the first author) was used to compute the p-values for both the ANOVA F-test and the FB test. Since cell sample sizes are equal, values of the F and FB statistics are identical. Notice that although the sample sizes are small ( $n = 3$ ), there is very little

difference between the p-values associated with the two methods, and only a strong effect of gender is evident from the data.

Source of variation	F	p-value	FB	p-value
Fats	2.593	0.146	2.593	0.153
Gender	41.969	<0.001	41.969	<0.001
Fats*Gender	0.630	0.450	0.630	0.454

#### Example 2.

This example utilizes data presented in Kuehl (2000), page 224. It is a 3x2 factorial experiment involving 3 levels of alcohol and two levels of base. Note that the data are unbalanced in terms of the number of replications per treatment combination.

Because the cell sample sizes are not equal, the calculated test statistics are not the same for the two methods, although the conclusions might be the same for both methods depending upon the level of significance the researcher adopted. The FB statistic gives stronger evidence for effects of interaction and main effects.

	Alcohol		
Base	1	2	3
1	90.7	89.3	89.5
	91.4	88.1	87.6
		90.4	88.3
			90.3
Mean	91.05	89.27	88.93
Std Dev	0.49	1.15	1.21
2	87.3	94.7	93.1
	88.3		90.7
	91.5		91.5
Mean	89.03	94.7	91.77
Std Dev	2.19	---	1.22

Source of variation	F	p-value	FB	p-value
Alcohol	1.931	0.195	4.297	0.053
Base	7.167	0.023	12.858	0.006
Alcohol*Base	7.357	0.011	14.087	0.002

## References

- Akritis, M.G. (1990). The rank transform method in some two-factor designs. *Journal of the American Statistical Association*, 85, 73-78.
- Akritis, M.G., & Arnold, S.F. (1994). Fully nonparametric hypotheses for factorial designs I: Multivariate repeated-measures designs. *Journal of the American Statistical Association*, 89, 336-343.
- Box, G.E.P. (1954). Some theorems on quadratic forms applied in the study of analysis of variance problems, I: Effect of inequality of variance in the one-way classification. *The Annals of Mathematical Statistics*, 25, 290-302.
- Brunner, E., Dette, H., & Munk, A. (1997). Box-type approximations in nonparametric factorial designs. *Journal of the American Statistical Association*, 92, 1494-1502.
- Kuehl, R. O. (2000). *Design of experiments: Statistical principles of research design and analysis*. (2<sup>nd</sup> ed.) Pacific Grove, CA: Brooks/Cole.
- Milliken, G.A., & Johnson, D.E. (1992). *Analysis of messy data, Volume 1: Designed experiments*. New York: Chapman and Hall.
- Patnaik, P.B. (1949). The noncentral  $\chi^2$  and F-distributions and their applications. *Biometrika*, 36, 202-232.
- Sokal, R. R., & Rohlf, F. J. (1995). *Biometry: The principles and practices of statistics in biological research*, New York: W. H. Freeman and Company.
- Thompson, G.L. (1991). A unified approach to rank tests for multivariate and repeated measures designs. *Journal of the American Statistical Association*, 86, 410-419.
- Weerahandi, S. (1995). ANOVA under unequal error variances. *Biometrics*, 51, 589-599.
- Welch, B.L. (1951). On the comparison of several mean values. *Biometrika*, 38, 330-336.

## Modeling Correlated Time-Varying Covariate Effects In A Cox-Type Regression Model

Mourad Tighiouart  
Department of Mathematics and Statistics  
Utah State University

---

In this paper, I extend the proposed model by McKeague and Tighiouart (2000) to handle time-varying correlated covariate effects for the analysis of survival data. I use the conditional predictive ordinates (CPO's) for model comparison and the methodology is illustrated by an application to nasopharynx cancer survival data. A reversible jump MCMC sampler to estimate the CPO's will be presented.

Key words: Correlated time-varying covariate effects, Right censoring; Reversible Jump MCMC; Pseudo-Bayes factors

---

### Introduction

The proportional hazards model of Cox (1972) is considered to be the most popular approach to the analysis of time-to-event data. In the past three decades, many authors have proposed variants of this model to relax the somehow restrictive proportional hazards assumption and to analyze multivariate survival data, see Andersen et al. (1992) and Ibrahim, et al. (2001).

In this paper, I use the local characteristics of Gaussian Markov random fields to describe the prior information of the conditional hazard function for right-censored survival data. McKeague and Tighiouart (2000) modeled the conditional hazard function (given covariates  $z$ )  $h(t|z)$  as a product of conditionally independent stochastic processes, corresponding to (1) a baseline hazard function  $h_0(t)$ , and (2) a regression function  $\exp(\mathbf{b}(t)'z)$  representing the effects of covariates:

$$h_0(t) = \sum_{i \geq 1} I(\mathbf{t}_i < t \leq \mathbf{t}_{i+1}) h_i \quad (1)$$

$$h(t|z) = \sum_{i \geq 1} I(\mathbf{t}_i < t \leq \mathbf{t}_{i+1}) h_i \exp(\mathbf{b}_i' z) \quad (2)$$

A discretized version of model (2) in which the jump times  $\mathbf{t}_1, \mathbf{t}_2, \dots, \mathbf{t}_k$  are fixed and the levels  $h_1, h_2, \dots, h_{k-1}$  form a first order autoregressive process has been considered by Gamerman (1991) and West (1992). Arjas and Gasbarra (1994) and McKeague and Tighiouart (2002) extended model (1) by allowing the jump times to be random and McKeague and Tighiouart (2000) considered a dynamic version of model (2) in which the log-levels  $\mathbf{I}_i = \log(h_i)$  and covariate effects  $\mathbf{b}_i, i=1,2,\dots$  form a Gaussian Markov random field. A related Markov random field model for the prior intensity of a non-homogenous Poisson process was introduced by Arjas and Heikkinen (1997), but was not studied in the survival analysis context and adjustment for covariate effects was not considered.

The class of priors used by McKeague and Tighiouart (2000) for  $\mathbf{b}(t)$  implies independence between the covariate effects, an assumption that may not be true in practice. For instance, in a case study of nasopharynx cancer survival data, West (1992) and McKeague and Tighiouart (2000) showed a clear correlation between the posterior

---

Mourad Tighiouart is an Assistant Professor in the Department of Mathematics and Statistics, Utah State University, 3900 Old Main Hill, Logan UT 84322-3900. E:mail: mourad@math.usu.edu.

mean effects of the two measures of the extent of the cancer, which was not accounted for in the model.

In this paper, I extend the model proposed by McKeague and Tighiouart (2000) by implementing a correlation structure between some of the covariate effects in the prior. I use the pseudo-Bayes factor for model selection, and calculation of the conditional predictive ordinates (CPO's) are performed using the output from the Metropolis-Hastings-Green (MHG) algorithm (Metropolis et al., 1953; Hastings, 1970; Green (1995). The analysis indicates that the null hypothesis of no correlation between the effects of the two measures of the extent of the cancer is rejected and a correlated prior process should be used to estimate conditional survival probabilities.

### Methodology

Let  $T_1, \dots, T_n$  be nonnegative independent random variables with associated  $p$ -dimensional covariate vectors  $z_j, j = 1, \dots, n$ . Assume that the data may be subject to right censoring, i.e., we observe  $(X_1, \mathbf{d}_1, z_1), \dots, (X_n, \mathbf{d}_n, z_n)$  where  $X_j = \min(T_j, U_j)$ ,  $U_j$  being the censoring time for the  $j$ -th individual, and  $\mathbf{d}_j = I\{T_j \leq U_j\}$ . The conditional hazard function is given by (2), where  $I\{\cdot\}$  is the indicator function,  $0 = \mathbf{t}_1 < \mathbf{t}_2 < \mathbf{t}_3 < \dots$  is an increasing sequence of jump times, the  $h_i$ 's represent the levels of the baseline hazard function  $h_0(t)$ , and  $\{\mathbf{b}_i, i \geq 1\} = \{(\mathbf{b}_{i1}, \dots, \mathbf{b}_{ip})', i \geq 1\}$  is a  $p$ -dimensional process describing the effect of covariate vector  $z$ .

Let  $\mathbf{t}_{\max} = \max\{X_j, 1 \leq j \leq n\}$ . The Bayesian approach consists of putting a prior distribution on the  $p$  covariate effects and the unknown baseline hazard function. The jump times  $\mathbf{t}_2, \mathbf{t}_3, \dots$  form a time-homogeneous Poisson process with rate  $\mathbf{g}$ . The prior distributions of the remaining parts of the model are specified conditionally given the number  $m$  of  $\mathbf{t}_i, i \geq 1$  in the interval  $[0, \mathbf{t}_{\max}]$ , as follows.

#### Covariate Effects Prior

I specify  $\mathbf{b}_m = \{\mathbf{b}_{kj}: k=1, \dots, m, j=1, \dots, p\}$  to be a Gaussian Markov random field with a neighborhood system  $\{\mathcal{I}(k, j)\}$  of the following form:  $\mathcal{I}(k, j) = \{(k-1, j), (k+1, j), (k, l), \hat{\mathcal{I}}(j)\}$ , where  $\{\hat{\mathcal{I}}(j), j = 1, \dots, p\}$  is a given neighborhood system for the covariate effects. This means that

interactions in time are only permitted between the same components of the covariate effects. The model then amounts to:

$$\begin{aligned} \mathbf{b}_{kj} | \{ \mathbf{b}_{il}, (i,l) \neq (k,j) \} \\ = \mathbf{b}_{kj} | \{ \mathbf{b}_{il}, (i,l) \in \partial(k,j) \}. \quad (3) \\ \sim N(\mathbf{n}_{kj}, \mathbf{s}_{kj}^2) \end{aligned}$$

In addition, I assume only pairwise interaction between the covariate effects. It follows that the conditional mean  $\mathbf{n}_{kj}$  is given by

$$\begin{aligned} \mathbf{n}_{kj} = \mathbf{m}_{kj} + s_{kj}(\mathbf{b}_{k-1,j} - \mathbf{m}_{k-1,j}) \\ + r_{kj}(\mathbf{b}_{k+1,j} - \mathbf{m}_{k+1,j}) + \\ \sum_{l \in \partial(j)} \mathbf{r}_{kl}(\mathbf{b}_{kl} - \mathbf{m}_{kl}), \end{aligned}$$

see Cressie (1993, Ch. 6).

The hyperparameters  $\mathbf{m}_j = E(\mathbf{b}_{kj}), k = 1, \dots, m$  represent the trend in the levels of the  $j$ -th covariate effect,  $s_{kj}, r_{kj}$  are used to smooth the  $j$ -th covariate effect, and  $\mathbf{r}_{kl}, l \in \mathcal{I}(j)$  measure the correlation between  $\mathbf{b}_{kj}$  and  $\mathbf{b}_{kl}, l \in \mathcal{I}(j)$ . The distribution of  $\mathbf{b}_m$  is completely determined by its local characteristics provided the hyperparameters satisfy the following conditions:  $s_{kj}, r_{kj}, \mathbf{r}_{kl}, \hat{\mathcal{I}}(j)$  are nonnegative with

$$\begin{aligned} s_{kj} + r_{kj} + \sum_{l \in \partial(j)} \mathbf{r}_{kl} < 1, \quad r_{kj} \mathbf{s}_{k+1,j}^2 = s_{k+1,j} \mathbf{s}_{kj}^2, \quad j=1, \dots, \\ p \text{ and } \mathbf{r}_{kj} \mathbf{s}_{kj}^2 = \mathbf{r}_{kl} \mathbf{s}_{kl}^2 \text{ for } j \in \hat{\mathcal{I}}(l), \end{aligned}$$

see, e.g., Besag and Kooperberg (1995). McKeague and Tighiouart (2000) introduced a way of controlling the hyperparameters by the length of adjacent time intervals and I can adapt their approach to the present setting as follows:

$$\begin{aligned} r_{kj} = \frac{(\Delta_k + \Delta_{k+1})c_j}{\Delta_{k-1} + 2\Delta_k + \Delta_{k+1}}, \quad s_{kj} = \frac{(\Delta_{k-1} + \Delta_k)c_j}{\Delta_{k-1} + 2\Delta_k + \Delta_{k+1}}, \\ \mathbf{s}_{kj}^2 = \frac{2\mathbf{s}_j^2}{\Delta_{k-1} + 2\Delta_k + \Delta_{k+1}}, \quad \mathbf{r}_{kj} = \mathbf{r}\mathbf{s}_j^2, \end{aligned}$$

where  $\Delta_k = \Delta_{k+1} - \Delta_k$  is the gap between the  $k$ -th and  $(k+1)$ -st jump times,  $2 \leq k \leq m-1$ , and the parameters  $c_j, \mathbf{s}_j > 0$ , and  $\mathbf{r} \geq 0$  satisfy

$$c_j + \mathbf{r} \sum_{l \in \partial(j)} \mathbf{s}_l^2 < 1, j=1, \dots, p. \quad (4)$$

The parameter  $\mathbf{g}$  controls the rate of jump times,  $c_j$  controls the nearest neighbor interaction between the levels of the  $j$ -th covariate effect,  $\mathbf{s}_j$  represents the precision of the prior information of the  $j$ -th covariate effect, and  $\mathbf{r}$  controls the dependency structure between neighboring covariate effects: higher values of  $\mathbf{r}$  signify greater correlation, and  $\mathbf{r} = 0$  gives rise to the conditionally independent time-varying covariate effects model analyzed by McKeague and Tighiouart (2000). For simplicity of presentation, I restrict attention to the case  $\mathbf{m}_j = \mathbf{m}$  which indicates constant prior levels in the mean of the  $j$ -th covariate effect.

The distribution of  $\mathbf{b}_m$  is Gaussian with mean vector  $\mathbf{m}_{bm}$  and covariance matrix  $(\mathbf{I}_{mp} - \mathbf{C}_1)^{-1} \mathbf{M}_1$ , where  $\mathbf{m}_{bm} = \{\mathbf{m}_j: k=1, \dots, m, j=1, \dots, p\}$ ,  $\mathbf{C}_1$  is an  $mp \times mp$  matrix defined as follows. For  $j=1, \dots, p$  and  $i=m(j-1), \dots, mj$ ,  $c_{i,i+1} = r_{ij}$ ,  $c_{i,i-1} = s_{ij}$ ,  $c_{i,i+ml} = \mathbf{r}_{il}$ , for  $l \in \mathcal{I}(j)$ ,  $c_{i+ml,i} = \mathbf{r}_{i+ml,i}$  for  $i \in \mathcal{I}(j)$ ,  $c_{lk} = 0$  otherwise,  $\mathbf{M}_1 = \text{diag}(\mathbf{s}_{kj}^2, k=1, \dots, m, j=1, \dots, p)$ , and  $\mathbf{I}_{mp}$  is the identity matrix.

### Baseline Hazard Prior

Let  $\mathbf{I}_i = \log(h_i)$ . The prior distribution for the levels of the log-baseline hazard  $\mathbf{I}_1, \dots, \mathbf{I}_m$  is taken to be the same as the prior for the  $j$ -th covariate effect when  $p=1$ . Denote by  $\mathbf{m}_k = E(\mathbf{I}_k)$  the trend in the levels of the baseline hazard function,  $\mathbf{s}_k^2$  the conditional variance of  $\mathbf{I}_k$  given  $\mathbf{I}_i, i \neq k$ , and  $s_k, r_k$  the influences of the left and right neighbors of  $\mathbf{I}_k$ , respectively. The corresponding nearest neighbor interaction and precision of the prior information parameters will be denoted by  $c$  and  $\mathbf{s}$ , respectively.

In what follows, I denote by  $\mathbf{I}_m$  both the random vector  $(\mathbf{I}_1, \dots, \mathbf{I}_m)$  and the last log-level of the baseline hazard function. The joint distribution of  $\mathbf{I}_m = (\mathbf{I}_1, \dots, \mathbf{I}_m)$  is Gaussian with mean vector  $\mathbf{m}_m$  and covariance matrix  $(\mathbf{I}_m - \mathbf{C})^{-1} \mathbf{M}$ , where  $\mathbf{m}_m = (\mathbf{m}_1, \dots, \mathbf{m}_m)$ ,  $\mathbf{C} = (c_{ij})_{1 \leq i, j \leq m}$ ,  $c_{i,i+1} = r_i$ ,  $c_{i,i-1} = s_i$ ,  $\mathbf{M} =$

$\text{diag}(\mathbf{s}_1^2, \dots, \mathbf{s}_m^2)$ , and  $\mathbf{I}_m$  is the identity matrix. Again, I will assume that  $\mathbf{m} = \mathbf{m}$  indicating a constant prior level in the mean of the log-baseline hazard function.

### Data Likelihood and Posterior

For  $k = 1, \dots, p$ , and  $i = 1, \dots, m$ , let  $N_i$  be the number of observed deaths in the interval  $(\mathbf{t}_i, \mathbf{t}_{i+1}]$ ,  $W_{ik} = \sum_{\{j: \mathbf{t}_i < X_j \leq \mathbf{t}_{i+1}, d_{j=1}\}} z_{jk}$ , and  $W_i = (W_{i1}, \dots, W_{ip})$  with  $\mathbf{t}_{m+1} = \mathbf{t}_{\max}$ . Assuming that the censoring mechanism is non-informative, the likelihood is proportional to the product form

$$\begin{aligned} & \prod_{j=1}^n [(h(X_j | z_j)]^{d_j} \prod_{j=1}^n \exp \left\{ - \int_0^{X_j} h(s | z_j) ds \right\} \\ & = \exp \left\{ \sum_{i=1}^m (N_i \mathbf{I}_i + \mathbf{b}_i' W_i) \right. \\ & \quad \left. - \int_0^{\mathbf{t}_{\max}} \left[ \sum_{j=1}^n I(X_j \geq s) h(s | z_j) \right] ds \right\}. \end{aligned}$$

Let  $\mathbf{t}_m = (\mathbf{t}_1, \dots, \mathbf{t}_m)$ , and  $\mathbf{I}_m = (\mathbf{I}_1, \dots, \mathbf{I}_m)$ , then the posterior density of the parameter  $(\mathbf{t}_m, \mathbf{I}_m, \mathbf{b}_m)$  is proportional to the product of the prior and likelihood

$$\begin{aligned} & \mathbf{g}^m (2p)^{-3m/2} |\mathbf{A}|^{1/2} \exp \left\{ - \frac{1}{2} (\mathbf{I}_m - \mathbf{m}_m)' \mathbf{A} (\mathbf{I}_m - \mathbf{m}_m) \right\} \\ & |\mathbf{A}_1|^{1/2} \exp \left\{ - \frac{1}{2} (\mathbf{b}_m - \mathbf{m}_{bm})' \mathbf{A}_1 (\mathbf{b}_m - \mathbf{m}_{bm}) \right\} \\ & \times \exp \left\{ \sum_{i=1}^m (N_i \mathbf{I}_i + \mathbf{b}_i' W_i) \right. \\ & \quad \left. - \int_0^{\mathbf{t}_{\max}} \left[ \sum_{j=1}^n I(X_j \geq s) h(s | z_j) \right] ds \right\}, \end{aligned}$$

where  $\mathbf{A} = \mathbf{M}^{-1} (\mathbf{I}_m - \mathbf{C})$  and  $\mathbf{A}_1 = \mathbf{M}_1^{-1} (\mathbf{I}_{2m} - \mathbf{C}_1)$ .

I use a reversible jump MCMC algorithm to extract features from this posterior distribution, see the appendix.

### Model Comparison

In this section, I test the null hypothesis  $H_0: \mathbf{r} = 0$  against the alternative  $H_1: \mathbf{r} > 0$ . This is equivalent to selecting between the conditionally independent time-varying covariate effects model



$M_1$  analyzed by McKeague and Tighiouart (2000) and model  $M_2$ , in which the covariate effects satisfy (3). Pseudo-Bayes factor is used to select the best model (Gelfand et al. (1992)), and its calculation uses the output of the MHG sampler.

Let  $X = (X_1, \dots, X_n)$  denote the data vector, and  $\mathbf{q} = (\mathbf{I}(t), \mathbf{b}(t))$  be the model parameter. The predictive density is  $f(X) = \int f(X | \mathbf{q}, z) \mathbf{p}(\mathbf{q}) d\mathbf{q}$ , where  $\mathbf{p}(\mathbf{q})$  denotes the prior density of  $\mathbf{q}$  and the conditional predictive ordinate (CPO) is given by

$$f(X_i | X_{(i)}) = \frac{f(X)}{f(X_{(i)})}$$

$$= \int f(X_i | X_{(i)}, \mathbf{q}, z) \mathbf{p}(\mathbf{q} | X_{(i)}) d\mathbf{q},$$

where  $X_{(i)}$  is the data vector  $X$  with  $X_i$  deleted. The pseudo-Bayes factor is given by

$$B = \frac{\prod_{i=1}^n f(X_i | X_{(i)}, M_1)}{\prod_{i=1}^n f(X_i | X_{(i)}, M_2)}$$

and model selection proceeds by choosing  $M_1$  ( $M_2$ ) according to  $B > (<) 1$ . For a complete discussion and justification of this technique, see Geisser and Eddy (1979), Box (1980), Gelfand et al. (1992), and Gelfand and Mallick (1995).

Exact calculation of  $B$  is not possible, however Monte Carlo estimates of the CPO's can be obtained using the output of the MHG sampler  $\mathbf{q}_1, \dots, \mathbf{q}_N$  and the idea of importance sampling density, see Gelfand and Dey (1994). The approximation is given by

$$f(X_i | X_{(i)}) \approx N \left[ \sum_{j=1}^N \frac{1}{f(X_i | \mathbf{q}_j, z)} \right]^{-1}.$$

For a censored observation, I compute the conditional survival function  $S(X_i | X_{(i)}, M_j), j=1, 2$ .

Results

West (1992) and McKeague and Tighiouart (2000) studied data on 181 nasopharynx cancer patients whose cancer careers, culminating in either death (127 cases) or censoring (54 cases) are recorded to the nearest month, ranging from 1 to 177 months.

The analyses were based on five covariates: (1) Sex of the patient (0 for male, 1 for female); (2) Age of the patient at time  $t = 0$ , the start of monitoring of the cancer career of that patient (standardized to have zero mean and unit standard deviation across all patients in the study); (3) Dose1, an average measure of the extent of radiotherapy treatment to which the patient has been subjected (also standardized, as with age); (4) Tumor1, a measurement of the extent of the cancer (in terms of an estimate of the number of cancerous cells), taking value 1, 2, 3 or 4; (5) Tumor2, a measure similar to Tumor1, taken from a different X-ray section, again taking values 1, 2, 3 or 4.

The right hand side of Figure 1 (following page) shows the posterior mean effects for tumor1 and tumor2 obtained by McKeague and Tighiouart, and the left hand side the estimates obtained by West. The similar pattern of the posterior mean effects of tumor1 and tumor2 suggests that a correlated prior process for the two effects is more realistic. I therefore fitted model  $M_2$  with  $\mathbf{r} = 1/2$  and compared it with model  $M_1$ , fitted by McKeague and Tighiouart, which corresponds to  $\mathbf{r} = 0$ . The remaining hyperparameters were the same for both models, and can be found in McKeague and Tighiouart (2000). The logarithm of the pseudo-Bayes factor is found to be

$$\text{Log}(B) = \text{Log} \left[ \frac{\prod_{i=1}^n f(X_i | X_{(i)}, M_2)}{\prod_{i=1}^n f(X_i | X_{(i)}, M_1)} \right] = 4.56$$

and

$$\frac{\prod_{i=1}^n S(X_i | X_{(i)}, M_2)}{\prod_{i=1}^n S(X_i | X_{(i)}, M_1)} = 1.53$$

suggesting that a time-varying correlated covariate effect should be used to estimate conditional survival probabilities.

### Conclusion

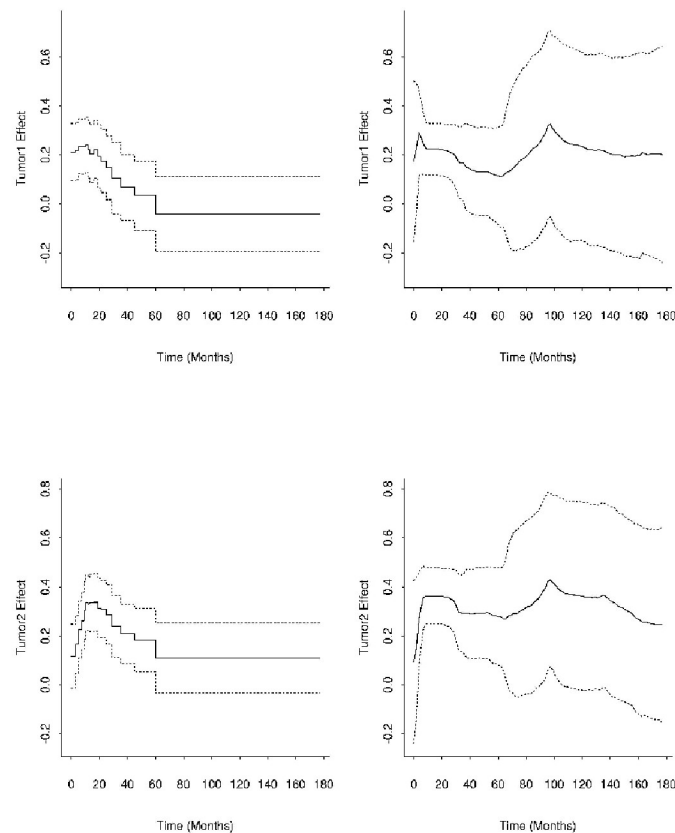
I have presented a complete nonparametric Bayesian approach to inference from right-censored survival data. The methodology is an extension of the model proposed by McKeague and Tighiouart (2000) in the sense that the Bayesian model accounts for any correlation structure between some of the time-varying covariate effects in the prior. Except for the constraints (4), direct specification of the parameter controlling the amount of correlation between the covariate effects is not possible. A second stage prior can easily be placed on the hyperparameter  $\rho$ ; I did not pursue this hierarchy

here because my goal is to simplify the presentation of this methodology.

The computational method used to extract features of the posterior distribution is similar to the one used in McKeague and Tighiouart (2000). The only difference is the extra term involved in the prior ratio of the correlated covariate effects. This is very convenient when writing the codes of this sampler.

The methodology was illustrated by an analysis of a nasopharynx cancer survival data set. The class of prior processes defining the Bayesian model was flexible enough to detect a correlation structure between some of the time-varying covariate effects; in particular, pseudo-Bayes factors were calculated to support this evidence.

Figure 1.



## References

- Andersen, P. K., Borgan, O., Gill, R. D., & Keiding, N. (1992). *Statistical Models Based on Counting Processes*. New York: Springer-Verlag.
- Arjas, E., & Gasbarra, D. (1994). Nonparametric Bayesian Inference for Right-Censored Survival Data, Using the Gibbs Sampler. *Statistica Sinica*, 2, 505-524.
- Arjas, E., & Heikkinen, J. (1997). An Algorithm for Nonparametric Bayesian estimation of a Poisson intensity. *Journal of Computational Statistics*, 12, 385-402.
- Besag, J. E., & Kooperberg, C. (1995). On Conditional and Intrinsic Autoregressions. *Biometrika*, 82, 733-746.
- Cox, D. R. (1972). Regression Models and Life-Tables (with discussion). *Journal of the Royal Statistical Society*, B 34, 187-220.
- Cressie, N. (1993). *Statistics for Spatial Data*. New York: Wiley.
- Gamerman, D. (1991). Dynamic Bayesian Models for Survival Data. *Applied Statistics*, 40, 63-79.
- Gelfand, A. E., Dey, D. K., & Chang, H. (1992). Model Determination using Predictive Distribution with Implementation via Sampling-Based Methods. *Bayesian Statistics 4*, 147-167.
- Gelfand, A. E., & Dey, D. K. (1994). Bayesian Model Choice: asymptotics and exact calculations. *Journal of the Royal Statistical Society*, B 56, 501-514.
- Gelfand, A. E., & Mallick, B. K. (1995). Bayesian Analysis of Proportional Hazards Models Built from Monotone Functions. *Biometrics*, 51, 843-852.
- Green, P. J. (1995). Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination. *Biometrika*, 82, 711-732.
- Hastings, W. K. (1970). Monte Carlo Sampling Methods using Markov Chains and their Applications. *Biometrika*, 57, 97-109.
- Ibrahim, J. G., Chen, M.-H., & Sinha, D. (2001). *Bayesian Survival Analysis*. New York: Springer-Verlag.
- McKeague, I. W., & Tighiouart, M. (2002). Nonparametric Bayes Estimators for Hazard Functions Based on Right Censored Data. *Tamkang Journal of Mathematics*, 33, No 2, 173-189.
- McKeague, I. W., & Tighiouart, M. (2000). Bayesian Estimators for Conditional Hazard Functions. *Biometrics*, 56, 1007-1015.
- Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of State Calculations by Fast Computing Machines. *Journal of Chemical Physics*, 21, 1087-1092.
- West, M. (1992). *Modelling Time-Varying Hazards and Covariate Effects*. Survival Analysis: State of the Art, J. P. Klein, P. K. Goel, eds. Kluwer, Boston, 47-62.

## Appendix

To simplify the description of the algorithm, I will assume that  $p=2$  and will denote by  $\mathbf{a}(t)$  and  $\mathbf{b}(t)$  the two time-varying correlated covariate effects and  $\mathbf{m}_a$ ,  $\mathbf{m}_b$  their constant prior means, respectively. The constant prior mean of the log-baseline hazard function  $\mathbf{I}(t)$  will be denoted by  $\mathbf{m}_I$ . The procedure for calculating features of the posterior distribution of  $(\mathbf{t}_m, \mathbf{I}_m, \mathbf{a}_m, \mathbf{b}_m)$  (note that here  $m$  is random) consists of running a reversible Markov chain on the state space  $S = \bigcup_{i \geq 1} S_i$ , where  $S_i = D_i \times \mathbb{R}^{3i}$ , and  $D_i = \{(x_1, x_2, \dots, x_i) : 0 = x_1 < x_2 < \dots < x_i < \mathbf{t}_{\max}\}$ , using the Metropolis-Hastings-Green algorithm.

A transition from  $(\mathbf{t}_m, \mathbf{I}_m, \mathbf{a}_m, \mathbf{b}_m)$  to  $(\mathbf{t}'_m, \mathbf{I}'_m, \mathbf{a}'_m, \mathbf{b}'_m)$  is accomplished by randomly selecting one of five types of moves ( $H_0, H_a, H_b, B, D$ ): a change of height of a randomly selected level of the baseline hazard rate, change of height of a randomly selected level of the covariate effect  $\mathbf{a}(t)$ , change of height of a randomly selected level of the covariate effect  $\mathbf{b}(t)$ , birth of a new jump time at a randomly selected location in  $(0, \mathbf{t}_{\max})$ , and death of a randomly selected jump time, respectively.

When selecting moves of type  $H_0, H_a, H_b$ , the acceptance probability is the same as in the classical Metropolis-Hastings algorithm:

$$\min \{1, (\text{likelihood ratio}) \times (\text{prior ratio}) \times (\text{proposal ratio})\},$$

whereas if moves of type  $B$  or  $D$  are selected, the current state  $(\mathbf{t}_m, \mathbf{I}_m, \mathbf{a}_m, \mathbf{b}_m)$  is mapped onto  $(\mathbf{t}'_m, \mathbf{I}'_m, \mathbf{a}'_m, \mathbf{b}'_m)$  by a one-to-one transformation  $\mathbf{t}$ . The acceptance probability then takes the form:

$$\min \{1, (\text{likelihood ratio}) \times (\text{prior ratio}) \times (\text{proposal ratio}) \times J(\mathbf{t})\},$$

where  $J(\mathbf{t})$  is the Jacobian of the transformation  $\mathbf{t}$ . Except for the expressions of the prior ratios in the moves of type  $H_a, H_b, B$ , and  $D$ , a complete description of the types of moves, transformation  $\mathbf{t}$ , expressions of the likelihood and proposal ratios, and

the Jacobian can be found in McKeague and Tighiouart (2000).

Move of type  $H_\alpha$ :

An index  $k$  is uniformly selected from  $\{1, 2, \dots, m\}$  and  $V$  is generated uniformly in the interval  $(-\delta_\alpha, \delta_\alpha)$ , where  $\delta_\alpha$  is a sampler parameter. The proposed new level for the covariate effect  $\alpha(t)$  is  $\alpha'_k = \alpha_k + V$ . The proposed new point is  $(\tau'_m, \lambda'_m, \alpha'_m, \beta'_m)$  with  $\tau'_m = \tau_m$ ,  $\lambda'_m = \lambda_m$ ,  $\beta'_m = \beta_m$ , and  $\alpha'_i = \alpha_i$  for  $i \neq k$ .

The prior ratio is:

$\exp \{ -\Delta_H(A_1, \alpha_m, \alpha'_m, \beta_m) / 2 \}$ , where for a  $2m \times 2m$  matrix  $A$  and  $m$ -dimensional vectors  $\alpha, \alpha', \beta$ ,

$$\begin{aligned} \Delta_H(A, \alpha, \alpha', \beta) &= a_{kk}(\alpha'_k - \alpha_k)(\alpha_k + \alpha'_k - 2\mu_\alpha) \\ &\quad + 2a_{k,k-1}(\alpha_{k-1} - \mu_\alpha)(\alpha'_k - \alpha_k) \\ &\quad + 2a_{k,k+1}(\alpha_{k+1} - \mu_\alpha)(\alpha'_k - \alpha_k) \\ &\quad + 2a_{k,m+k}(\alpha'_k - \alpha_k)(\beta_k - \mu_\beta). \end{aligned}$$

A similar expression holds for a move of type  $H_\beta$ .

Move of type B:

A new jump time  $\tau^*$  is drawn uniformly in the interval  $(\tau_1, \tau_{\max})$ . Suppose that  $\tau^* \in (\tau_{k-1}, \tau_k)$ . This new jump time induces two new levels for the log-baseline hazard rate  $\lambda'_{k-1}$  and  $\lambda'_k$ , two new levels for the covariate effect  $\alpha(t)$ ,  $\alpha'_{k-1}$ ,  $\alpha'_k$ , and two new levels for the covariate effect  $\beta(t)$ ,  $\beta'_{k-1}$ ,  $\beta'_k$  using the transformation  $\tau$  described in McKeague and Tighiouart.

The prior ratio is :

$$\begin{aligned} &\gamma(2\pi)^{-3/2} \left( \frac{|A' || A'_1|}{|A || A_1|} \right)^{1/2} \\ &\exp \left\{ \frac{1}{2} (\Delta_B(A, A', \lambda'_m, \lambda'_m) + \Delta_B(A_{11}, A'_{11}, \alpha'_m, \alpha'_m) \right. \\ &\quad \left. + \Delta_B(A_{22}, A'_{22}, \beta'_m, \beta'_m) + \Delta_{BC}(A_1, A'_1, \alpha'_m, \beta'_m, \alpha'_m, \beta'_m)) \right\}, \end{aligned}$$

where  $A_{11}, A_{22}$  are  $m \times m$  matrices such that

$$A_1 = \begin{pmatrix} A_{11} & \times \\ \times & A_{22} \end{pmatrix},$$

and for an  $m \times m$  matrix  $A$  and  $m$ -dimensional vector  $\lambda_m$ ,

$$\begin{aligned} \Delta_B(A, A', \lambda'_m, \lambda'_m) &= (\lambda_{k-2} - \mu_\lambda)^2 (a_{k-2, k-2} \\ &\quad - a'_{k-2, k-2}) \\ &\quad + (\lambda_k - \mu_\lambda)^2 (a_{kk} - a'_{kk}) \\ &\quad + a_{k-1, k-1} (\lambda_{k-1} - \mu_\lambda)^2 \\ &\quad - a'_{k-1, k-1} (\lambda'_{k-1} - \mu_\lambda)^2 \\ &\quad - a'_{k+1, k+1} (\lambda_k - \mu_\lambda)^2 \\ &\quad - 2(\lambda_{k-2} - \mu_\lambda) [a'_{k-2, k-1} \\ &\quad (\lambda'_{k-1} - \mu_\lambda) \\ &\quad - a_{k-2, k-1} (\lambda_{k-1} - \mu_\lambda)] \\ &\quad - 2(\lambda_k - \mu_\lambda) [a'_{k, k+1} (\lambda'_k - \mu_\lambda) \\ &\quad - a_{k-1, k} (\lambda_{k-1} - \mu_\lambda)] \\ &\quad - 2a'_{k-1, k} (\lambda'_{k-1} - \mu_\lambda) \\ &\quad (\lambda'_k - \mu_\lambda). \end{aligned}$$

Also, for a  $2m \times 2m$  matrix  $A_1$ , and  $m$ -dimensional vectors  $\alpha_m, \beta_m$ ,

$$\begin{aligned} \Delta_{BC}(A_1, A'_1, \alpha'_m, \beta'_m) &= -2(\alpha_{k-2} - \mu_\alpha)(\beta_{k-2} - \mu_\beta) \\ &\quad - (a'_{k-2, m+k-2} - a_{k-2, m+k-2}) \\ &\quad - 2(\alpha_k - \mu_\alpha)(\beta_k - \mu_\beta) \\ &\quad - (a'_{k+1, m+k+1} - a_{k, m+k}) \\ &\quad - 2a'_{k-1, m+k-1} \\ &\quad - (\alpha'_{k-1} - \mu_\alpha)(\beta'_{k-1} - \mu_\beta) \\ &\quad - 2a'_{k, m+k} \\ &\quad - (\alpha'_k - \mu_\alpha)(\beta'_k - \mu_\beta) \\ &\quad + 2a_{k-1, m+k-1} \\ &\quad - (\alpha_{k-1} - \mu_\alpha)(\beta_{k-1} - \mu_\beta). \end{aligned}$$

Move of type D:

An index  $k$  is uniformly selected from  $\{2, 3, \dots, m\}$  corresponding to the removal of the jump time  $\tau_k$ . The proposed new point is  $(\tau'_m, \lambda'_m, \alpha'_m, \beta'_m)$ , with  $\lambda'_i = \lambda_i$ ,  $\alpha'_i = \alpha_i$ ,  $\beta'_i = \beta_i$  for  $i \leq k-2$ ,  $\lambda'_j = \lambda_{j+1}$ ,  $\alpha'_j = \alpha_{j+1}$ ,  $\beta'_j = \beta_{j+1}$ ,  $\tau'_j = \tau_{j+1}$  for  $j \geq k$ , and  $\tau'_i = \tau_i$  for  $i \leq k-1$ .

The likelihood, prior, proposal ratios, and the Jacobian for this type of move are the inverse ratios of the ones for the move of type  $B$  with the proper labeling of the jump times and the log-hazard levels.

## A More Efficient Way Of Obtaining A Unique Median Estimate For Circular Data

B. Sango Otieno  
Department of Statistics  
Virginia Tech

C. M. Anderson-Cook  
Department of Statistics  
Virginia Tech

---

The procedure for computing the sample circular median occasionally leads to a non-unique estimate of the population circular median, since there can sometimes be two or more diameters that divide data equally and have the same circular mean deviation. A modification in the computation of the sample median is suggested, which not only eliminates this non-uniqueness problem, but is computationally easier and faster to work with than the existing alternative.

Key words: Preferred direction, circular median, uniqueness, robustness, local averaging

---

### Introduction

Two common choices for summarizing the preferred direction are the mean direction and the median direction. (Fisher 1993, p. 30-36). The notion of preferred direction in circular data is analogous to the “center” of a distribution for data on a linear scale. The sample mean direction is frequently preferred for moderately large samples, because when combined with a measure of sample dispersion, it acts as a summary of the data suitable for comparison and amalgamation with other such information. An alternative, the sample median, can be thought of as balancing the number of observations on two halves of the circle.

Because there is no natural preferred direction for data that are uniformly distributed around the circle, it is natural and desirable that any measures of preferred direction are undefined if the sample data are equally spaced around the circle. In this paper, we consider estimating the preferred direction for a sample of unimodal circular data.

Ko and Guttorp (1988) showed that for a very wide class of families of distributions on  $S^{p-1}$ , the mean has infinite standardized gross error sensitivity; i.e., the asymptotic effect of a small contamination can be large compared with the dispersion. Hence, for the purposes of robust estimation, it is desirable to have a version of the sample median for circular data. As a nonparametric and robust estimate for the preferred direction of a distribution, the circular median has a different character from the sample circular mean as illustrated by different breakdown properties.

The sample median direction  $\hat{q}$  of angles  $q_1, \dots, q_n$  is defined to be the point  $P$  on the circumference of the circle that satisfies the following two properties: (a) The diameter  $PQ$  through  $P$  divides the circle into two semi-circles, each with an equal number of observed data points and, (b) the majority of the observed data is closer to  $P$  than to the anti-median  $Q$ , See Mardia (1972, p. 28-30) or Fisher (1993, p. 35-36), for further details. For odd size samples, the median is an observation, while for even sized samples, the median is the midpoint of two adjacent observations. Observations directly opposite each other do not contribute to the preferred direction, since these observations balance each other for all possible choices of medians. The procedure for finding the circular median has the flexibility to find a balancing point for situations involving ties,

---

B. Sango Otieno is a Visiting Assistant Professor in the Department of Mathematics and Computer Science at The College of Wooster. E:mail: sango@vt.edu. C. M. Anderson-Cook is an Associate Professor in the Department of Statistics at Virginia Tech. Email: candcook@vt.edu

by mimicking the midranking idea for linear data. Potential median values are shown in Figure 1. For even samples, the candidate values are the midpoints of all neighboring observations, as shown in Figure 1a. For odd samples, the candidate values are the observations themselves, as in Figure 1b.

The circular median is rotationally invariant as shown by Ackermann (1997). Lenth (1981), and, Wehrly and Shine (1981) studied the robustness properties of both the circular mean and median using influence curves, and revealed that the circular mean is quite robust, in contrast to the sample mean on the real line. Durcharme and Milasevic (1987), show that in the presence of outliers, the circular median is more efficient than the mean direction. Many authors, including He and Simpson (1992), advocate the use of circular median as an estimate of preferred direction especially in situations where the data are not from the von Mises distribution.

A strategy to deal with non-unique circular median estimates is desired, especially for small samples, which are commonly encountered in circular data as is the case described below.

Consider the Frog data, given in Table 1 and shown in Figure 2, which relates the homing ability of Northern cricket frog, *Acris crepitans*, (Ferguson, et. al., 1967). For this data set, it is thought that the preferred direction for the population is  $121^\circ$  (where  $0^\circ$  is taken to be true North, and angles are measured in a clockwise direction), Collett (1980). The sample appears to be consist of a single modal group, with one observation which can be classified as an outlier. We wish to obtain the median as the point estimate of the preferred direction.

Notice that diameters  $P_1Q_1$  and  $P_2Q_2$  both divide the data evenly between the two semicircles, and hence both  $P_1(133^\circ)$  and  $P_2(140.5^\circ)$  satisfy the definition of a circular median. This implies that the median for this data set is not unique. A method for dealing with this non-uniqueness is the focus of this paper.

## Methodology

To find a unique estimate of median, it is suggested to select the angle satisfying the median definition, such that it has the smallest circular mean deviation (Fisher, 1993, p. 35-36). The circular mean deviation is given by

$$d(\tilde{\mathbf{q}}) = \mathbf{p} - \frac{1}{n} \sum_{i=1}^n \left| \mathbf{p} - \left| \mathbf{q}_i - \tilde{\mathbf{q}} \right| \right|, \text{ where } \tilde{\mathbf{q}} \text{ is the}$$

estimate of the preferred direction, and it is used as a measure of dispersion. Computing the circular median proposed by Mardia (1972, p. 28,31), henceforth referred to as ‘‘Mardia Median’’, occasionally leads to a non-unique estimate of the circular median since there can sometimes be two or more diameters that divide the data equally and have the same circular mean deviation.

In this section, we adapt the existing definition of circular median and propose that the estimate of the population circular median be the average (circular mean) of all angles satisfying the definition of median. This gives a unique estimate of the median, henceforth referred to as ‘‘New Median’’.

For the Frog data above,  $P_1(133^\circ)$  and  $P_2(140.5^\circ)$  are the two candidate sample medians. That is, the point estimate of the preferred direction based on Mardia Median can be taken to be either  $P_1(133^\circ)$  or  $P_2(140.5^\circ)$ , since both have equal circular mean deviation of 0.650759. However, based on the new procedure, the point  $P(136.75^\circ)$  in Figure 2 is the circular mean of the two sample medians ( $P_1$  &  $P_2$ ). We conjecture that  $P$  will be more robust to rounding and will be a unique estimate since it involves local averaging, Cabrera et.al. (1994). Note that in this procedure, we eliminate the step of computing the circular mean deviation of candidate medians.

However, it is important to point out that if we treat  $P_1(133^\circ)$  and  $P_2(140.5^\circ)$  as equally good choices of median, since they have the same circular mean deviation, the circular mean deviation of  $P(136.75^\circ)$  is also 0.650759, hence it is the unique median. S-Plus functions for computing the circular mean direction, the Mardia Median and the New Median are given in the Appendix.

Figure 1: Original Observation  $o$ , Potential Median  $p$

Figure 1a: Even sample size

Figure 1b: Odd sample size

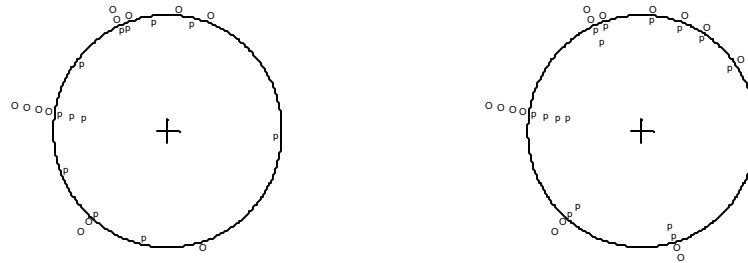
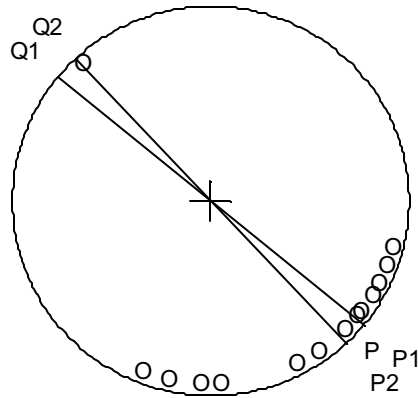


Table 1: Frog Data-Angles in degrees measured due North.

104	110	117	121	127	130	136
144	152	178	184	192	200	316

Figure 2: Homing Ability of Northern Cricket Frog



Results

Comparison of Mardia Median & New Median

To determine the relative performance of Mardia Median and the New Median, data was simulated from a von Mises (VM) distribution with probability density function  $f(\mathbf{q}) = [2pI_0(\mathbf{k})]^{-1} \exp[\mathbf{k} \cos(\mathbf{q} - \mathbf{m})]$ ,  $0 \leq \mathbf{q}, \mathbf{m} < 2\pi$  and  $0 \leq \mathbf{k} < \infty$ , Where  $\mathbf{m}$  is the mean direction,  $\mathbf{k}$  is the concentration parameter and

$$I_0(\mathbf{k}) = (2\pi)^{-1} \int_0^{2\pi} \exp[\mathbf{k} \cos(\mathbf{f})] d\mathbf{f} = \sum_{j=0}^{\infty} -\frac{\mathbf{k}^{2j}}{4^j j^2}$$

is the modified Bessel function of order zero.

Without loss of generality, the center of all the distributions considered was  $\mathbf{m} = 0$ . Ten thousand samples each of sizes between 5 & 20 from the distributions with 6 dispersion values ranging from  $\mathbf{k} = 0.5$  to 10 were obtained. The choice of sample size and dispersion values was based on the fact that non-uniqueness problems of the circular median are most common for small samples and large dispersions, so that is what we studied. For each sample, the sample circular medians (both Mardia Median and New Median) were computed.

The results were summarized using the following measures: 1) Circular mean ( $\hat{\mathbf{m}}$ ); and 2) circular variance  $(1 - \hat{\mathbf{r}})$  of the 10000 estimates obtained by solving the equations

$$\frac{1}{n} \sum_{i=1}^n \cos \mathbf{q}_i = \hat{\mathbf{r}} \cos(\hat{\mathbf{m}}), \quad \frac{1}{n} \sum_{i=1}^n \sin \mathbf{q}_i = \hat{\mathbf{r}} \sin(\hat{\mathbf{m}}),$$

where  $\hat{\mathbf{r}}$  is the sample resultant length; 3) the 95% Empirical Confidence Interval or the central 95% of the 10000 values; 4) Circular Mean Deviation (CMD) and 5) Circular Median Absolute Deviation (CMAD) given by  $Median \left[ |\mathbf{q}_1 - \tilde{\mathbf{q}}|, \dots, |\mathbf{q}_n - \tilde{\mathbf{q}}| \right]$ . Some of the simulation results are given in Tables 2 and 3.

Table 2, illustrates the effect of sample size on the two measures for  $\mathbf{k} = 2$ . The measures appear unbiased, since the average of the point estimates is very close to zero, the true expected value. The confidence bands for the two medians are very similar and would be interchangeable for

most required precision levels and become narrower as sample size increases for the two measures. The circular variances of the two medians, which could range between 1 for maximum variability to 0 for no variability, are consistently close over the whole range of sample sizes considered. Similarly, both the circular mean deviation (CMD), and the circular median absolute deviation (CMAD) are nearly the same for the two measures. These results were similar for other concentration parameters studied as well.

The effect of changing the concentration parameter on the two measures of preferred direction is illustrated in Table 3 for  $n = 20$ . Again, the two measures appear unbiased, and their confidence bands are very similar. The confidence bands become narrower as the concentration parameter increases for the two measures. The remaining measures for both medians are nearly identical for all possibilities. These results were similar for other sample sizes studied as well.

Note that computationally, the new procedure for obtaining the circular median is faster and simpler, since it eliminates the step of computing the circular mean deviation of each candidate median as opposed to Mardia Median. From the above results, we observe that the new procedure results in an estimate which minimizes the circular mean deviation relative to its counterpart, utilizing the benefits of local averaging.

Conclusion

For a fixed sample size and concentration, the Mardia Median and New Median give remarkably consistent results for all combinations of sample sizes and concentrations studied. Most strikingly, the two estimators, Mardia Median and New Median are approximately identical, which implies that either of the two can be used as an estimate of preferred direction. Computationally, the new measure is easier and faster to work with. Both Mardia Median and New Median are robust alternatives to the mean.



References

Ackermann, H. (1997). A note on circular nonparametrical classification. *Biometrical Journal*, 5, 577-587.

Cabrera, J., Maguluri, G. & Singh, K. (1994). An odd property of the sample median. *Statistics & Probability Letters* 19, 349-354.

Collett, D. (1980). Outliers in circular data. *Applied Statistics*, 29, 50-57.

Durcharme, G.R. & Milasevic, P. (1987). Some asymptotic properties of the circular median. *Communications in Statist. Theory and Methods*, 16, 163-169.

Ferguson, D.E., Landreth, H.F. & McKeown, J.P. (1967). Sun compass orientation of northern cricket frog, *Acris crepitans*. *Anim. Behav.*, 15, 45-53.

Fisher, N.I. (1993). *Statistical analysis of circular data*. Cambridge University Press, Cambridge.

He, X. & Simpson, D.G. (1992). Robust direction estimation. *Annals of Statistics*, 20, 1, 351-369.

Ko, D. & Guttorp, (1988). Robustness of estimators for directional data. *Annals of Statistics*, 16, 609-618.

Lenth, R.V. (1981). Robust measures of location for directional data. *Technometrics*, 23, 77-81.

Mardia, K.V. (1972). *Statistics of Directional Data*. London: Academic Press.

Wehrly, T. & Shine, E.P. (1981). Influence curves for directional data. *Biometrika*, 68, 334-33.

Table 2. Mardia Median and New Median for VM(0, 2).

Sample Size	Measure	Point Estimate	Lower & Upper Confidence Limits	Circular Variance	Mean Deviation	Median Absolute Deviation
5	Mardia	0.001206	(-0.914198, 0.884683)	0.098107	0.559813	0.461589
	New	0.001347	(-0.913211, 0.889418)	0.098065	0.559152	0.461589
6	Mardia	-0.002618	(-0.77354, 0.790136)	0.075744	0.593154	0.484028
	New	-0.002350	(-0.774848, 0.787038)	0.075065	0.592542	0.484028
7	Mardia	0.004926	(-0.773052, 0.776042)	0.075079	0.597941	0.499424
	New	0.004867	(-0.771782, 0.778294)	0.075053	0.597611	0.499424
8	Mardia	-0.003863	(-0.700625, 0.658065)	0.059276	0.612813	0.507610
	New	-0.004103	(-0.699964, 0.65746)	0.058872	0.612625	0.507610
9	Mardia	-0.006341	(-0.69237, 0.673193)	0.059405	0.615008	0.515896
	New	-0.006230	(-0.693563, 0.668901)	0.059312	0.614815	0.515896

10	Mardia	-0.001831	(-0.62134, 0.631115)	0.049014	0.626990	0.524162
	New	-0.001734	(-0.619628, 0.631212)	0.048872	0.626892	0.524162
15	Mardia	0.000521	(-0.53107, 0.515293)	0.035605	0.641045	0.540889
	New	0.000580	(-0.531013, 0.515249)	0.03559	0.641003	0.540889
20	Mardia	0.000071	(-0.45413, 0.457305)	0.02582	0.651075	0.548252
	New	0.000010	(-0.453727, 0.455789)	0.025815	0.651067	0.548252

Table 3: Mardia Median and New Median for  $VM(0, \mathbf{m})$ ,  $n = 20$ .

$k$	Measure	Point Estimate	Lower and Upper Confidence Limits	Circular Variance	Mean Deviation	Median Absolute Deviation
0.5	Mardia	-0.005483	(-1.796451, 1.664871)	0.265584	1.189068	1.044356
	New	-0.010259	(-1.787609, 1.647442)	0.263658	1.178366	1.044356
1	Mardia	-0.002878	(-0.775017, 0.777624)	0.075995	0.959626	0.815823
	New	-0.003105	(-0.777569, 0.777397)	0.076126	0.958215	0.815823
2	Mardia	0.000071	(-0.45413, 0.457305)	0.02582	0.651075	0.548252
	New	0.000010	(-0.453727, 0.455789)	0.025815	0.651067	0.548252
4	Mardia	-0.000058	(-0.296221, 0.285816)	0.010901	0.415821	0.350094
	New	-0.000058	(-0.296221, 0.285816)	0.010901	0.415821	0.350094
8	Mardia	0.000323	(-0.191746, 0.200085)	0.005015	0.280498	0.236698
	New	0.000323	(-0.191746, 0.200085)	0.005015	0.280498	0.236698
10	Mardia	-0.000812	(-0.176491, 0.169498)	0.003927	0.249753	0.211066
	New	-0.000812	(-0.176491, 0.169498)	0.003927	0.249753	0.211066

## Appendix

## A.1 cmed()

This function calculates circular median “New Median”. It is a main program, one that the user will need to run. Input: data vector, x.

```
cmed<- function(x){
lenx <- length(x)
sx <- sort(x)
difsin <-c()
numties <-c()
if(lenx/2 == round(lenx/2)) {
# Checks if sample size is odd or even
# Computes median if sample size is even
posmed<- checkeven(x)
for(i in 1:length(posmed)) {
newx <- sx - posmed[i]
difsin[i] <-sum(round(sin(newx),10)> 0) - sum(round(sin(newx),10) < 0)
numties[i] <- sum(round(newx, 10) == 0)}
}
else
# Computes median if sample size is odd
posmed <- checkodd(x)
for(i in 1:length(posmed)) {
newx <- sx - posmed[i]
difsin[i] <- sum(round(sin(newx),10) > 0) - sum(round(sin(newx),10) < 0)
numties[i] <- sum(round(newx, 10) == 0)}
}
# Checks for ties
cm <- c(posmed[round(difsin, 10) == 0 | abs(difsin) > numties])
circmed <- ave.ang(cm)
}
#takes into account if possible circmed are equidistant from mean
direction
circmed}
```

## A.2 cmedM()

This function calculates Mardia Median. It is a main program, one that the user will need to run.

Input: data vector, x.

```
cmedM <- function(x) {
lenx <- length(x)
sx <- sort(x)
sx2 <- c(sx[2:lenx], sx[1])
# Determines closest neighbors of a fixed observation
posmed <- rep(0, lenx)
difsin <- rep(0, lenx)
numties <- rep(0, lenx)
med <- c()
if(lenx/2 == round(lenx/2)) {
\# Checks if sample is odd or even
posmed <- posmedf(x)
```

```

# Computes median if sample size is even
for(i in 1:length(posmed)) {
newx <- sx - posmed[i]
difsin[i]<- sum(round(sin(newx),10) > 0) - sum(round(sin(newx),10) < 0)
numties[i]<- sum(round(newx, 10) == 0)}
}
else {
# Computes median if sample size is even
posmed <- checkodd(x)
for(i in 1:length(posmed)) {
newx <- sx - posmed[i]
difsin[i]<- sum(round(sin(newx),10) > 0) - sum(round(sin(newx),10) < 0)
numties[i]<- sum(round(newx, 10) == 0) }
}
# Checks for ties
cm <- c(posmed[round(difsin, 10) == 0 | round(abs(difsin),10) < numties])
for (i in 1:length(cm)) {
# Computes the circular mean deviation for candidate medians
med[i] <- meandev(x,cm[i]) }
circmed <- ave.ang(cm[round(med,10) == round(min(med),10)])
}
# Chooses the candidate medians with smallest circular mean deviations
and takes circular mean of them if more that one.

```

#### A. 3 ave.ang()

This function calculates circular mean direction. It is an internal function needed for the main programs. Input: data vector a.

```

ave.ang <- function(a) {
y <- sum(sin(a))
x <- sum(cos(a))
ifelse(round(x, 10) == 0 & round(y, 10) == 0, 9999, atan(y, x))}
# If both x and y are zero, then no circular mean exists, so assign it a
large number (9999).

```

#### A. 4 posmedf()

This function calculates all potential medians for even samples It is an internal function needed for the main programs. Input: data vector x.

```

posmedf <- function(x){
lenx <- length(x)
sx <- sort(x)
sx2 <- sx[c(2:lenx,1)]
# Determines closest neighbors of a fixed observation
posmed <- c()
for(i in 1:lenx) {
posmed[i]<- ave.ang(c(sx[i],sx2[i]))}
# Computes circular mean of two adjacent observations
posmed <- posmed[posmed != 9999]
posmed }

```

## A.5 checkeven()

This function checks if the number of possible medians is even. It is an internal function for the main programs. Input: data vector x.

```
checkeven<-function(x){
lenx <- length(x)
sx <- sort(x)
check <- c()
# Computes possible medians
posmed<- posmedf(x)
for(i in 1:length(posmed)){
#Takes posmed[i] as the center, i.e. draws diameter at posmed[i] and
counts observations on either side of the diameter
newx <-sx-posmed[i]
check[i]<-ifelse(sum(round(cos(newx),10)>0)<lenx/2, 9999,posmed[i])}
nposmed<- check[check≠ 9999]
nposmed }
```

## A.6 checkodd()

This function checks if the number of possible medians is odd. It is an internal function needed for the main programs. Input: data vector x.

```
checkodd <- function(x) {
lenx <- length(x)
sx <- sort(x)
check <- c()
posmed <- sx
# Each observation is a possible median
for (i in 1:length(posmed)) {
newx <- sx-posmed[i]
#Takes posmed[i] as the center, i.e. draws diameter at posmed[i] and
counts observations on either side of the diameter
check[i] <- ifelse(sum(cos(newx) > 0) > (lenx-1)/2, 9999,posmed[i]) }
nposmed <- check[check ≠9999]
nposmed }
```

## A.7 meandev()

This function calculates circular mean deviation. It is an internal function needed for the main programs. Input: data vector x.

```
meandev <- function(x, teta) {
# Checks if circular mean exists
ifelse(teta == 9999, 9999, (pi - mean(round(abs(pi -
(abs(rangeang( x - teta))), 10))))})
```

## A.8 rangeang()

This function puts data in  $(-p, p)$  range. It is an internal function needed for the main programs. Input: data vector x.

```
rangeang <-function(x) {
ang <-ifelse(x < - pi, x + 2 * pi, x)
ang2<- ifelse(ang > pi, ang - 2 * pi, ang)
return(ang2)
```

## Using Multinomial Logistic Models To Predict Adolescent Behavioral Risk

Chao-Ying Joanne Peng  
School of Education  
Indiana University-Bloomington

Rebecca Naegle Nichols  
School of Health, Physical Ed. and Recreation  
Indiana University-Bloomington

---

Multinomial logistic regression was applied to data comprising 432 adolescents' self reports of engagement in risky behaviors. Results showed that gender, intention to drop from the school, family structure, self-esteem, and emotional risk were effective predictors collectively. Three methodological issues were highlighted: (1) the use of odds ratio, (2) the absence of an extension of the Hosmer and Lemeshow test for multinomial logistic models, and (3) the missing data problem. Psychologists and educators can utilize findings to plan prevention programs, as well as to apply the versatile and effective logistic technique in psychological, educational, and health research concerning adolescents.

Key words: Adolescent behavior, self-esteem, behavioral risk, emotional risk, family structure, multinomial logistic model, logistic modeling

---

### Introduction

Adolescence is a very influential time in the life of a young person. It is a time of change and possible insecurity, accompanied by low self-esteem and emphasis on peer approval (Bergman & Scott, 2001; Brack, Orr, & Ingersoll, 1988; McGee & Williams, 2000). This may be the reason that many risky health habits are developed during adolescence. One example is smoking. A study conducted by Everett and Husten (1999) revealed that 81% of college aged students who reported ever being daily smokers began smoking before the age of 18. Furthermore, they found that among those who ever smoked a whole cigarette, 43.0% did so for the first time at the age of 14 or younger; 23.7% at age 15 or 16. Other researchers have come to similar conclusions regarding the adoption of risky health habits during adolescents (Bergman & Scott, 2001; McGee & Williams, 2000; Orr, Wilbrandt, Brack, Rauch, & Ingersoll, 1989).

Because many health-endangering behaviors are engaged in for the first time during adolescence, one goal of health education is to reduce the initiation of health-endangering behaviors. These behaviors include, but are not limited to, unsafe sexual activity (Orr, et al., 1989) and the use of alcohol, tobacco, and marijuana (McGee & Williams, 2000). It is essential that health educators identify those youth at greatest risk so that effective programs may be targeted specifically toward minimizing or eliminating these behaviors. In this paper, we demonstrate the utility of multinomial logistic regression model in identifying adolescents at greatest health risk from their personal as well as family characteristics. Psychologists and educators can utilize findings to plan prevention programs, as well as to apply the versatile logistic regression technique in psychological, educational, and health research concerning adolescents.

Logistic regression is a promising statistical technique that can be used to predict the likelihood of a categorical outcome variable. It has found widespread use in the epidemiological literature, where often the dependent variable is presence or absence of a disease state. This technique has also proven useful in broader areas — social sciences (e.g., Chuang, 1997; Janik and Kravitz, 1994; Tolman and Weisz, 1995) and education, especially higher education (Austin,

---

Send correspondence to Joanne Peng, Department of Counseling and Educational Psychology, 4050 201 N. Rose Ave., Bloomington, IN 47405-1006 or email peng@indiana.edu. E-mail Rebecca Naegle at rnaegle@yahoo.com. We wish to thank Gary M. Ingersoll for the use of the data.

Yaffee, & Hinkle, 1992, Cabrera, 1994; Peng, So, Stage, & St. John, 2002) — than the typical epidemiological situation. To profile adolescents who are at greatest risk of participation in risky health behaviors, multinomial logistic regression was applied to data comprising 432 adolescents' self reports of engagement in risky behaviors. Results are interpreted in terms of substantive and methodological implications. The remainder of this paper is divided into four sections: (1) Methodology, (2) The Multinomial Logistic Regression Model, (3) Interpreting and Assessing Multinomial Logistic Regression Results, and (4) Conclusion.

### Methodology

Self-reported health behavior data were collected from 517 adolescents enrolled in two junior high schools (grades 7 through 9) in the fall of 1988. Parents were notified by mail that the survey was to be conducted. Both the parents and the students were assured of their rights to optional participation and confidentiality of students' responses. Written parental consent was waived with the approval of the school administration and the university Institutional Review Board (Ingersoll, Grizzle, Beiter, & Orr, 1993). Among the 517 students, 85 did not complete all questions. Thus, the final sample size was 432 (83.4% were Whites and the remaining Blacks or others) with a mean age of 13.9 years and nearly even composition of girls ( $n=208$ ) and boys ( $n=224$ ). The problem with missing data is addressed later in a section titled Missing Data.

Health Behavior Questionnaire (HBQ; Ingersoll & Orr, 1989; Resnick, Harris, & Blum, 1993) and Rosenberg's self esteem inventory (Rosenberg, 1965) were administered on the same day to all students in all math classes (a mandatory subject). The HBQ asked adolescents to indicate whether they engaged in specific risky health behaviors (Behavioral Risk Scale) or had experienced selected emotions (Emotional Risk Scale). Examples of behavioral risk items were "I use alcohol (beer, wine, booze)," "I use pot," and "I have had sexual intercourse/gone all the way." These items measured frequency of adolescents' alcohol and drug use, sexual activity, and delinquent behavior. They were responded to on a 4-point ordinal scale (1=never, and 4=about once a

week). Emotional risk items measured adolescents' quality of relationship with others, and management of emotions (e.g., "I have attempted suicide," "I have felt depressed," etc.). Cronbach's alpha reliability (Nunnally, 1977) was 0.84 for the Behavioral Risk Scale and 0.81 for the Emotional Risk Scale.

Adolescents' self esteem was assessed using Rosenberg's self esteem inventory (Rosenberg, 1965). Self-esteem scores ranged from 9.79 to 73.87 with a mean of 49.97 and standard deviation of 10.09. Furthermore, among the 432 adolescents, 12.27% (or 53) indicated an intention to drop out of school; 44.68% (or 193) were from intact families, 22.69% (or 98) were from families with one step-parent, and 32.63% (or 141) were from families headed by a single parent.

For the present data, we were interested in identifying adolescents at the greatest behavioral risk from their gender, intention to drop out from school, family characteristics, emotional risks, and self-esteem scores. In addition to identifying youth at the greatest behavioral risk, we were also interested in differentiating adolescents at medium level of risk from those at low risk so that psychologists and educators could utilize findings to design appropriate prevention programs to help adolescents with different needs. Given the objective of this study, the research hypothesis posed to the data was stated as follows: "the likelihood that an adolescent is at high, medium, or low behavioral risk is related to his/her gender, intention to drop out of school, family structure, emotional risk, and self esteem." The dependent variable was students' risk level on the Behavioral Risk Scale of the HBQ; it is hereafter referred to as the RISK variable. The explanatory variables included gender, intention to drop out of school, type of family structure, emotional risk, and self-esteem scores.

Scores on the Behavioral Risk Scale of the HBQ ranged from 40.44 to 66.81 with a mean of 47.69 and a standard deviation of 10.89. Adolescents at highest behavioral risk ( $n=29$ ) were identified to be those scored at least one standard deviation above the mean, i.e., 60 or higher. Those scored between 45 and 59 were identified to be at medium behavioral risk ( $n=170$ ), and those scored between 44 and 40 were at low behavioral risk ( $n=233$ ). The cutoff used to separate those at

medium risk from those at low risk was the median of the distribution (between 44 and 45), given the positive skewness of the scores on the Behavioral Risk Scale and the 4 point scale used for each item. Those classified as at low behavior risk were adolescents who answered, on the average, between “never”, coded as 1, and “once in a while”, coded as 2.

The relationship between the RISK dependent variable and each of the three categorical explanatory variables is shown in Tables 1A through 1C. According to Table 1A, boys were classified into high or medium behavioral risk groups more frequently than girls while the trend was reversed for the low risk group. Table 1B revealed that adolescents intending to drop out of school were more likely to exhibit high or medium behavioral risk than those without such an intention. As for the relationship between family structures and behavioral risk, a majority of adolescents from either intact or step-parent families exhibited a low level of behavioral risk whereas a majority of those from single-parent families showed a medium level of behavioral risk (Table 1C).

Table 1A. Distribution of Gender and Three Levels of Behavioral Risk.

Behavioral Risk Levels	Gender		Total
	Girls=0	Boys=1	
High Risk	5	24	29
Medium Risk	66	104	170
Low Risk	137	96	233
Total	208	224	432

Table 1B. Distribution of Dropout and Three Levels of Behavioral Risk.

Behavioral Risk Levels	Dropout		Total
	No=0	Yes=1	
High Risk	15	14	29
Medium Risk	137	33	170
Low Risk	227	6	233
Total	379	53	432

Table 1C. Distribution of Family Structure and Three Levels of Behavioral Risk.

Behavioral Risk Levels	Family Structure			Total
	Intact=1	Step=2	Single=3	
High Risk	8	7	14	29
Medium Risk	62	38	70	170
Low Risk	123	53	57	233
Total	193	98	141	432

#### The Multinomial Logistic Regression Model

Logistic regression is well suited for describing and testing hypotheses about relationships between a categorical dependent variable and one or more categorical or continuous explanatory variables. Specifically, multinomial logistic regression was chosen to answer the research question for two reasons. First, multinomial logistic regression provides an effective and reliable way to obtain the estimated probability of belonging to a specific population (e.g., high risk adolescents) and the estimate of odds ratio of adolescents' characteristic on their behavioral risk (Peng, Lee, & Ingersoll, 2002; Peng, Manz, & Keck, 2001; Scott, Mason, & Chapman, 1999).

Second, multinomial logistic regression is a procedure by which estimates of the net effects of a set of explanatory variables on the dependent variable can be obtained (Morgan & Teachman, 1988). Even though logistic regression has been used in health research, the use of multinomial logistic regression is rare. In this section, we will first describe the general logic behind the multinomial logistic regression model. This is followed by the specification of a multinomial logistic model for the present data in order to answer the research question.

The simplest form of the multinomial logistic regression model involves one categorical dependent variable  $Y$  (e.g., three levels of behavioral risk) and one explanatory variable,  $X$  (e.g., emotional risk score). Let  $p_1$  = the probability of high behavioral risk ( $Y=3$ ),  $p_2$  = the probability of medium behavioral risk ( $Y=2$ ), and  $p_3$  = the probability of low behavioral risk ( $Y=1$ ). The simplistic multinomial logistic regression model relates the log of odds (or logit) of  $Y$  to the explanatory variable,  $X$ , in a linear form:



$$\text{Logit}(p_1) = \text{naturallog}(\text{odds}) = \ln\left(\frac{p_1}{1-p_1}\right) = \mathbf{a}_1 + \mathbf{b}X \quad (1)$$

$$\text{Logit}(p_1 + p_2) = \text{naturallog}(\text{odds}) = \ln\left(\frac{p_1 + p_2}{1-p_1-p_2}\right) = \mathbf{a}_2 + \mathbf{b}X. \quad (2)$$

Note both equations (1) and (2) constitute one multinomial logistic model with the constraint that  $\Sigma p_i = 1$ . They model the cumulative probabilities with a common slope parameter ( $\mathbf{b}$ ) but different  $Y$  intercepts ( $\alpha_1, \alpha_2$ ). The two  $Y$  intercepts are two constants in the multinomial logistic model; they are not a function of the predictor  $X$ .

The predictor,  $X$ , can be categorical or continuous while the outcome ( $Y$ ) is always categorical. Parameters,  $\alpha_1, \alpha_2$ , and  $\beta$ , are typically estimated by the maximum likelihood (ML) method. The ML method is designed to maximize the likelihood of reproducing the data given their parameter estimates (Peng, Lee, et al., 2002). The value of the coefficient  $\beta$  reveals the direction of the relationship between  $X$  and the logit of  $Y$ . When  $\beta$  is greater than 0, larger (or smaller)  $X$  values are associated with larger (or smaller) logits of  $Y$ , and the curve will resemble an increasing sigmoid (or  $S$ -shape). Conversely, if  $\beta$  is less than 0, larger (or smaller)  $X$  values are associated with smaller (or larger) logits of  $Y$ . Such a relationship is often shown in data in the form of a reverse sigmoid curve. In other words, an increase in  $X$  is associated with a decrease in logits of  $Y$  and vice versa.

Within the framework of inferential statistics, the null hypothesis states that  $\beta$  equals zero in the population. Rejecting such a null hypothesis implies that a linear relationship exists between  $X$  and the logit of  $Y$ . If an explanatory variable is binary, such as gender in Table 1A and dropout in Table 1B, the  $\beta$  coefficient can also be interpreted as an odds ratio which numerically equals  $e$  (the natural logarithm base) raised to the exponent of  $\beta$  (i.e.,  $e^\beta$ ).

If two or more explanatory variables are included in the model (say  $X_1$ = gender and  $X_2$ = emotional risk score), one may construct a complex logistic regression for the logit of  $Y$  (high, medium, or low levels of behavioral risk) as follows:

$$\text{Logit}(p_1) = \text{natural log}(\text{odds}) = \ln\left(\frac{p_1}{1-p_1}\right) = \mathbf{a}_1 + \mathbf{b}_1X_1 + \mathbf{b}_2X_2 \quad (3)$$

$$\text{Logit}(p_1) = \text{naturallog}(\text{odds}) = \ln\left(\frac{p_1 + p_2}{1-p_1-p_2}\right) = \mathbf{a}_2 + \mathbf{b}_1X_1 + \mathbf{b}_2X_2. \quad (4)$$

As noted before, equations (3) and (4) constitute one complex multinomial logistic model with the constraint that  $\Sigma p_i = 1$ . They model the cumulative probabilities with common slope parameters ( $\beta_1$  and  $\beta_2$ ) but different  $Y$  intercepts ( $\alpha_1, \alpha_2$ ). The two  $Y$  intercepts are two constants in the multinomial logistic model; they are not a function of the explanatory variables. Explanatory variables,  $X_1$  and  $X_2$ , can be categorical or continuous while the dependent variable ( $Y$ ) is always categorical. Parameters,  $\alpha_1, \alpha_2, \beta_1$ , and  $\beta_2$ , are estimated by the maximum likelihood (ML) method, as in the simple multinomial model. Data are entered into the analysis as 1, 2, or 3 coding for the trichotomous dependent variable, continuous values for continuous explanatory variables, and dummy coding (e.g., 0 or 1) for categorical explanatory variables.

The null hypothesis underlying the complex multinomial logistic model states that all  $\beta$ 's equal zero. Rejecting this null hypothesis implies that at least one  $\beta$  does not equal 0 in the population. The interpretation of  $\beta$  is rendered using odds ratios. If  $\beta_j$  represents the regression coefficient for predictor  $X_j$ , exponentiating  $\beta_j$  yields the odds ratio ( $e^{\beta_j}$ ). When all other explanatory variables are held at a constant, odds ratio is the change in the odds of  $Y$  given a unit change in  $X_j$ .

For the behavioral risk data, we hypothesized the following linear relationship might exist:

$$\text{Logit}(p_1) = \ln\left(\frac{p_1}{1-p_1}\right) = \mathbf{a}_1 + \mathbf{b}_1X_1 + \mathbf{b}_2X_2 + \mathbf{b}_3X_3 + \mathbf{b}_4X_4 + \mathbf{b}_5X_5, \quad (5)$$

$$\text{Logit}(p_1 + p_2) = \ln\left(\frac{p_1 + p_2}{1-p_1-p_2}\right) = \mathbf{a}_2 + \mathbf{b}_1X_1 + \mathbf{b}_2X_2 + \mathbf{b}_3X_3 + \mathbf{b}_4X_4 + \mathbf{b}_5X_5, \quad (6)$$

where  $p_1$ = the probability of high behavioral risk ( $Y=3$ ),  $p_2$ = the probability of medium behavioral risk ( $Y=2$ ), and  $p_3$ = the probability of low behavioral risk ( $Y=1$ ),  $X_1$ =GENDER (boys=1, girls=0),  $X_2$ =intention to drop out of school (DROPOUT, yes=1, no=0),  $X_3$ =family structure (FAMILY, intact family=1, step-family =2, and

single-parent family=3),  $X_4$ =emotional risk score (EMOTION), and  $X_5$ =self-esteem score (ESTEEM).

Alternatively, one can express the same functional relationship by taking the antilog function of Equations (5) and (6) to obtain a direct estimate of the probabilities of behavioral risk:

$$p_1 = \frac{e^{a_1+b_1X_1+b_2X_2+b_3X_3+b_4X_4+b_5X_5}}{1+e^{a_1+b_1X_1+b_2X_2+b_3X_3+b_4X_4+b_5X_5}} \quad (7)$$

$$p_1 + p_2 = \frac{e^{a_2+b_1X_1+b_2X_2+b_3X_3+b_4X_4+b_5X_5}}{1+e^{a_2+b_1X_1+b_2X_2+b_3X_3+b_4X_4+b_5X_5}} \quad (8)$$

where  $e=2.71828$  is the base of the system of natural logarithms. Equation (7) defines  $p_1$  directly, whereas  $p_2$  and  $p_3$  are derived by subtraction; i.e.,  $p_2 = (p_1 + p_2) - p_1 =$  equation 8 – equation 7, and  $p_3 = 1 - (p_1 + p_2) = 1 -$  equation 8. As previously defined,  $p_1$  = the probability of high behavioral risk ( $Y=3$ ),  $p_2$  = the probability of medium behavioral risk ( $Y=2$ ), and  $p_3$  = the probability of low behavioral risk ( $Y=1$ ).

#### Interpreting and Assessing Multinomial Logistic Regression Results

Equations (7) and (8) were fitted to data using SAS® PROC LOGISTIC (Version 8e, SAS Institute Inc., 1999) in order to support/refute the research hypothesis posed earlier that “the likelihood that an adolescent is at high, medium, or low behavioral risk is related to his/her gender, intention to drop out of school, family structure, emotional risk, and self esteem.” The result showed that

$$\text{Predicted logit (Y1=High RISK)} = -0.6211 + (1.1070)*\text{GENDER} + (2.1818)*\text{DROPOUT} + (0.4135)*\text{FAMILY} + (0.00738)*\text{EMOTION} + (-0.0488)*\text{ESTEEM}, \quad (9)$$

and

$$\text{Predicted logit (Y1+ Y2 =High + Medium RISK)} = 2.5220 + (1.1070)*\text{GENDER} + (2.1818)*\text{DROPOUT} + (0.4135)*\text{FAMILY} + (0.00738)*\text{EMOTION} + (-0.0488)*\text{ESTEEM} \quad (10)$$

The  $\chi^2$  test of proportional odds assumption was insignificant ( $df=5$ ;  $p=0.6548$ ), indicating that there was no need to fit a second model with distinct  $\beta$  parameters (Peterson & Harrell, 1990). Hence, Equations (9) and (10) will be hereafter referred to as the MLR model.

#### Interpreting Multinomial Logistic Regression Results

According to the MLR model, the log of the odds of an adolescent’s behavioral risk level was positively related to gender ( $p<.0001$ , Table 2), intention to drop out of school ( $p<.0001$ ), and family structure ( $p<.001$ ); it was negatively related to self-esteem ( $p<.0001$ ), and insignificantly related to emotional risk ( $p = 0.5211$ ). The positive coefficient (1.1070) associated with GENDER in the MLR model implied that boys were more likely, than girls, to be at high behavioral risk, holding all other explanatory variables constant. In fact, the odds of a boy being at high behavioral risk were 3.025 ( $= e^{1.1070}$ , Table 2) times greater than the odds for a girl. The same trend was observed with the dichotomous variable of DROPOUT from school. The odds of teen-age students engaging in high or medium risk of behavior, than not, were 8.8622 times higher for students intending to drop out than students without such an intention. This relationship can also be seen in Table 1B in which the majority of those intending to not stay in school were placed in high or medium level of behavioral risk, compared to those with intentions to stay in school.

Regarding the third categorical variable family structure, interpretation should be based on the reference group of intact families. Thus, the higher the score on FAMILY, the less stability in the family structure and the greater is the behavioral risk for adolescents. This interpretation was rendered by the positive coefficient associated with FAMILY. As a family’s structure changed from 1 (intact family) to 2 (step family) or from 2 to 3 (single family), the odds increased by 1.5121 for adolescents to be at a higher behavioral risk than medium or low risk.

The coefficient for self-ESTEEM indicated that the decrease in log odds of risky behavior corresponded to one unit increase in self-esteem scores. In other words, the higher the self-esteem score, the less likely an adolescent would

be at high behavioral risk. For each point increase on the self esteem score, the odds of participating in risky behavior, compared to the odds of not participating, decreased from one to 0.952 ( $= e^{-0.0488}$ , Table 2). If the increase on the self-esteem score was 10 points, the odds decreased from one to 0.6139 [ $= e^{10*(-0.0488)}$ ].

Combining the four explanatory variables that were found to be statistically significant in the MLR model, a profile emerged for a youth at the greatest predicted behavioral risk: a male who intended to drop out of school, came from a single parent household, scored low on the self-esteem measure, and possibly high on the emotional risk measure (based on the positive correlation between behavioral risk and emotional risk) — this last characteristic did not reach statistical significance in the MLR model.

#### Assessing Multinomial Logistic Regression Results

How effective was the MLR model expressed in Equations (9) and (10)? How can a health educator assess the soundness of a multinomial model? To answer these questions, we attended to (a) overall model evaluations, (b) statistical tests of each explanatory variable, (c) goodness-of-fit statistics, and (d) validations of predicted probabilities. These evaluations are discussed below based on Equations (9) and (10), or the MLR model.

(a) Overall model evaluations. The Likelihood Ratio, Score, and Wald tests were examined to determine the improvement of the MLR model over the intercept-only model (also called the null model). According to Peng, Lee, and Ingersoll (2002, p.6), “An intercept-only model serves as a good baseline because it contains no predictors; consequently all observations would be predicted to belong in the largest outcome category, according to this model.” All three tests yielded similar results ( $p < .0001$ , Table 2), namely, the MLR Model was more effective than the null model. It was therefore inferred that at least one explanatory variable was a significant predictor of adolescents’ behavioral risk. After splitting the sample randomly 5 times, resulting in 10 random sub-

samples, we applied the same multinomial model to the sub-samples. The overall significance of the MLR model was replicated in all 10 sub-samples.

(b) Statistical tests of individual predictors. The individual  $\beta$  coefficients were tested using the Wald  $\chi^2$  statistic (Table 2). All variables except for EMOTION were significant predictors of adolescents’ risk for self-injurious behaviors ( $p < .001$ ). Two predictors (GENDER, and ESTEEM) were cross-validated to be significant; one predictor (EMOTION) was replicated to be statistically insignificant, all with 10 random sub-samples. FAMILY structure and intention to DROPOUT were confirmed to be statistically significant predictors in 9 out of 10 cross-validation random samples. It was not necessary to statistically test the intercepts for the two constants (CONSTANTS 1 and 2 in Table 2) as the test result merely indicates if intercepts should be included in a logistic model (Peng, Lee, & Ingersoll, 2002).

(c) Goodness-of-fit statistics. Goodness-of-fit statistics assess the fit of a logistic model against actual classifications, i.e., high, medium, or low level of behavioral risk. Two descriptive measures of goodness-of fit are presented in Table 2 for the MLR model:  $R^2$  indices defined by Cox and Snell (1989) and Nagelkerke (1991), respectively. These two measures were similar for the MLR model (24.67% and 29.78%). According to Peng, Lee, and Ingersoll (2002), these indices are variations of the  $R^2$  concept defined for the ordinary least squares (OLS) regression model.

Even though the  $R^2$  has a clear definition in OLS regression, there have been no equivalents of this concept devised by methodologists for multinomial logistic models that render the meaning of variance explained; none correspond to predictive efficiency, and none can be tested in an inferential framework (Mendard, 2000). For these reasons, a researcher may treat these two  $R^2$  indices reported in Table 2 as supplementary to other, more useful evaluative indices such as the overall evaluation of the model, tests of individual regression coefficients, and the inferential test of the goodness-of-fit suggested by Begg and Gray (1984) for multinomial logistic models.

Table 2. Multinomial Logistic Regression Analysis of Adolescent's Self-inflicting Behavior Risk by SAS® PROC LOGISTIC (version 8).

Predictor	<i>b</i>	<i>SE b</i>	Wald's $\chi^2$ ( <i>df</i> =1)	<i>p</i>	<i>e<sup>b</sup></i> ( <i>odds ratio</i> )
CONSTANT 1 ( <i>Y</i> <sub>1</sub> )	-0.6211	1.0627	0.3416	0.5589	Not necessary
CONSTANT 2 ( <i>Y</i> <sub>1</sub> + <i>Y</i> <sub>2</sub> )	2.5220	1.0723	5.5317	0.0187	Not necessary
<u>GENDER</u> (boys=1,girls=0)	1.1070	0.2111	27.5060	<0.0001	3.0253
<u>DROPOUT</u> (yes=1, no=0)	2.1818	0.3287	44.0618	<0.0001	8.8622
<u>FAMILY</u>	0.4135	0.1179	12.2979	<0.001	1.5121
<u>EMOTION</u>	0.0074	0.0115	0.4118	0.5211	1.0074
<u>ESTEEM</u>	-0.0488	0.0118	16.9867	<0.0001	0.9524

Overall Model Evaluation

Tests	$\chi^2$	<i>df</i>	<i>p</i>
Likelihood Ratio Test	122.38	5	<0.0001
Score test	110.47	5	<0.0001
Wald test	97.87	5	<0.0001

*Notes.* Cox and Snell *R* squared=0.2467. Nagelkerke *R* squared (Max rescaled *R* squared)=0.2978. Kendall's Tau-*a* = 0.297. Goodman-Kruskal's Gamma= 0.548. Somers' *D*<sub>xy</sub>= 0.539. *c*-statistic = 0.769.

SAS® Programming Codes

```
PROC LOGISTIC DATA=risk432
  MODEL risk= gender dropout family emotion esteem;
  OUTPUT out=probs predicted=prob xbeta=logit;
RUN;
```

According to Begg and Gray (1984, cited in Hosmer & Lemeshow, 2001, p. 281), the goodness-of-fit test of a multinomial model may be carried out by applying the Hosmer and Lemeshow (H-L) test to two of the three outcome categories, then integrating the test results descriptively. For the logistic model comparing low risk adolescents with medium risk adolescents, the H-L test yielded a  $\chi^2$  of 5.8011 with 8 degrees of freedom. For the logistic model comparing low risk adolescents with high risk adolescents, the H-L test yielded a  $\chi^2$  of 8.2925, also with 8 degrees of freedom. Both test results were statistically insignificant ( $p > .40$ ) indicating that both models fit the data well. In other words, the null hypothesis of a good model fit to data was tenable.

(d) Validations of predicted probabilities. As was explained previously, the MLR model predicts the logit of high and medium levels of behavioral risk from a set of explanatory variables. Since logit is the natural log of the odds [or probability/ (1-probability)], it can be transformed back to the probability scale, according to Equations (7) and (8). Once the predicted probability of behavioral risk is calculated, it can be compared with the actual risk behavior to determine if high probabilities are associated with the high level of behavioral risk, low probabilities with the low level of behavioral risk, and middle-range probabilities with the medium level of behavioral risk.

SAS® PROC LOGISITC (version 8) provides four measures of association for logistic regression models. These are: Kendall's Tau-*a*, Goodman-Kruskal's Gamma, Somers' *D* statistic, and the *c* statistic (Table 2). Kendall's Tau-*a* is a rank-order correlation coefficient without adjustments for ties; for the MLR model, it equaled 0.287. Goodman-Kruskal's Gamma equaled 0.548. According to Peng, Lee, and Ingersoll (2002), it is a more useful and appropriate measure than Tau-*a* when there are ties on both dependent variable categories and predicted probabilities (the present data had 923 ties — approximately 1.8% of all pairs). This measure is interpreted as 54.8% fewer errors made in predicting which of two adolescents would be at a greater behavioral risk by utilizing the estimated probabilities, than by chance alone (Demaris,

1992). Some caution is advised in using the Gamma statistic since: (1) it has a tendency to overstate the strength of association between estimated probabilities and outcomes (Demaris), and (2) a value of zero does not necessarily imply independence when the data structure exceeds a 2 by 2 format (Siegel & Castellan, 1988).

Somers' *D* is a preferred extension of Gamma whereby one variable is designated as the dependent variable and the other the independent variable (Siegel & Castellan, 1988). For the MLR model, Somers' *D* was 0.539 (Table 2). There are two asymmetric forms of Somers' *D* statistic:  $D_{yx}$  and  $D_{xy}$ . Only  $D_{yx}$  correctly represents the degree of association between the behavioral risk level (*y*), designated as the dependent variable, and the estimated probability (*x*), designated as the independent variable (Demaris, 1992).

Unfortunately, SAS® computes only  $D_{xy}$ , although this index can be corrected to  $D_{yx}$  in SAS® (Peng & So, 1998). For the present model, the *c* statistic was 0.769 (Table 2). This means that for 76.90% of all possible pairs of adolescents, one at a greater risk (e.g., high or medium level) than the other (e.g., medium or low level), the MLR model correctly assigned a higher probability to those measured by HBQ at greater behavioral risk. Thus the model worked better than assigning observations randomly into categories of high, medium, or low behavior risk. The *c* statistic ranges from 0.5 to 1.

A 0.5 value means that the model is no better than assigning observations randomly into categories of the dependent variable. A value of 1 means that the assignment of probabilities matches perfectly with the ordered categories of the dependent variable (e.g., high with high, medium with medium, and low with low). If several models were fitted to the same data, the model chosen as the "best" model should be associated with the highest *c* statistic. Thus, the *c* statistic provides a basis for comparing different models fitted to the same data, or the same model fitted to different data sets.

#### Reporting Multinomial Logistic Regression Results

In addition to Tables 1 and 2, it is helpful to profile adolescents with certain characteristics and relate these characteristics to the predicted

probability of engaging in high, medium, or low level of risky behaviors. For this purpose, several boys and girls, from either an intact, step-parent, or single-parent home, were selected from the data base. These characteristics, along with their indication to stay in or drop out of school and their emotional risk and self-esteem measure, are shown in Table 3 (following References section) to be related to their predicted probability of engaging in various levels of risky behaviors. It is noted in Table 3 that 8 cases (#6, 12, 19, 22, 30, 31, 34, and 36) did not exist in the data. These cases may be explained by their refusal to participate, missing data (to be addressed in the next section), and the improbable likelihood of locating these adolescents in the population (e.g., case #30, 31, 34, and 36).

Among boys from the intact family (cases #1 to #5), the probability of engaging in low-level of risky behaviors (#3) was associated with a very low emotional risk score and no intention to drop out of school. Likewise, girls from the intact family (cases #7 to #11), who were predicted to engage in low-level of risky behaviors, did not intend to drop out from school and were measured low on emotional risk.

Boys from the step-parent family (#13 to #18), were predicted to engage in medium to high level of risky behaviors. The higher the emotional risk score, the greater was the probability of being associated with high-risk behaviors (#18). For girls from step-parent families (#20, 21, 23, and 24), those with no intention to drop out of school (#20 and #21) were predicted to engage in lower levels of risky behaviors than those with an intention to drop out of school.

Among boys from the single-parent home (#25 to #29), engaging in high-level risky behaviors was predicted for the boy with an intention to drop out of school (#29), whereas low-level was predicted for the boy who had no intention to drop out of school, scored low on the emotion risk measure, and high on the self-esteem test (#26). Among girls from the single-parent home (#32, 33, and 35), all were predicted to engage in medium level of risky behaviors. Though cases #32 and #33 did not intend to drop out of school, they scored high on emotional risk and low on self-esteem. Case #35 intended to drop out of school; she was measured comparatively low on emotional risk and high on self-esteem.

### Missing Data

It is important to point out the problem with missing data encountered in the multinomial logistic modeling, especially for the explanatory variable emotional risk (EMOTION). Descriptive analyses of the data suggested one plausible explanation for the insignificant relationship between emotional risk and behavioral risk (Table 2). Of the 85 cases with missing data, 77 were missing behavioral risk data, 34 were missing emotional risk data, and six were missing drop-out scores. It was noted that the range (34.21 to 82.03), mean (50.11), and standard deviation (10.94) for the 51 (=85-34) emotional risk scores not included in the analysis, were slightly higher than those used in the analysis. Furthermore, 25 (or 49.02%) of the 51 emotional risk scores were above the overall sample mean of 48.72. It would be important to ascertain why adolescents with slightly higher emotional risk scores chose not to complete the behavioral risk assessment. Thus, missing data on the dependent variable might not be missing completely at random (Little and Rubin, 1987).

To answer this question statistically, we imputed all missing values using the EM method installed in the MVA (missing value analysis) module of SPSS Version 11.01. The complete data set with imputed values ( $N=517$  observations) contained 255 adolescents at low behavioral risk, 228 at medium risk, and 34 at high risk. The complete data set was submitted to SAS® PROC LOGISTIC (Version 8e) for multinomial logistic regression modeling. Results were very similar to those in Table 2, namely, gender, intention to drop out from school, family structure, and self-esteem were statistically significant at  $p<.0001$ . The emotional risk variable was again not a statistically significant predictor. An examination of correlations between the behavioral risk level and the five predictors showed that the positive correlation between emotional risk scores and the behavioral risk level, though positive, was not as high as the correlation between self-esteem scores and behavioral risks. And there was a strong negative correlation between emotion risk and self-esteem (Pearson  $r = -.494$ ). Based on these results, we concluded that the missing data did not bias the interpretations given earlier for the MLR model.

Conclusion

In this article, we applied multinomial logistic regression to data based on 432 adolescents' self-reported measures of behavioral risk, emotional risk, self-esteem, intention to drop out of school, and their gender and family structure to test a research hypothesis. The research hypothesis stated that, "the likelihood that an adolescent child is at high, medium, or low level of self-injurious behavioral risk is related to his/her gender, intention to drop out of school, family structure, emotional risk, and self esteem." Logistic regression results supported the statistical significance of four explanatory variables.

Specifically, the likelihood of an adolescent participating in risky behaviors was negatively related to his/her self-esteem scores, but positively related to intention to drop out of school, family structure, and gender. If all other explanatory variables were held as constants, adolescents with the following profiles were more likely, than their counterparts, to engage in risky behaviors: boys, intending to drop out of school, living in a single-parent household, and having low self-esteem. The effectiveness of the multinomial logistic model was supported by multiple indices, including the model's overall test of all explanatory variables, statistical significance test of each explanatory variable, the predictive power of the model, and its interpretability.

Three methodological issues encountered during the logistic regression analysis were highlighted and treated in our discussion of the results. These included (1) the use of odds ratio in interpreting results obtained from MLR models, (2) the absence of an extension of the Hosmer and Lemeshow goodness-of-fit test for multinomial logistic models, and (3) the missing data problem.

From the standpoint of modeling categorical outcomes, logistic regression is more flexible and less restrictive than discriminant function analysis, log-linear models, or modified probability models (Peng, Manz, & Keck, 2001). While logistic regression is gaining popularity in health and social sciences research (Peng, Lee, & Ingersoll, 2002; Peng, So, Stage, & St. John, 2002), there are few studies that demonstrate a preferred pattern of the application of multinomial logistic regression methods. It is hoped that this paper has demonstrated that multinomial logistic

regression is an effective technique for profiling those youth at greatest risk for participation in risky health behaviors. Psychologists and educators can utilize findings to plan prevention programs, as well as to apply the versatile logistic technique in psychological, educational, and health research concerning adolescents.

References

Austin, J. T., Yaffee, R. A., & Hinkle, D. E. (1992). Logistic regression for research in higher education. *Higher education: Handbook of theory and research, Vol. VIII*, 379-410.

Begg, C. B., & Gray, R. (1984). Calculation of polychotomous logistic regression parameters using individualized regressions. *Biometrika*, 71, 11-18.

Brack, C. J., Orr, D. P., & Ingersoll, G. (1988). Pubertal maturation and adolescent self-esteem. *Journal of Adolescent Health Care*, 9, 280-285.

Cabrera, A. F. (1994). Logistic regression analysis in higher education: An applied perspective. *Higher education: Handbook of theory and research, Vol. X*, 225-256.

Chuang, H. L. (1997). High school youth's dropout and re-enrollment behavior. *Economics of Education Review*, 16(2), 171-186.

Demaris, A. (1992). *Logit modeling: Practical applications*. Newbury Park, CA: Sage.

Everett, S. & Husten, C. G. (1999). Smoking initiation and smoking patterns among US college students. *Journal of American College Health*, 48(2), 55-61.

Hosmer, D. W., Jr., & Lemeshow, S. (2000). *Applied logistic regression* (2nd ed.). New York: John Wiley & Sons, Inc.

Ingersoll, G. M. & Orr, D. P. (1989). Behavioral and emotional risk in early adolescents. *Journal of Early Adolescence*, 9, 392-408.

Ingersoll, G. M., Grizzle, K., Beiter, M. & Orr, D. P. (1993). Frequent somatic complaints and psychosocial risk in adolescents. *Journal of Early Adolescence*, 13(1), 67-78.

Janik, J., & Kravitz, H. M. (1994). Linking work and domestic problem with police suicide. *Suicide and Life Threatening Behavior*, 24(3), 267-274.

- Little, R. J., & Rubin, D. B. (1987). *Statistical analysis with missing data*. New York: Wiley.
- Menard, S. (2000). Coefficients of determination for multiple logistic regression analysis. *The American Statistician*, 54(1), 17-24.
- Morgan, S. P., & Teachman, J.D. (1988). Logistic regression: Description, examples, and comparisons. *Journal of Marriage and the Family*, 50, 929-936.
- Nunnally, J. (1977). *Psychometric theory* (2<sup>nd</sup> ed.). New York: McGraw-Hill.
- Orr, D. P., Wilbrandt, M. L., Brack, C. J., Rauch, S. P., & Ingersoll, G. M. (1989). Reported sexual behaviors and self-esteem among young adolescents. *American Journal of Diseases of Children*, 143, 86-90.
- Peng, C. Y., & So, T. S. (1998). If there is a will, there is a way: Getting around defaults of PROC LOGISTIC in SAS. *Proceedings of the Midwest SAS Users Group 1998 Conference*, 243-252. (<http://php.indiana.edu/~tso/articles/mwsug98.pdf>)
- Peng, C. Y., Manz, B. D., & Keck, J. (2001). Modeling categorical variables by logistic regression. *American Journal of Health Behavior*, 25(3), 278-284.
- Peng, C. Y., Lee, K. L., & Ingersoll, G.M. (2002). An introduction to logistic regression analysis and reporting. *Journal of Educational Research*, 96(1), 3-14.
- Peng, C. Y., So, T. S. H., Stage, F. K., & St. John, E. P. (2002). The use and interpretation of logistic regression in higher education journals: 1988-1999. *Research in Higher Education*, 43(3), 259-293.
- Peterson, B. & Harrell, F. (1990). Partial proportional odds models for ordinal response variables. *Applied Statistics*, 39, 205-217.
- Resnick, M. D., Harris, L. J., & Blum R.B. (1993). The impact of caring and connectedness on adolescent health and well-being. *Journal of Pediatrics Child Health*, 29(1), 53-59.
- Robins, P. K., & Dickinson, K. P. (1985). Child support and welfare dependence: A multinomial logit analysis. *Demography*, 22(3), 367-380.
- Rosenberg, M. (1965). *Society and the adolescent self-image*. Princeton, NJ: Princeton University Press.
- Rouse, K. A. G., Ingersoll, G. M., & Orr, D. P. (1998). Longitudinal health endangering behavior risk among resilient and nonresilient early adolescents. *Journal of Adolescent Health*, 23, 297-302.
- SAS Institute Inc. (1995). *Logistic regression examples: Using the SAS®, version 6, first ed.* Cary, NC: SAS Institute Inc.
- SAS Institute Inc. (1999). *SAS/STAT® user's guide, version 8, volume 2*. Cary, NC: SAS Institute Inc.
- Scott, K. G., Mason, C. A., & Chapman, D. A. (1999). The use of epidemiological methodology as a means of influencing public policy. *Child Development*, 70(5), 1263-1272.
- Siegel, S., & Castellan, N. J. (1988). *Nonparametric statistics for the behavioral science* (2nd ed.). New York: McGraw-Hill.
- Tabachnick, B. G., & Fidell, L. S. (1996). *Using multivariate statistics* (3rd ed.). New York: Harper Collins.
- Tolman, R.M., & Weisz, A. (1995). Coordinated community intervention for domestic violence: the effects of arrest and prosecution on recidivism of woman abuse perpetrators. *Crime and Delinquency*, 41(4), 481-495.
- Yarandi, H. N., & Simpson, S. H. (1991). The logistic regression model and the odds of testing HIV positive. *Nursing Research*, 40(6), 372-373.



Table 3. Predicated Probability of Participating in Self-injurious Behavior for 36 Children.

Case No.	SEX β=1.107 1=boy 0=girl	DROPOUT β=2.1818 1=yes 0=no	FAMILY β=0.4135 1=intact, 2=step, 3=single	EMOTION β=0.0074	ESTEEM β=-0.0488	Intercept 1 α <sub>1</sub> =-0.6211	Intercept 2 α <sub>2</sub> =2.522	Predicted probability of participating in self-injurious behavior			Actual Behavior risk, 1=high, 2=med, 3=low (score on HBO, M=47.69, SD=10.89)
								p <sub>1</sub> (high)	p <sub>2</sub> (medium)	p <sub>3</sub> (low)	
1	1	0	1	62.39	32.68	-0.6211	2.5220	.0818	.5921	.3261	1 (60.40)
2	1	0	1	80.74	32.68	-0.6211	2.5220	.0926	.6102	.2972	2 (52.77)
3	1	0	1	32.07	71.58	-0.6211	2.5220	.0106	.1878	.8016	3 (42.65)
4	1	1	1	72.72	46.41	-0.6211	2.5220	.3038	.6062	.0900	1 (95.21)
5	1	1	1	63.07	37.25	-0.6211	2.5220	.3885	.5479	.0636	2 (50.00)
6	1	1	1	----	----	-0.6211	2.5220	---	---	---	3 (-----)
7	0	0	1	47.29	41.83	-0.6211	2.5220	.0166	.2645	.7189	1 (61.53)
8	0	0	1	45.78	44.12	-0.6211	2.5220	.0147	.2422	.7431	2 (47.07)
9	0	0	1	42.05	21.24	-0.6211	2.5220	.0425	.4643	.4932	3 (42.70)
10	0	1	1	51.37	34.97	-0.6211	2.5220	.1772	.6559	.1669	1 (70.23)
11	0	1	1	56.77	37.25	-0.6211	2.5220	.1670	.6559	.1771	2 (53.27)
12	0	1	1	----	----	-0.6211	2.5220	---	---	---	3 (-----)
13	1	0	2	41.36	50.98	-0.6211	2.5220	.0451	.4776	.4773	1 (72.83)
14	1	0	2	46.14	50.98	-0.6211	2.5220	.0467	.4848	.4685	2 (45.84)
15	1	0	2	36.11	41.83	-0.6211	2.5220	.0663	.5559	.3778	3 (40.44)
16	1	1	2	38.59	57.85	-0.6211	2.5220	.2269	.6449	.1282	1 (92.50)
17	1	1	2	54.87	46.41	-0.6211	2.5220	.3665	.5641	.0694	2 (46.99)
18	1	1	2	70.35	34.97	-0.6211	2.5220	.5312	.4321	.0367	3 (43.52)
19	0	0	2	---	---	-0.6211	2.5220	---	---	---	1 (-----)
20	0	0	2	34.21	44.12	-0.6211	2.5220	.0203	.3041	.6756	2 (45.78)
21	0	0	2	50.18	53.27	-0.6211	2.5220	.0147	.2421	.7432	3 (40.44)
22	0	1	2	----	---	-0.6211	2.5220	---	---	---	1 (-----)
23	0	1	2	54.84	50.98	-0.6211	2.5220	.1326	.6473	.2201	2 (48.64)
24	0	1	2	50.18	46.41	-0.6211	2.5220	.1559	.6547	.1894	3 (43.08)
25	1	0	3	63.52	23.52	-0.6211	2.5220	.2432	.6384	.1184	1 (67.90)
26	1	0	3	32.07	67.00	-0.6211	2.5220	.0296	.3848	.5856	2 (56.69)
27	1	0	3	50.18	48.70	-0.6211	2.5220	.0786	.5854	.3360	3 (40.44)
28	1	1	3	43.54	48.70	-0.6211	2.5220	.4184	.5250	.0566	1 (85.49)
29	1	1	3	56.74	44.12	-0.6211	2.5220	.4979	.4604	.0417	2 (54.31)
30	1	1	3	---	---	-0.6211	2.5220	---	---	---	3 (-----)
31	0	0	3	---	---	-0.6211	2.5220	---	---	---	1 (-----)
32	0	0	3	64.12	28.10	-0.6211	2.5220	.0786	.5856	.3358	2 (48.41)
33	0	0	3	60.08	39.54	-0.6211	2.5220	.0453	.4781	.4766	3 (44.41)
34	0	1	3	---	---	-0.6211	2.5220	---	---	---	1 (-----)
35	0	1	3	43.63	48.70	-0.6211	2.5220	.1922	.6543	.1535	2 (46.34)
36	0	1	3	---	---	-0.6211	2.5220	---	---	---	3 (-----)

## Bayesian Analysis Of Poverty Rates: The Case Of Vietnamese Provinces

Dominique Haughton  
Bentley College, USA

Nguyen Phong  
General Statistics Office, Vietnam

---

This paper presents a Bayesian analysis of poverty rates in urban Ho Chi Minh City and rural Nghe An province in Vietnam. Using mixtures of beta distributions as priors for the poverty rates, we find that, when the prior is reasonably informative, our approach yields more accurate estimated poverty rates than a frequentist approach. On the other hand, we find that, in the presence of poor/non-poor misclassification, average probabilities of posterior credible intervals for poverty rates can fall well short of .95 even with sample sizes such as 2000 or 3000 when the width of the interval is for example four percentage points. In general, we suggest reporting prior and posterior means and standard deviations along with traditional frequentist measures. Our results rely on techniques due to Nandram and Sedransk (1993) and Rahme, Joseph and Gyorkos (2000), and make use of the software WINBUGS.

Key words: Vietnamese poverty, Bayesian analysis, WINBUGS

---

### Introduction

The problem of estimating the binomial parameter has attracted a lot of attention among statisticians and others in the business of estimating proportions. It is widely known that, informally speaking, large sample sizes are needed to get acceptable accuracies when estimating proportions.

Sample size estimations are often based on classical computations of confidence intervals, sometimes adjusted to take into account special survey designs. Recent work of Brown (2001) has focused attention on the shortcomings of such confidence intervals, notably on the fact that “95% confidence intervals” have less than 95% coverage in a number of cases.

In the context of the estimation of poverty rates, we are led to the estimation of a binomial parameter, since the poverty rate is in general defined as the proportion of households whose annual expenditure per capita falls below a given poverty line. In most of this paper we will assume that this poverty line is non-random, and that the classification poor/non-poor is known accurately. We will discuss the implications of an inaccurately known poverty line in the latter part of the paper.

---

Dominique Haughton is Professor of Mathematical Sciences at Bentley College. Her research interests include model selection, statistics applied to marketing and the analysis of living standards in Vietnam. With Jonathan Haughton and Nguyen Phong, she is a co-editor of the book “Living standards during an economic boom: the case of Vietnam”. Email her at [DHAUGHTON@bentley.edu](mailto:DHAUGHTON@bentley.edu). Nguyen Phong is the director of the Social and Environmental Department in the General Statistics Office in Hanoi, Vietnam. He leads the implementation of large nationwide surveys on living standards. With Dominique Haughton and Jonathan Haughton, he is a co-editor of the book “Living standards during an economic boom: the case of Vietnam”. Email him at [nphong@gso.gov.vn](mailto:nphong@gso.gov.vn).

The estimation of poverty rates for Vietnamese provinces lends itself very well to a Bayesian analysis: informative prior information is frequently available; moreover sample sizes tend to be fairly small, since surveys are expensive and prone to non-sampling errors. Sampling statisticians and others involved in the design and analysis of such surveys (in Vietnam or elsewhere) have to date not performed a Bayesian analysis of poverty rates (see, for example, Glewwe & Yansaneh, 2001, for an exposition of a typical

analysis in this context).

We will show in this article that a gain in accuracy is obtained when a reasonably informative prior is used, and when the poverty line is assumed known. We will illustrate this result with a wealthier urban sample (urban Ho Chi Minh City), and a poorer rural sample (rural Nghe An). However, to qualify these results, one should keep in mind that when poor/non-poor misclassification occurs, as it almost certainly does, the average coverage of four-percentage-point-wide probability intervals does not reach .95, even asymptotically in large sample sizes, while it is likely to do so for an eight-percentage-point-wide probability interval.

Methodology

Bayesian Estimation Of Poverty Rates When The Poverty Line Is Known

In urban Ho Chi Minh City, our sample from the Vietnam Living Standards Survey of 1998 has 433 households, 2 of which are poor. Frequentist weighted (according to sampling weights) computations yield a poverty rate of .00462, with a standard deviation of .00334 (yielding a coefficient of variation of about .7). In order to perform the Bayesian analysis, we use a mixture of beta distributions as a prior for the unknown poverty rate as suggested in Nandram and Sedransk. This is justified by the work of Dalal and Hall (1983), who showed that any prior can be approximated by such a mixture. We then apply the closed form formulas of Nandram and Sedransk for the posterior mean and posterior standard deviation of the poverty rate for a two-stage cluster sample design.

In our case, we assume that a commune is randomly selected, then a household randomly selected from the commune; in reality there is an additional step in the sampling design – a village is randomly selected from the commune – and then a household is randomly selected from the village. We expect to address the issue of three-level clustering in future work; no closed form formula is available in this case for the posterior mean and standard deviation of the poverty rate. The present analysis is a close approximation of reality, though; we don't expect the addition of the third level to make a large difference. We then simulate the posterior distribution using

WINBUGS, with the code published in Congdon (2001; example 5.18 p. 196). In addition to the data on poor/non-poor households from surveyed communes, the analysis makes use of the number of households in each commune of urban Ho Chi Minh City and rural Nghe An respectively; the model specifies an individual poverty rate for each commune and then combines these poverty rates into an overall poverty rate for the province.

Results

In Table 1 and Figure 1, we present the results from four different priors for urban Ho Chi Minh City. In Table 2 and Figure 2, we present the results from two different priors for rural Nghe An. The posterior means and standard deviations are those of the overall poverty rate for the whole area (urban Ho Chi Minh City and rural Nghe An, respectively). The mixture of beta distributions used for the prior for a vector  $\theta$  of N poverty rates for N communes is given by Nandram and Sedransk (1993) as:

$$\pi(\theta | \tau) = \sum_{r=1}^R \omega_r B(a_r, \tau - a_r)^{-N} \prod_{k=1}^N \theta_k^{a_r - 1} (1 - \theta_k)^{\tau - a_r - 1}$$

where  $\theta_k$  is the poverty rate for the  $k^{th}$  province, and B denotes the Beta function. The values of  $\omega_r$ ,  $a_r$  and  $\tau$  must be chosen when eliciting the prior. Note that the means of the beta distributions in the mixture are  $a_r/\tau$ , and that the value of  $\tau$  controls the standard deviation of the beta distributions; the higher  $\tau$  is, the smaller the standard deviation.

The two first priors for urban Ho Chi Minh City are based loosely on poverty rates and their standard deviations for Vietnamese provinces estimated in Baulch et al. (2002), using data from the Census of 1999 and regression equations based on VLSS data. These estimates were used to define 4 bins centered at the values indicated in the column "Mean" in Table 1 for each of 4 components, and prior probabilities of .07, .43, .43 and .07 for each of the 4 bins. Note that the value of 4 for R was chosen somewhat arbitrarily for convenience and flexibility. Priors 1 and 2 differ by the value of  $\tau$ , and thus by the standard deviations.

The components are less separated in prior 2, as seen on Figure 1. The results from both priors are close, a posterior poverty rate of about .01, with a standard deviation of about .005, an improvement (coefficient of variation of about .5) over the frequentist estimation. Figure 1 shows the two posterior densities from priors 1 and 2 to be close, and to give most of the posterior probability to two components, conceivably corresponding to more and less affluent communes. Prior 3 corresponds to a prior elicited from the expert opinion that “we are 95% certain that the poverty rate for urban Ho Chi Minh City is between .01 and .03”. As for priors 1 and 2, 4 bins were created for prior 3, centered at values given in Table 1 and with widths consistent with the expert prior belief. The summary statistics for the posterior poverty rate are quite similar to those for priors 1 and 2. Prior 4 is a very diffuse prior, and in this case, the posterior poverty rate is not accurate (standard

deviation of .008) as can be expected.

In this case, we have both closed form expressions for the posterior means and standard deviations, as well as the option of using WINBUGS to generate a sample from the posterior. The results from both analyses should be, and are, close. We note here that we have found that if the beta components are too well separated or if one of the components is too close to 0, it can happen that the MCMC chain in WINBUGS gets “stuck” in a component, and gives an incorrect posterior mean. This in fact is not surprising to the authors of WINBUGS (N. Best, personal communication), and could be remedied by checking the WINBUGS results against the closed form formulas for a two-level cluster sample design for a given prior, and then moving on to more complex survey designs if desired.

Table 1: Prior And Posterior Means And Standard Deviations; Ho Chi Minh City Urban

		Prior 1, $\tau = 200$		Prior 2 $\tau, = 80$	
$\omega$		<i>Mean</i>	<i>St. Dev.</i>	<i>Mean</i>	<i>St. Dev.</i>
.07	Comp. 1	.005	.005	.005	.008
.43	Comp. 2	.015	.009	.015	.014
.43	Comp. 3	.045	.015	.045	.023
.07	Comp. 4	.075	.019	.075	.029
	Overall	.031	.023	.031	.027
		Post. Mean	Post. St. Dev	Post. Mean	Post. St. Dev.
	Closed form	.009872	.004982	.010765	.004911
	Winbugs	.009664	.004964	.010611	.004910

Table 1 (continued)

		Prior 3, $\tau = 80$		Prior 4, $\tau = 40$	
$\omega$		<i>Mean</i>	<i>St. Dev.</i>	<i>Mean</i>	<i>St. Dev.</i>
.07	Comp. 1	.009	.010	.005	.011
.43	Comp. 2	.016	.014	.025	.024
.43	Comp. 3	.024	.017	.080	.042
.07	Comp. 4	.031	.019	.140	.054
	Overall	.020	.017	.055	.051
		Post. Mean	Post. St. Dev	Post. Mean	Post. St. Dev.
	Closed form	.013684	.004561	.008841	.007801
	Winbugs	.013530	.004508	.010130	.008632

FIGURE 1: PRIOR DENSITIES AND POSTERIOR KERNEL DENSITIES;  
HO CHI MINH CITY URBAN

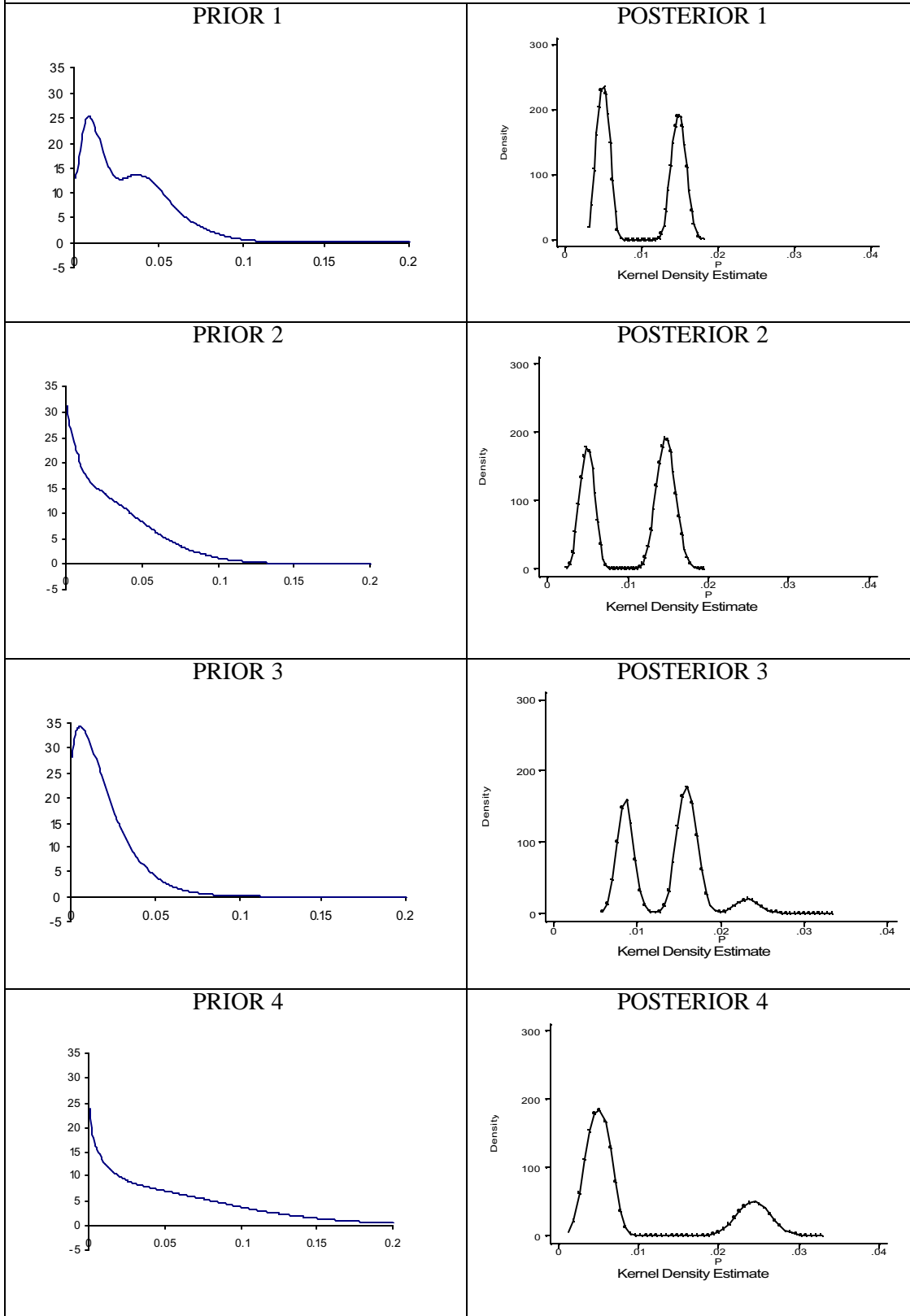


Table 2: Prior And Posterior Means And Standard Deviations; Nghe An Rural

$\alpha$		Prior 1, $\tau = 40$		Prior 2, $\tau = 30$	
		<i>Mean</i>	<i>St. Dev.</i>	<i>Mean</i>	<i>St. Dev.</i>
.07	Comp. 1	.225	.065	.050	.039
.43	Comp. 2	.375	.076	.125	.059
.43	Comp. 3	.525	.078	.275	.080
.07	Comp. 4	.675	.073	.425	.089
	Overall	.450	.133	.205	.122
		Post. Mean	Post. St. Dev	Post. Mean	Post. St. Dev.
	Closed form	.499810	.055138	.424697	.008203
	Winbugs	.503400	.051560	.424500	.009934

For rural Nghe An, we have 225 sampled households, among which 110 are poor. Weighted frequentist estimations give an estimated poverty rate of .489, with a standard deviation of .104. Prior 1 is again based loosely on the estimations in Baulch et al. (2002); it yields a posterior mean for the poverty rate of about .5, with a posterior standard deviation of .05, an improvement in accuracy over the frequentist analysis.

Prior 2 is based on an estimated poverty rate of about .2 from MOLISA (Ministry of Labour, Invalids and Social Affairs), used to create 4 bins of about the same width as in prior 1. The prior poverty rate of .2 is probably too low, and it is interesting to see how the Bayesian analysis uses the data to correct this prior information: the MCMC chain concentrates almost exclusively on one higher component to yield a posterior mean of .42 with a standard deviation of about .01 for the poverty rate.

#### Bayesian Estimation Of Sample Sizes In The Presence Of Misclassification

We now consider the case where it is in fact not known exactly which households are poor and which are not. Poverty lines are difficult to establish, in large part because of the difficulty in getting accurate data on the prices of basic goods. So the problem of identifying poor households is similar to the problem of diagnosing a disease on the basis of an imperfect test.

We use here work of Rahme et al. (2000) where Bayesian sample size determinations are performed for the binomial parameter subject to

misclassification, and applied to a situation in the medical area. In this context, the test for poverty has a sensitivity (probability of a poor household being classified as poor) and a specificity (probability of a non-poor household being classified as non-poor), both with a beta prior distribution following Rahme et al. (2000), and the prevalence of poverty (the poverty rate) is also given a beta prior distribution.

We illustrate this approach in the case of rural Nghe An. We define a prior distribution of a beta with parameters  $\alpha = 70.32$  and  $\beta = 77.1$  for the poverty rate, on the basis of the estimates for the poverty rate and its standard deviation in Baulch et al. (2002), and elicit beta distributions as priors for the sensitivity and specificity of the poor/non-poor classification from the opinion that the mean sensitivity (and specificity) is about .95 and that we are 95% certain that the sensitivity (or specificity) is between .9 and 1. This opinion yields the values for the beta parameters given in Table 3.

The table gives average coverages of probability intervals for two different interval widths and three different sample sizes, calculated from an S-plus program made available by Rahme (2000) et al. It is clear that the coverage will not attain .95 for a width of 4 percentage points, even with very large sample sizes. Such a coverage might be feasible with an interval of width .08, with large sample sizes. However, we note that the techniques in Rahme et al. (2002) assume i.i.d. samples, so the situation is likely to be somewhat worse in a situation where a more complex survey design was used. We also note that less

informative priors on the poverty rates and/or the sensitivity and the specificity of the poor/non-poor

classification would be likely to yield even smaller average coverage probabilities.

Table 3: Average Coverage Of Probability Intervals For Poverty Rates For Nghe An Rural Assuming I.I.D. Samples

$\alpha_{sens} = \alpha_{spe\ c} = 71.25; \beta_{sens} = \beta_{spe\ c} = 3.75; \alpha = 70.32; \beta = 77.1$		
Width of interval	Sample size	Prob. coverage
.04	1000	.6439
.04	2000	.6924
.04	3000	.6995
.08	1000	.9261
.08	2000	.9471
.08	3000	.9587

Conclusion

We have shown in this paper the benefits of a Bayesian approach to the estimation of poverty rates. Poverty rates are often calculated – and reported – as sample proportions. In some cases, a measure of accuracy such as standard deviation is reported as well.

In our analyses, the use of sensible prior information has provided a significant improvement in the accuracy of the poverty rates, as measured by their posterior standard deviation, provided that the poverty line is known exactly. Results tend to be robust with respect to the choice of a sensible prior.

Our Bayesian analysis has also shown that whenever there is uncertainty in the poor/non-poor classification, the accuracy of poverty rates as measured by the width of posterior credible intervals is significantly negatively affected. For example, coverage probabilities of about 95% may require interval widths of about 8 percentage points, implying poverty rates known only up to four percentage points.

In general we suggest that posterior means and standard deviations be reported along with more traditional measures, and that a discussion of the accuracy of poverty lines accompany poverty rates reports.

References

Baulch, B. & Minot, N. (2002). The spatial distribution of poverty in Vietnam and the potential for targeting. *World Bank working paper* 2829, Washington, DC.

Brown, L. D., Cai, T. T., & DasGupta, A. (2001). Interval estimation for a binomial proportion. *Statistical Science*, 16(2), 101-133.

Congdon, P (2001). *Bayesian statistical modelling*. NY:Wiley.

Dalal, S. R., & Hall, W. J. (1983). Approximating priors by mixtures of natural conjugate priors. *Journal of the Royal Statistical Society*, B45(2), 287-286.

General Statistical Office (1999). *Vietnam living standards survey*, Hanoi, Vietnam.

Glewwe, P., & Yansaneh, I. (2001). *Mission report: Recommendations for multi-purpose household surveys from 2002 to 2010*. World Bank, Washington, DC.

Nandram, B., & Sedransk, J. (1993). Bayesian predictive inference for a finite population proportion: Two-stage cluster sampling. *Journal of the Royal Statistical Society*, B55(2), 399-408.

Rahme, E., Joseph L., & Gyorkos, T. (2000). Bayesian sample size determination for estimating binomial parameters from data subject to misclassification. *Applied Statistics*, 49, 119-128.

## Comparison Of Estimates Of Proprietary And Syndicated Methods In Auto Industry Surveys

Daniel X. Wang  
Department of Mathematics  
Central Michigan University

---

Proprietary and syndicate surveys are often used in assessing appeal and initial quality of new vehicles for automobile manufactures. This study discusses the difference between the two types of studies, and proposes a computer simulation based method for checking the appropriateness of the comparisons.

Key words: J.D. Power, sample base, pp100 score, initial quality

---

### Introduction

Quality and assessing quality becomes more and more important issues to the modern automotive industry. The customer survey of J.D. Power and Associates was founded in 1968 as an independent professional information provider for management and it has been considered the most important source for assessing marketing, quality and customer satisfaction.

As one of the important J. D. Power auto surveys, the *Initial Quality Study 2 (IQS2)* contains comprehensive and analytically rich information that can help auto manufacturers position their image and products. Consumers of new vehicles are surveyed regarding problems they experienced after 90 days of vehicle ownership. All the problems are weighted equally and the result is summarized with problems per 100 vehicles. The pp100 scores are compared across models and platforms, by manufacturer and assembly plants. The survey contains 135 problems (since 1998) and over nine categories.

Auto manufacturers highly regard the results of J. D. Power auto surveys as a measure of their performance in terms of quality, service and customer's satisfaction. Toyota considers that J. D. Power and Associates is the most respectable name in auto consumers' minds and its IQS has been the industry standard benchmark for vehicle quality since 1987. Auto manufacturers would like to mention their achievement recognized through the surveys by J.D. Power and Associates. For example "Corolla was the highest ranked Compact Car in the J.D. Power and Associates' 2000 Initial Quality Study. Study based on a total of 47,909 consumer responses indicating owner-reported problems during the first 90 days of ownership (Spring 2002 [www.toyota.com](http://www.toyota.com))". "Expedition shines when it comes to Initial Quality. The Expedition ranked as the Best Full-Size Sport Utility Vehicle in Initial Quality in the J.D. Power and Associates 2001 Initial Quality Study based on a total of 54,565 consumer responses indicating owner reported problems during the first 90 days of ownership (Spring 2002, [www.ford.com](http://www.ford.com))."

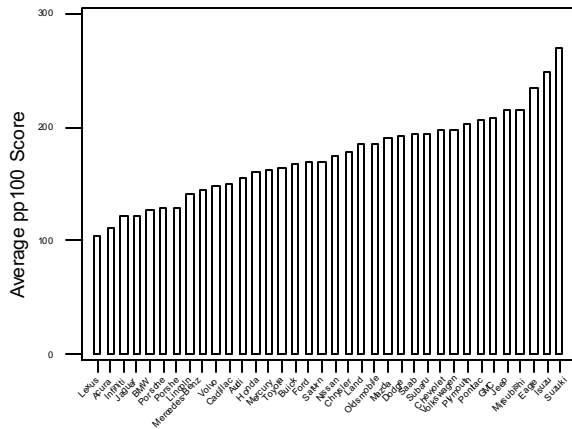
Figure 1 is an example of IQS results, which give the industrial performance for the total of 36 manufacturers (Spring 2002, [www.auto.com](http://www.auto.com)).

---

Daniel Wang is an Assistant Professor of Statistics at Central Michigan University. He graduated from the University of Alabama with a Ph. D. in Applied Statistics in 1999. He worked at the Biostatistics Unit at the University of Alabama as a Ph.D. fellow and was a statistical consultant at Mercedes Benz, US International. E-mail him at: [Wang1dx@cmich.edu](mailto:Wang1dx@cmich.edu).



Figure 1  
Example of J. D. Power IQS2 by Mark with 175 Models



In order to monitor the continuous quality improvement and to forecast the IQS results, manufacturers often conduct proprietary studies similar to the IQS study through J.D. Power & Associates monthly or quarterly. However, due to the effect of many factors of sampling methods, the comparison of the two types of studies is questionable. For example a random sample is used for the IQS study while a stratified random sample is used for the IQS study. Other factors may include different periods for reporting problems, sample size of vehicles, complete sample base and incomplete sample base. A valid comparison cannot be made without appropriately addressing these issues. This article focuses on comparing the results using two different sampling methods. Concerns about partial sample base and complete sample base are also discussed.

### Methodology

The two different sampling methods used in auto surveys of J. D. Power and Associates are introduced in this section with the notations and derived estimates.

#### Syndicated Study and Proprietary Study

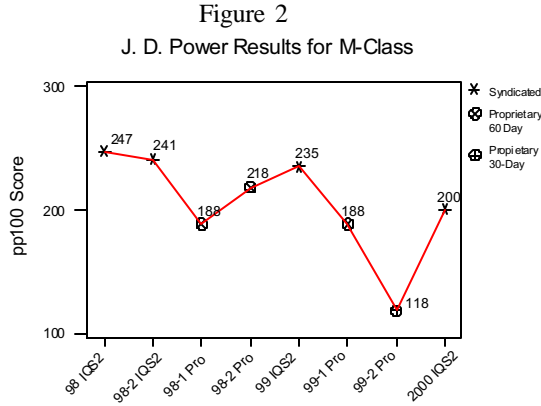
The syndicated survey is a number of studies of automobiles conducted by J.D. Power and Associates independently. The results of these studies are published and the detail results for a specific model may be sent to the manufacturer. The detail results can be analyzed for quality and customer's satisfaction improvement, especially

for manufacturers who believe the philosophy that customer should determine what they want and what they like. The Appeal Study by J.D. Power Associates is also used for assessing customer's satisfaction.

Proprietary survey is the studies, which are usually similar to J.D. Power study conducted by J.D. Power and Associates but upon the request of a manufacturer. In addition to the syndicated studies, the proprietary studies are considered as a continuous monitor of the product performance. Further the results are used for forecasting the future J.D. Power score. Instead of the three month time period for reporting problems for customers in the syndicated study, the time period for the proprietary study may vary. For example it could be one month or two months depending on the manufacture's interest.

Two different sampling methods have been used in the two types of auto survey. For the syndicated survey such as IQS study, stratified samples are drawn from the same model of vehicles, because minimum sample size is required for a model. Usually about 30% of the registrations for the total leased vehicles are not available for J.D. Power. Therefore using a stratified sample can help to obtain a desired number of vehicles in the sample, which include both purchased and leased vehicles. On the other hand, for the proprietary survey the manufacturers usually provide all possible registrations for the purchased and the leased vehicles. So a random sample is used for the proprietary study. Figure 2 gives a typical example of the IQS2 scores sampled using different methods in different time periods for a type of vehicle.

When a result of proprietary study is compared to the syndicated study, there are some concerns about how the difference of the sampling bases, and different sample methods and different time period for reporting problems. This study focuses on the discussion of comparing the two sampling methods given the same sample base, then discusses the results for the case of having different sample bases, which simulates the situation of the syndicated study without part of leased vehicles versus the proprietary study with full sample base.



Estimates

Suppose a stratified sample is drawn for the syndicated study with size  $n$ , then  $n = n_p + n_L$  where  $n_p$  is the number of purchased vehicles and  $n_L$  is the number of leased vehicles in the sample. The estimated pp100 score the estimated as

$$\hat{S}_{pp100} = \frac{100 \left( \frac{N_p}{n_p} \sum_{i=1}^{n_p} X_i^P + \frac{N_L}{n_L} \sum_{j=1}^{n_L} X_j^L \right)}{N} = \frac{100(N_p \bar{X}^P + N_L \bar{X}^L)}{N}$$

where  $N = N_p + N_L$  is the total number of vehicles sampling from,  $N_p$  and  $N_L$  are the total numbers for the purchased and the leased.  $\frac{N_p}{n_p}$  and

$\frac{N_L}{n_L}$  are the weights for the number of problems

for the purchased vehicle  $X_i^P$  and the leased vehicle  $X_j^L$ . Suppose the true average number of problems per vehicle for the purchased and the leased are  $\mu_p$  and  $\mu_L$ , then the estimate of the true pp100 score is

$$E(\hat{S}_{pp100}) = \frac{100(N_p E(\bar{X}^P) + N_L E(\bar{X}^L))}{N} = \frac{100(N_p \mu_p + N_L \mu_L)}{N}$$

which is the weighted true pp100 score for the vehicles. The variance of  $\hat{S}_{pp100}$  is

$$\text{Var}(\hat{S}_{pp100}) = \text{Var} \left( \frac{100(N_p (\bar{X}^P) + N_L (\bar{X}^L))}{N} \right) = \frac{100^2 (N_p^2 \sigma_p^2 + N_L^2 \sigma_L^2)}{N^2}$$

For the proprietary study, suppose a random sample is drawn with size  $n$ . The pp100 score is notated as follows using the same type of notation.

$$\hat{S}_{pp100} = \frac{100 \left( \sum_{i=1}^{n_p} X_i^P + \sum_{j=1}^{n_L} X_j^L \right)}{n} = \frac{100}{n} (n_p^* \bar{X}^P + n_L^* \bar{X}^L) = 100\bar{X}$$

and

$$E(\hat{S}_{pp100}) = E(100\bar{X}) = 100\mu = \frac{100(N_p \mu_p + N_L \mu_L)}{N}$$

where  $\mu$  is the true average number of problems per vehicle for all vehicles including both purchased and the leased. Since this is a random sample, both sample sizes for purchased  $n_p^*$  and leased group  $n_L^*$  are also random and they are correlated, because  $n = n_p^* + n_L^*$ . Therefore given the same sample base, both estimates of pp100 scores for the two studies have the same mean, and they are unbiased. For the proprietary study, the variance can be denoted as

$$\text{Var}(\hat{S}_{pp100}) = \text{Var}(100\bar{X}) = \frac{100^2 \sigma^2}{n}$$

where  $\sigma^2$  is the true variance for the number of problems per vehicle for all vehicles. This means the two studies give the unbiased estimates with different variances.

If 30% of leased vehicles are excluded from the sample base due to certain reason, for

example the registration information is not available at the sampling time period, the parameters of the sample base  $\mu_L$ ,  $\sigma_L$ ,  $\mu$  and  $\sigma$  are affected. So the estimates of pp100 scores will depend on how the samples are excluded for the leased base.

Results

It is clear that theoretically comparing the results of the two different surveys is impossible since too many assumptions have to be made about the unknown parameters. Especially for the proprietary sampling, the sample sizes for the purchased  $n_p^*$  and for the leased group  $n_L^*$  are random and they are correlated, but in the syndicated sampling they are both constant. Based on the discussion in the previous section, applied approaches are proposed to investigate the two sampling methods.

For a specific model of vehicle, a computer simulation is used with a simulated sample base. The sample base can be built using existing J.D. Power data as a good approximation to the real sample base. Then exclude 30% or as desired portion of vehicles from the full base to obtain an approximation to sample base similar to the one used in the syndicated study. The next step is to write computer programs or macros for the syndicated study and the proprietary study, then apply them a large number of times to the sample bases built. Comparisons for two studies can be made based on the simulated results.

A sample base for the proprietary study can be built using existing information, which could be from a published source or data for a model of vehicles if the study is conducted for an auto manufacturer. First chose the size of sample base  $N$  with  $N_L$  for the leased and  $N_p$  for the

purchased. Then determine the proportions for the vehicles to have 0 problems to 12 problems, which is the maximum number of problems used in the IQS2 of J.D. Power and Associates. The problems can be also attributed to the nine different categories. Finally form the sample base for the syndicated study by excluding a proportion, for example 30% of leased vehicles from the sample base for the proprietary study.

As an example, using the IQS2 1998 (J.D. Power, 2001 Knight Ridder Inc.) result for M-Class, a sample base with following characteristics (see Table 1), where the mean is the mean number of problems per vehicle. The above sample base is for the proprietary study and it can be considered a good approximation of the M-Class registered during the sampling period of 1998 J.D. Power IQS2 study. Now randomly exclude 30% of vehicles from the leased vehicles, the sample base for the syndicate study of J.D. Power is made with the following statistical summaries (see table).

The structure of the sample base is hypothetical to allocate the proportion of the number of problems from 0 to 12 to all vehicles. The proportions for a specific model of vehicle can be obtained from the actual J. D. Power survey.

Two Minitab macros, one for the syndicated study and the other for the proprietary study are created for the simulations. 5,000 simulations are run for each of the two sampling methods, and for each of the full and partial sample bases. For each combination of sampling method and sampling base, the weighted and not-weighted pp100 scores are reported. The results are shown in Table 3 (on next page) and are also presented as in the following distribution dot plots on the same scale. See figure 3 (next page).

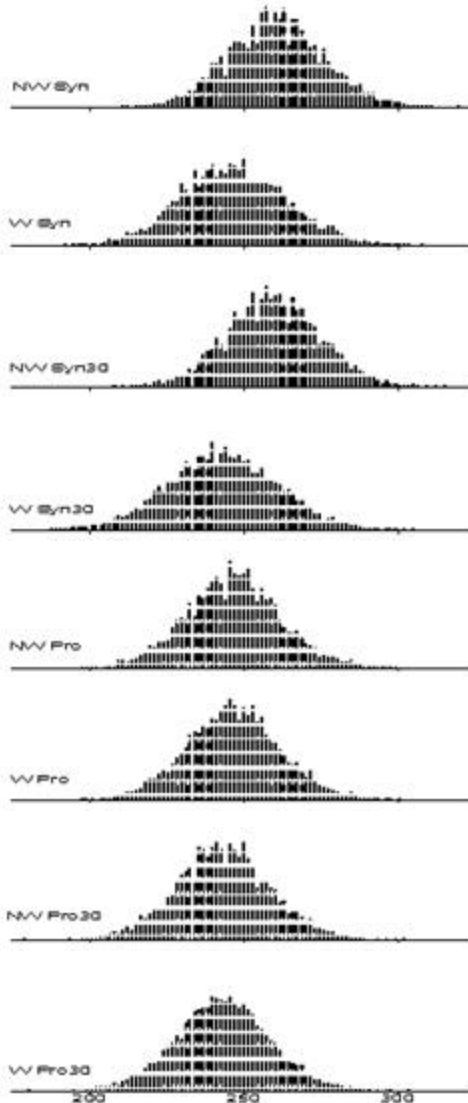
Table 1  
Descriptive Statistics for the Full Sample Base

Variable	N	Mean	Median	StDev	Minimum	Maximum
Purch.&Leas	7760	2.4647	2.0000	2.5267	0.000	12
Purchased	5807	2.3337	2.0000	2.4395	0.000	12
Leased	1953	2.8541	2.0000	2.6965	0.000	11

Table 2  
Descriptive Statistics for Partial Sample Base  
without 30 % of Leased Vehicles

Variable	N	Mean	Median	StDev	Minimum	Maximum
Purch.&Leased	7174	2.4295	2.0000	2.4874	0.000	12
Purchased	5807	2.3337	2.0000	2.4395	0.000	12
Leased	1367	2.8361	2.0000	2.6440	0.000	11

Figure 3



Conclusion

Based on the discussions and the results the computer simulations with the examples in

previous sections, the comments and recommendations are made as the following.

For the same sample bases, both the syndicate and proprietary studies give the same accurate estimates of the true pp100 score on the average. But the syndicate sampling method tends to have larger variation for the estimated score. This means that the syndicated sampling method introduces extra variation into the sample scores.

Table 3  
Summary of the Simulations  
W/NW: Weighted/Not weighted  
F/P: Full Sample Base/Partial Sample Base

	W/ NW	F/P	Simulated pp100 Score	True pp100 Score	StDev
Syndicated Study	N	F	260.47	246.47	15.23
	W	F	246.50	246.47	16.90
	N	P	259.42	242.95	14.96
	W	P	242.91	242.95	17.56
Proprietary Study	N	F	246.41	246.47	14.91
	W	F	246.40	246.47	14.89
	N	P	243.03	242.95	14.72
	W	P	243.02	242.95	14.76

Therefore the syndicate study gives a less accurate sample score than that of the proprietary method. Even though this simulation does not provide in general by how much the variation is between the two sampling methods, it does provide informative details for comparing the results from different sampling methods for a particular model of an auto manufacturer. For example when the manufacturer compares the results from two sampling methods, the variation due to using

syndicated sampling can be assessed with the simulation results.

The 30% exclusion of leased vehicles has some impact on the average score and this impact is significant depending on number and the way vehicles are excluded. The partial sample base introduces additional variation into the syndicated study. In general this is expected, but the simulation gives specific results. If the manufacturer has some knowledge about excluding the leased vehicles, then that can be put into the simulation to get more details about the effect of using partial sample base.

For the proprietary study, both the weighted and the not-weighted scores are the same since random samples are used. But for the syndicated study they are different because stratified random samples are used. This helps the management of an auto manufacturer to understand the "weight" used in syndicated studies of J. D. Power and Associates.

Finally, when comparing the syndicate and the proprietary studies, it is necessary to consider the effect of the variation due to using different sampling methods and different sample bases, especially for monitoring the on-going performance of an auto manufacturer through J.D. Power auto survey. The proposed simulation method can be adapted to a particular model for which both syndicated and proprietary surveys are available. The computer macros can be easily modified for carrying out the simulations. After assessing the variation attributed to the sampling methods and sample base, manufacturers can appropriately compare the pp100 scores of their products.

Clearly, it would be better for the manufacturers to have the proprietary study conducted in the same way as the syndicated study. Although different sample bases are used for the two studies, the extra variation in estimating the pp100 score will be coming from just one source instead of two sources. It is important to get as many details as possible for the proprietary study.

Comparing the pp100 scores with different reporting time periods is worth further study. The reason for auto manufacturers to have one or two month surveys is because the short time studies provide quick response. If the proprietary study is conducted using different time

periods for reporting problems (one or two months), comparing the pp100 scores to that of two months or three months is more complicated because an extra source of variation is introduced. Manufacturers multiply a weight to the one-month or two-month proprietary scores and then compare them to the three-month scores. The weight may be obtained from J. D. Power, for example 70% percent of problems associated with new vehicles are usually reported in the first two months.

#### References

- Alwin, D. F. (1991). Research on survey quality. *Sociological Methods and Research*, 20, 3-29.
- Babbie, E. R. (1973). *Survey research methods*. Belmont, CA: Wadsworth.
- Biemer, P., & R. Caspar. (1994). Continuous quality improvement for survey operations: Some general principles and applications. *Journal of Official Statistics*, 10, 307-326.
- Breiman, L. (1994). The 1991 census adjustment: Undercount or bad data? *Statistical Science*, 9, 458-475.
- Brewer, K. R. W. (1963). Ratio estimation and finite populations: Some results deducible from the assumption of an underlying stochastic process. *Australian Journal of Statistics*, 5, 93-105.
- Bureau of Labor Statistics, *Handbook of Methods*, vols. I & II, (1982). Washington, DC: U.S. Department of Labor.
- Campbell, C., & Joiner, B. (1973). How to get the answer without being sure you asked the question. *American Statistician*, 27, 229-231.
- Cornfield, J. (1944). On samples from finite populations. *Journal of the American Statistical Association*, 39, 236-239.
- Dalenius, T. E. (1981). The survey statistician's responsibility for both sampling and measurement errors. In D. Krewski, R. Platek, & J. N. K. Rao (Eds.) *Current topics in survey sampling*, 17-29. New York, Academic Press.
- Deming, W. E. (1950). *Some theory of sampling*. NY, Dover.

- Deming, W. E., & Stephan, F. F. (1940). On a least squares adjustment of a sample frequency table when the expected marginal totals are known. *Annals of Mathematical Statistics*, *11*, 427-444.
- Godambe, V. P. (1955). A unified theory of sampling from finite populations. *Journal of the Royal Statistician Society, Ser B.*, *17*, 269-278.
- J. D. Power & Associates. *Explore our studies and awards*. [www.jdpa.com/studies/](http://www.jdpa.com/studies/)
- Jones, H. L. (1956). Investigation of the properties of a sample mean by employing random subsample means. *Journal of the American Statistical Association*, *51*, 54-83.
- Kish, L. (1965). *Survey sampling*. NY: Wiley.
- Kover, J. G., Rao, J. N. K., & Wu, C. F. J. (1988). Bootstrap and other methods to measure errors in survey estimates. *Canadian Journal of Statistics* *16 (suppl)*, 25-45.
- Parten, M. (1950). *Surveys, polls, and samples: Practical procedures*. NY: Harper & Brothers.
- Raj, D. (1968). *Sampling theory*. NY: McGraw-Hill.
- Rao, J. N. K., Hartley, H. O., & Cochran, W. G., (1962). A simple procedure for unequal probability sampling without replacement. *Journal of the Royal Statistician Society, Ser. B.*, *24*, 482-491.
- Sudman, S. (1976). *Applied Sampling*. NY: Academic Press.
- Toyota Quality – Satisfaction. (2002). [www.toyota.com/html/shop/distinct/satisfaction/index.html](http://www.toyota.com/html/shop/distinct/satisfaction/index.html)
- Waksberg, J. (1978). Sampling methods for random digit dialing. *Journal of the American Statistical Association*, *73*, 40-46.
- Williams, B. (1978). *A Sampler on Sampling*. NY: Wiley.
- Yates, F. (1946). A review of recent statistical developments in sampling and sampling surveys. *Journal of the Royal Statistical Society*, *109*, 12-30.

## Steady State Analysis Of An M/D/2 Queue With Bernoulli Schedule Server Vacations

Kailash C. Madan      Walid Abu-Dayyeh      Firas Tayyan

Department of Statistics, Faculty of Science  
Yarmouk University, Irbid, Jordan

---

We examine an M/D/2 queue with Bernoulli schedules and a single vacation policy. We have assumed Poisson arrivals waiting in a single queue and two parallel servers who provide identical deterministic service to customers on first-come, first-served basis. We consider two models; in one we assume that after completion of a service both servers can take a vacation while in the other we assume that only one may take a vacation. The vacation periods in both models are assumed to be exponential. We obtain steady state probability generating functions of system size for various states of the servers.

Key words: Two parallel servers, Bernoulli schedules, single vacation policy, deterministic service

---

### Introduction

Vacation Queues have been studied by numerous researchers including Kleinrock (1983), Keilson and Servi (1986), Baba (1986), Doshi (1986,1990), Cramer (1989), Choi & Park (1990), Borthakur & Choudhury (1997), Madan (1992, 1999, 2001), to mention a few. Most of these authors have investigated single server queues assuming Bernoulli schedules or exhaustive service or generalized vacations among several other vacation policies with a single or multiple vacations. Madan and Saleh (2001, 2001, 2001) have studied a single server queue with exponential service and deterministic vacations, deterministic service with exponential vacations and deterministic service with deterministic vacations, assuming Bernoulli schedules.

Those articles considered single server vacation models. Here, we study a queueing system with two parallel servers providing identical deterministic service assuming Bernoulli schedule server vacations with a single vacation policy.

We consider two models. In model A we assume that after a service completion both servers may take a vacation of identical exponential duration and in model B, we assume that only one of the servers can take a vacation of exponential duration. In both models, we assume a single vacation policy which means that whenever a vacation period of a server ends, then he must join the system irrespective of whether there are customers waiting for service or not. That is, the server must join the system even if he finds the system empty on return. The following assumptions briefly describe our models:

Model A: Both Servers Can Take A Vacation.  
The Underlying Assumptions:

A - Customers arrive at the system one by one and their arrivals follow a Poisson distribution with mean arrival rate  $\lambda$ , ( $\lambda > 0$ ).

B - Both servers provide identical deterministic (constant) service with constant service time of length  $b$ , ( $b > 0$ ).

C - After every service, both servers together may take a vacation with probability  $p$  or continue to stay in the system with probability  $1-p$ . The vacation times follow an exponential distribution with mean vacation time  $1/\beta$ , ( $\beta > 0$ ).

---

Send correspondence to Kailash C. Madan, Department of Statistics, Faculty of Science, Yarmouk University, Irbid, Jordan. E-mail him at kailashmadan@hotmail.com.

D- All stochastic processes involved in the system are independent of each other.

Definitions and Notations

Define:

$B_n(t)$ : as the probability that at time  $t$  both servers are available in the system providing service and there were  $n(0)$  customers in the system when the current service started.

$N_n(t)$ : as the probability that at time  $t$  there are  $n(0)$  customers in the system and both servers are on vacation.

$P_n(t)$ : as the probability that at time  $t$  there are  $n(0)$  customers in the system without regardless of the state of the servers

$k_i$ : as the probability of  $i$  arrivals during a service period of constant length  $b$ .

Steady State Forward Equations of the System

Assuming that the steady state exists, let

$$\lim_{t \rightarrow \infty} B_n(t) = B_n, \quad \lim_{t \rightarrow \infty} N_n(t) = N_n, \quad \text{and}$$

$\lim_{t \rightarrow \infty} P_n(t) = P_n$ . Thus,  $B_n$ ,  $N_n$  and  $P_n$  denote the corresponding steady state probabilities. Then applying the usual probability reasoning we obtain the following set of steady state forward equations:

$$B_n = (1-p)k_n[B_0 + B_1 + B_2] + (1-p)\sum_{i=3}^{n+2} B_i k_{n+2-i} + \mathbf{b}N_n, \quad n \geq 0, \quad (1)$$

$$(\mathbf{I} + \mathbf{b})N_n = \mathbf{I}N_{n-1} + pk_n[B_0 + B_1 + B_2] + p\sum_{i=3}^{n+2} B_i k_{n+2-i}, \quad n \geq 1, \quad (2)$$

$$(\mathbf{I} + \mathbf{b})N_0 = pk_0[B_0 + B_1 + B_2], \quad n=0. \quad (3)$$

Steady State Probability Generating Functions for the System Size

We define the following probability generation functions:

$$B(z) = \sum_{n=0}^{\infty} B_n z^n, \quad N(z) = \sum_{n=0}^{\infty} N_n z^n, \quad P(z) = \sum_{n=0}^{\infty} P_n z^n, \quad (4a)$$

$$K(z) = \sum_{n=0}^{\infty} K_n z^n = \sum_{n=0}^{\infty} \frac{e^{-1b}(\mathbf{I}b)^n z^n}{n!} = e^{-1b} \sum_{n=0}^{\infty} \frac{(\mathbf{I}bz)^n}{n!} = e^{-1b(1-z)}, \quad |z| \leq 1. \quad (4b)$$

We multiply (1) by  $z^{n+2}$  and add for all  $n = 0, 1, 2, \dots$ . Then we have

$$\begin{aligned} z^2 \sum_{n=0}^{\infty} B_n z^n &= (1-p)B_0 z^2 \sum_{n=0}^{\infty} k_n z^n + (1-p)B_1 z^2 \sum_{n=0}^{\infty} k_n z^n + (1-p)B_2 z^2 \sum_{n=0}^{\infty} k_n z^n \\ &+ (1-p) \sum_{n=0}^{\infty} \sum_{i=3}^{n+2} B_i k_{n+2-i} z^{n+2} + \mathbf{b}z^2 \sum_{n=0}^{\infty} N_n z^n, \quad n \geq 0 \end{aligned} \quad (5)$$



Then using (4) we obtain from (5)

$$B(z) = \frac{\mathbf{b}z^2N(z) + (1-p)B_0(z^2-1)e^{-Ib(1-z)} + (1-p)B_1z(z-1)e^{-Ib(1-z)}}{z^2 - (1-p)e^{-Ib(1-z)}} \tag{6}$$

Similarly we multiply (2) by  $z^{n+2}$  and (3) by  $z$  and add them for all  $n = 0, 1, 2, \dots$ . Then we have

$$\begin{aligned} (I + \mathbf{b})z^2 \sum_{n=0}^{\infty} N_n z^n &= Iz^3 \sum_{n=0}^{\infty} N_n z^n + pB_0z^2 \sum_{n=0}^{\infty} k_n z^n + pB_1z^2 \sum_{n=0}^{\infty} k_n z^n \\ &+ pB_2z^2 \sum_{n=0}^{\infty} k_n z^n + p \sum_{n=0}^{\infty} \sum_{i=3}^{n+2} B_i k_{n+2-i} z^{n+2}, n \geq 0 \end{aligned} \tag{7}$$

Using (4) we obtain from (7)

$$N(z) = \frac{pB(z)e^{-Ib(1-z)} + pB_0(z^2-1)e^{-Ib(1-z)} + pB_1z(z-1)e^{-Ib(1-z)}}{(I - Iz + \mathbf{b})z^2} \tag{8}$$

Then we solve (6) and (8) simultaneously for  $B(z)$  and  $N(z)$  and obtain on simplifying

$$\begin{aligned} B(z) &= \left[ \frac{[(1-p)(z^2-1)B_0e^{-Ib(1-z)} + (1-p)z(z-1)B_1e^{-Ib(1-z)}][(I - Iz + \mathbf{b})z^2]}{[z^2 - (1-p)e^{-Ib(1-z)}][(I - Iz + \mathbf{b})z^2] - p\mathbf{b}z^2e^{-Ib(1-z)}} \right. \\ &\left. + \frac{\mathbf{b}z^2[p(z^2-1)B_0e^{-Ib(1-z)} + pz(z-1)B_1e^{-Ib(1-z)}]}{[z^2 - (1-p)e^{-Ib(1-z)}][(I - Iz + \mathbf{b})z^2] - p\mathbf{b}z^2e^{-Ib(1-z)}} \right], \end{aligned} \tag{9}$$

$$\begin{aligned} N(z) &= \left[ \frac{[p(z^2-1)B_0e^{-Ib(1-z)} + pz(z-1)B_1e^{-Ib(1-z)}][z^2 - (1-p)e^{-Ib(1-z)}]}{[z^2 - (1-p)e^{-Ib(1-z)}][(I - Iz + \mathbf{b})z^2] - p\mathbf{b}z^2e^{-Ib(1-z)}} \right. \\ &\left. + \frac{pe^{-Ib(1-z)}[(1-p)(z^2-1)B_0e^{-Ib(1-z)} + (1-p)z(z-1)B_1e^{-Ib(1-z)}]}{[z^2 - (1-p)e^{-Ib(1-z)}][(I - Iz + \mathbf{b})z^2] - p\mathbf{b}z^2e^{-Ib(1-z)}} \right]. \end{aligned} \tag{10}$$

Hence, adding (9) and (10) we have

$$P(z) = B(z) + N(z). \tag{11}$$

Now we have to determine the unknown probabilities  $B_0$  and  $B_1$  which appear in the numerators of the right hand sides of equations (9), (10) and (11). For this purpose we use Rouché's theorem as follows. Let

$$\begin{aligned} f(z) &= [z^2 - (1-p)e^{-Ib(1-z)}][(I - Iz + \mathbf{b})z^2], \\ g(z) &= -p\mathbf{b}z^2e^{-Ib(1-z)}. \end{aligned}$$

Note that both  $f(z)$  and  $g(z)$  are regular on and inside  $|z|=1$ . We aim to prove that  $|f(z)| \geq |g(z)|$  on  $|z|=1$ . Now, on  $|z|=1$ ,

$$\begin{aligned} |f(z)| &= \left| \left[ z^2 - (1-p)e^{-Ib(1-z)} \right] \left[ (I - Iz + \mathbf{b})z^2 \right] \right| \\ &= \left| z^2 - (1-p)e^{-Ib(1-z)} \right| \left| (I - Iz + \mathbf{b})z^2 \right| \\ &\geq \left[ |z^2| - |(1-p)e^{-Ib(1-z)}| \right] \left[ |(\lambda - \lambda z + \beta)| |z^2| \right] \\ &= \left[ 1 - (1-p)e^{-Ib(1-z)} \right] \left[ I - I + \mathbf{b} \right] = p\mathbf{b}e^{-Ib(1-z)} = |g(z)|. \end{aligned}$$

Because  $|f(z)| \geq |g(z)|$ , therefore by Rouché's theorem  $f(z) + g(z)$  has the same of zeros inside or on  $|z|=1$  as that of  $f(z)$ . Now, it is easy to show that  $f(z)$  has four zeros on or inside  $|z|=1$ . Therefore,  $f(z) + g(z)$ , i.e., the denominator of the right hand side of (11) has four zeros on or inside  $|z|=1$ . For each of these four zeros the numerator of the right hand side of (11) must vanish, thus giving us four linear equations in the two unknowns  $B_0$  and  $B_1$ . Then two of these four equations are sufficient to determine the two unknowns, whereas the other two may just be redundant. Hence, the probability generating functions  $B(z)$ ,  $N(z)$  and  $P(z)$  obtained in (9), (10) and (11) can be completely determined.

Next, we shall use normalizing condition

$$P(1) = B(1) + N(1) = 1. \quad (12)$$

At  $z=1$ ,  $P(1) = \frac{\text{zero}}{\text{zero}}$ , therefore using L'Hopital's rule we have from (11)

$$P(1) = \lim_{z \rightarrow 1} P(z) = \frac{(p + \mathbf{b})(2B_0 + B_1)}{2\mathbf{b} - I\mathbf{b}\mathbf{b} - pI} = 1, \quad (13)$$

which gives

$$(p + \mathbf{b})(2B_0 + B_1) = 2\mathbf{b} - I\mathbf{b}\mathbf{b} - pI. \quad (14)$$

Equation (14) will hold only if  $2\mathbf{b} - I\mathbf{b}\mathbf{b} - pI > 0$  which gives the steady state condition

$$\frac{I(\mathbf{b}\mathbf{b} + p)}{2\mathbf{b}} < 1. \quad (15)$$

Note that when there are no server vacations, then with  $p=0$ ,  $N_n=0$  for all  $n \geq 0$ , equation (10) yields  $N(z)=0$  as it should be. Further, equations (9) and (15) respectively give

$$B(z) = \frac{\left[ (z^2 - 1)B_0 e^{-Ib(1-z)} + z(z-1)B_1 e^{-Ib(1-z)} \right]}{\left[ z^2 - e^{-Ib(1-z)} \right]}, \quad (16)$$

$$I\mathbf{b} < 2. \quad (17)$$

Note that (16) and (17) are the known results for the M/D/c queue for c=2. (See Kashyap and Chaudhury, 1988, p. 60-61.)

**Model B: Only One Server At A Time Can Take A Vacation**

The Underlying Assumptions.

In this case, the assumptions (a), (b) and (d) in section 2.1 for the previous case are the same. However, assumption (c) is different in this case under which we assume that after every service completion, only one server may take a vacation with probability  $p$  or continue to stay in the system with probability  $1-p$ . The vacation times follow an exponential distribution with mean vacation time  $1/b$ , ( $b > 0$ ).

**Definitions and Notations**

We define:

$B_n(t)$ : as the probability that at time  $t$  both servers are available in the system providing service and there were  $n (\geq 0)$  customers in the system when

the current service started.

$O_n(t)$ : as the probability that at time  $t$  only one server is available in the system providing service and there were  $n (\geq 0)$  customers when the current service started.

$P_n(t)$ : as the probability that at time  $t$  there are  $n (\geq 0)$  customers in the system regardless of the state of the servers.

**Steady State Forward Equations of the System**

Assuming that steady state exists, we let

$$\lim_{t \rightarrow \infty} B_n(t) = B_n, \quad \lim_{t \rightarrow \infty} O_n(t) = O_n \quad \text{and}$$

$\lim_{t \rightarrow \infty} P_n(t) = P_n$ . Thus  $B_n$ ,  $O_n$  and  $P_n$  denote the corresponding steady state probabilities. Then we obtain the following set of steady state equations:

$$B_n = (1-p)k_n [B_0 + B_1 + B_2] + (1-p) \sum_{i=3}^{n+2} B_i k_{n+2-i} + bO_n, \quad n \geq 0, \tag{18}$$

$$(1+b)O_n = O_0 k_n + O_1 k_n + \sum_{j=2}^{n+1} O_j k_{n+1-j} + p k_n [B_0 + B_1 + B_2] + p \sum_{i=3}^{n+2} B_i k_{n+2-i}, \quad n \geq 0. \tag{19}$$

**Steady State Probability Generating Functions for the System Size**

In addition to the probability generating functions defined in (4a) and (4b) in section 2.4, we define the following probability generation function:

$$O(z) = \sum_{n=0}^{\infty} O_n z^n, \quad |z| \leq 1. \tag{20}$$

We multiply both sides of equation (16) by  $z^{n+2}$  and add for all  $n = 0, 1, 2, \dots$ . Thus we have

$$\begin{aligned} z^2 \sum_{n=0}^{\infty} B_n z^n &= (1-p)B_0 z^2 \sum_{n=0}^{\infty} k_n z^n + (1-p)B_1 z^2 \sum_{n=0}^{\infty} k_n z^n + (1-p)B_2 z^2 \sum_{n=0}^{\infty} k_n z^n \\ &+ (1-p) \sum_{n=0}^{\infty} \sum_{i=3}^{n+2} B_i k_{n+2-i} z^{n+2} + b z^2 \sum_{n=0}^{\infty} O_n z^n, \quad n \geq 0. \end{aligned} \tag{21}$$

Then using (4a), (4b) and (20) we obtain from (21)

$$B(z) = \frac{\mathbf{b}z^2 O(z) + (1-p)B_0(z^2 - 1)e^{-1b(1-z)} + (1-p)B_1 z(z-1)e^{-1b(1-z)}}{z^2 - (1-p)e^{-1b(1-z)}}. \quad (22)$$

Similarly, we multiply both sides of (19) by  $z^{n+2}$  and add them for all  $n = 0, 1, 2, \dots$ . Then we have

$$\begin{aligned} (1+\mathbf{b})z^2 \sum_{n=0}^{\infty} O_n z^n &= O_0 z^2 \sum_{n=0}^{\infty} k_n z^n + O_1 z^2 \sum_{n=0}^{\infty} k_n z^n + \sum_{n=0}^{\infty} \sum_{j=2}^{n+1} O_j k_{n+1-j} z^{n+2} \\ &+ pB_0 z^2 \sum_{n=0}^{\infty} k_n z^n + pB_1 z^2 \sum_{n=0}^{\infty} k_n z^n + pB_2 z^2 \sum_{n=0}^{\infty} k_n z^n \\ &+ p \sum_{n=0}^{\infty} \sum_{i=3}^{n+2} B_i k_{n+2-i} z^{n+2}, \quad n \geq 0. \end{aligned} \quad (23)$$

Then using (4a), (4b) and (18) we obtain from (23)

$$O(z) = \frac{p\mathbf{B}(z)e^{-1b(1-z)} + pB_0(z^2 - 1)e^{-1b(1-z)} + pB_1 z(z-1)e^{-1b(1-z)} + z(z-1)O_0 e^{-1b(1-z)}}{(1+\mathbf{b})z^2 - ze^{-1b(1-z)}}. \quad (24)$$

Then, we solve equations (22) and (24) simultaneously for  $B(z)$  and  $O(z)$  and obtain

$$\begin{aligned} B(z) &= \left[ \frac{[(1-p)(z^2 - 1)B_0 e^{-1b(1-z)} + (1-p)z(z-1)B_1 e^{-1b(1-z)}][(1+\mathbf{b})z^2 - ze^{-1b(1-z)}]}{[z^2 - (1-p)e^{-1b(1-z)}][(1+\mathbf{b})z^2 - ze^{-1b(1-z)}] - p\mathbf{b}z^2 e^{-1b(1-z)}} \right. \\ &\left. + \frac{\mathbf{b}z^2 [p(z^2 - 1)B_0 e^{-1b(1-z)} + pz(z-1)B_1 e^{-1b(1-z)} + z(z-1)O_0 e^{-1b(1-z)}]}{[z^2 - (1-p)e^{-1b(1-z)}][(1+\mathbf{b})z^2 - ze^{-1b(1-z)}] - p\mathbf{b}z^2 e^{-1b(1-z)}} \right], \end{aligned} \quad (25)$$

$$\begin{aligned} O(z) &= \left[ \frac{[p(z^2 - 1)B_0 e^{-1b(1-z)} + pz(z-1)B_1 e^{-1b(1-z)} + z(z-1)O_0 e^{-1b(1-z)}][z^2 - (1-p)e^{-1b(1-z)}]}{[z^2 - (1-p)e^{-1b(1-z)}][(1+\mathbf{b})z^2 - ze^{-1b(1-z)}] - p\mathbf{b}z^2 e^{-1b(1-z)}} \right. \\ &\left. + \frac{pe^{-1b(1-z)} [(1-p)(z^2 - 1)B_0 e^{-1b(1-z)} + (1-p)z(z-1)B_1 e^{-1b(1-z)}]}{[z^2 - (1-p)e^{-1b(1-z)}][(1+\mathbf{b})z^2 - ze^{-1b(1-z)}] - p\mathbf{b}z^2 e^{-1b(1-z)}} \right]. \end{aligned} \quad (26)$$

Then adding (25) and (26), we obtain

$$P(z) = B(z) + O(z). \quad (27)$$

The unknowns  $B_0$ ,  $B_1$  and  $O_0$  can be determined by applying Rouché's theorem as before. Hence, the probability generating functions  $B(z)$ ,  $O(z)$  and  $P(z)$  can be completely determined.

Further, we use the normalizing condition

$$P(1) = B(1) + O(1) = 1. \quad (28)$$

At  $z = 1$ , because  $P(1) = \frac{\text{zero}}{\text{zero}}$ , and hence, using L'Hopital's rule we have from (27)

$$P(1) = \lim_{z \rightarrow 1} P(z) = \frac{(p + \mathbf{b})(2B_0 + B_1 + O_0)}{2\mathbf{b} - \mathbf{1}b\mathbf{b} + p - p\mathbf{1}b} = 1, \quad (29)$$

which gives

$$(p + \mathbf{b})(2B_0 + B_1 + O_0) = 2\mathbf{b} - \mathbf{1}b\mathbf{b} + p - p\mathbf{1}b. \quad (30)$$

Equation (30) will hold only if  $2\mathbf{b} - \mathbf{1}b\mathbf{b} + p - p\mathbf{1}b > 0$  which yields the steady state condition

$$\frac{\mathbf{1}b(\mathbf{b} + p)}{2\mathbf{b} + p} < 1. \quad (31)$$

Again note that when there are no server vacations, then with  $p = 0$  and  $O_n = 0$  for all  $n \geq 0$ , equation (26) yields  $O(z) = 0$  as it should be. Further, (25) and (31) respectively give

$$B(z) = \left[ \frac{[(z^2 - 1)B_0 e^{-\mathbf{1}b(1-z)} + z(z-1)B_1 e^{-\mathbf{1}b(1-z)}]}{[z^2 - e^{-\mathbf{1}b(1-z)}]} \right], \quad (32)$$

$$\mathbf{1}b < 2. \quad (33)$$

Note that (32) and (33) are the same known results for the M/D/c queue for  $c=2$  as in section 2.4. (See Kashyap & Chaudhury, 1988, p. 60-61.)

#### References

- Baba, Y. (1986): On the queue with vacation time. *Operations Research Letters*, 5, 93-98.
- Borthakur, A. & Choudhury, G. (1997). On a batch arrival Poisson queue with generalized vacation. *Sankhya Ser. B.*, 59, 369-383.
- Choi, B. D., & Park, K.K. (1990): The M/G/1 retrial queue with Bernoulli schedules, *Queueing Systems*, 7, 219-228.
- Cramer, M. (1989). Stationary distribution in a queueing system with vacation times and limited service, *Queueing Systems Theory and Applications*, 4, 1, 57-68.
- Doshi, B.T. (1986). Queueing systems with vacations - a survey. *Queueing Systems*, 1, 29-66.
- Doshi, B.T. (1990). Conditional and Unconditional distributions for the M/G/1 queue with vacations. *Questa*, 7, 229-252.
- Gross, C., & Harris, C.M. (1985): *The fundamentals of queueing theory*. NY: John Wiley and Sons.
- Kashyap, B. R. K., & Chaudhry, M. L. (1988). *An introduction to queueing theory*. Ontario, Canada: A and A Publications.

Keilson, J., & Servi, L.D. (1986): Oscillating random walk models for GI/G/1 vacation systems with Bernoulli schedules. *Journal of Applied Probability*, 23, 790-802.

Kleinrock, L. (1983). On the M/G/1 queue with rest periods and certain service independent queueing disciplines. *Operations Research*, 31(4), 705-719.

Madan, K. C. (1992). An M/G/1 queueing system with compulsory server vacations. *Trabajos de Investigacion Operativa*, 7(1), 105-115.

Madan, K. C. (1999). An M/G/1 queue with optional deterministic server vacations. *Metron*, LVII, 3-4, 83-95.

Madan, K.C. (2001). On a single server queue with two-stage heterogeneous service and deterministic server vacations. *Intern. J. Systems Science*, 32(7), 837-844.

Madan, K. C., & Saleh, M. F. (2001). On M/D/1 queue with general server vacations. *International Journal of Man. & Inform. Sc.*, 12(2), 25-37.

Madan, K. C. & Saleh, M. F. (2001). On Single server vacation with deterministic service or deterministic vacations. *Calcutta Statistical Association Bulletin*, 51, 203-204; 225-241.

Madan, K. C., & Saleh, M. F. (2001). On M/D/1 queue with deterministic vacations. *Systems Science*, 27(2), 107-118.

Medhi, J. (1982). *Stochastic Processes*. NY: Wiley Eastern.

## Homogeneous Markov Processes For Breast Cancer Analysis

Ricardo Ocaña-Riola    Emilio Sanchez-Cantalejo    Carmen Martinez-Garcia

Escuela Andaluza de Salud Pública  
Granada (Spain)

---

Sometimes, the introduction of covariates in stochastic processes is required to study their effect on disease history events. However these types of models increase the complexity of analysis, even for simpler processes, and standard software to analyse stochastic processes is limited. In this paper, a method for fitting homogeneous Markov models with covariates is proposed for analysing breast cancer data. Specific software for this purpose has been implemented.

Key words: Stochastic processes, Markov processes, cancer, covariates

---

### Introduction

Multi-state Markov processes have been introduced recently in health sciences in order to study the evolution of patients through different states or stages before death, even in cases where exact transition times are not known (Kay, 1986). This type of model has been mainly applied in AIDS (De Gruttola & Lagakos, 1989; Frydman, 1992; Mariotto et al., 1992), cancer (Kay, 1986), and psychiatric research (Keiding & Andersen, 1989), employing different methodologies depending on the particular conditions of each study. In practice, it is often useful to use a homogeneous Markov process to model disease history events because generally they are easy to interpret and the assumption that the process is homogeneous simplifies the methods used to fit the model.

In multivariate studies, the use of models that incorporate covariates allows analysis of the effect of these variables on the outcome variable. When multi-state models are used, it is also possible to study the effect of these covariates on different transitions between states throughout the patient's disease history.

Some authors have worked on the introduction of covariates in multi-state processes and particularly in homogeneous Markov processes (Kalbfleisch & Lawless, 1985; Pastorello, 1993); however, they mentioned the increased complexity of analysis in this sort of model where an added problem is the shortage of standard software. In spite of these problems, the introduction of covariates in stochastic processes is required to explain the effect of these factors on disease history events.

In this paper we present a breast cancer study where two transient states and a death state have been defined. In this study, observation is continuous, i.e., information on exact transition times between transient states is available; in this context, the main objectives of this paper are:

- a) To propose a method, computationally tractable, to estimate homogeneous Markov models with covariates in continuous time.
- b) To study the evolution of patients diagnosed with breast cancer in Granada province (South of Spain).

---

Correspondence should be sent to Ricardo Ocaña-Riola, Escuela Andaluza de Salud Pública, C/ Cuesta del Observatorio, 4 Apdo de Correos 2070 18080 Granada (Spain). E-mail: ricardo@easp.es. This research was developed at the *Escuela Andaluza de Salud Pública* and financed by grant number IN92-D24255738 from the *Programa Nacional de Formación de Personal Investigador en España* of the *Ministerio de Educación y Ciencia*. The authors would like to thank Dr. Jacques Estève and Angela Maldonado García.

### Methodology

The study was carried out with 241 women with breast cancer diagnosed in 1985-86 who received radical treatment and had a period free of symptoms. The follow-up ended on 31 of December 1990 (Ocaña-Riola, 2002). Data originated from the Granada Cancer Registry (South of Spain).

The variables T, N and Hormonal Status (HS) on the disease history of individuals have been recorded. The definition of T and N was taken from the Classification of Malignant Tumours (Sobin and Wittekind, 1997), where these variables are two components of the TNM system for describing the anatomical extent of disease. Variable M was not considered because there were no patients with distant metastasis. Additional numbers on TNM components indicates the extent of the malignant tumour as follows:

a) T: The extent of primary tumour; T0: No evidence of primary tumour; T1: Tumour 2 cm or less in greatest dimension; T2: Tumour more than 2 cm but not more than 5 cm in greatest dimension; T3: Tumour more than 5 cm in greatest dimension; T4: Tumour of any size with direct extension to chest wall or skin.

b) N: The absence or presence and extent of regional lymph node metastasis; N0: No regional lymph node metastasis; N1: Metastasis to movable ipsilateral axillary nodes(s); N2: Metastasis to ipsilateral axillary node(s) fixed to one another or to other structures; N3: Metastasis to ipsilateral internal mammary lymph node(s).

It is considered to be a three-state Markov model with two transient states and one absorbing (Chiang, 1968). These states are “With symptoms “ (state 1), “Without symptoms “ (state 2) and “Death “ (state 3) where the possible transitions are represented in Figure 1 in appendix.

We consider the transition intensity matrix:

$$Q(x) = \begin{pmatrix} -(q_{12}(x) + q_{13}(x)) & q_{12}(x) & q_{13}(x) \\ q_{21}(x) & -(q_{21}(x) + q_{23}(x)) & q_{23}(x) \\ 0 & 0 & 0 \end{pmatrix}$$

where each transition intensity is dependent on a vector of covariates; that is:

$$q_{ij}(x) = \exp(x \mathbf{b}_{ij}) \quad i \neq j$$

$$q_i(x) = \sum_{j \neq i} \exp(x \mathbf{b}_{ij}) \quad ,$$

where  $x = (x_0, \dots, x_b)$ ,  $x_0 = 1$ , is a vector of covariates and  $\mathbf{b}_{ij} = (\mathbf{b}_{ij0}, \dots, \mathbf{b}_{ijb})$  is a vector of unknown parameters.

In order to estimate the model an approximate method was used (Ocaña-Riola, 2002). The Likelihood Ratio Statistic (LRS) was used in a backward analysis to test the signification of regression parameters (De Groot, 1986). Moreover, the LRS test was used for the goodness of fit of the final model (Kalbfleisch & Lawless, 1985). When the transition intensity matrix is estimated, the estimated transition probability matrix is  $P(u; x) = \exp(Q(x)u)$ ,  $u > 0$ .

### Results

In order to estimate the model, we used a partition of the time using 35 intervals which extent was between 0.002 and 0.260 years (Figure 2 in appendix). Because of shortage of subjects in the groups N2 and N3 (Table 1), the variable N has been transformed in a binary variable as  $N_i = 0$  if  $N=0$  and  $N_i = 1$  if  $N=1, N=2$  or  $N=3$ .

There were not transitions from state 2 to state 3 in Non-menopause patients, however there are some in the Menopause group; if we interpret 2-3 as the transition to other causes of death, the transitions observed in Menopause group could be due to an age effect because older women heavily weight this group. For this reason we propose the following model:



$$q_{23} = \exp(\beta_{230} + \beta_{231} T_2 + \beta_{232} T_3 + \beta_{233} T_4 + \beta_{234} N_1) \text{ if } HS = 1$$

$$q_{23} = 0 \text{ if } HS = 0$$

$$q_{23} = \exp(\beta_{230} + \beta_{231} T_2 + \beta_{232} T_3 + \beta_{233} T_4 + \beta_{234} N_1) \text{ if } HS = 1$$

$$q_{23} = 0 \text{ if } HS = 0$$

where  $T_2, T_3, T_4$  are dummy variables from  $T$  ( $T_1$  is the category of reference).

A backward analysis using LRS test showed that variable N is not statistically significant when T and HS are into the model ( $P=0.482$ ). Besides, there is no evidence ( $P=0.370$ ) against the codification of T in only two categories: patients with a better prognosis (T1 or T2) and patients with a bad prognosis (T3 or T4). Therefore it was considered a new covariable, TR, with value 0 for T1 or T2 and value 1 for T3 or T4. The final model is shown in Table 2. MLE's for transition intensities in different groups of covariates are in Table 3.

Figures 3 and 4 show these transition probabilities by groups of covariates. These graphs show a notable difference between T1-T2 and T3-T4. A LRS test for the goodness of fit of the final model shows that there is no evidence against a homogeneous Markov process ( $p=0.177$ ).

### Conclusion

Multi-state Markov models offer some advantages over traditional survival models for studying disease history events, making it possible to estimate the probability that a subject could be in different states at any time in the future. Homogeneous processes are the simplest of Markov models but in some studies it is possible to find evidence against this sort of model. The absence of homogeneity in time could be the result of the absence of homogeneity between people. In this case, the use of covariates could improve the fit of the model and homogeneous Markov models with covariates are an interesting option.

However, the incorporation of covariates in a stochastic process increases the complexity of analysis, even on simple processes. Because of that and the shortage of standard software to analyse

Markov process with incomplete observations, many researchers refuse to use these multi-state models. In spite of these problems, some authors worked on the inclusion of covariates in a homogeneous Markov process (Andersen, 1988; Pastorello, 1993; Tuma & Robins, 1990). The more used methods are based on the extended Kalbfleisch and Lawless algorithm to incorporate covariates (Kalbfleisch & Lawless, 1985).

In this article we have used a particular partition of the time when observation is continuous. In this situation an approximate method has been proposed in order to introduce covariates and to estimate the intensity matrix in a homogeneous Markov process (Ocaña-Riola, 2002). MLE's obtained from this method are not computationally costly and, in practice, the algorithm converges to very similar estimates of parameters given by other methods when the length of the intervals  $u_k$  tends to be small (Ocaña-Riola, 2002). Moreover, covariates can easily be introduced in the model.

The method proposed here consider only categorical covariates because this is the sort of variables analysed in the breast cancer study. Continuous covariables, as age, could be introduced in the analysis using different categories for them. This idea has been used in some research about stochastic processes and in practice it is the most used (Tuma & Robins, 1980; Pastorello, 1993).

In this breast cancer study, incorporation of variables T, N and Hormonal Status in the model have allowed us to evaluate its effects on disease history. However, covariate information was missing for 36 women not included in the analysis. In general, it is not a good statistical practice to leave out patients with missing values; therefore different statistical methods have been published recently in order to incorporate these patients into the analysis. Some authors have shown that using a Bayesian approach implemented via Markov Chain Monte Carlo it is possible to obtain a suitable regression model for the missing values (Raghuathan & Siscovick, 1996).

Due to the complexity of the Bayesian analysis in a Markov process with covariates, we have not implemented this method. However, it would an interesting research into stochastic processes.

Along these lines, Volinsky et al. (1997) applied Bayesian Model Averaging to the selection of variables in Cox proportional hazard models.

Their investigations into the risk factors for strokes using this model improve the results obtained by traditional stepwise, forward and backward selection methods, which have poor properties (Miller, 1990). Again, the implementation of Bayesian Statistics into Markov processes could yield interesting results, although some theoretical research is needed before using these methods in practice.

In a traditional backward analysis a relationship was found between T and Hormonal Status and the evolution of patients diagnosed with breast cancer. Non-menopausal women with a tumour T1 or T2 have the best prognosis since recurrence probability and death probability are the smallest. In the same way and using traditional survival models, other population base studies have found that both, T and Hormonal Status, are important factors in order to predict survival and recurrence probability in breast cancer (Coebergh et al., 1995). Other analytic factors and hormonal data, not included in this study, could explain to a great extent part of breast cancer survival and recurrence. Vascular and lymphatic invasion of cancer cells, type of histology, age, site of first recurrence, female sex steroid receptors and ploidy measurements have been reported in some articles as prognostic factors for breast cancer recurrence (Blanco G et al., 1990; Murayama et al., 1986). In this way, a prospective study could be interesting in order to analyse the effect of all these variables on the evolution of patients through different states of their disease, obtaining a complete and detailed study on breast cancer history.

In this study, the interpretation of the transition from "without symptoms" to "death" is difficult in menopausal women. Older women heavily weight this group and perhaps the effect of age can explain this situation. It might be interesting to consider a fourth state "death from other causes" in order to know the proportion of patients dying from the direct or indirect consequences of breast cancer but unfortunately this information is rarely available in the Granada Cancer Registry.

From this paper's findings, it will be possible to estimate the proportions of patients who shall be in each disease state in the future; therefore we will be able to obtain highly relevant information for health planning services. Furthermore, the proposed method can easily be used for other situations in cancer and other disciplines such as

public health, economics, sociological research or medical sciences.

#### References

- Andersen, P. K. (1988). Multistate models in survival analysis: A study of nephropathy and mortality in diabetes. *Stat. Med.*, 7, 661-670.
- Blanco, G., Holli, K., Heikkinen, M., Kallioniemi, O. P., & Taskinen P. (1990). Prognostic factors in recurrent breast cancer. *Br. J. Cancer*, 62, 142-146.
- Chiang, C. L. (1968). *An introduction to stochastic processes and their applications*. New York: John Wiley & Sons.
- Coebergh, J. W., Van der Heijden, L. H., & Janssen, M. L. (1995). *Cancer incidence and survival in the southeast of the Netherlands, 1955-1994*. Eindhoven: IKZ.
- De Groot, M. H. (1986). *Probability and Statistics*. New York: Addison-Wesley.
- De Gruttola, V., & Lagakos, S. W. (1989). Analysis of doubly-censored survival data, with application to AIDS. *Biometrics*, 45, 1-11.
- Frydman, H. (1992). A nonparametric estimation procedure for a periodically observed three-state Markov process, with application to AIDS. *J. Roy. Stat. Soc. B Met.*, 54, 853-866.
- Kalbfleisch, J. D., & Lawless, J. F. (1985). The analysis of panel data under a Markov assumption. *J. Am. Stat. Assoc.*, 80, 863-871.
- Kay, R. (1986). A Markov model for analysing cancer markers and disease states in survival studies. *Biometrics*, 42, 855-865.
- Keiding, N., & Andersen, P. K. (1989). Nonparametric estimation of transition intensities and transition probabilities: a case study of a two-state Markov process. *Appl. Statist.*, 38, 319-329.
- Mariotto, A. B., Mariotti, S., Pezzotti, P., Rezza, G., & Verdecchia, A. (1992). Estimation of the Acquired Immunodeficiency Syndrome incubation period in intravenous drug users: A comparison with male homosexual. *Am. J. Epidemiol.*, 135, 428-437.
- Murayama, Y., Mishima, Y., & Ogimura, H. (1986). Determination of discriminatory power of prognostic factors for recurrence of breast cancer. *Cancer Detect. Prev.*, 9, 449-453.
- Miller, A. J. (1990). *Subset selection in regression*. London: Chapman and Hall.
- Ocaña-Riola, R. (2002) Two methods to estimate homogeneous Markov processes. *Journal of Modern Applied Statistical Methods*, 1, 131-138.

Pastorello, S. (1993). La mobilità nel mercato del lavoro: un'analisi econometrica con osservazioni in tempo discreto. *Statistica*, 53, 185-206.

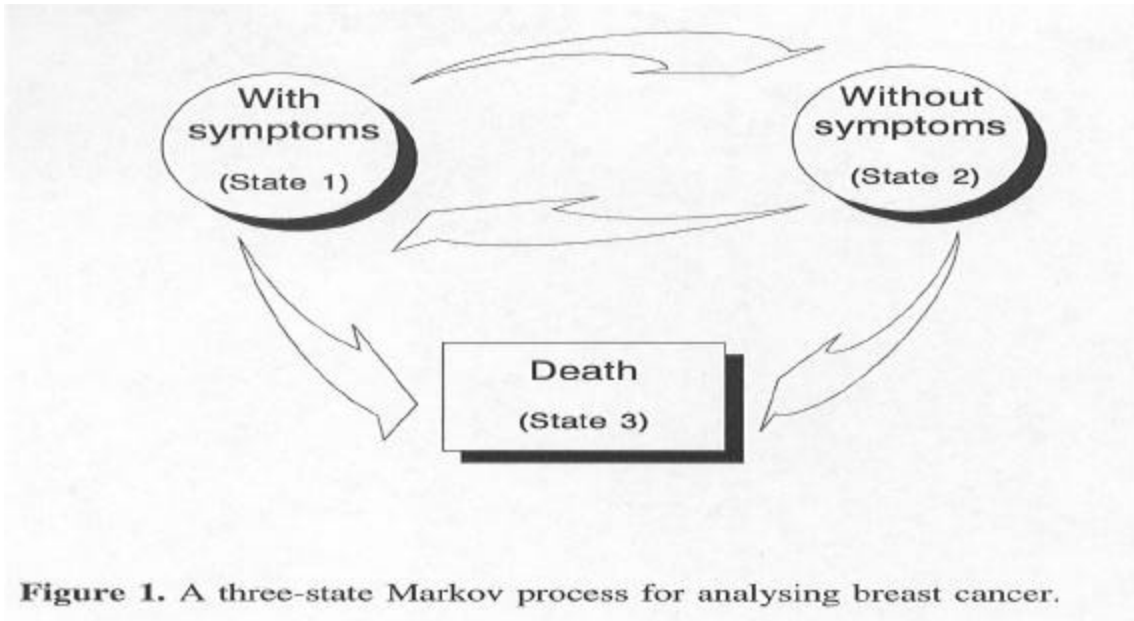
Raghunathan, T. E., & Siscovick, D. S. (1996). A multiple-imputation analysis of a case-control study of the risk of primary cardiac arrest among pharmacologically treated hypertensives. *Appl. Statist.*, 45, 335-352.

Tuma, N. B., & Robins, P. K. (1980). A dynamic model of employment behavior: an application to the Seattle and Denver income maintenance experiments. *Econometrica*, 48, 1031-1052.

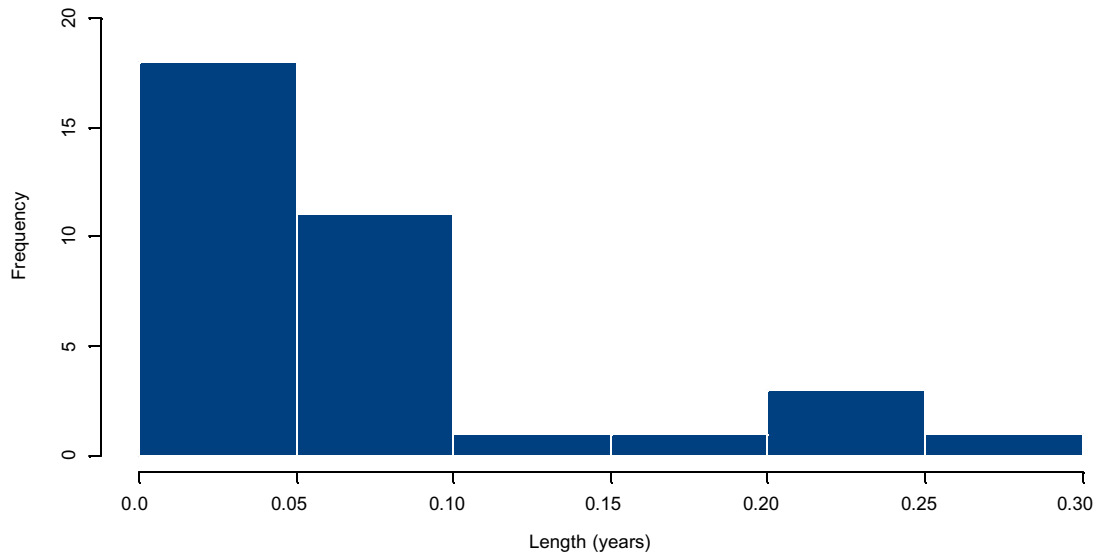
Sobin, L. H., & Wittekind, C. H. (1997). *TNM Classification of Malignant Tumours*. New York: John Wiley & Sons.

Volinsky, C. T., Madigan, D., Raftery, A. E., & Kronmal, R. A. (1997) Bayesian Model Averaging in proportional hazard models: assessing the risk of a stroke. *Appl. Statist.*, 46, 433-448.

## Appendix



**Figure 1.** A three-state Markov process for analysing breast cancer.



**Figure 2.** Length of the intervals that give a partition of the follow-up time

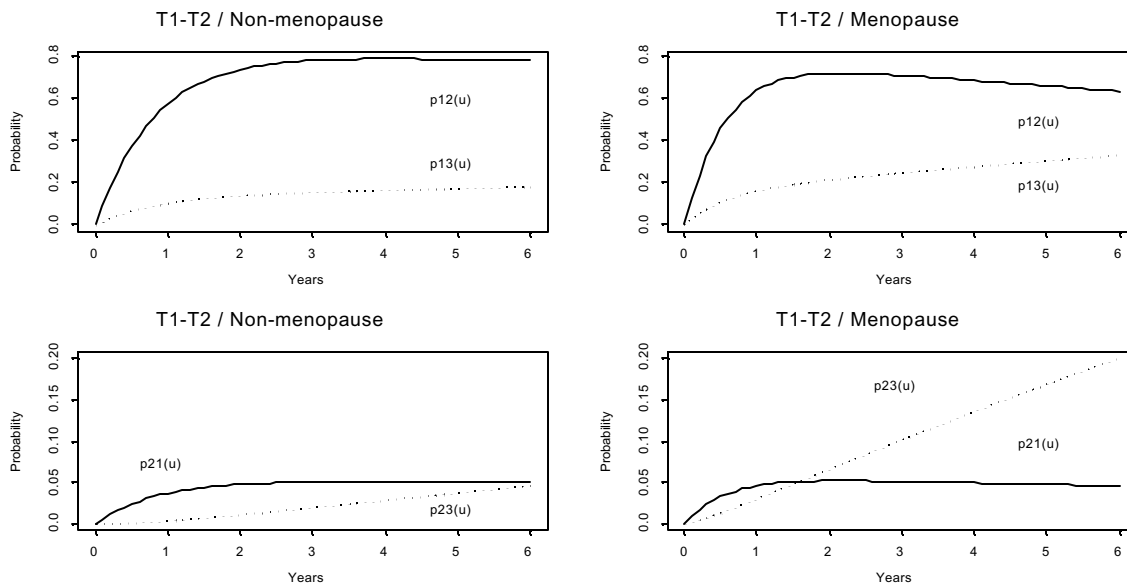


Figure 3. Estimated transition probabilities for T1-T2 and hormonal status

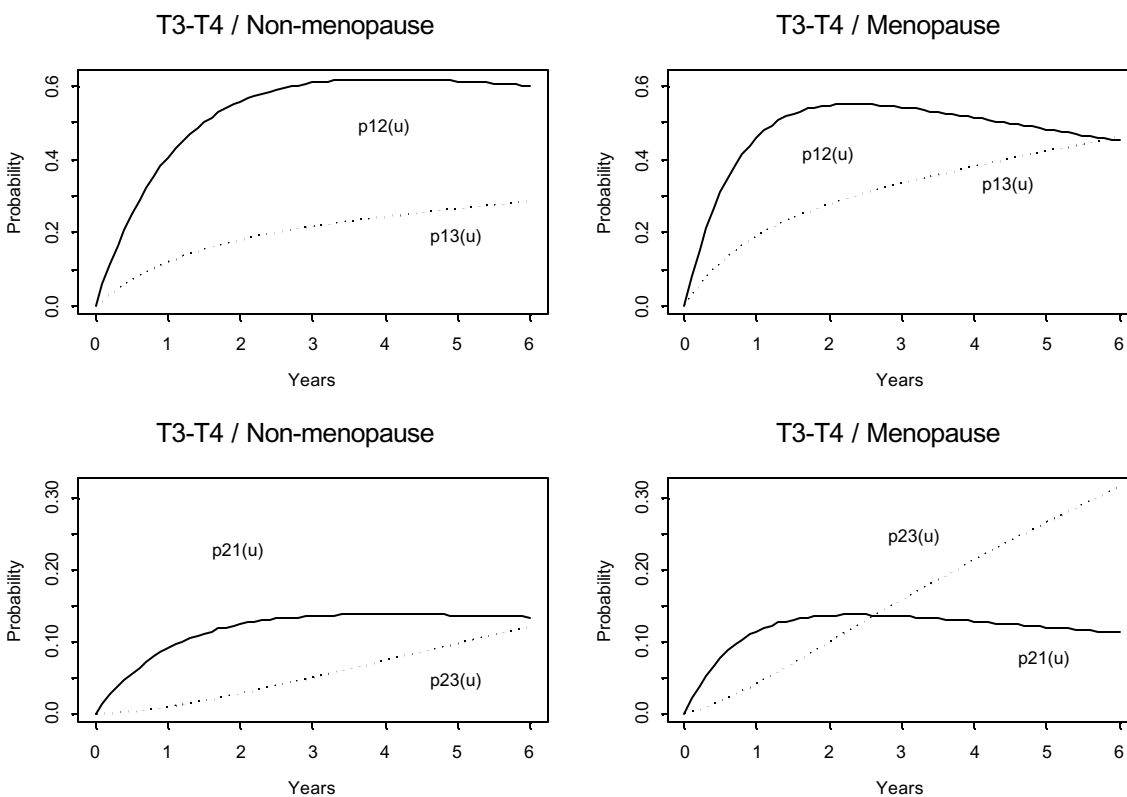


Figure 4. Estimated transition probabilities for T3-T4 and hormonal status

Table 1. Breast cancer data. Granada Cancer Registry, 1985-1986.

	Non-menopause					Menopause				
	N0	N1	N2	N3	Total	N0	N1	N2	N3	Total
T1	15	8	0	0	23	29	6	0	0	35
T2	20	7	1	0	28	41	21	3	0	65
T3	1	5	1	0	7	7	4	1	2	14
T4	2	7	0	0	9	4	15	2	3	24
Total	38	27	2	0	67	81	46	6	5	138

Note: There were 36 patients with missing values.

Table 2. MLE's estimates for breast cancer data (standard error in brackets)

Transition ( $ij$ )	Constant ( $\mathbf{b}_{ij0}$ )	TR ( $\mathbf{b}_{ij1}$ )	Hormonal Status ( $\mathbf{b}_{ij2}$ )
1 - 2	*	-0.4665 (0.0108)	0.3321 (0.0067)
1 - 3	-1.7584 (0.0203)	*	0.5802 (0.0235)
2 - 1	-2.7298 (0.0169)	0.7570 (0.0166)	0.4442 (0.0183)
2 - 3	-3.7965 (0.0219)	*	No included

(\*) Null statistical significance for  $\mathbf{a} = 0.05$

Table 3. Estimated transition intensities for breast cancer data.

T	Hormonal Status	$\hat{q}_{12}$	$\hat{q}_{13}$	$\hat{q}_{21}$	$\hat{q}_{23}$
T1 or T2	Non-menopause	1.0000	0.1723	0.0652	0
T1 or T2	Menopause	1.3939	0.3078	0.1017	0.0224
T3 or T4	Non-menopause	0.6272	0.1723	0.1391	0
T3 or T4	Menopause	0.8742	0.3078	0.2168	0.0224

## *Invited Debate: Target Article* You Think You've Got Trivials?

Shlomo S. Sawilowsky  
Educational Evaluation & Research  
Wayne State University University

---

Effect sizes are important for power analysis and meta-analysis. This has led to a debate on reporting effect sizes for studies that are not statistically significant. Contrary and supportive evidence has been offered on the basis of Monte Carlo methods. In this article, clarifications are given regarding what should be simulated to determine the possible effects of piecemeal publishing trivial effect sizes.

Key words: Trivial effect sizes, meta-analysis, Monte Carlo, simulation, Monte Carlo simulation

---

### Introduction

“It would seem that power analysis has arrived” (Cohen, 1988, p. xiii). This was the conclusion of the late Jacob Cohen in reviewing twenty-six years of the literature since he brought the importance of effect size (and sample size) to the attention of behavioral and social science researchers (Cohen, 1962). The explosion of meta-analyses being published, which followed Gene Glass’ presidential address to the American Educational Research Association (AERA) in April of 1976, also depends on the proliferation of effect sizes.

Researchers and editors, after neglecting power analyses in the past, or to provide raw materials for future meta-analyses, are now being asked to report effect sizes associated with statistically non-significant results. A recent motivating example of this call was made by Thompson (1996, 1999), who recommended effect sizes “can and should be reported and interpreted in all studies, regardless of whether or not statistical tests are reported” (1996, p. 29), and “even [for] non-statistically significant effects” (1999, p. 67).

Robinson and Levin (1997; see also Levin & Robinson, 1999) gave a reasoned approach to the reporting of effect sizes. On the basis of a thought experiment, they concluded that it is better to “First convince us that a finding is not due to chance, and only then, assess how impressive it is” (p. 23). Knapp and Sawilowsky (2001) added additional heuristic arguments against the practice.

Sawilowsky and Yoon (2001, 2002) conducted a Monte Carlo simulation to provide rigor for this position. Their results indicated that “effect sizes should not be reported or interpreted in the absence of statistical significance” (Sawilowsky & Yoon, 2002, p. 144). In contrast, Roberts and Henson’s (2002) Monte Carlo study came to the opposite conclusion. The purpose of this paper is to bring resolution to these opposing results.

### High Quality Monte Carlo Simulation & Sampling With Replacement

It is necessary to preface with a brief discussion of (a) simulation, (b) Monte Carlo, (c) Monte Carlo simulation, (d) sampling *with* vs *without* replacement, and (e) characteristics of a high quality Monte Carlo simulation. This will clarify the study conducted by Sawilowsky and Yoon (2001, 2002), and explicate the flaws in the design and conclusion of the study conducted by Roberts and Henson (2002). It will also serve as a brief review of Monte Carlo simulation methods. (For more complete coverage of the Monte Carlo simulation method, see Sawilowsky & Fahome, 2003).

---

Shlomo S. Sawilowsky is Professor of Educational Evaluation and Research (EER), College of Education, Wayne State University, Detroit, MI. He is the program coordinator of (EER), and Wayne State University Distinguished Faculty Fellow. Email: shlomo@wayne.edu. The title of this article is based on Gerrold (1973).

### Simulation

A simulation “mimics important elements” (Roberts, et. al, 1983, p. xi) of a system or phenomenon. It is “a representation ...in simplified form to study its behavior” (p. 452). Negoita and Ralescu (1987) noted that “In science... ‘simulation’ is forming an abstract model from a real situation in order to understand the impact of modifications and the effect of introducing various” (p. 29) interventions.

Norlén (1975) stated that simulation can be viewed as a “numerical technique for the carrying out of experiments” (p. 15). As an example, consider simulating the tossing of a fair die. This may be accomplished by accessing an uniform pseudo-random number generator that produces a value on the interval [0,1]. Draw a variate from the generator. Suppose it is .1770 (rounding to four significant digits, or to as many significant digits as desired). Using the assignment in Table 1 below, this process results in the simulation of throwing a fair die and having two spots surface.

Table 1. Simulation of a fair die using uniform variates on the interval [0,1].

Outcome	Assignment
.0000 - .1666	1 spot
.1667 - .3333	2 spots
.3334 - .5000	3 spots
.5001 - .6666	4 spots
.6667 - .8333	5 spots
.8334 - 1.000	6 spots

### Monte Carlo

Monte Carlo, in the sense it is being used in this article, is of rather recent origin (Metropolis & Ulam, 1949). Its usage appeared over a half century ago in reference to the gaming establishments of previous centuries of a famous city in the Monaco principality. It is an explicit reference to the use of *repetition* as a method of discovery of the long run outcome of an event.

More technically, it is the “use of stochastic techniques to solve... a deterministic problem” (Moshman, 1967, p. 250). As such, “one of the simplest and most direct applications of the Monte Carlo methods is to the evaluation of integrals” (Kahn, 1966, p. 249-250), or the area of any geometric figure, but particularly those *irregular* in shape. (The first moment of the uniform distribution over the interval [0,1] can be obtained via the calculus:

$$\int_0^1 x dx = .5 .$$

This result could be estimated via Monte Carlo methods by drawing a large number of variates from a uniform pseudo-random number generator and computing the mean, but *usually* there is little point in doing so.)

As an example, consider the problem of determining the area of an irregular closed figure that is unwieldy to the calculus. Inscribe the figure within a unit square. Draw two variates from the uniform pseudo-random number generator to represent Cartesian coordinates for the ordered pair (x, y), and plot them accordingly. Repeat the previous step many times. The area of the irregular geometric figure is estimated (as accurately as desired) by the ratio of the number of dots that fall within the figure, divided by the total number of repetitions (i. e., pairs of dots created). Note, however, that no system or phenomenon was simulated.

A famous example of the Monte Carlo method was undertaken in 1908 by William Sealy Gosset (Student, 1908a, 1908b), a chemist working for the Guinness brewing company. He bolstered his analytical expression of the distribution of the Pearson product-moment correlation coefficient on small samples via a Monte Carlo conducted by hand. Similarly, he supported the derivation of the t statistic with a Monte Carlo demonstration of the sampling distribution of t.

### Monte Carlo Simulation

Statistical historians (e.g., Hald, 1998, p. 196 - 201) noted that multinomial outcomes, such as tossing a fair die with equiprobability of one through six spots surfacing, was determined mathematically by Laplace in 1774. As an



alternative to the mathematical approach, the Monte Carlo simulation approach arose with Buffon in 1777, who tossed a coin 2,048 times and recorded the results. The distribution of outcomes indicated an expectation of heads to occur in 50.693% of the tosses. In 1837, Poisson determined  $0.48468 < p < 0.52918$  to be what he called the 99.555% interval of the probability “p” representing the chance of a heads occurring.

A famous Monte Carlo simulation was reported in 1900 by the eugenicist, Karl Pearson. His zoologist colleague and co-founder of *Biometrika*, Walter Frank Raphael Weldon, tossed twelve dice at the same time, recorded the results, and repeated the process 26,306 times. Pearson (1900) procured this data set and applied his newly developed goodness of fit  $\chi^2$  test to demonstrate the frequency of obtained outcomes were as expected due to combinatorial analysis.

Norlén noted (1975) ‘the advent and use of computers... freed the method from manual calculations... and... afford richer possibilities for the creation of complex, dynamic, and multivariate’ (p. 20) problems. Thus, the modern Monte Carlo simulation obviates the physical tossing of a die (or flipping of a coin). The combination of assignment in Table 1 (simulation) with many repetitions (Monte Carlo) via computer software and hardware results in the Monte Carlo simulation of the probability of outcomes in tossing a fair die with far more accuracy than could be achieved with the manual methods used by Buffon or Weldon.

The richness of possibilities for Monte Carlo simulation are truly amazing. Some examples include annealing, electromagnetism, image processing, and genetic linkage (Robert & Casella, 1999); inventory control, queuing systems at a two-minute car wash, expected waiting times, management planning, short-term forecasting, consumer behavior of switching brands, and customer product ordering behavior, (McMillan & Gonzalez, 1968); mass-supply systems, and quality and reliability of products (Sobol, 1974); growth of yeast in a sugar solution, cooling temperature of coffee, development of ability to perform pushups, estimating migration patterns, material or time delays, ecology of the Kaibab Plateau on the rim of the Grand Canyon, urban growth, sale and consumption of commodities, controlling dam water, projection of discovery of

natural gas reserves, and heroin addiction’s impact on a community (Roberts et. al, 1983); and studying random neutron diffusion in fissile material in the development of the atom bomb during World War II.

#### Sampling *With* vs *Without* Replacement

Sampling via Monte Carlo simulations can be conducted *with* or *without* replacement. In the examples using dice or coins, the correct sampling technique is *with* replacement. Once the result for the experiment has been recorded, the value obtained from the uniform pseudo-random number generator is returned to the repository of values that may again be drawn. This is because the spots don’t leave the dice after being tossed and the heads don’t leave the coin after being flipped.

Conversely, sampling *without* replacement would be appropriate in simulating the turning of cards. Once the Queen of Hearts has been turned, it is no longer in the deck, and cannot reappear. The Queen of Hearts must be prevented from further assignment. The choice of which technique to use in a Monte Carlo simulation is determined by what is being simulated.

The matter of sampling *with* vs *without* replacement is practically irrelevant when drawing variates from the continuous uniform distribution, which is represented by an infinite number of real numbers, each in turn with an infinite string of digits. Furthermore, this consideration is often moot with asymptotically large data sets. However, Monte Carlo simulation based on discrete and bounded distributions, and even more so with small sample data sets, may lead to different results based on which sampling technique is used.

#### Characteristics Of A High Quality Monte Carlo Simulation

There are a variety of factors that must be attended to in order to assure a Monte Carlo simulation is correct and useful. Some of these factors are as follows:

- the pseudo-random number generator has certain characteristics (e. g. a long “period” before repeating values)
- the pseudo-random number generator produces values that pass tests for randomness

- the number of repetitions of the experiment is sufficiently large to ensure accuracy of results
- the proper sampling technique is used
- the algorithm used is valid for what is being modeled
- the study simulates the phenomenon in question

Sawilowsky and Yoon (2001, 2002) vs Roberts and Henson (2002)

The Monte Carlo *simulation* by Sawilowsky and Yoon (2001, 2002) was conducted with

A Fortran 95 program “written to randomly draw variates from a de Moivreian (i. e., normal) distribution and then randomly assigned to two groups ( $n_1 = n_2 = 10$ ), with the first group designated the treatment group and the second the control. A two-sided two independent samples t test was conducted with nominal  $\alpha = 0.05$ . 10,000 repetitions were conducted. (p. 143).

Under the truth of the null hypothesis, the results indicated that the average of the absolute values of the effect size, Cohen’s  $d$ , was not near zero, but rather, was approximately what Cohen (1988) categorized as a small treatment effect. Thus, the conclusion of their brief report was the publishing of the constituent effects sizes would be misleading.

The Monte Carlo *study* by Roberts and Henson (2002) was designed to examine the “amount of bias in the effect size” (p. 241). They used an S-Plus macro to

generate two normally distributed populations of 1 million cases... the factors in this simulation study included the size of Cohen’s  $d$  in the population, the standard deviation of the two populations, and the sample sizes of the two groups... A total of 5,000 pairs of

samples were drawn from the populations within each condition of the simulation study. (p. 245)

The results of their study found “the amount of bias in  $d$  remained small under most conditions of consideration” (p. 247). Because the “average across samples tended to more closely approximate zero” under the truth of the null hypothesis, meaning “Cohen’s  $d$  does not appear to be biased in practical terms” (p. 252), they concluded the opposite of Sawilowsky and Yoon (2001). Therefore, they supported the reporting of effect sizes for results that are not statistically significant.

#### Criticism of Roberts and Henson’s (2002) Study Nine Minor Criticisms

(1) Roberts and Henson (2002) claimed that “effect sizes can serve a valuable function to help evaluate the magnitude of a difference or relationship” (p. 241). Although effect sizes do quantify the magnitude of a difference or relationship, they do not evaluate it. Content knowledge of the research question is required to decide if the difference or relationship is of theoretical, clinical, or practical importance.

(2) Their Monte Carlo study was written in a recent albeit dated version of S Plus, which is a superb statistical package. There are advantages of using statistical packages over programming languages, such as ease of use. There have been bugs, however, in this software’s pseudo-random number generator (e.g., see the discussion at [www.insightful.com/support/faqdetail.asp?FAQID=137&IsArchive=0](http://www.insightful.com/support/faqdetail.asp?FAQID=137&IsArchive=0)).

On the positive side, if a glitch due to this bug occurred it should have produced an observable error message. The built generator has an excellent period length (i. e.,  $2^{64} - 2^{32}$ ) compared with most other statistical packages, but the algorithm it is based on fails at least four DIEHARD tests of randomness (available at <http://stat.fsu.edu/~geo/>). The default option requires the programmer to reset the seed, which was not mentioned by Henson and Roberts (2002). Otherwise, the two “populations” of 1 million values would be identical. The current version of S-Plus eliminated these potential concerns.

(3) The entry of .0611 for the maximum  $r^2$  when  $d = .00$  and  $n_1 = n_2 = 10$  in Table 2 is obviously a typographical error.

(4) They presented “descriptive statistics” (p. 247), including the minimum and maximum  $d$ , in Tables 1 - 3. Roberts and Henson (2002) mistakenly labeled and considered the strongest negative effect size as a “minimum”. Although mathematically it is a “minimum”, in the context of effect sizes, the minimum  $d$  is, of course, defined as zero.

(5) Whereas Sawilowsky and Yoon (2001, 2002) used 10,000 replications and reported results to three significant digits, Roberts and Henson (2002) used 5,000 repetitions, but reported results to four significant digits. The number of repetitions was likely due to the limitations of using an S-Plus macro instead of Fortran, as the latter is far more flexible to program and faster in terms of execution. (It is not uncommon to use millions of repetitions to gain precision.)

(6) Roberts and Henson (2002) conducted their study on “5,000 pairs of samples” that “were drawn from the populations” ( p. 245). Thus, they used sampling *without* replacement. This is incorrect if the intent was to simulate the occurrence of test scores, group means, p values, or effect sizes. For example, the appearance of an IQ score of 107.5 as one sample mean should not preclude another sample from having the same mean. Each sample mean of a pair must be returned to the population, with the chance of being drawn again being equal to every other possible sample mean. This is accomplished by sampling *with* replacement.

(7) Because the study was conducted on Cohen’s  $d$  (and  $r^2$ ), which is a standardized value, there was no need for Roberts and Henson (2002) to include three different population standard deviations, and hence, two-thirds of their study (i. e., Tables 2 - 3) is redundant.

(8) There is little justification for publishing Monte Carlo work when results can be computed easily and directly. The bias in  $d$  can be computed analytically under population normality, which is the only distribution Roberts and Henson (2002) examined. Cohen (1988) noted:

It has been shown by Hedges (1981) and Kraemer (1983), in the context of the use of  $d$  in meta-

analysis that the absolute value of  $d_s$  is positively biased by a factor of approximately  $(4df - 1)/(4df - 4)$ , which is of little consequence except for small samples. (p. 66)

Their Monte Carlo results for the bias of Cohen’s  $d = .2, .5, \text{ and } .8$  in Table 2 for  $n_1 = n_2 = 10$  differ from  $(4df - 1)/(4df - 4)$  by only .005, -.014, and -.013, respectively. The results should converge as the number of repetitions in their Monte Carlo study increase.

(9) Roberts and Henson (2002) cited literature reviews indicating authors inadequately documented effect sizes. They cited editors who promoted citing effect sizes. They cited the same list of journals previously given by Thompson (2001, p. 83), whose editors require reporting of effect sizes. Their point is well taken, despite the apparent recanting of this form of persuasion by Thompson (2002), who cautioned “headcounts of views are not perfect indicators of truth” (p. 85). Nevertheless, Roberts and Henson’s (2002) Monte Carlo study did not present any compelling reason to report effect sizes *when the null hypothesis remains tenable*.

Major Criticism

Sawilowsky and Yoon (2001, 2002) never “argued that small effects can in some cases be due solely to sampling error” (Roberts & Henson, 2002, p. 245), as claimed by Roberts and Henson and which was the premise of their counter-study. Instead, Sawilowsky and Yoon (2002) demonstrated the trouble with reporting effect sizes for studies that were not statistically significant by simulating the process and examining the false impression that would subsequently be created in the literature. The following fabricated data sets (Data Set A and Data Set B) represent two possible patterns of results in terms of effect sizes when the null hypothesis is tenable.

Table 2. Hypothetical Effect Sizes (e. g., Cohen’s  $d$ ) For Data Sets A & B Over Six Replications.

A	.001	-.004	.003	.008	-.003	-.005
B	.23	.12	-.07	.17	-.27	-.17

To appreciate the impact of the information (hypothetical results) in the table above, consider the following vignette. First, consider Data Set A. Readers of the literature will see an effect size of .001 published in a study of interest, -.004 in the subsequent study, and so forth. If the reader has a good memory, it would be remembered that the typical positive effect size averaged .004, and the typical negative effect size averaged -.004. The sign of the effect size, to be discussed further below, depends on the context of the study. Prior to the sought-after and highly prized meta-analysis, what message will have formed in the mind of the reader of the literature? Most likely, there isn't much here.

Now consider Data Set B. The effect size for the first study was .23. Although marginally respectable, the study was published to publicize a subtle, yet detectable treatment effect in education or psychology research. A year later, a replication study appeared in the literature. The magnitude of the effect size was only .12. Explanations were given for the reduction (e. g., the reliability estimate was lower, the sampling plan was inadequate, the period of treatment was reduced). After another year passed and the next replication appeared in the literature, serious questions regarding the veracity of the intervention arose. This was because the effect size for the third non-statistically significant study was only -.07.

This impression dissipated somewhat with the appearance of the fourth study and its effect size of .17. After the fifth and six studies, however, readers of the literature were thoroughly confused on the effectiveness of the intervention. What message might be formed in their minds? A reader with a good memory may recall the magnitude of the effect sizes averaged approximately .2, indicating there was a small but important treatment effect. Readers who (a) recalled the oldest studies maintained the direction was positive, or (b) recalled the newest studies maintained the direction was negative.

When the readers are presented with the published meta-analysis on the series of non-statistically significant studies, they will realize they have been misled. In the absence of a Type I error, the meta-analytic synthesis will determine the studies conducted over the past half-decade are not statistically significant. The meta-analysis, and the misconceptions it clarified, would have been

obviated initially had effect sizes for non-statistically significant studies not been published in the first place.

The Sawilowsky and Yoon (2002) Monte Carlo was a simulation designed to determine which type of data set should readers of the literature expect to see under the truth of the null hypothesis. Are the magnitudes clustered about 0.0? The absolute value was taken, and it was determined that the typical magnitude expected is not near zero, but rather, what Cohen (1988) labels a small treatment outcome. Their simulation showed readers should expect to see results such as that depicted by Data Set B, not Data Set A. In contrast, Roberts and Henson's (2002) work was a Monte Carlo study of the bias of  $d$ , which does not relate to the process being simulated.

(Without remarking on it, Roberts and Henson, 2002, with slightly different study parameters, found the strongest effect sizes to be -2.31 and 2.06 for negatively and positively signed  $d$ 's, respectively. You think you've got trivials? These huge results occurred with a treatment modeled by random numbers! Publishing specious effect sizes of such astronomically high magnitude (i.e.,  $\pm 2.19$ ) could wreak havoc in the literature. Sawilowsky and Yoon, 2001, 2002, considered reporting results in this fashion. It was decided, however, that to be realistic, the simulation should depict the typical magnitude expected, not extrema.)

### Conclusion

Consider the chaotic fashion in which meta-analyses are currently being conducted. One researcher is not the holder of results from many tightly integrated experiments, publishing only the final meta-analysis. If that were the case, the presence of effect sizes for non-statistically significant results, duly noted and preserved as they occurred, would never become a misleading menace to the public.

Therefore, Sawilowsky and Yoon's (2001, 2002) brief report was based on taking the absolute value of Cohen's  $d$  to determine the typical magnitude expected when an intervention was random numbers. Roberts and Henson's (2002) argument against taking the absolute value was "in real experiments, it is known which group received the intervention" (p. 244). Is their

position correct as far as readers of the literature are concerned? In some treatment vs control studies, the effect of the treatment is demonstrated when the mean of the treatment group is *higher* than that of the control group; in other contexts when the mean of the treatment group is *lower* than that of the control group. For example, the *same* intervention might be used to increase self-esteem scores (treatment group mean is greater than control group mean), and reduce the number of times per week the bed was wet (treatment group mean is lower than control group mean).

The direction (sign of + or -) of that same intervention is entirely arbitrary. The sign depends on the context of the use of the intervention. If one researcher held all of the interim results, then the interpretation could safely rest on the meta-analysis, as the context would be known. However, the reader of the literature, who is getting these results piecemeal, will have the nigh impossible task of making sense of the contexts of a series of independently conducted studies published sporadically over time.

In addition to the above vignette, consider using a compound designed to block the serotonin uptake pump in a treatment one vs treatment two study on patients at risk for suicide. Suppose 30 mg, a common dosage for depression, was being compared with 70 mg, a common dosage for trichotillomania and other obsessive-compulsive disorders. Which dosage is the intervention? Clearly, the resulting direction (sign of + or -) is arbitrary. Thus, both the *magnitude* and the *sign* of published effect sizes for non-statistically significant studies mislead the public.

Cohen (1988) noted the researcher "hardly needs convincing of the centrality of the concept of effect size (ES) to the determination of power or necessary sample size in research design" (p. 531). "It is, after all, what science is all about" (p. 532). Yet, Cohen (1988, p. 10) opined that of all the factors in research design, behavioral scientists understand effect size the least. "Whatever the manner of representation of a phenomenon ... the null hypothesis always means the effect size is zero...[but] when the null hypothesis is false, it is false to some specific degree, i.e., *the effect size (ES) is some specific nonzero value in the population*" (Cohen, 1988, p. 10). Thompson (1996, 1999), supported by Roberts and Henson (2002), called for publishing specific nonzero

values under the truth of the null hypothesis. According to Cohen (1988), however, "the ES serves as an index of degree of departure *from* the null hypothesis" (p. 10, italics added for emphasis).

#### References

- Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology*, 65, 145-153.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. (2<sup>nd</sup> ed.) Hillsdale, NJ: Erlbaum.
- Gerrold, D. (1973). *The story behind a Star Trek show! "The trouble with tribbles": The birth, sale, and final production of one episode*. NY: Ballantine.
- Hald, A. (1998). *A history of mathematical statistics: From 1750 to 1930*. NY: John Wiley & Sons.
- Kahn, H. (1966). Multiple quadrature by Monte Carlo methods. In A. Ralston & H. S. Wilf, (Eds.) *Mathematical methods for digital computers*. Vol. 1, 249 - 257.
- Knapp, T. R., & Sawilowsky, S. S. (2001). Constructive criticisms of methodological and editorial practices. *The Journal of Experimental Education*, 70, 65-79.
- Levin, J. R., & Robinson, D. H. (1999). Further reflections on hypothesis testing and editorial policy for primary research journals. *Educational Psychology Review*, 11, 143-155.
- McMillan, C., & Gonzalez, R. F. (1968). *Systems analysis: A computer approach to decision models*. (Revised ed.). Homewood, IL: Irwin.
- Metropolis, N., & Ulam, S. (1949). The Monte Carlo method. *Journal of the American Statistical Association*, 44, 335-351.
- Moshman, J. (1967.) Random number generation. In A. Ralston & H. S. Wilf, (Eds.) *Mathematical methods for digital computers*. Vol. 2, 249 - 263.
- Negoita, C. V., & Ralescu, D. (1987). *Simulation: Knowledge-based computing, and fuzzy statistics*. NY: Van Nostrand Reinhold Co.
- Norlén, U. (1975). *Simulation model building: A statistical approach to modelling in the social sciences with the simulation method*. NY: John Wiley & Sons.

Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine*, 50, 157-175.

Roberts, J. K., & Henson, R. K. (2002). Correction for bias in estimating effect sizes. *Educational and Psychological Measurement*, 62, 241-253.

Roberts, C. P., & Casella, G. (1999). *Monte Carlo statistical methods*. NY: Springer-Verlag.

Roberts, N., Andersen, D., Deal, R., Garet, M., & Shaffer, W. (1983). *Introduction to computer simulation: A system dynamics modeling approach*. Portland, OR: Productivity Press NY: John Wiley & Sons, Inc.

Robinson, D. H., & Levin, J. R. (1997). Reflections on statistical and substantive significance, with a slice of replication. *Educational Researcher*, 26, 21-26.

Sawilowsky, S. S. (April, 1996). *Encyclopedia of educational and psychological effect sizes*. Annual Meeting of the American Educational Research Association, Division D, Measurement and Research Methodology, NY, NY.

Sawilowsky, S. S., & Fahoome, G. (2003). *Statistics through Monte Carlo simulation via Fortran*. Rochester Hills, MI: SS, Inc.

Sawilowsky, S. S., & Yoon, J. (2001). *The trouble with trivials ( $p > .05$ )*. Paper presented at the 53<sup>rd</sup> session of the International Statistical Institute, Seoul, South Korea.

Sawilowsky, S. S., & Yoon, J. (2002). The trouble with trivials ( $p > .05$ ). *Journal of Modern Applied Statistical Methods*, 1, 143-144.

Sobol, I. M. (1974). The Monte Carlo Method. (Translated and adapted from the second Russian edition by R. Messer, J. Slone, and P. Fortini. ERIC No. ED 184845.

Student. (1908a). On the error of counting a haemocytometer. *Biometrika*, 5, 351-360.

Student. (1908b). The probable error of a mean. *Biometrika*, 6, 1-25.

Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. *Educational Researcher*, 25, 26-30.

Thompson, B. (1999). Five methodology errors in educational research: A pantheon of statistical significance and other faux pas. In B. Thompson (Ed.), *Advances in social science methodology*, 5, 23-86.

Thompson, B. (2001). Significance, effects sizes, stepwise methods, and other issues: Strong arguments move the field. *Journal of Experimental Education*, 70, 80-93.

Thompson, B. (2002). "Statistical," "practical," and "clinical": How many kinds of significance do counselors need to consider? *Journal of Counseling and Development*, 80, 64-71.

*Invited Debate: Response*  
Not All Effects Are Created Equal: A Rejoinder To Sawilowsky

J. Kyle Roberts  
University of North Texas

Robin K. Henson  
University of North Texas

In the continuing debate over the use and utility of effect sizes, more discussion often helps to both clarify and syncretize methodological views. Here, further defense is given of Roberts & Henson (2002) in terms of measuring bias in Cohen's  $d$ , and a rejoinder to Sawilowsky (2003) is presented.

Key words: Effect size, Cohen's  $d$ , bias, simulation

### Introduction

Under a spirit of collegiality and zeal to further the field of research, dialogues like this play an important role in discussing areas where researchers both agree and disagree. Through open-ended dialogue, it is hoped that readers will continue to see the benefit in debate about important topics.

In this brief rejoinder to Sawilowsky (2003), we will provide discussion to the nine minor criticisms and one major criticism point by point. Although the first portion of his paper is lengthy, it does not bear comment on because it was expertly written and we do not disagree with any of the substance laid therein.

As we respond to each of the criticisms, however, we feel it important to note two things. First, the point of our paper was to show whether or not Cohen's  $d$  contains any amount of bias and is therefore in need of a correction to account for this bias.

Dr. Roberts is an Assistant Professor of educational research whose research interests include hierarchical linear modeling and measurement. Correspondence can be sent to [kroberts@unt.edu](mailto:kroberts@unt.edu). Dr. Henson is an Assistant Professor of educational research. His areas of research include applied statistics, measurement, reliability generalization, and self-efficacy theory. Correspondence can be sent to [rhenson@unt.edu](mailto:rhenson@unt.edu).

For all practical purposes, our answer to this question was NO. As we stated in our article, "the amount of bias in  $d$  remained small under most conditions of consideration . . . [and the] incredibly small amount of difference between the population  $d$  and the average sample  $d$  leads us to believe that  $d$  is in fact not biased in terms of practical differences" (p. 247, 251).

Second, we examined Thompson's (2002) proposed correction of  $d$  for accuracy and to see whether or not the correction was even necessary. In response to this proposed correction, we state, "although this correction of  $d$  seems to make sense theoretically, it overcorrects for the actual amount of bias" (p. 251).

As we begin our reply, we would like to note that NOWHERE in the rebuttal does Sawilowsky (2003) refute either of these findings. Instead, the arguments fall into two categories: minor criticisms that are mostly methodological, and one major criticism that has to do with the publishing of reported effect sizes. Once again, it bears mentioning that none of these criticisms, once having addressed and clarified the methodological issues, directly calls into suspect the findings of Roberts and Henson (2002).

### Responses to Minor Criticisms

#### Criticism 1: Effect sizes help evaluate

Although we agree with Sawilowsky's statement that effect sizes do not evaluate the effect of a difference or relationship, we want to note that we pointed out in our paper that the purpose of the effect size is to "*help* evaluate the magnitude of a difference" (emphasis ours, p. 241); for judgments are of course made by people.

As Sawilowsky (2003) quoted this very statement, we do not see any point of disagreement here.

$$d = \frac{200 - 225}{35} = -0.714. \quad (1)$$

#### Criticism 2: S-PLUS Random Number Generator

As Sawilowsky makes a good point about resetting the random number seed, it should be pointed out that this seed was reset for both populations so that they weren't identical. Concerning the random number generator (RNG) in S-PLUS, however, we feel that the criticisms are unwarranted. The DIEHARD tests for randomness were designed to work on RNGs that assume 32 random bits. The RNG for S-PLUS is 31 bit. As a result it should be assumed that the RNG will fail some of the tests that are 32 bit based. If there is a need for a 32 bit RNG, then S-PLUS users can install a patch that will paste together 16 bits from each of two consecutive numbers and then the S-PLUS RNG will pass all of the DIEHARD tests. Also, the bug which Sawilowsky speaks of only applies to the Chi-Square distribution function when  $X$  is large (e.g.,  $10^{13}$ ). (Our thanks to Tim Hesterberg from Insightful Corporation for his guidance concerning the RNG).

#### Criticism 3: Typo!!

The entry of .0611 for the maximum  $r^2$  when  $d = .00$  and  $n_1 = n_2 = 10$  in Table 2 should read .611.

#### Criticism 4: Negative values for $d$

Although Sawilowsky (2003) disagrees, there are instances when a minimum  $d$  is actually less than zero. Consider the directional hypothesis t-test where we are comparing the effects of a diet pill on 100 people. We randomly assign people to one of two groups; experimental and control. The point of the study is to show the effect of the diet pill on the experimental group. Let's suppose that when we compare the mean weights of the people at the beginning of the study and note that both group means are 200, and then again at the end of the study and note  $\bar{X}_{\text{exp}} = 225$  and the  $\bar{X}_{\text{control}} = 200$ . If we were to consider that the  $\sigma = 35$ , then we could compute the  $d$  for this study as:

Consider that it would be incorrect to interpret the absolute value of this formula (Cohen, 1988, formula 2.2.2) because we are witnessing an actual negative effect of the diet pill (e.g., people who took the diet pill actually gained weight). If we were to follow the logic of Sawilowsky, we would either interpret this as a positive effect or simply assume the effect is zero. In this case, interpreting a negative effect is *important*. It means that the diet pill worked worse than if we had done nothing at all! Sawilowsky also mistakenly states that the minimum effect (or  $d$ ) should be defined as zero when in fact this is not true (c.f., Cohen, 1988, formula 2.2.1, p. 20).

As this formula applies to our study, we explicitly stated in our manuscript (p. 247) that the design of the study was to test this specific effect with a directional hypothesis where the expected effect was that the experimental group would have a larger mean in the population than did the control group (except for the case where  $d = .00$ ).

#### Criticism 5: Repetitions

Although Sawilowsky and Yoon (2001) used 10,000 replication, we felt that 5,000 was plenty to obtain generalizability. This was not a limitation due to using a macro in S-PLUS as S-PLUS is a programming language and changing the number of replications is as simple as typing a new number into the script file. However, since Sawilowsky posited this as a criticism of the study, we re-ran all analysis with 10,000 replications and noticed that even under extreme conditions, estimates typically did not differ until the 1000<sup>th</sup> decimal place!

#### Criticism 6: Sampling without replacement

We feel that we may have been misleading with our statement, "5,000 pairs of sample data were randomly drawn without replacement at the specified sample sizes" (Roberts & Henson, 2002, p. 246). What would have been better stated is that we sampled without replacement *within* each given replication. After people were drawn from the population for the replication, they were then re-inserted into the population at the completion of that replication.



We chose this method because it seemed counterintuitive to allow for the inclusion of the same person twice within each study (although the probability for being chosen twice is less than 1% for  $n = 100$ ). We should have been clearer in pointing out that we sampled *with* replacement across the replications, just not inside each replication.

#### Criticism 7: Redundancy is reinforcement!!

Although Sawilowsky points out that there was no need for 2/3 of our study since there was no change in the standardized values, we felt it important to further reinforce the point that the spread of the data make simply a marginal difference in effecting the bias (or lack thereof) in both  $d$  and  $r^2$ . We would argue that if the results really were redundant then we would see exactly the same values in each of the tables, which we in fact did not. Therefore the inclusion of all three tables serves to reinforce the point that under multiple conditions,  $d$  shows practically no bias.

#### Criticism 8: Results that shouldn't be published?

This criticism probably should have been labeled under the "major criticisms" because it states "there is little justification for publishing Monte Carlo work when results can be computed easily and directly." As per our manuscript, we would again point out that the purpose of it was two-fold: to see if  $d$  contained bias and to see if Thompson's (2002) correction formula should be applied. If nothing else than to show that Thompson's formula "overcorrects for the actual amount of bias" (Roberts & Henson, 2002, p. 251), then the manuscript has merit. Furthermore our study shows that even though the correction cited by Sawilowsky may apply to meta-analysis, it seems of little concern to attempt to correct  $d$  in directional hypothesis settings.

#### Criticism 9: Compelling reasons to report effect sizes

We might restate that it was not the purpose of our study to present a "compelling reason to report effect sizes *when the null hypothesis remains tenable*." Our purpose was to investigate the bias in  $d$ . However, having said that we would like to add that in any given study, we may obtain a result in which the null hypothesis is tenable, *but that doesn't mean that the effect is not*

*real!* We will deal more thoroughly with this in the next section.

#### Response to Major Criticism

##### *Is the Effect Trivial or Not?*

Sawilowsky (2003) suggests that he and Yoon (2001) never "argued that small effects can in some cases be due solely to sampling error" as we summarized (Roberts & Henson, 2002, p. 245). Nevertheless, in their paper Sawilowsky and Yoon (2001) noted that reporting their simulated average Cohen's  $d$  effect of .17 would be "misleading because these effect sizes are specious" (p. 2). In their conclusion, the authors claimed: "It was shown that effect sizes should not be reported or interpreted in the absence of statistical significance" (Sawilowsky & Yoon, 2001, p. 4). (It should be noted as well that only the Sawilowsky & Yoon [2001] paper was referenced in our original article. Sawilowsky and Yoon's 2002 article resulting from this paper was not in print during our manuscript development, and therefore was not considered in our article.)

If Sawilowsky is not arguing that these effect sizes could be solely due to sampling error, then why not report and interpret them? Indeed, the average  $d$  of .17 was presented as a case when a non-zero effect was obtained from purely random numbers. Surely the logic of this conclusion suggests that small effects can be obtained even when the null hypothesis remains tenable under a statistical significance test. If the significance test is to be trusted over the small effect size, then from whence must the researcher conclude the effect originated? Under this logic, the effect must have been a function of sampling error.

#### Confused vs Informed Methodology and Readership

Sawilowsky (2003) proceeds in his major criticism by presenting two literatures of effect sizes (A and B). He supposes that after reading one of these literatures, a reader may be "thoroughly confused on the effectiveness of the intervention" (p. 223) because of the presence of non-statistically significant results mixed with other, presumably, statistically significant results. We agree that interpretation of such a literature may present certain challenges. Nevertheless, we would be hopeful that a *more informed use of*

*statistics* would be the solution to this difficulty rather than avoidance of potential confusion by replacing it with another source of misleading information.

(As a caveat, we would also be hopeful that even a modestly informed consumer of research would be able to determine the expected directionality of an effect, and whether the experimental group is expected to outperform or underperform the control on relevant outcomes. This assumes, perhaps, at least a modestly effective job at communication from the authors.)

It is at this point that we fundamentally disagree with Sawilowsky (2003). It is perhaps very appealing to some to employ statistical significance as a gatekeeper for reporting and interpreting meaningful outcomes. As we cited previously, Robinson and Levin (1997) and Levin and Robinson (2000) propose a reasoned argument for just such a two-stage process, where a finding must be deemed statistically significant before evaluation of the effect size is permitted. Of course, this would work only to the extent that the gatekeeper is effective in performing its duties.

This process also will only work when (a) the readership of the article understands fully the factors impacting statistical significance tests and the elements of power that underlie them and (b) the author understands and communicates these issues directly. Unfortunately, empirical studies have demonstrated that there are a great number of misconceptions about statistical significance testing (cf., Nelson, Rosenthal, Rosnow, 1986; Oakes, 1986; Rosenthal & Gaito, 1963; M. Zuckerman, Hodgins, A. Zuckerman, & Rosenthal, 1993), and so neither of these outcomes is likely on a widespread basis. Is this the method's fault or our own? We would suggest, of course, primarily the latter. Unfortunately, statistical significance testing has come to be treated among many researchers as a truly dichotomous outcome that relates directly to result importance. This interpretation is a result of many factors, none of which make the misinterpretation any more correct. As Sawilowsky (2003) correctly indicated, the context of the study is critical when interpreting both statistical significance and effect size outcomes.

It is of course very true that a small effect size may be due to sampling error. It is also just as true that the same small effect size may be a *real*

effect in spite of it not being statistically significant due to a lack of power. The arguments presented by Sawilowsky (2003) simply do not discount the possibility (and yes, historical truth) that some very real effects may exist but be at risk of not being discovered due to a lack of statistical significance. Meta-analytically speaking, however, when these small but non-statistically significant effects are examined across studies, a more meaningful outcome may be discovered. While it is very easy for methodologists to say that these studies should have had more power, it is much more difficult to attain sufficient power for every study in all applied situations. Should we pay more attention to power? Yes, of course. Should we also recognize that some small effects may indeed be reasonable outcomes not due entirely to sampling error? Absolutely!

A better approach to this issue, in our view, would not just result in discussion of whether statistical significance should be the gatekeeper, or even whether small effects should necessarily be reported and/or interpreted, but rather how methodologists and applied researchers can seek a more informed understanding and use of both of these statistics for what they are.

### Conclusion

Effect sizes are not final determinants regarding whether a result is meaningful any more than statistical significance tests are, and if we interpret effect sizes with the same rigidity that we have historically interpreted statistical significance testing, we are guilty of committing the same error yet again. Instead, researchers ought to view their studies in context with prior literature, make comparisons between their outcomes and those from prior studies, attend to power issues, and interpret the findings to the readership for what they are.

Is a small yet non-statistically significant effect important? Maybe, maybe not. We certainly would not know for sure without replication and some form of meta-analysis. We certainly could not do either of these, at least in a world where Type II error exists as much as its Type I counterpart, unless these same small effects were reported.

## References

- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.
- Levin, J. R., & Robinson, D. H. (2000). Rejoinder: Statistical hypothesis testing, effect-size estimation, and the conclusion coherence of primary research studies. *Educational Researcher*, 29, 34-36.
- Nelson, N., Rosenthal, R., & Rosnow, R. L. (1986). Interpretation of significance levels and effect sizes by psychological researchers. *American Psychologist*, 41, 1299-1301.
- Oakes, M. (1986). *Statistical inference: A commentary for the social and behavioral sciences*. New York: Wiley.
- Roberts, J. K., & Henson, R. K. (2002). Correction for bias in estimating effect sizes. *Educational and Psychological Measurement*, 62(2), 241-253.
- Rosenthal, R., & Gaito, J. (1963). The interpretation of level of significance by psychological researchers. *Journal of Psychology*, 55, 33-38.
- Sawilowsky, S. S. (2003). You think you've got trivials? *Journal of Modern Applied Statistical Methods*, 2(1), 218-225.
- Sawilowsky, S. S., & Yoon, J. (2001, August). *The trouble with trivials ( $p > .05$ )*. Paper presented at the 53rd session of the International Statistical Institute, Seoul, South Korea.
- Sawilowsky, S. S., & Yoon, J. (2002). The trouble with trivials ( $p > .05$ ). *Journal of Modern Applied Statistical Methods*, 1, 143-144.
- Thompson, B. (2002). "Statistical," "practical," and "clinical": How many kinds of significance do counselors need to consider? *Journal of Counseling and Development*, 80, 64-71.
- Zuckerman, M., Hodgins, H. S., Zuckerman, A., & Rosenthal, R. (1993). Contemporary issues in the analysis of data: A survey of 551 psychologists. *Psychological Science*, 4, 49-53.

*Invited Debate: Comment*  
**The Trouble With Interpreting Statistically  
Nonsignificant Effect Sizes in Single-Study Investigations**

Joel R. Levin  
University of Arizona

Daniel H. Robinson  
University of Texas at Austin

---

In this commentary, we offer a perspective on the problem of authors reporting and interpreting effect sizes in the absence of formal statistical tests of their *chanceness*. The perspective reinforces our previous distinction between single-study investigations and multiple-study syntheses.

Key words: Hypothesis testing, effect sizes, conclusion coherence

---

### Introduction

Yes, everybody has troubles, and not just with trivials (Sawilowsky, 2003). We adopt a different perspective on the Sawilowsky vs. Roberts-Henson debates about appropriate methodologies for, and interpretations of, their respective Monte Carlo investigations (Roberts & Henson, 2002; Sawilowsky & Yoon, 2002).

Although we have decided biases concerning the rights and wrongs of that particular debate, we also have decided not to jump into the fray for two related reasons: (1) Knapp (2003) considers a number of general issues that need to be considered in the context of Monte Carlo simulation studies; and (2) because we regard such issues more as background to certain more fundamental research-related effect-size-reporting foreground issues, we elected to forego additional hammering on the former so that we might nail down the latter.

### Single-Study Investigations vs Multiple-Study Syntheses

The major argument promoted here is one that we have presented elsewhere (e.g., Levin, 1998; Levin & Robinson, 2000; see also Onwuegbuzie & Levin, 2003). It can be sum-

marized as follows: Research conductors and consumers need to be more attentive to the different purposes/functions of an educational research article. Is it: (a) to report the results of an *individual* empirical study (a single-study investigation) or is it (b) to summarize a *set* of empirical studies (a meta-analytic multiple-study synthesis)?

If *a*, then we contend that hypothesis testing should be a critical precursor to effect-size estimation in telling the researcher's story; whereas if *b*, then effect-size reporting should play a more prominent role. In that context, a critical point of contention concerns whether the effect sizes associated with a single-study investigation should be interpreted in the absence of statistical significance. We have cast our nay votes on (and justifications for) this issue elsewhere (e.g., Levin, 1993; Levin & Robinson, 1999; Robinson & Levin, 1997; Robinson, Funk, Halbur, & O'Ryan, in press; Wainer & Robinson, in press) and will summarize our stance here.

Almost without exception, introductory statistics textbooks present examples based on single-study investigations. And, of course, a good number of single-study investigations are published in educational-research scholarly journals. Authors are forced to interpret the results of statistical inference tests – and this is where most of the troubles begin. In our previous writings, we have argued that statistical significance should serve a gatekeeper function to screen out effects whose direction has not been determined probabilistically. What may appear to be an interesting or important effect worth talking

---

Correspondence concerning this article should be addressed to Joel R. Levin, Department of Educational Psychology, University of Arizona, Tucson, AZ 85721. Email: jrlevin@u.arizona.edu.

about can easily be a chance finding, or one that is attributable solely to sampling error. In that case, by screening out spurious effects through a formal statistical test, an author protects the reader from erroneously interpreting the effects as if they were real.

Let us insert an important comment that has rarely been mentioned in relation to the so-called “significance-testing controversy.” It is simply that under the truth of the null hypothesis, testing the hypothesis that, say, two population means are equal or that the correlation between two variables is zero *is equivalent to testing the hypothesis that the effect size is equal to zero*. This may be readily appreciated when inferences about correlation coefficients are desired (because the correlation coefficient itself is an effect-size measure), though not as readily appreciated in the mean-difference situation.

Yet, it becomes apparent when one realizes that if the two population means are equal, then  $\mu_1 - \mu_2 = 0$ , and the corresponding population Cohen’s  $d$  effect-size measure is  $0/\sigma = 0$ . Thus, if a researcher applies a formal statistical test and then proceeds to report/interpret the sample effect size regardless of the test’s outcome, the question arises: What function did the statistical test serve, and why was it even conducted in the first place? That conclusion coherence issue (Levin & Robinson, 2000) is one that Roberts and Henson (2003) need to reconcile.

#### Another Troubling, Yet Telling, Hypothetical Example

As a sequel to a perplexing example (Levin & Robinson, 2000, p. 34-35; see also Levin’s, 1993, p. 379), let us consider an instructional intervention study with  $n = 2$  participants in each of two conditions, where Condition 1’s scores are both 5 and Condition 2’s scores are both 6. For this example, a nondirectional permutation test would indicate that there is not sufficient evidence to conclude that the two populations are statistically different ( $p = 2/6 = .333$ , which far exceeds the conventional .05 level of statistical significance).

On the other hand, if an effect-size measure were computed and reported, it would likely be communicated as gigantic or even infinitely large, for in fact, in this particular instance  $d$  is equal to  $\infty$ . Alternatively, with effect

size defined as a squared point-biserial correlation coefficient, one would conclude that there is perfect prediction of scores from knowledge of condition, with no score variability left to be explained, for  $r^2$  turns out to be 1.00 here. Never mind that the study included only a couple participants per condition and that a valid statistical test performed on these data indicates a nonsurprising event associated with an outcome this or more extreme (i.e.,  $p = .333$ ), assuming that the population-identity hypothesis is true. Moreover, even if each condition were to include a third participant (resulting in  $n = 3$ ) who produced the same scores of 5 and 6 for Conditions 1 and 2, respectively, the associated significance probability would be only  $p = 2/20 = .10$ , still above the conventional .05 level.

Although this particular example may sound extreme, far fetched, or even ridiculous, consider the myriad experiments in the educational research literature that involve a comparison of two different instructional approaches each based on three teachers, classrooms, or schools. With those teachers/classrooms/schools representing the appropriate data-analysis units (e.g., Levin & O’Donnell, 1999) and with the aggregated data equal to the values just described, the above significance probability of .10 applies.

This example also serves to clarify an oft-made argument that statistically nonsignificant effects are invariably associated with small or trivial effect sizes. Yes, a large-scale study (e.g.,  $N = 100$ ) with trivial effects (e.g.,  $d = .10$ ) can produce nonsignificant results, but so can a very small-scale study with huge effects (as was just illustrated). Conscientious conclusion-coherent researchers should refrain from interpreting such effects as either real (in both cases) or important (in the second case).

Our example leads to consideration of a converse situation as well, which was earlier discussed by Robinson and Levin (1997). The following question is regularly posed by one of us on Ph.D. qualifying examinations: “What is wrong with a researcher’s claim that ‘although the anticipated outcome did not quite reach statistical significance in this study, it would have if only a few more participants had been included’?” This claim is reminiscent of the substance of Thompson’s (e.g., 1989, 1996) proposed “what if” analyses and something toward which Roberts and

Henson (2003) tread dangerously close. (We are also troubled by the researcher's use of the term "quite" in the qualifying-examination question, as will be reflected in our concluding paragraph.) Thus, in our above amended example based on  $n = 3$  participants per condition (for which  $p = .10$ ), can it be claimed that if only one more participant were added to each condition the difference between conditions would have been statistically significant (since with  $n = 4$ ,  $p = 2/70 = .029$  according to a two-sample permutation test)? Well, could it?

Only if you are willing also to add that the outcome produced by the two additional scores (resulting in  $n = 4$  participants per condition) mimicked exactly what was present in the original data. In the case of a two-sample permutation test, just as all three Condition 2 participants had higher scores than all three Condition 1 participants in the actually conducted study, only if the additional participant in each condition maintained that situation would there be a statistically significant difference at the .05 level. In contrast, if either the additional Condition 1 participant were to score higher than any Condition 2 participant or the additional Condition 2 participant were to score lower than any Condition 1 participant, then  $p < .05$  statistical significance would not be attained (see, for example, Fisher, 1960, pp. 11-15).

The key to answering the qualifying-examination question is recognizing that one cannot simply *assume* that the mean difference or pattern will stay exactly the same with the addition of a few more participants. That is precisely the reason why one needs to collect actual data and conduct the analysis, rather than sitting around thinking in hypothetical "what if?" terms. Robinson, Fouladi, Williams, and Bera (2002) provide empirical data bearing on "what if" pondering and Hoenig and Heisey (2001) discuss an equally troubling related issue, post hoc or observed power analyses.

But we have other fish to fry. In Roberts and Henson's (2003) concluding paragraph, it is implied that researchers would be unable either to conduct replication studies or to perform meta-analyses unless authors calculate and report *all* effect sizes – including statistically nonsignificant ones. Let us consider each of the two implied components (replication studies and meta-analyses) of this contention in turn.

Is Effect-Size Information A Necessity For Independent Replication Studies?

First, the replication component. If a researcher chooses to replicate an experiment, knowledge of the specific magnitude of a nonsignificant outcome from that experiment is not a prerequisite. The forefather of experimental design and statistical hypothesis testing, Sir Ronald Fisher, certainly could – and did – replicate his agricultural experiments without betting the farm on a single study's effect sizes. Indeed, Fisher believed that the direction of an effect was only established if he could produce consistent results based on several replications.

As investigators who have collected our share of primary research data, our replication philosophy is similar to Fisher's. And the difference between that philosophy and the one apparently held by Roberts and Henson basically comes down to the difference between the publication of single-shot (one-experiment) studies (their conception of published educational research) and multiple-experiment replication-and-extension studies (our conception). In fact, we contend that much of the fury that characterizes the debates between those who wish to do away with statistical hypothesis testing and those who defend the essence of it (see, for example, Harlow, Mulaik, & Steiger, 1997) would dissipate if researchers refrained from publishing and interpreting single-shot studies.

Results that are statistically significant permit two conclusions. First, they provide evidence that the hypothesis under test (of which the null hypothesis is a special case) is not supported. Second, and less trivially (e.g., Cohen, 1994), they provide evidence of the direction of the difference or relationship. For example, a statistically significant *t*-test comparing the mean scores of a treatment and control group tells us that it is likely that the treatment group outperformed the control group in the sampled-from populations. Results that are not statistically significant do not permit either of these conclusions.

On the other hand, it is also possible that certain statistically nonsignificant effects are real but too small or fragile to be detected within the parameters of the initial study. In that case, the researcher must decide whether or not the effect is worth pursuing. If so, a replication study is in

order, which may involve changing/tweaking one or more of the initial study's features to make the statistical test of the treatment effect more sensitive – such as by incorporating a larger, more homogeneous, or differently defined sample, strengthening the treatment and/or its implementation, modifying the experimental design and analysis in some way (e.g., through blocking or by including a relevant antecedent variable in the analysis), or improving the psychometric properties of the outcome measure. If the replication study finds the effect to be statistically significant, and if that replication is followed by additional successful replications, then the initially spurned statistically nonsignificant effect will be resurrected.

#### Is Explicit Effect-Size Reporting A Necessity For Meta-Analytic Literature Syntheses?

Roberts and Henson (2003, p. 226-230) argue (again, at least implicitly) that if multiple-study syntheses are to be conducted, then reporting effect sizes for each experiment allows a meta-analyst to compute an average effect size, as well as to see how the size of the effect may vary as a function of design changes. The argument has been made that single-study investigations should always include effect sizes, even for statistically nonsignificant outcomes, so that meta-analysts will be able to ply their trade using that study's effect-size estimate. What is ignored in this argument is that a meta-analyst does not need the primary researcher to provide explicit effect-size information. As long as the researcher provides sufficient statistics (in the form of either means, variances/covariances, and sample sizes or the associated test statistics) then a competent meta-analyst will be able to calculate the standardized effect-size measures required for multiple-study syntheses (see, for example, Robinson & Levin, 1997).

It is important to note here that we also differ from Roberts and Henson (2003, p. 227-230) in our view of whether research syntheses should consist mostly of meta-analyses or of programmatic replication-and-extension studies. We opt mainly for the latter. We do not disagree that meta-analysis, as conceived by Gene Glass (1976) more than a generation ago, holds great potential for revealing potentially important findings that are shrouded in a literature where

studies are classified only in terms of significant and nonsignificant (see also Hunt, 1997). However, much of what we have witnessed as passing for meta-analyses in the educational and psychological literature since Glass coined the term may be more masking than revealing. For example, certain meta-analytic studies consider all the research on, say, visual aids in learning from text (Robinson, 2002) or phonics/phonemic instruction in beginning reading (Ehri, Nunes, Willows, Schuster, Yaghoub-Zadeh, & Shanahan, 2001) without attending to the type and quality of the materials or the specifics of the instruction. Reporting the average effect sizes in such global meta-analyses may inadvertently misinform the reader.

Finally, we believe that there is another plausible meta-analytic reason to favor single-study authors reporting sufficient summary data rather than the effect-size measures that can be derived from them. It is because (at least in our experience) that it is not unusual for authors to derive effect-size measures incorrectly – in the case of  $d$ , often with respect to the particular standard deviation selected for the specific design (e.g., between-subjects, within-subjects, ANCOVA) or question being asked, and in the case of  $r^2$ , by not distinguishing between (or confusing) unconditional and conditional proportions of variance explained (see, for example, Olejnik & Algina, 2000).

This could easily lead an incautious, or unchecking, meta-analyst down the wrong estimation path. Meta-analysts are generally more skilled in the nuances of effect-size types and variations and are less prone to calculating effect sizes incorrectly. Therefore, might it not even be a more judicious research practice/recommendation that meta-analysts routinely calculate effect sizes themselves based on a researcher's provided summary statistics?

#### Conclusion

In summary, and in contrast to Roberts and Henson's (2003) research philosophy, we argue that in the context of single-study investigations statistically nonsignificant effect sizes should not be reported or interpreted. That is because such reporting/interpreting may lead readers to believe – unwarrantedly – that evidence has been provided

concerning the direction of the effect. Reporting and interpreting effect sizes (with corresponding confidence intervals) in multiple-experiment studies where the effect of interest is replicated (i.e., its direction is confirmed) may provide readers with more useful information concerning the believability and magnitude of the effect, along with the consistency with which it can be produced. Additionally, when a multiple-experiment study is programmatic in nature (i.e., where the design is cumulatively extended to estimate the effect under differing contextual and procedural variations), then reporting effect sizes may be helpful in pinpointing the conditions under which the effect is strongest.

We hope that editors of educational research journals will encourage authors to report work consisting of multiple-experiment studies that replicate and extend initial findings. This is routine procedure in many behavioral science disciplines; and as a clear illustration of editorially practicing what we are preaching, see Levin (1991, p. 5-6). For each experiment conducted, *a priori*  $\alpha$  levels, *a posteriori* *p*-values, sample-size and power information, and sufficient statistics should be reported.

In terms of summarizing the multiple experiments, an author may wish to quantify replicated effects, if that serves to inform practitioners who are considering adopting the intervention. At the same time, we are not so naive as to believe that a journal-policy change of this kind will happen overnight. Thus, until the practice of publishing single-shot, non-replicated findings changes, at least we hope that statistically nonsignificant results will be regarded as evidence that the direction of an effect of interest remains undetermined and further research is needed before a more definitive conclusion can be made. Single-study investigators should not routinely provide effect-size estimates for statistically nonsignificant outcomes.

Multiple-study synthesizers can capture those effect sizes from the sufficient statistics reported. Finally, single-study authors should not persist in interpreting or promoting a statistically nonsignificant effect (which includes use of the terms “not quite significant,” “almost significant,” or “approaching significance”), due to the risk of consumers regarding the effect as having been formally screened as believable – when, in fact, no

formal evidence to that effect has been provided. With editorial changes such as these, we strongly suspect that many of educational research’s analysis-and-reporting troubles would simply burst like bubbles!

#### References

- Cohen, J. (1994). The earth is round ( $p < .05$ ). *American Psychologist*, *49*, 997-1003.
- Ehri, L. C., Nunes, S. R., Willows, D. M., Schuster, B. V., Yaghoub-Zadeh, A., & Shanahan, T. (2001). Phonemic awareness instruction helps children learn to read: Evidence from the National Reading Panel’s meta-analysis. *Reading Research Quarterly*, *36*, 250-283.
- Fisher, R. A. (1960). *The design of experiments*. New York: Hafner Publishing Company.
- Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher*, *5*(10), 3-8.
- Harlow, L. L., Mulaik, S. A., & Steiger, J. A. (Eds.). (1997). *What if there were no significance tests?* Mahwah, NJ: Erlbaum.
- Hoenig, J. M., & Heisey, D. M. (2001). The abuse of power: The pervasive fallacy of power calculations for data analysis. *American Statistician*, *55*, 19-24.
- Hunt, M. (1997). *How science takes stock: The story of meta-analysis*. New York: Russell Sage.
- Knapp, T. R. (2003). Was Monte Carlo necessary? *Journal of Modern Applied Statistical Methods*, *2*(1), 237-241.
- Levin, J. R. (1991). Editorial. *Journal of Educational Psychology*, *83*, 5-7.
- Levin, J. R. (1993). Statistical significance testing from three perspectives. *Journal of Experimental Education*, *61*, 378-382.
- Levin, J. R. (1998). What if there were no more bickering about statistical significance tests? *Research in the Schools*, *5*(2), 43-53.
- Levin, J. R., & O’Donnell, A. M. (1999). What to do about educational research’s credibility gaps? *Issues in Education: Contributions from Educational Psychology*, *5*, 177-229.
- Levin, J. R., & Robinson, D. H. (1999). Further reflections on hypothesis testing and editorial policy for primary research journals. *Educational Psychology Review*, *11*, 143-155.



- Levin, J. R., & Robinson, D. H. (2000). Statistical hypothesis testing, effect-size estimation, and the conclusion coherence of primary research studies. *Educational Researcher*, 29(1), 34-36.
- Olejnik, S., & Algina, J. (2000). Measures of effects size for comparative studies: Applications, interpretations, and limitations. *Contemporary Educational Psychology*, 25, 241-286.
- Onwuegbuzie, A. J., & Levin, J. R. (2003). Without supporting statistical evidence, where would reported measures of substantive importance lead? To no good effect. *Journal of Modern Applied Statistical Methods*, 2(1), 133-151.
- Roberts, J. K., & Henson, R. K. (2002). Correction for bias in estimating effect sizes. *Educational and Psychological Measurement*, 62, 241-253.
- Roberts, J. K., & Henson, R. K. (2003). Not all effects are created equal: A rejoinder to Sawilowsky. *Journal of Modern Applied Statistical Methods*, 2(1), 226-230.
- Robinson, D. H. (Ed.). (2002). Spatial text adjuncts and learning. Special issue of *Educational Psychology Review*, 14(1).
- Robinson, D. H., Fouladi, R. T., Williams, N. J., & Bera, S. J. (2002). Some effects of including effect size and "what if" information. *Journal of Experimental Education*, 70, 365-382.
- Robinson, D. H., Funk, D. C., Halbur, D., & O'Ryan, L. (in press). The .05 level of significance in educational research: Traditional, arbitrary, sacred, magical, or simply psychological? *Research in the Schools*.
- Robinson, D. H., & Levin, J. R. (1997). Reflections on statistical and substantive significance, with a slice of replication. *Educational Researcher*, 26, 21-26.
- Sawilowsky, S. S. (2003). You think you've got trivials? *Journal of Modern Applied Statistical Methods*, 2(1), 218-225.
- Sawilowsky, S. S., & Yoon, J. S. (2002). The trouble with trivials ( $p > .05$ ). *Journal of Modern Applied Statistical Methods*, 1, 143-144.
- Thompson, B. (1989). Statistical significance, result importance, and result generalizability: Three noteworthy but somewhat different issues. *Measurement and Evaluation in Counseling and Development*, 22, 2-6.
- Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. *Educational Researcher*, 25, 26-30.
- Wainer, H., & Robinson, D. H. (in press). Shaping up the practice of null hypothesis significance testing. *Educational Researcher*.

## *Invited Debate: Comment* Was Monte Carlo Necessary?

Thomas R. Knapp  
Kailua-Kona, Hawaii

---

In the critique that follows, I have attempted to summarize the principal disagreements between Sawilowsky and Roberts & Henson regarding the reporting and interpreting of statistically non-significant effect sizes, and to provide my own personal evaluations of their respective arguments.

Keywords: Non-significant effect sizes; Monte Carlo investigations

---

### Introduction

There are three principal matters to consider. They are (in my opinion) in order of decreasing importance:

The Reporting and Interpreting of Non-significant Statistics.

I think that the matter of reporting, interpreting, publishing, etc. statistically non-significant effect sizes can be argued without appealing to the results of any Monte Carlo investigations. Indeed, that matter has been debated almost ad nauseam over the last half-century, as the reference to Melton (1962) in the exchange between Knapp & Sawilowsky (2001) and Thompson (2001) indicates.

Consider, for example, a researcher who draws a simple random sample from a population, assumes linearity and bivariate normality, calculates a Pearson product-moment correlation coefficient (one of the simplest and most important effect size measures) between two variables for the sample, tests it for statistical significance, and gets a p-value of .03.

---

Thomas R. Knapp, Ed. D. (Harvard, 1959) is Professor Emeritus of Education and Nursing, University of Rochester and The Ohio State University. Email him at [tknapp5@juno.com](mailto:tknapp5@juno.com)

Should that correlation be reported? Of course; the correlation between those two variables in that sample is \_\_\_\_\_. Should it be interpreted? Of course; that correlation is not statistically significant at the .01 level, is statistically significant at the .05 level, etc. (depending upon the value of alpha chosen at the outset of the study). [Or, if interval estimation is preferred, one's confidence is .99 (or .95, or whatever) that the interval from \_\_\_\_\_ to \_\_\_\_\_ covers the population correlation.]

Should that study be published? Aye, there's the rub. Melton wouldn't have (he insisted that p be less than .01); I presume Sawilowsky & Yoon wouldn't either; and I further presume that Roberts & Henson would – all other things being equal (good theory, design, measurement, etc.). If statistically non-significant findings are not published occasionally, the literature will have an imbalance of Type I errors.

### One-sided vs Two-sided Inference

If I'm wrong and if one does need Monte Carlo evidence in order to decide whether or not a statistically non-significant effect size is of interest, should the focus be on one-sided inference or two-sided inference? Sawilowsky & Yoon (2002) chose two-sided inference and concentrated on the absolute value of Cohen's d. Roberts & Henson (2002) chose one-sided inference by concentrating on d's that were greater than or equal to 0 (with the alternative hypothesis taken to be that the experimental mean is greater than the control mean). I agree with Roberts & Henson, since that better reflects the more typical

research hypothesis and is also simpler (it involves only two sampling distributions rather than three).

#### Technical Aspects of Monte Carlo Investigations

Sawilowsky & Yoon (2002) carried out one kind of Monte Carlo investigation. Roberts & Henson (2002) carried out another kind of Monte Carlo investigation. The particular details (number of replications, Fortran vs S-Plus, etc.) also differed. I have no idea who's right and who's wrong there.

#### Specific Comments

##### Sawilowsky & Yoon (2002)

1. They chose sample sizes of 10 and 10, and a power level of .2 "to mimic applied research" (p. 143). Those sample sizes strike me as too small for typical educational experiments, and there is a considerable amount of evidence (see, for example, Aberson, et al., 2002) that the average a priori power for published studies in education is approximately .5, not .2.

2. Their Monte Carlo investigation revealed an average obtained absolute effect size of .169 for statistically non-significant results when comparing the means of two samples of size 10 drawn from normal populations in which the population effect size was zero. I believe such a result could have been determined analytically (mathematically), and I also believe that .169 is actually too low. In the Appendix to this critique I have outlined a proof of those beliefs.

I conclude that the Sawilowsky & Yoon (2002) research was not necessary.

##### Roberts & Henson (2002)

Roberts & Henson (2002) were reacting to Sawilowsky & Yoon (2001), not Sawilowsky & Yoon (2002), but those two papers are almost identical.

1. In their opening sentence, Roberts & Henson (2002) referred to a controversy between "the role and function of effect sizes" and the use of "statistical significance tests" (p. 241). That is a false comparison. People who use statistical significance tests have almost always calculated some sorts of sample effect sizes *before* they carry

out the significance tests (see the Pearson  $r$  example, above).

The general controversy involves whether or not significance tests should be prohibited; the specific controversy between Sawilowsky and Roberts & Henson involves whether or not statistically non-significant sample effect sizes should be reported and interpreted.

2. They (Roberts & Henson) went through an elaborate discussion of Thompson's (2002) recommendation of converting  $d$  to  $r$ , Friedman's (1968) formula for converting  $r$  to  $d$ , Ezekiel's (1930) correction formula, etc. That is unnecessary. All one needs to do is algebraically re-solve the  $d$ -to- $r$  formula given by Cohen (1988) for  $r$  in terms of  $d$  (but see Aaron, Kromrey, & Ferron, 1998 regarding that formula - it only works for equal and large  $n$ 's) and/or appeal to the work of Hedges (1981), Kraemer (1983), and Hedges & Olkin (1985) concerning the amount of bias in Cohen's  $d$ .

3. They then went on to report in three separate tables the results of their Monte Carlo investigation, for various values of Cohen's  $d$  in the population, various values of the population standard deviation (the mean for the control group was taken to be 100), and various sample sizes, including the  $n_1 = n_2 = 10$  case that was of interest to Sawilowsky & Yoon. Several of those results are already reasonably well known.

The expected value (mean) of a sample  $r^2$  is equal to  $1/(N-1)$  when the population  $r^2$  is equal to zero (see, for example, Marascuilo & Levin, 1983, p. 97), so the small differences between that expected value for  $n_1 = n_2 = 10$  (an  $N = 20$ ), i.e., .0526315..., and the mean sample  $r^2$  for a population  $r^2$  of 0 in their tables are all attributable to Monte Carlo sampling variation. Formulas for the expected value and sampling variance for Cohen's  $d$  can be found in Hedges (1981), in Kraemer (1983), and in Hedges & Olkin (1985, pp. 78-81), so their results for  $d$  differ from those derived mathematically also because of Monte Carlo sampling variation.

Some of the other results are a bit baffling. For example, why isn't the Bias row for  $d$  in each of those tables equal to the difference between the mean sample  $d$  and the  $d$  in the population? [Is it because of the discrepancies between the desired

population  $d$ 's and the Monte Carlo population  $d$ 's to which they referred on page 247?] And how can the bias for the sample  $d$  for a population  $d$  of .20 be *greater* for  $n$ 's of 100 than for  $n$ 's of 50 in both Table 1 and Table 2?

4. In their concluding section Roberts & Henson (2002) claimed that "...replication of a given study is the only true way to evaluate possible generalizability" (p. 252). I agree (by definition). They went on to say that "Statistically nonsignificant effects may be fully replicable." Of course; if nothing is going on, nothing will keep getting replicated, but that doesn't help their argument.

I equally regretfully conclude that the Roberts & Henson (2002) research was also not necessary.

#### Sawilowsky (2003)

1. He drew several distinctions among simulation, Monte Carlo, Monte Carlo simulation, sampling with replacement vs. sampling without replacement, and characteristics of a "high quality Monte Carlo simulation" (p. 218) The first three and the fifth are apparently important to make in any Monte Carlo investigation (I leave that to others to decide). The fourth distinction (sampling with replacement vs. sampling without replacement) is of course always important to make, especially when it comes to sampling within sample and sampling between samples.

Under "Monte Carlo" he properly acknowledged that there are some situations, such as finding the definite integral from 0 to 1 of  $f(x) = x$ , where the Monte Carlo approach could be used but should not be. However, under "Sampling With vs. Without Replacement" he claimed that sampling without replacement is appropriate when sampling from a deck of cards. I disagree; such sampling can be either with or without replacement within sample - it all depends upon whether or not a sampled card gets replaced in the deck prior to the sampling of a subsequent card - but sampling must be with replacement between samples or you soon run out of cards to sample.

2. The remainder of his response is the heart of his paper (in my opinion). He first listed what he called "Nine Minor Criticisms" of Roberts & Henson (2002). I would have identified at least

two of those (# 7 and # 8) as major criticisms. Why Roberts & Henson bothered with three different tables is beyond me (their rationale on page 246 is interesting but irrelevant, given that both  $d$  and  $r^2$  are scale free); and I have already indicated above in my Comments #2 and #3 regarding their study that the bias in  $d$  had already been addressed analytically by Hedges (1981), by Kraemer (1983), and by Hedges & Olkin (1985).

Sawilowsky's "Major Criticism" apparently has to do with the kinds of results one might obtain when sampling from populations with  $d$ 's of 0, and with the order in which the results appear. I found that section rather difficult to follow. I guess the point he's making is that the findings in Data Set B are more likely to be obtained and will look more impressive than the findings in Data Set A, but the obtained effect sizes in both data sets could easily be attributable to chance.

#### Roberts & Henson (2003)

1. At the beginning of their paper, Roberts & Henson (2003) stated that the first portion of Sawilowsky's (2003) paper "does not bear comment on" (p. 226). Although I don't particularly care for Monte Carlo investigations, Roberts and Henson apparently do (since their research was such an investigation), and Sawilowsky's claims concerning how a good Monte Carlo simulation should be carried out deserved a response. (They did comment on some technical Monte Carlo features in their responses to Sawilowsky's minor criticisms.)

2. They then went on to address all of Sawilowsky's minor criticisms. I have already implied my lack of interest in #2 and #5. And they appear to have accepted criticisms #1, 3, and 6. Where they disagreed most with Sawilowsky is with respect to criticisms #4, 7, 8, and 9. I shall accordingly concentrate on those matters.

As indicated above, I agree with them regarding negative values of  $d$  (#4). But I take exception to their responses to those last three criticisms. Their paragraph (regarding #7) that bears the heading "Redundancy is reinforcement!" (with an exclamation point yet) is bizarre. As Sawilowsky (2003) pointed out, and as I argued above, there was no good reason for including all three tables. Their sentence "We would argue that if

the results were redundant then we would see exactly the same values in each of the tables, which we in fact did not.” (p. 229) shows a lack of understanding of Monte Carlo. It is inherent in the method that you do not get “exactly” anything; it is subject to sampling variation just like any sample statistic is. And they missed the point regarding #8. The published work on the bias of  $d$  obviated the need for Monte Carlo. (I’m not sure what point they were trying to make regarding #9, other than the fact that Type II errors are possible.)

3. They concluded their paper by responding to Sawilowsky’s (2003) major criticism. They may have been even more confused than I was by that section of Sawilowsky’s paper, because they seemed to be talking about Type II error all over again, introducing several citations to the literature on misconceptions regarding significance testing, etc. rather than directly addressing Sawilowsky’s examples of data sets that could be realized and how they should be interpreted.

#### References

- Aaron, B., Kromrey, J.D. & Ferron, J.M. (November, 1998). Equating  $r$ -based and  $d$ -based effect size indices: Problems with a commonly recommended formula. Paper presented at the annual meeting of the Florida Educational Research Association, Orlando, FL. (ERIC Document Reproduction Service No. ED 433 353)
- Abersson, C.L., Berger, D.E., Healy, M.R., & Romero, V.L. (2002). An interactive tutorial for teaching statistical power. *Journal of Statistics Education*, 10 (3) [Online].
- Ezekiel, M. (1930). *Methods of correlational analysis*. New York: Wiley.
- Friedman, H. (1968). Magnitude of experimental effect and a table for its rapid estimation. *Psychological Bulletin*, 70, 245-251.
- Hedges, L.V. (1981). Distribution theory for Glass’s estimator of effect size and related estimators. *Journal of Educational Statistics*, 6, 107-128.
- Hedges, L.V. & Olkin, I. (1985). *Statistical methods for meta-analysis*. Orlando, FL: Academic Press.
- Knapp, T.R., & Sawilowsky, S.S. (2001). Constructive criticisms of methodological and editorial practices. *Journal of Experimental Education*, 70 (1), 65-79.
- Kraemer, H.C. (1983). Theory of estimation and testing of effect sizes: Use in meta-analysis. *Journal of Educational Statistics*, 8, 93-101.
- Marascuilo, L.A., & Levin, J.R. (1983). *Multivariate statistics in the social sciences*. Monterey, CA: Brooks/Cole.
- Melton, A.W. (1962). Editorial. *Journal of Experimental Psychology*, 64, 553-557.
- Roberts, J.K., & Henson, R.K. (2002). Correction for bias in estimating effect sizes. *Educational and Psychological Measurement*, 62 (2), 241-253.
- Roberts, J.K., & Henson, R.K. (2003). Not all effects are created equal: A rejoinder to Sawilowsky. *Journal of Modern Applied Statistical Methods*, 2(1), 227-231.
- Sawilowsky, S.S. (2003). You think you’ve got trivials? *Journal of Modern Applied Statistical Methods*, 2(1), 218-226.
- Sawilowsky, S.S., & Yoon, J.S. (August, 2001). The trouble with trivials ( $p > .05$ ). Paper presented at the 53rd Session of the International Statistical Institute, Seoul, South Korea.
- Sawilowsky, S.S., & Yoon, J.S. (2002). The trouble with trivials ( $p > .05$ ). *Journal of Modern Applied Statistical Methods*, 1 (1), 143-144.
- Thompson, B. (2001). Significance, effect sizes, stepwise methods, and other issues: Strong arguments move the field. *Journal of Experimental Education*, 70 (1), 80-93.
- Thompson, B. (2002). “Statistical”, “practical”, and “clinical”. How many kinds of significance do counselors need to consider? *Journal of Counseling and Development*, 80, 64-71.

#### Appendix

It can be shown (personal communication from Ingram Olkin, May 5, 2003) that the expected value of the absolute value of Cohen’s  $d$ , i.e.,  $E(|d|)$ , can be expressed as an infinite series in terms of gamma functions of the two sample sizes and in terms of the population effect size. If the population effect size is equal to zero and  $n_1 = n_2 = 10$  (the case of particular interest to Sawilowsky and one of the cases of interest to Roberts & Henson),  $E(|d|)$  is approximately .3726.

Kraemer (1983) showed that  $d$  follows the  $t$  sampling distribution with  $n_1 + n_2 - 2$  degrees of freedom and provided a formula for calculating the percentiles of that distribution. From the 97.5<sup>th</sup> percentile it can be determined that the cut-off point for the .05 significance level is approximately .940 for the absolute value of  $d$ .

And, from the middle 95% of that distribution it can be determined that the mean of the “non-rejectable” absolute values of  $d$  is approximately .336 (not .169). By appealing to the formula for a weighted mean it can be further determined that the mean of the “rejectable” absolute values of  $d$  is approximately 1.076 (not .508).

*Invited Debate: Rejoinder*  
Trivials: The Birth, Sale, And Final Production Of Meta-Analysis

Shlomo S. Sawilowsky  
Educational Evaluation & Research  
Wayne State University

---

The structure of the first invited debate in *JMASM* is to present a target article (Sawilowsky, 2003), provide an opportunity for a response (Roberts & Henson, 2003), and to follow with independent comments from noted scholars in the field (Knapp, 2003; Levin & Robinson, 2003). In this rejoinder, I provide a correction and a clarification in an effort to bring some closure to the debate. The intension, however, is not to rehash previously made points, even where I disagree with the response of Roberts & Henson (2003).

Key words: Effect size, meta-analysis, Monte Carlo simulation, trivials

---

### Introduction

Many such techniques were developed throughout the half-century before Gene Glass gave meta-analysis its modern name in 1976. Twenty-four years later, despite considerable developments in the field, Glass (2000) lamented the use of meta-analysis. Nevertheless, there remain powerful lobbyists for meta-analysis, including those who use their editorial position to coerce statistical policy to ensure its survival.

The question arises: Has the advent of meta-analysis in social and behavioral sciences in the past quarter century increased the ability to synthesize and evaluate research, as compared with – for example – traditional scholarly analysis? Or, perhaps has meta-analysis become the favored tool in the hunt for Type I errors? When professional associations and learned societies are lobbied to require their journals report and interpret effect sizes, the coin of the realm of meta-analysis, “in all studies, regardless of whether or not statistical tests are reported” (Thompson, 1996, p. 29) even for “non-statistically significant effects” (Thompson, 1999, p. 67), the answer to the initial question will be negative, and the latter question will be positive.

This was the point I made in Knapp & Sawilowsky (2001), and Sawilowsky and Yoon (2001, 2002). A Monte Carlo simulation was conducted to determine what magnitude of effect sizes should be expected if studies, whose results were obtained under the truth of the null hypothesis, were published piecemeal for the sake of meta-analysis. The Monte Carlo simulation indicated that effect sizes near zero should not be expected. Hence, publishing effect sizes for nonstatistically significant study results are ill advised.

Roberts & Henson (2002)

Subsequently, Roberts and Henson (2002) demurred, and the battle was joined. They advanced the following argument: Sawilowsky and Yoon’s Monte Carlo simulation (2001) must imply that the bias associated with effect sizes is large under the truth of the null hypothesis. Hence, Sawilowsky and Yoon (2001) cautioned against the publication of effect sizes in the absence of statistical significance. Yet, Roberts and Henson’s (2002) Monte Carlo study indicated the bias was near zero. Therefore, the publication of such effect sizes should not be suppressed.

The purpose of the target article (Sawilowsky, 2003) in this debate was to illustrate this is a straw-person argument. The bias associated with effect sizes under population normality is easily determined, and indeed its

---

Email the author at [shlomo@wayne.edu](mailto:shlomo@wayne.edu). The title of this article is based on Gerrold (1973).

average is near zero. This result was known two decades prior to the Roberts and Henson (2002) Monte Carlo study (Cohen, 1988, p. 66). This does not, however, detract from the main pronouncement of Sawilowsky and Yoon (2001, 2002). The expected magnitudes (i. e., absolute value) of the constituent effect sizes are *not* near zero. Publicizing these non-near zero values, for the sake of meta-analysis, will wreak havoc in the literature.

Levin & Robinson (2003)

Levin and Robinson's (2003) comments are very insightful. A premise of Sawilowsky and Yoon (2001, 2002) is that scientific research is by definition comprised of multiple-study investigations, regardless of who actually conducts the experiment.

Knapp (2003)

Knapp's (2003) comments prompt a (1) correction and a (2) clarification.

(1) Material in Knapp's (2003) appendix correctly estimates the non-near zero magnitudes of the effect sizes to be approximately  $|\bar{d}| = .34$ , not .17 as indicated in Sawilowsky and Yoon (2001, 2002). I reran the Monte Carlo simulation and got approximately the same value reported by Knapp (2003). I cannot find the errant value in my lab notes, so I must conclude that by some error I halved the result to present the value as a " $\pm$ " when setting the table for publication. Nevertheless, the correct result *doubles* the warning raised by Sawilowsky and Yoon (2001, 2002), as .34 is situated half-way between what Cohen (1988) loosely defines as a "small" and a "moderate" effect size.

(2) Knapp (2003) estimated the correct value via formulas provided by Kraemer (1983), and thus, he argued that Monte Carlo methods were not necessary. He amplified this with remarks on the general utility of Monte Carlo in the presence of mathematical statistics. As the latter comment goes to the issue of one of the three missions of *JMASM*, it demonstrates to me that the message of the power of Monte Carlo methods requires further demonstration and publicity.

As noted in the target article (Sawilowsky, 2003), there usually is no need to invoke Monte Carlo methods when results may be obtained easily, conveniently, and accurately via mathematical statistics. For example, the statistical properties of the t test, under asymptotic conditions, can easily be determined through an expansion of moments. The question in applied statistics, however, pertains to the small samples properties of this test, and, its properties under departures from underlying assumptions, especially for real data sets. Here, asymptotic mathematical statistics have utterly failed, and have misled the discipline. Monte Carlo methods, however, have been used successfully and convincingly to set the record straight regarding the properties of the t and other statistics.

#### Methodology

Sawilowsky and Yoon (2001, 2002) was remiss in not explaining that in Monte Carlo work, (1) should desirable results be obtained when underlying assumptions are met, it is still necessary to proceed to when underlying assumptions are not met, but, (2) should undesirable results be obtained when underlying assumptions are met, there is little point in proceeding to when underlying assumptions are not met. Thus, when non-near zero results were obtained under normality, the remainder of the Monte Carlo simulation results obtained became irrelevant and were not presented in Sawilowsky and Yoon (2001, 2002). However, to respond to Knapp's criticism against appealing to the use of Monte Carlo methods, these results are provided below.

#### Results

Table 1 contains the Type I error rates of the two independent samples t test under the De Moivre distribution for the purpose of demonstrating the viability of the algorithms used. The  $|\bar{d}|$  for fail to reject  $H_0$  is shown to be about .34 for  $\alpha=.05$ , and about .38 for  $\alpha=.01$ , when the sample size is 10. The 95% bracketed interval for  $|\bar{d}|$  is [.2841489 - .4107949] for  $\alpha=.05$ , and [.2968488 - .4601668] for  $\alpha=.01$ .



Because Knapp was concerned about this sample size, new results are presented below for samples of size 20 and 30. To address concerns regarding the number of repetitions, it was increased from 10,000 to ten million. Additional precision was obtained by using critical values to six decimals. The warning of Sawilowsky and Yoon (2001, 2002) remains fully supported by these new results.

Table 1. Two Independent Samples t Test Type I Error Rates,  $\overline{d}$  (Fail To Reject  $H_0$ ),  $\overline{d}$  (Reject  $H_0$ ); For De Moivre (Normal) Distribution, And Various Sample Sizes And  $\alpha$  Levels.

Statistic	$\alpha=.050000$	$\alpha=.010000$
	$n_1=n_2=10$	
Type I Error Rate	.0499992	.0099861
Fail to Reject $H_0$	.3474719	.3785078
Reject $H_0$	1.217658	1.571810
	$n_1=n_2=20$	
Type I Error Rate	.0499181	.0099800
Fail to Reject $H_0$	.2348740	.2547229
Reject $H_0$	.7940045	1.001228
	$n_1=n_2=30$	
Type I Error Rate	.0500528	.0099930
Fail to Reject $H_0$	.1891833	.2053082
Reject $H_0$	.6326703	.7928227

Notes: Critical t Taken To Six Decimals. Each Cell Entry Is Based On 10,000,000 Repetitions.

Knapp (2003) obtained approximately  $\overline{d} = .34$  without appealing to a Monte Carlo procedure. (Indeed, in e-mail correspondence, he delivered yet another method to obtain these results. It was a less satisfying solution 3, as it depended on the simulation of values with unknown characteristics by hand, instead of values with known characteristics by machine.) However, Sawilowsky and Yoon (2001, 2002) was not a Monte Carlo *study* to determine this value; it was a Monte Carlo *simulation* designed to determine the magnitude of effect sizes expected under the truth of the null hypothesis. In retrospect, perhaps the use of  $\overline{d}$  to communicate the study results obscured the objective.

Indeed, it takes a Monte Carlo simulation to determine the values in Table 2, which are the first 20 of ten million from the first run of the Fortran program that produced the value of .3474719 in Table 1. The simulation results are understood as follows. The first study to appear in the literature regarding a certain outcome, that is not statistically significant, will publicize a large effect size of .9. The second study to appear in the literature will be about .24, followed by a study that obtained an effect size of about -.18. The subsequent study will follow with an effect size of .31, and so forth.

Table 2. First Twenty Of 10,000,000 Simulated Values of  $\overline{d}$  For (Fail To Reject  $H_0$ ) For De Moivre (Normal) Distribution,  $n_1=n_2=10$ ,  $\alpha=.05$ .

#	ES	#	ES
1	.902532	11	-.214086
2	.239664	12	-.386423
3	-.184106	13	.100410
4	.311091	14	-.682867
5	.291022	15	.305013
6	-.204143	16	-.537210
7	-.105137	17	-.410020
8	.662463	18	-.330778
9	.111973	19	.168260
10	-.366065	20	.202596

The objective of Sawilowsky and Yoon (2001, 2002) was to have proponents of publishing these effect sizes imagine the incorrect message this will promote in the literature. After all, these are effect sizes obtained for an intervention modeled as random numbers! Clearly, the magnitudes of these values are non-near zero. (It should be recognized that the interpretation of the simulation results can begin at any arbitrary point within the 10 million effect sizes.)

Roberts and Henson (2002) indicated the maximum effect sizes obtained in their simulation. It was so huge that it prompted the title of Sawilowsky (2003). The maximum effect sizes obtained here for  $n_1=n_2=10$ , when there was a fail to reject decision under the truth of the null hypothesis, was  $\max \overline{d}_{\alpha=.05} = .9942942$  and  $\max$

$\overline{|d|}_{\alpha=.01} = 1.56907$  for the De Moivre distribution.

This means that an intervention modeled by random numbers can produce an effect size as large as  $d = \pm .99$  or  $d = \pm 1.6$ , for  $\alpha = .05$  and  $.01$ , respectively! Why would the members of any committee on statistical practices and reporting empowered by their professional association or learned society give credence to the position of the lobbyist who promotes the piecemeal publication of apparently huge albeit trivial effect sizes?

It is likely possible, although difficult, to obtain mathematical solutions for  $\overline{|d|}$  for small samples under population nonnormality for certain theoretical distributions. It is easy, however, to obtain results via the Monte Carlo method, as indicated in Table 3. It is impossible, however, to obtain solutions for  $\overline{|d|}$  using mathematical statistics for the populations represented by real data sets. The results are easily obtained, however, via Monte Carlo methods, as indicated in Table 4.

Table 3.  $\overline{|d|}$  (Fail to Reject  $H_0$ ) For Various Theoretical Distributions, Sample Sizes, And  $\alpha$  Levels.

Distribution	$\alpha=.050000$	$\alpha=.010000$
	$n_1=n_2=10$	
Uniform	.3439692	.3748572
Mixed Normal	.4028708	.4149501
Cauchy	.4047977	.4177936
	$n_1=n_2=20$	
Uniform	.2336624	.2535020
Mixed Normal	.2713618	.2781797
Cauchy	.2766581	.2851480
	$n_1=n_2=30$	
Uniform	.1885313	.2046196
Mixed Normal	.2133092	.2209231
Cauchy	.2228022	.2299003

Notes: Critical t Taken To Six Decimals. Each Cell Entry Is Based On 10,000,000 Repetitions. The Mixed Normal distribution is comprised of two distributions: (1)  $Z(0,1)$  with frequency of 95%, (2)  $Z(22,10)$  with frequency of 5%.

Table 4.  $\overline{|d|}$  (Fail to Reject  $H_0$ ) For Various Psychology/Education Data Sets, Sample Sizes, And  $\alpha$  Levels.

Data Set	$\alpha=.050000$	$\alpha=.010000$
	$n_1=n_2=10$	
Bimodal (P)	.3408427	.3716145
Asymmetry (P)	.3594031	.3877410
Mass At Zero (E)	.3646502	.3864528
	$n_1=n_2=20$	
Bimodal (P)	.2314171	.2512609
Asymmetry (P)	.2372115	.2572745
Mass At Zero (E)	.2355214	.2562985
	$n_1=n_2=30$	
Bimodal (P)	.1877642	.2036923
Asymmetry (P)	.1902705	.2064020
Mass At Zero (E)	.1909938	.2073510

Notes: Critical t Taken To Six Decimals. Each Cell Entry Is Based On 10,000,000 Repetitions. P = psychometric instrument, A = education test.

### Conclusion

As Knapp (2003) pointed out, “Kraemer (1983) showed that  $d$  follows the  $t$  sampling distribution with  $n_1 + n_2 - 2$  degrees of freedom” (p. 242). From this statement alone it should be obvious that the publishing of effect sizes should be handled the same as  $p$  values associated with the  $t$  statistic in hypothesis testing (as opposed to so-called significance testing, which in my view is outside the boundary of the scientific method).

A nonsignificant obtained  $t$  is interpreted, based on the samples, as the difference in means between the two groups are not statistically significantly different from zero. More formally, there is no evidence that the two samples were drawn from populations with different values of  $\mu$ . For this reason, it is the policy at many journals that  $p$  values for nonsignificant  $t$  statistics are suppressed from publication. (Typically, the author supplies an “\*” in tabled statistical material to indicate the result was not significant at the a priori specified  $\alpha$  level.)

The same should hold true for  $d$ . When the  $t$  is not statistically significant, the effect size (regardless of its magnitude) is not statistically

significantly different from zero. Unfortunately, this type of argument has not been compelling to the meta-analysis lobby.

The purpose, therefore, for the Monte Carlo simulation by Sawilowsky and Yoon (2001, 2002), was to provide another type of demonstration that the publicizing of effect sizes associated with nonstatistically significant results are an invitation to disaster in the literature. One has but to consider the effects of the proliferation of trivials (e.g., such as those in Table 2) to reject the position of lobbyists seeking to promote the piecemeal publishing of effect sizes for meta-analysis in a fashion never envisioned by its developers.

#### References

- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. (2<sup>nd</sup> ed.) Hillsdale, NJ: Erlbaum.
- Gerrold, D. (1973). *The story behind a Star Trek show! "The trouble with tribbles": The birth, sale, and final production of one episode*. NY: Ballantine.
- Knapp, T. R. (2003). Was Monte Carlo necessary? *Journal of Modern Applied Statistical Methods*, 2(1), 238-242.
- Knapp, T. R., & Sawilowsky, S. S. (2001). Constructive criticisms of methodological and editorial practices. *The Journal of Experimental Education*, 70, 65-79.
- Kraemer, H.C. (1983). Theory of estimation and testing of effect sizes: Use in meta-analysis. *Journal of Educational Statistics*, 8, 93-101.
- Levin, J. R., & Robinson, D. H. (2003). The trouble with interpreting statistically nonsignificant effect sizes in single-study investigations. *Journal of Modern Applied Statistical Methods*, 2(1), 232-237.
- Roberts, K. J., & Henson, R. K. (2003). Not all effects are created equal: a rejoinder to Sawilowsky. *Journal of Modern Applied Statistical Methods*, 2(1), 227-231.
- Sawilowsky, S. S. (2003). You think you've got trivials? *Journal of Modern Applied Statistical Methods*, 2(1), 218-226.
- Sawilowsky, S. S., & Yoon, J. (2001). *The trouble with trivials (p > .05)*. Paper presented at the 53<sup>rd</sup> session of the International Statistical Institute, Seoul, South Korea.
- Sawilowsky, S. S., & Yoon, J. (2002). The trouble with trivials (p > .05). *Journal of Modern Applied Statistical Methods*, 1, 143-144.
- Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. *Educational Researcher*, 25, 26-30.
- Thompson, B. (1999). Five methodology errors in educational research: A pantheon of statistical significance and other faux pas. In B. Thompson (Ed.), *Advances in social science methodology*, 5, 23-86.

## *Early Scholars* Improved Multiple Comparisons With The Best In Response Surface Methodology

Laura K. Miller    Ping Sa  
University of North Florida

---

A method to construct simultaneous confidence intervals about the difference in mean responses at the stationary point and at  $\mathbf{x}$  for all  $\mathbf{x}$  within a sphere with radius  $R_f$  is proposed. Results of an efficiency study to compare the new method and the existing method by Moore and Sa (1999) are provided.

Key words: Comparison with the best, response surface methodology, bounding algorithm.

---

### Introduction

Response surface methodology uses a polynomial response function to explain and analyze the relationship between a response variable  $y$  and several predictor variables  $\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_k)'$ . Usually the  $\mathbf{x}_i$  will be converted to coded variables  $x_i$  by  $x_i = (\mathbf{x}_i - \mathbf{x}_{i0}) / (sc)_i$ , where  $\mathbf{x}_{i0}$  is a centering constant and  $(sc)_i > 0$  is a scaling constant,  $i = 1, 2, \dots, k$ . The mean response at  $\mathbf{x} = (x_1, x_2, \dots, x_k)'$ ,  $E(y | \mathbf{x})$ , can be approximated using the quadratic polynomial model with  $k$  predictor variables

$$y = \mathbf{b}_0 + \sum_i \mathbf{b}_i x_i + \sum_i \mathbf{b}_{ii} x_i^2 + \sum_{i < j} \mathbf{b}_{ij} x_i x_j + \mathbf{e},$$

where  $\mathbf{b}_0, \mathbf{b}_i, \mathbf{b}_{ij}$  are unknown constants,  $i, j = 1, 2, \dots, k$  and random error  $\mathbf{e} \sim NID(0, \mathbf{s}^2)$ .

The mean response is optimized at the stationary point that may be a minimum, maximum, or a saddle point. After determining the levels of the predictor variables where the mean response is optimized, it is possible that this point is not a reasonable option due to practical considerations, such as expense. In this situation, multiple comparisons can be performed with other points in the region to determine if some other points provide responses that are not significantly different from the optimal point.

This problem will subsequently be referred to as multiple comparisons with the best (MCB) in response surface methodology (RSM). The MCB problem was first approached by Hsu (1984) in design of experiments where he considered the problem of comparing the treatment means under study with the "best" treatment mean. Moore and Sa (1999) first approached the MCB problem in the RSM setting. There has also been other substantial work on related problems within the field of response surface methodology. Sa and Edwards (1993) and Merchant, McCann, and Edwards (1998) investigated the multiple comparisons with the control (MCC) problem.

Sa and Edwards (1993) first addressed the MCC in RSM problem by constructing simultaneous confidence intervals for  $\mathbf{d}_c(\mathbf{x}) = E(y | \mathbf{x}) - E(y | \mathbf{0})$  for all  $\mathbf{x}$  within a

---

Laura K. Miller received her Master degree in Statistics from the University of North Florida. She was also the recipient of the Most Outstanding Statistics Graduate Student and the Most Outstanding Mathematics Undergraduate Student for University of North Florida in the year 2000 and 1998, respectively. Ping Sa is an Associate Professor of the Department of Mathematics and Statistics at the University of North Florida. She received the Ph.D. in Statistics from the University of South Carolina in 1990. She has published 15 papers. Her recent scholarly activities have involved research in multiple comparisons and quality control. Email address for correspondence regarding this article is psa@unf.edu.

pre-specified distance  $R_I$  of the origin, such that for all  $\mathbf{x}$ ,  $\mathbf{x}'\mathbf{x} = \sum_{i=1}^k x_i^2 \leq R_I^2$ , where  $R_I$  is the “radius of inference.” They showed that for a rotatable design, the bounds of  $\mathbf{d}_C(\mathbf{x}) \in \hat{\mathbf{d}}_C(\mathbf{x}) \pm (rF_{a,r,u})^{1/2}s(\mathbf{x})$  can be improved using a result of Casella and Strawderman (1980) where the Scheffé critical point,  $(rF_{a,r,u})^{1/2}$  can be replaced by a smaller value  $c_a$  depending on  $\mathbf{a}$ ,  $\mathbf{u}$ , and the nature of the predictor constraints as summarized by two other constraints, an integer  $m$  and a distance  $q^2 > 0$ .

Because the design used in practice is often not rotatable, Merchant, McCann, and Edwards (1998) introduced a new method which combined the Bonferroni method and the McCann and Edwards (1996) algorithm for two or more predictors that gives much sharper intervals than the Scheffé and also the Sa and Edwards (1993) adaptation of the Casella and Strawderman method. Merchant, McCann, and Edwards’ method does not require a rotatable design and allows for one-sided bounds for  $\mathbf{d}_C(\mathbf{x})$ . They generated a critical point  $d$  with simultaneous confidence bounds for

$$\mathbf{d}_C(\mathbf{x}) = E(y | \mathbf{x}) - E(y | \mathbf{0})$$

for all  $\mathbf{x}$  within a specified distance of  $\mathbf{0}$  via a bounding algorithm that requires only a few seconds to a few minutes of computer time.

Closely related and within the field of RSM, Moore and Sa (1999) addressed the MCB problem. They constructed confidence intervals about the difference in mean responses at the stationary point and alternate points over the entire  $k$  dimensional hyperplane based on a theory that does not depend on the design of the experiment. To solve the MCB problem, they utilized the delta method to approximate the variance of the estimated difference for

$$\begin{aligned} \mathbf{d}_B(\mathbf{x}) &= \mathbf{d}_B(\mathbf{x}, \mathbf{b}) \\ &= E(y | \mathbf{x}_0) - E(y | \mathbf{x}) \\ &= -\mathbf{x}'\mathbf{b} - \mathbf{x}'\mathbf{B}\mathbf{x} - \frac{1}{4}\mathbf{b}'\mathbf{B}^{-1}\mathbf{b}, \end{aligned}$$

where  $\mathbf{d}_B(\mathbf{x})$  represents the difference between the mean response at the stationary point  $\mathbf{x}_0 = -\frac{1}{2}\mathbf{B}^{-1}\mathbf{b}$  and the mean response at any other point  $\mathbf{x}$ ,  $\mathbf{b} = (\beta_1, \beta_2, \dots, \beta_k)'$ , and

$$\mathbf{B} = \begin{bmatrix} \beta_{11} & \frac{1}{2}\beta_{12} & \cdots & \frac{1}{2}\beta_{1k} \\ & \beta_{22} & \cdots & \frac{1}{2}\beta_{2k} \\ & & \ddots & \vdots \\ sym & & & \beta_{kk} \end{bmatrix}.$$

This confidence interval is useful in determining whether an alternate point could be substituted for the stationary point as an optimizer. Furthermore, it provides how much loss in the mean response can be expected if  $\mathbf{x}$  is moved away from  $\mathbf{x}_0$ . They investigated both Bonferroni and Scheffé type confidence intervals for the MCB problem. They also investigated Scheffé’s F-projection method of constructing conservative confidence intervals. However, the delta method is much less conservative than the F-projection method and of course, much easier to use.

It is the purpose of this article to address the MCB problem in RSM, but instead of considering the entire  $k$ -dimensional space, it would be more realistic to restrict the region to provide confidence bounds for  $\mathbf{d}_B(\mathbf{x})$  within a sphere with radius  $R_I$  for all  $\mathbf{x}$  such that  $\mathbf{x}'\mathbf{x} \leq R_I^2$ . The method proposed by Merchant, McCann, and Edwards (1998) for the MCC problem should be adaptable to the MCB problem since the requirement for using this method is that the covariance matrix of the estimators must be known.

The delta method will be used to approximate the variance of  $\hat{\mathbf{d}}_B(\mathbf{x})$  for the MCB problem. The next section explains the theory and the bounding algorithm used to generate the critical point for the MCB problem. The algorithm is design free, that is, it does not depend on the design of the experiment and should therefore provide consistent results regardless of the design.

### Theory Behind the Method

The method proposed by Merchant, McCann, and Edwards (1998) will be adapted to solve the MCB problem. The goal is to generate an improved critical point  $d$  with simultaneous upper confidence bounds of the form

$$\hat{\mathbf{d}}_B(\mathbf{x}) + ds(\mathbf{x}) \quad (1)$$

where

$$\hat{\mathbf{d}}_B(\mathbf{x}) = \hat{\mathbf{d}}_B(\mathbf{x}, \hat{\mathbf{b}}) = -\mathbf{x}'\hat{\mathbf{b}} - \mathbf{x}'\hat{\mathbf{B}}\mathbf{x} - \frac{1}{4}\hat{\mathbf{b}}'\hat{\mathbf{B}}^{-1}\hat{\mathbf{b}},$$

such that  $\hat{\mathbf{b}} = (\hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_k)'$  where the  $\hat{\beta}_i$ 's are the least square estimators for  $\beta_i$ 's and  $\hat{\mathbf{B}}$  is the matrix such that  $\hat{\beta}_{ij}$  is substituted into the matrix  $\mathbf{B}$ . The estimated standard error of  $\hat{\mathbf{d}}_B(\mathbf{x})$  is  $s(\mathbf{x}) = s(\mathbf{l}'\boldsymbol{\Sigma})^{1/2}$  derived by Moore and Sa (1999) where  $s^2$  is the mean square error which satisfies  $\mathbf{u}s^2/\mathbf{s}^2 \sim \mathbf{c}^2(\mathbf{u})$  for integer  $\mathbf{u} > 0$  and is independent of all  $\hat{\beta}_i$ 's,  $\hat{\mathbf{a}}$  is the  $(\mathbf{X}'\mathbf{X})^{-1}$  matrix with the first row and the first column deleted and  $\mathbf{l}$  is the vector of partial derivatives of  $\delta_B(\mathbf{x}, \beta)$  with respect to  $\mathbf{b}$  such that

$$\mathbf{l} = \left(-\frac{1}{2}m_1 - x_1, \dots, -\frac{1}{2}m_k - x_k, \frac{1}{4}m_1^2 - x_1^2, \dots, \frac{1}{4}m_k^2 - x_k^2, \frac{1}{4}m_1m_2 - x_1x_2, \dots, \frac{1}{4}m_{k-1}m_k - x_{k-1}x_k\right)'$$

where  $\mathbf{m} = \mathbf{B}^{-1}\mathbf{b}$  such that  $m_i$  is the  $i$ th component of  $\mathbf{m}$  and  $\mathbf{x} = (x_1, x_2, \dots, x_k)'$  is any point satisfying  $\mathbf{x}\boldsymbol{\alpha} \leq R_j^2$ .

In order to approximate the entire set of interest, we adapt the fine grid of inference  $G_I$ , suggested by Merchant, McCann, and Edwards (1998) of individual  $\mathbf{x}_j$  points for  $j = 1, 2, \dots, p$  in the region. This grid is constructed by user defined multiples of these  $\mathbf{x}$  values within a radius  $R_j$  radiating from the center  $\mathbf{0}$ . This matrix is defined as  $\mathbf{L}: p \times r$  whose  $j^{\text{th}}$  row is  $\mathbf{l}'_j = \mathbf{l}(\mathbf{x}_j)'$

where a simultaneous bound over this finite collection is calculated.

The critical point  $d$  must satisfy

$$P\left[\max_{\mathbf{x} \in G_I} \frac{\hat{\mathbf{d}}_B(\mathbf{x}) - \mathbf{d}_B(\mathbf{x})}{s(\mathbf{l}'\hat{\mathbf{a}}\mathbf{l})^{1/2}} \leq d\right] \geq 1 - \mathbf{a}. \quad (2)$$

For each  $\mathbf{x}$ ,  $\frac{\hat{\mathbf{d}}_B(\mathbf{x}) - \mathbf{d}_B(\mathbf{x})}{s(\mathbf{l}'\hat{\mathbf{a}}\mathbf{l})^{1/2}} \sim t_u$ , where

$t_u$  is the univariate- $t$  distribution with  $\mathbf{u}$  degrees of freedom. Equation (2) can be rewritten as

$$P[T_j \leq d, j = 1, 2, \dots, p] \geq 1 - \mathbf{a}$$

or

$$P[T_j > d, j = 1, 2, \dots, p] \leq \mathbf{a}$$

where  $T_1, T_2, \dots, T_p$  have a multivariate  $t$  distribution (Dunnnett & Sobel 1954) with  $\mathbf{u}$  degrees of freedom and underlying correlation matrix  $\mathbf{R}$  derived from  $\mathbf{s}^2\mathbf{L}\boldsymbol{\Sigma}\mathbf{L}'$ . The critical point  $d$  is then solved by the following equation,

$$\int_0^{1/d} P\{E(t)\} f_T(t) dt = \mathbf{a} \text{ for}$$

$$E(t) = \bigcup_{j=1}^p (\mathbf{a}'_j \mathbf{U} > td), \quad (\text{Brown, 1984})$$

where  $f_T$  is the probability density function of  $T$ , a random variable such that  $rT^2$  is distributed as  $F(\mathbf{u}, r)$ ;  $\mathbf{U}$  is a random vector independent of  $T$ , distributed uniformly on the  $r$ -dimensional sphere; and  $\mathbf{a}'_j$  are the rows of the full rank matrix  $\mathbf{A}: p \times r$  such that  $\mathbf{R} = \mathbf{A}\mathbf{A}'$ .

Finally, the probability  $P\{E(t)\}$  for the MCB problem can be calculated using the same bounding algorithm proposed by Merchant, McCann, and Edwards (1998) for the MCC problem. This bounding algorithm is a combination of Bonferroni method and the McCann and Edwards algorithm (1996) and is for upper bound only. If a lower bound is required

over this region, that is  $\hat{\mathbf{d}}_B(\mathbf{x}) - ds(\mathbf{x})$ , it can be computed by constructing the upper bound for  $-\mathbf{d}_B(\mathbf{x}) = \mathbf{E}(y | \mathbf{x}) - \mathbf{E}(y | \mathbf{x}_0)$  because an upper bound for  $-\mathbf{d}_B(\mathbf{x})$  is equivalent to a lower bound for  $\mathbf{d}_B(\mathbf{x})$ .

The critical points for the MCB problem were computed using a Fortran program and routines from the IMSL Fortran Numerical Libraries (1997). These included calling the routines DLINRG to calculate the inverse of a matrix, DLFTDS to compute the Cholesky factorization of a matrix, DFDF to evaluate the F distribution function, and DQDAGS to perform

the numerical integration. The Fortran program is available from the first author.

Examples and comparisons

Box and Draper (1987) give an example from an investigation by Derringer and Suich (1980) in which RSM is used to analyze the effects of  $\mathbf{x}_1$  = hydrated silica level in phr (parts per hundred) and  $\mathbf{x}_2$  = silane coupling agent level in phr on the elongation at break of a tire tread compound. One of the goals was to maximize  $y$  = elongation at break. Convert  $\mathbf{x}_1$  to the coded variable  $x_1 = (\mathbf{x}_1 - 1.2)/0.5$  and  $\mathbf{x}_2$  to the coded variable  $x_2 = (\mathbf{x}_2 - 50)/10$ . The design points  $\mathbf{x}_i$  and the responses  $y_i$  are listed in Table 1.

Table 1. Experimental Results: Elongation at break  $y$  of a tire tread compound Versus  $x_1$ =(phr silica - 1.2)/0.5 and  $x_2$ =(phr silane - 50)/10 (Source: Derringer 1980)

Run	$x_1$	$x_2$	$y$
1	-1	-1	900
2	1	-1	860
3	-1	1	800
4	1	1	2294
5	-1	-1	490
6	1	-1	1289
7	-1	1	1270
8	1	1	1090
9	-1.633	0	770
10	1.633	0	1690
11	0	-1.633	700
12	0	1.633	1540
13	0	0	2184
14	0	0	1784
15	0	0	1300
16	0	0	1300
17	0	0	1145
18	0	0	1090
19	0	0	1260
20	0	0	1344

The estimated polynomial response function is

$$\hat{y} = 1412.892 + 268.151x_1 + 246.503x_2 - 97.794x_1^2 - 139.044x_2^2 + 69.375x_1x_2 .$$

The vector  $\hat{\mathbf{b}} = (268.151, 246.503)'$ , the matrix  $\hat{\mathbf{B}} = \begin{pmatrix} -97.794 & 34.688 \\ 34.688 & -139.044 \end{pmatrix}$ , and the matrix

$$\hat{\mathbf{a}} = \begin{bmatrix} .075 & 0 & 0 & 0 & 0 \\ 0 & .075 & 0 & 0 & 0 \\ 0 & 0 & .075 & .005 & 0 \\ 0 & 0 & .005 & .075 & 0 \\ 0 & 0 & 0 & 0 & .125 \end{bmatrix}$$

are calculated.

The estimated stationary point for this surface where elongation at break ( $y$ ) is maximized is  $\hat{\mathbf{x}}_0 = (1.849, 1.348)$  yielding an estimated response of 1826.91. Figure 1 gives the estimated surface plot.

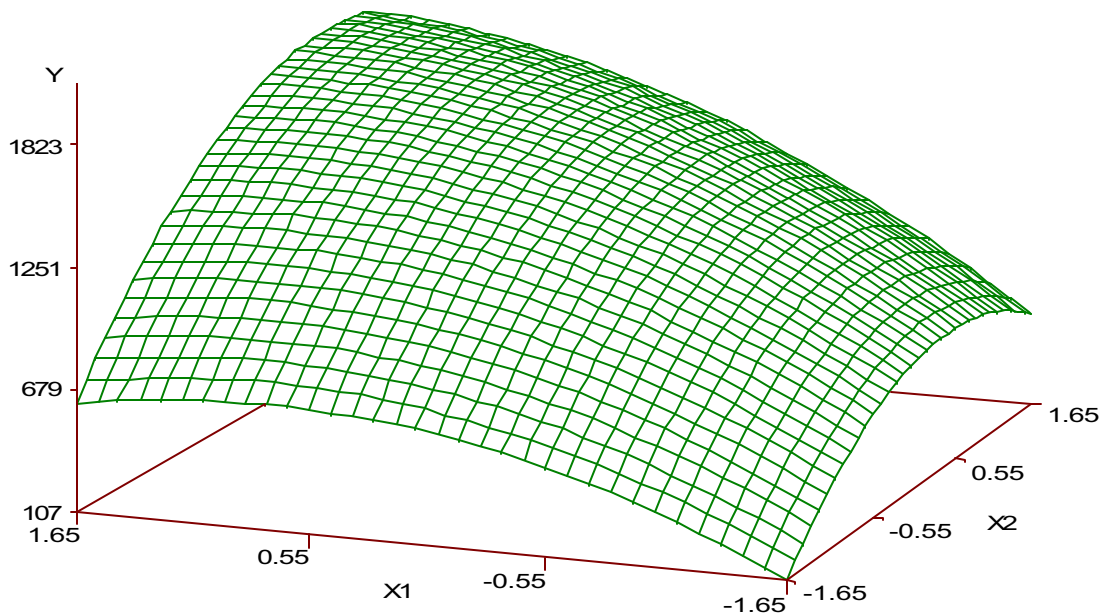


Figure 1. Estimated Surface Plot for the Tire Tread Compound Example.

As one can see, the stationary point is out of the experimental region and it may not be a reasonable option due to practical considerations or expense. Therefore, multiple comparisons can be performed with other points in a region to determine if any other point within the region of operability will produce a response that is not significantly different from the point that maximizes elongation at break ( $y$ ). Since the optimal point was a maximum, this suggests that lower bounds for  $\mathbf{d}_B(\mathbf{x}) = E(y | \mathbf{x}_0) - E(y | \mathbf{x})$  are more important than upper bounds.

Simultaneous 90% lower confidence bounds are constructed for  $\mathbf{d}_B(\mathbf{x})$  for all  $\mathbf{x}$  whose values are on the grid defined by multiples of .2 with a radius of  $R_f = \sqrt{2}$  radiating from the center of the region of interest. Figure 2 shows the contours for the estimated difference and the simultaneous lower confidence bounds  $L(\mathbf{x}) = \hat{\mathbf{d}}_B(\mathbf{x}) - ds(\mathbf{x})$  for generated  $d^2 = 6.871607$  by the method detailed in the previous section.



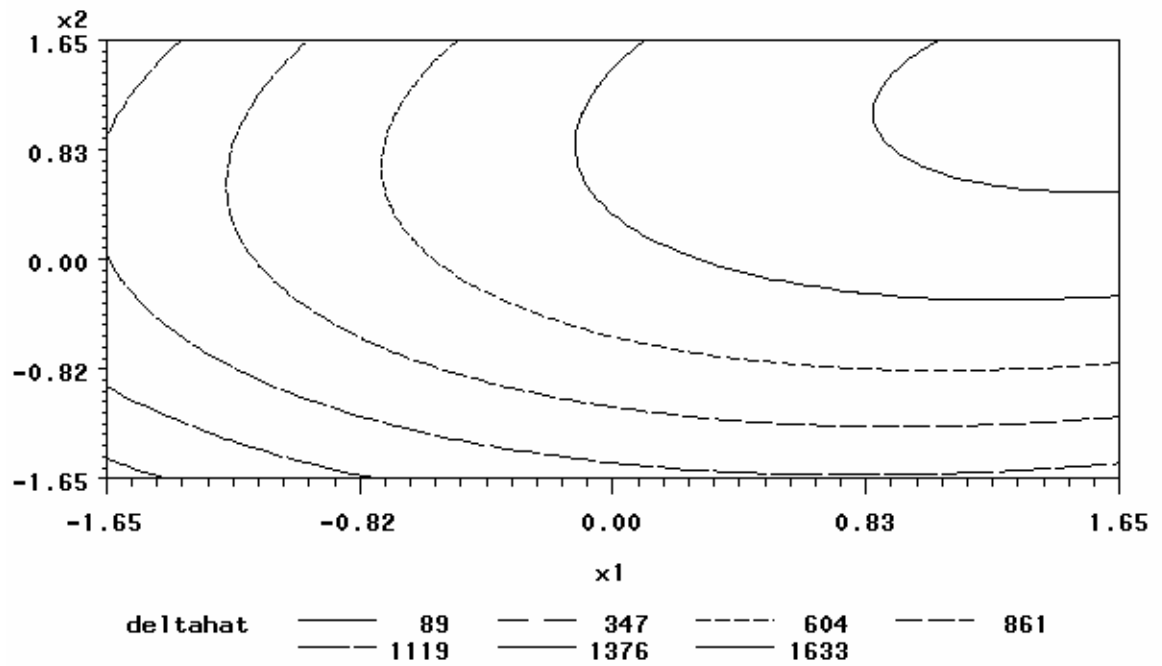
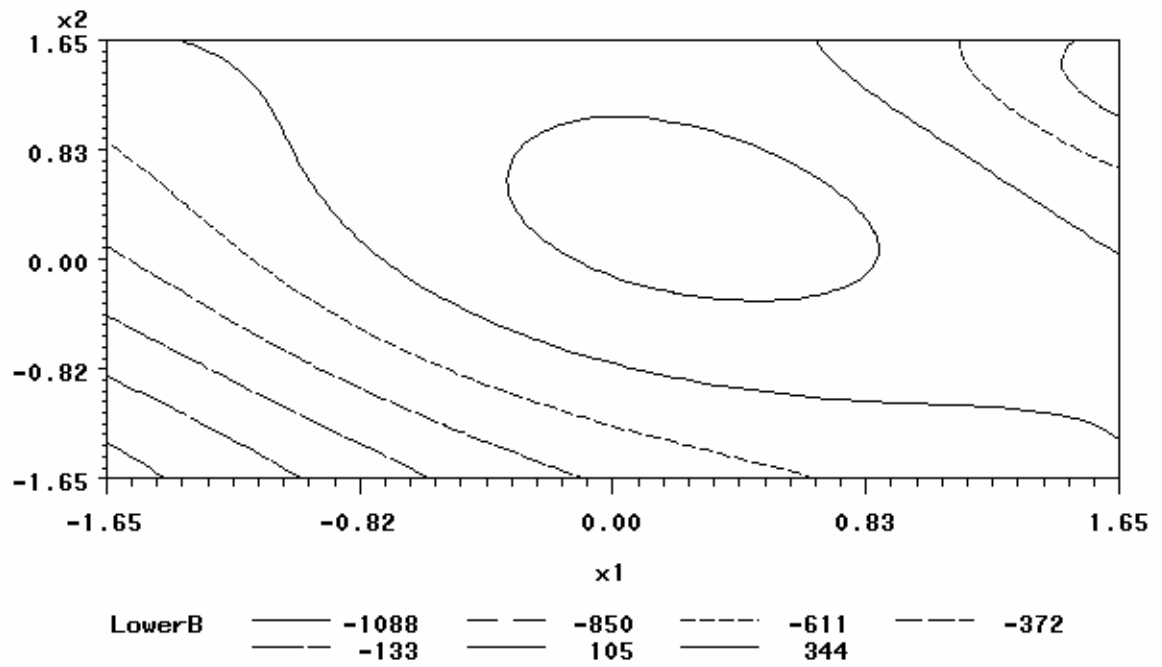


Figure 2. Contour Plots

In Figure 2 (above), the contour plots are for the estimated difference  $\hat{d}_B(\mathbf{x})$  (top) and the simultaneous 90% lower confidence bounds (bottom) for the Tire Tread Compound Example with generated  $d^2 = 6.871607$ . The points that lie inside the negative contour lines indicate possible alternate points that will produce responses that are not significantly different from the point that maximizes elongation at break ( $y$ ). The region inside the negative contour lines indicates possible region that will produce responses that are not significantly different from the point that maximizes elongation at break ( $y$ ).

The squared critical constant  $d^2 = 6.871607$  compares very favorably to that of the Scheffé method,  $rF_a(r, \mathbf{u}) = (\sqrt{5F_{.10}(5, 14)})^2 = 11.534702$ . Therefore, an experimenter using the Scheffé method would have to increase the experiment size by a factor of approximately  $(\sqrt{5F_{.10}(5, 14)})^2/d^2 = 11.534702 / 6.871607 = 1.6786$ , in other words, by 67.86% in order to achieve a precision of estimation (interval width) equal to what would be obtained using the adapted method's critical constant.

Next, three different designs will be used for an example using the bounding algorithm to generate improved critical points for the MCB problem where the sample-size savings will be compared to the Scheffé and Bonferroni critical points.

Khuri and Cornell (1987) provide an example in which they use RSM to investigate the effects of the amounts of two fertilizers,  $x_1$  and  $x_2$ , on the yield of peanuts measured in pounds per acre. For the purpose of the efficiency study, the estimated parameters from this example will be treated as the true parameters of an underlying model. The true quadratic response function is given by

$$y = 13.85 - .90x_1 + .56x_2 - 1.94x_1^2 - .78x_2^2 - .57x_1x_2 + \mathbf{e}$$

The vector  $\mathbf{b} = (.90, .56)'$ , the matrix

$$\mathbf{B} = \begin{bmatrix} -1.94 & -.285 \\ -.285 & -.78 \end{bmatrix}, \text{ and the matrix}$$

$$\hat{\mathbf{a}} = \begin{bmatrix} .125 & 0 & 0 & 0 & 0 \\ 0 & .125 & 0 & 0 & 0 \\ 0 & 0 & .144 & .019 & 0 \\ 0 & 0 & .019 & .144 & 0 \\ 0 & 0 & 0 & 0 & .25 \end{bmatrix} \text{ are found.}$$

The stationary point for this surface is  $\mathbf{x}_0 = (.189, .290)$  yielding a response of 13.676. Assume that this option is not a reasonable option, multiple comparisons are performed to determine if alternate points can substitute for the stationary point in terms of maximizing peanut yield. Therefore, the critical point  $d$  is required to perform these comparisons.

Three central composite designs were chosen. The three designs are a rotatable central composite design with uniform precision, a rotatable central composite design without uniform precision, and a central composite design with one centerpoint. These designs will be referred to as Design 1, 2, and 3 respectively.

Table 2 (following References section) shows the critical points that were generated for one and two replications of the three different designs using multiples of .2,  $R_t = 1$ , and  $\sqrt{2}$  for this example using Merchant, McCann, and Edwards' (1998) method in order to compare the critical values and the approximate sample-size savings for each design. Because the Bonferroni method is conservative due to the large number of comparisons, only the approximate sample-size savings vs the Scheffé method were calculated.

Considerable improvement (between 34% and 47%) over the Scheffé adaptation for all three designs is possible using the new method by choosing the radius of inference  $R_t = 1$ . For  $R_t = \sqrt{2}$  (which is near the limits of the experimental region for Designs 1 and 2), the sample-size savings are 26% to 33% over the Scheffé method. Also, as expected, the increased sample sizes produced by replicating the designs resulted in smaller critical values.

## Conclusion

In conclusion, this article has addressed the problem of multiple comparisons with the best in RSM via simultaneous confidence bounds for  $d_B(\mathbf{x}) = E(y | \mathbf{x}_0) - E(y | \mathbf{x})$  for all  $\mathbf{x}$  such that  $\mathbf{x}'\mathbf{x} \leq R_1^2$ . The method proposed by Merchant,

McCann, and Edwards (1998) for the MCC problem has been adapted to the MCB problem. It has provided confidence bounds for an example for two predictors where the critical values compare favorably to the Bonferroni and Scheffé methods as shown by Table 2 (following page).

This will also hold true for problems containing more than two predictor variables. For the example provided, this method has been shown to provide approximate sample-size savings of at least 25% for three different central composite designs. In fact, based on the theory behind the bounding algorithm, the Merchant, McCann, and Edwards' method for the MCB problem will always outperform the Scheffé and Bonferroni methods (Merchant, McCann, and Edwards, 1998).

## References

- Box, G. E. P. & Draper, N. R. (1987), *Empirical model-building and response surfaces*, New York: John Wiley.
- Brown, L. D. (1984), A note on the Tukey-Kramer procedure for pairwise comparisons of correlated means, in *Design of experiments: ranking and selection*, eds. T. J. Santner & A. C. Tamhane, New York: Marcel Dekker, pp.1-6.
- Casella, G., & Strawderman, W. E. (1980), Confidence bands for linear regression with restricted predictor variables, *Journal of the American Statistical Association*, 75, 862-868.
- Derringer, G. C. & Suich, R. (1980), Simultaneous optimization of several response variables, *Journal of Quality Technology*, 12, 214-219.
- Dunnett, C. W., & Sobel, M. (1954), A bivariate generalization of student's *t* distribution, with tables for certain special cases, *Biometrika*, 31, 153-169.
- Hsu, J. C. (1984), Constrained simultaneous confidence intervals for multiple comparisons with the best, *Annals of Statistics*, 12, 1136-1144.
- International Mathematical & Statistical Libraries, Inc. (1997), *Fortran routines for mathematical applications*, Visual Numerics, Inc.
- Khuri, A. I., & Cornell, J. A. (1987), *Response surfaces*, New York: Marcel Dekker.
- McCann, M., & Edwards, D. (1996), A path-length inequality for the multivariate *t* distribution, *Journal of the American Statistical Association*, 91, 211-216.
- Merchant, A., McCann, M., & Edwards, D. (1998), Improved multiple comparisons with a control in response surface analysis, *Technometrics*, 40, 297-303.
- Moore, L. J., & Sa, P. (1999), Comparisons with the best in response surface methodology, *Statistics and Probability Letters*, 44, 189-194.
- Sa, P. & Edwards, D. (1993), Multiple comparisons with a control in response surface methodology, *Technometrics*, 35, 436-445.

Table 2. Generated critical points using the improved method (one-sided bounds) and Scheffé and Bonferroni critical points with grid spacing = .2 for each design and approximate Sample-Size Savings of the New Method Versus the Scheffé method.

Design	$R_i$	reps	$u$	Improved Critical Point $d$	Scheffé Critical Point	Bonferroni Critical Point	Sample-Size Savings vs Scheffé
1	1	1	7	3.233020	3.794733	4.605120	37.78%
		2	20	2.803443	3.286335	3.460804	37.42%
	$\sqrt{2}$	1	7	3.368990	3.794733	5.207830	26.87%
		2	20	2.918108	3.286335	3.756539	26.83%
2	1	1	3	4.423153	5.152669	9.505157	35.71%
		2	12	2.980882	3.456877	3.813342	34.49%
	$\sqrt{2}$	1	3	4.576950	5.152669	12.008948	26.74%
		2	12	3.079915	3.456877	4.195280	25.98%
3	1	1	3	4.255092	5.152669	9.505157	46.64%
		2	12	2.875548	3.456877	3.813342	44.52%
	$\sqrt{2}$	1	3	4.482340	5.152669	12.008948	32.15%
		2	12	3.018781	3.456877	4.195280	31.13%

## A Semiparametric Regression Model For Oligonucleotide Arrays

Jianhua Hu  
Department of Biostatistics  
University of North Carolina

Guosheng Yin  
Department of Biostatistics  
M. D. Anderson Cancer Center

---

A semiparametric model incorporating the spline smoothing technique is proposed to study oligonucleotide gene expression data. No specific parametric functional form is assumed for mismatch probe intensities, which allows much more flexibility in the fitted model. The new approach improves the model fitting, hence the estimation of expression indexes. The method is applied to a data set of 18 HuGeneFL arrays.

Key words: Affymetrix, gene expression, microarray, semiparametric spline smoothing

---

### Introduction

DNA microarray technologies have been increasingly used and began to play an important role in many areas of biomedical research. There are two most popular types, namely cDNA microarrays and oligonucleotide arrays. The common advantages of them are to monitor the expression levels of very large numbers of genes simultaneously and repeatedly in cell lines, human tissues and a wide range of organisms. Microarrays have the potential and power to advance our knowledge and understanding at a genomic scale. In particular, the high-density oligonucleotide array has been shown to be very promising. Not only does it have the capability of monitoring all yeast genes, mouse and human genes, but it also can identify important genes and classify disease types or states reliably, due to its special design feature.

The distinctive feature of the oligonucleotide array technology is the effective utilization of the probe redundancy. Multiple oligonucleotides of different sequences are hybridized onto different regions of the same RNA that are complementary to the oligonucleotides.

It offers us the possibility to test and examine the stability and reliability of gene expression measurements (outlier detection), improve the accuracy of RNA quantification, reduce cross-hybridization effects, and thus reduce the measurement noise and false-positive percentages. Usually, a probe set of around 20 pairs of a particular length (25 nucleotides typically) represents a gene uniquely (Lockhart et al., 1996).

The other source of redundancy is that mismatch (MM) probes are used, which are identical to their correspondent perfect match (PM) except for a single base mutated at the central position (13th position typically). The MM probes can provide some information on background and cross-hybridization signals, and provide the ability to discriminate between “real” signals and those due to non-specific or semi-specific hybridization (Lipshutz et al., 1999). In other words, the design of oligonucleotide arrays with PM/MM probe sets can improve the differentiating ability over the cDNA arrays that use a single spot. It can help to distinguish whether a signal detected is really due to the hybridization onto the intended RNA region or it happens just by chance due to cross-hybridization or other measurement errors.

Obtaining an accurate gene expression index is essential and fundamental for further research and analysis of oligonucleotide arrays, such as differentiating important genes, classifying genes to co-regulated or anti-coregulated groups and categorizing samples. Hence, it is very

---

Correspondence should be sent to: Jianhua Hu, graduate student, Department of Biostatistics, McGavran-Greenberg Hall, CB# 7420, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599-7420. Phone: 919-966-7287. Email: [jhu@bios.unc.edu](mailto:jhu@bios.unc.edu).

important to develop some methodologies to estimate the gene expression indexes as accurately as possible.

In recent years, various statistical methods have been proposed for analyzing oligonucleotide arrays. For example, the GeneChip software computes the “average difference” (AD) (<http://www.genechip.org/index.affx>).

Affymetrix's average log ratio is based on  $\log(\text{PM}/\text{MM})$  where the log transformation may be helpful in reducing the skewness and the variation. Li and Wong (2001) proposed a parametric regression model to calculate the model-based expression indexes (MBEI) based on the difference (PM-MM). It can improve the fitness of hybridization intensity extracted from PM and MM, and model the probe effects explicitly. Also, MBEIs are closer to the underlying true gene expression indexes than those provided by most of other software. The way of dealing with the relationship between PM and MM for almost all the above methods is to subtract MM from PM or  $\log(\text{MM})$  from  $\log(\text{PM})$  directly. The model based on (PM-MM) assumes a linear relationship between PM and MM and the regression

coefficient of MM equals one. Although the old Affymetrix pre-5.0 algorithm claims that there is a linear relationship between most PM and MM probes, there are still a certain amount of probes with nonlinearity. Better fitting models to these genes are desired in order to avoid missing some important biological information.

In practice, the paired PM and MM probe expression levels may not be linearly correlated for a specific probe set (Schadt et al., 2001). As shown in Figure 1, we randomly chose the probe set 17 of Gene 2111 and obtained the scatter plot of PM versus MM intensity levels with a smoothing spline curve fitted after normalization. It is clear that the relationship between PM and MM is not simply linear and some curvature pattern needs to be addressed. For the same gene, we also plotted  $\log(\text{PM})$  versus  $\log(\text{MM})$  with a smoothing spline fit. Although the log transformation helps clarify the pattern between them, there is still a curve trend. Therefore, there may be some excess non-linearity that cannot be captured by the parametric model simply based on (PM-MM).

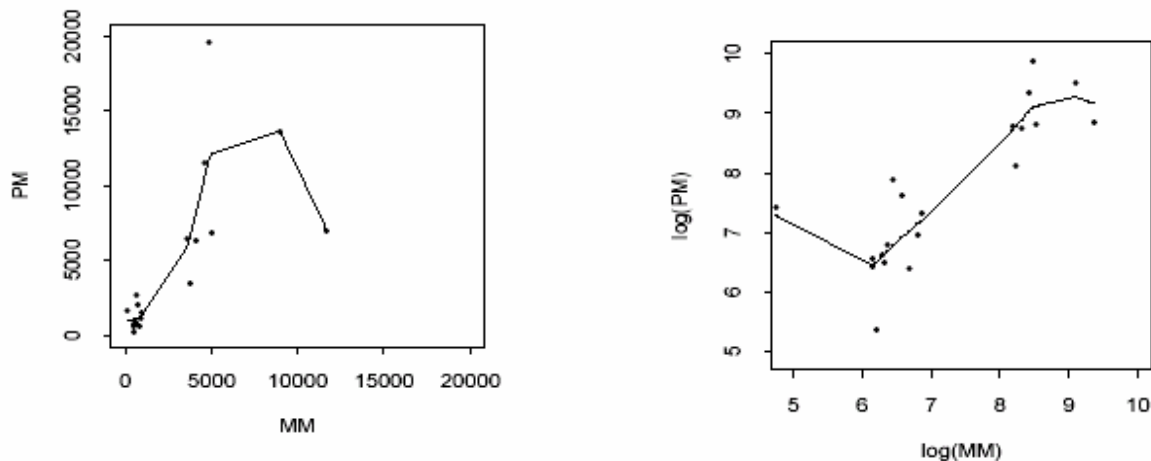


Figure 1: Smoothing spline fitting curves of PM versus MM and  $\log(\text{PM})$  versus  $\log(\text{MM})$  for probe set 17 of Gene 2111.

Another notable feature is that it is not rare for MM to be bigger than PM expression intensities after some are removed as outliers. The old Affymetrix pre-5.0 algorithm sets the expression levels of probes to be positive only if  $\text{PM}-\text{MM} \geq \text{SDT}$  or  $\text{PM}/\text{MM} \geq \text{SRT}$ , where SDT

is the statistical difference threshold and SRT is the statistical ratio threshold. By this brutal truncation, it throws away many probes such that some useful biological information might be lost. Current Affymetrix MAS 5.0 handles this situation by setting MM always lower than its paired PM, which is similar to the approach of truncation

(Irizarry et al., 2001). But in many situations, the phenomenon of intensities of MM larger than PM may be caused by some sensible biological variations. Thus researchers still want to keep the features in the data analysis. Moreover, the algorithm is not as flexible and adjustable as model-based approaches.

Li-Wong's reduced model has been proved to be simple, feasible and popular with collaborating biologists and have several aspects of superior behavior. It can produce better estimation for the gene expression indexes, which is one of the most critical steps for further analysis. Since MM probes are used to eliminate the background and hybridization noise as much as possible, the one of most interest to researchers is still PM probes. Validity and goodness-of-fit of a model is essential to obtain accurate parameter estimates and statistical inferences.

We propose a semiparametric regression model to study PM probes with adjustment for MM probes in this article. After normalizations and dropping outliers, we keep the original feature for each gene and seek to obtain a better model-fitting by capturing the nonlinear relationship between PM and MM probes with a semiparametric approach based on Li-Wong's reduced model. We do not assume any parametric functional form of MM while the multiplicative relationship between the gene expression index ( $q$ ) and the increasing rate (the probe sensitivity index,  $f$ ) is still kept as in Li-Wong's reduced model. The approach involves three stages and relaxes the restriction of the regression coefficient of PM on MM being one, which is completely data-driven. We apply the proposed method to the analysis of HuGeneFL oligonucleotide arrays for Antibody Stain CEL data (<http://thinker.med.ohio-state.edu/projects/fbss/index.html>).

### Methodology

Let  $\theta_i$  be the expression index for the gene in the  $i$ th sample which is the primary target of interest. The full model proposed by Li and Wong (2001) for each gene is given by

$$\begin{aligned} PM_{ij} &= \mathbf{n}_j + \mathbf{q}_i(\mathbf{a}_j + \mathbf{f}_j) + \mathbf{e}_{ij} \\ MM_{ij} &= \mathbf{n}_j + \mathbf{q}_i\mathbf{a}_j + \mathbf{e}_{ij}, \end{aligned} \quad (1)$$

where  $PM_{ij}$  and  $MM_{ij}$  are the PM and MM intensity values for the  $i$ th array and the  $j$ th probe pair for this gene,  $i=1, \dots, I; j=1, \dots, J$ . Note that  $\mathbf{n}_j$  is the reference response due to nonspecific hybridization,  $\mathbf{a}_j$  is the increasing rate of MM response,  $\mathbf{f}_j$  is the additional increasing rate of PM response, and  $\mathbf{e}_{ij}$  represents a random error. There are many parameters in the full model, whereas a parsimonious statistical model may be preferred with the smaller sample size. A simpler reduced model (LWR) for the difference PM-MM is strongly supported by collaborating biologists. The model is given by

$$PM_{ij} - MM_{ij} = \mathbf{q}_i\mathbf{f}_j + \mathbf{e}_{ij} \quad (2)$$

It states that the PM and MM intensity differences have a multiplicative relation between  $q$  and  $f$ .

For the purpose of identifiability, a constraint is set as  $\sum_j \mathbf{f}_j^2 = J$ . The error terms are assumed to be independent and identically normally distributed, i.e.  $\mathbf{e}_{ij} \sim N(0, \mathbf{s}^2)$ . Depending on the value of  $\mathbf{f}_j$ , the least square estimate for  $q_i$  is

$$\hat{\theta}_i = \sum_j \frac{(PM_{ij} - MM_{ij})f_j}{J} \quad (3)$$

and the approximate standard error is given by,

$$\begin{aligned} S.E.(\hat{\theta}_i) &= \sqrt{\sigma^2/J}, \\ \sigma^2 &= \sum_j (\text{fitted} - \text{observed})^2 / (J - 1) \end{aligned} \quad (4)$$

An iterative least square algorithm is carried out for the estimation of the parameters. A software DNA-Chip Analyzer (dChip) has been developed to fit the parametric regression model that Li and Wong proposed (<http://www.dchip.org/>).

However, Li-Wong's reduced model (LWR) is analogous to the usual regression model for the difference between the pre-treatment (baseline) and post-treatment effects in clinical trials. In some sense, it forces the regression parameter of MM to be one which is a very stringent restriction and may affect the goodness-of-fit of the model tremendously. Moreover, there

is strong evidence of a non-linear relationship between PM and MM intensities (see Figure 1). Therefore, we propose a semiparametric approach to model the expression intensity data for each gene. Inspired by the additive partially linear models (Heckman, 1986; Hastie & Tibshirani, 1990), we model MM based on a nonparametric spline smoothing technique (LWS), namely,

$$PM_{ij} = g(MM_{ij}) + \mathbf{q}_i \mathbf{f}_j + \mathbf{e}_{ij} \quad (5)$$

where  $g(\cdot)$  is an unknown smooth function and is estimated with the cubic spline smoothing method. In many instances, rather than modeling every covariate nonparametrically or parametrically, a semiparametric partially linear regression model is more desirable. The model specification for the oligonucleotide array data is particularly appealing since the gene expression index  $\mathbf{q}$  is the major interest, while the effects of MM are nuisance.

We can draw statistical inferences and estimate  $\mathbf{q}$  by making minimal assumptions about the effects of MM with a fully nonparametric function. LWR does not have the same computational issue (too many parameters for sample sizes of practical use) as Li-Wong's full model that involves too many parameters. Basically, we relax the relationship between PM and MM to get a better fitted model and expect to have a more accurate estimate of the expression indexes. Hence, it is practically applicable to oligonucleotide gene expression data analysis.

Our estimating procedure involving three stages of iterative algorithms is described as follows:

**Stage 1:** Take LWR estimates as the initial values of  $\mathbf{q}_i^{(0)}$  and  $\mathbf{f}_j^{(0)}$ . Note that LWR itself iteratively fits the sets of  $\mathbf{q}_i$  and  $\mathbf{f}_j$  while treating one of the two sets as known and fixed. We calculate the initial values using the dChip software.

**Stage 2:** Use the cubic spline smoothing technique to fit a nonparametric model with  $PM_{ij} - \mathbf{q}_i^{(0)} \mathbf{f}_j^{(0)}$  as the response and  $MM_{ij}$  as the predictor, and thereby get the predicted values of  $\hat{g}(MM_{ij})$ .

**Stage 3:** Calculate the updated PM values  $PM_{ij}^{\text{new}} = PM_{ij}^{\text{old}} - \hat{g}(MM_{ij})$ , then regress the new estimates of PM on  $\mathbf{q}$ 's and  $\mathbf{f}$ 's, namely,  $PM_{ij}^{\text{new}} = \mathbf{q}_i \mathbf{f}_j + \mathbf{e}_{ij}$ . The new estimates of  $\mathbf{q}$ 's and  $\mathbf{f}$ 's have been obtained. Go back to Stage 2, and continue till the prescribed convergence criteria are met.

Spline smoothing methods consisting of piecewise cubic polynomials are popular because they provide great flexibility for fitting the data and model non-linearities without specifying a functional form, with fewer parameters than higher-degree splines. To reduce the undesirable instability in the tails, one may restrict the function to be linear before the first knot and after the last knot. Fitting a cubic spline model which minimizes the residual sum of squares while

$$\sum_{i=1}^n \{y_i - g(x_i)\}^2 + \lambda \int \{g''(x)\}^2 dx \quad (6)$$

adjusting the smoothness of the resulting spline can be achieved by minimizing the penalized residual sum of squares

The smoothing parameter  $\lambda$  controls the trade-off between bias and variance and may be estimated by the cross-validation procedure. Excellent reviews of nonparametric regression and spline smoothing are available in the literature (Silverman, 1985; Eubank 1999).

## Results

### Description of Experiment and Data set

The data set is from an experiment conducted by the Division of Human Cancer Genetics at the Ohio State University (Lemon et al., 2002). There are 18 HuGeneFL arrays, each of which was loaded with 11 ug/200uL labeled cRNA. As shown by the graph in the Appendix, the process is described as the following. Human fibroblast cells were grown in media supplemented with 20% FBS for 5 passages (27 flasks) according to the distributor's recommendations. After 48 hours of placing cultures in serum-reduced media (0.1% FBS), 9 flasks (Stimulated) were returned to a 20% serum condition for 24 more hours and were then placed in RNA-Stat60. Cells from the other flasks (Starved) were placed in RNA-Stat60 directly after being placed in



serum-reduced media for 48 hours. Finally total RNA was extracted and purified according to a certain criterion. Based on the above steps, a set of stimulated and starved samples is produced. Another RNA sample was produced as a balanced mixture of simulated and starved samples, which is called the 50:50 sample.

For each condition (serum stimulated, serum starved and a 50:50 mixture of serum stimulated and starved), two aliquots of RNA were drawn and processed separately on three consequent days. Meanwhile, spiked-in genes were added in the following way: *Lys* and *Phe* RNAs at 0.08 ng/8 $\mu$ g total RNA were added to Stimulated RNA samples. The Starved samples received the same amount of *Dap* and *Thr* and all the four spiked-in genes at 0.04 ng/8 $\mu$ g were assigned to the 50:50 samples. Another set of control genes were added as well, which were *BioB*, *BioC*, *BioD* and *Cre* with final concentrations of 1.5, 5, 25 and 100 pM, respectively. For each group (Stimulated, Starved and 50:50), six replicated HuGeneFL arrays were produced. Eighteen arrays were produced in total. The technical variability was minimized through using a single fluidics station and a same lot for the 18 arrays. Multiple experiments or arrays for each gene allows researchers to evaluate the potentially different variability of genes.

There are 7129 probe sets in each array. Among them, a total of 149 genes are represented twice or more although they might not be in the same probe set. Most of the probe sets have 20 probe pairs. However, there are 330 probe sets with probe pairs less or more than 20. To compare Li-Wong's reduced model with our new proposal, the 330 probe sets were left out without losing any practical meaning.

The experimental design has very appealing features that the relationship among the arrays are known in advance and control genes are spiked in. Hence, it is suitable to use the data set to make comparisons among different estimation approaches.

#### Normalization, Variance and Goodness-of-fit

Because scanned images may have different overall brightness, it is important to

normalize arrays such that they have comparable brightness before any analysis on expression levels. A traditional Average Difference (AD) method analyzes one array at a time, thus normalization among the different arrays can be done after calculating the quantities of interest. Because the model-based expression index analysis involves different arrays simultaneously, the comparable brightness of the arrays needs to be assured. As a very important issue, normalization has been extensively discussed and studied in the literature, and it is still an active research area.

We use the normalization method based on an "invariant set" (Li & Wong, 2001; Schadt et al., 2002). Normalization is based on probe values of non-differentially expressed genes that are identified through an iterative procedure (called the "invariant set"). Keeping the array which has the median overall brightness (the baseline array) as the invariant one, all the other arrays are normalized to it. The two arrays are drawn on the y-axis and x-axis, respectively. A straight line through the origin or a curve (i.e. smoothing spline) is fitted to the scattered points, which shows the normalization relationship between the two arrays.

If the variance of the model based expression index is overestimated, it may be possible not to differentiate some important genes that are supposed to express significantly, especially for genes with low expression levels. Hence, the model which yields smaller variances of the estimated expression indexes is desired. On average, LWS reduces the standard error of  $\theta$  by 22% with respect to LWR. It indicates that LWS gives the more stable estimated expression index in terms of the 20 probe pairs than LWR. Figure 2 shows the histogram plots of standard errors of all the expression index estimates from both LWR and LWS. Obviously there are shifting differences between the distributions of S.E.'s from the two models (LWR and LWS). Most of the S.E.'s from LWS are within the range of (0, 500) while those from LWR even exceed beyond 1000.

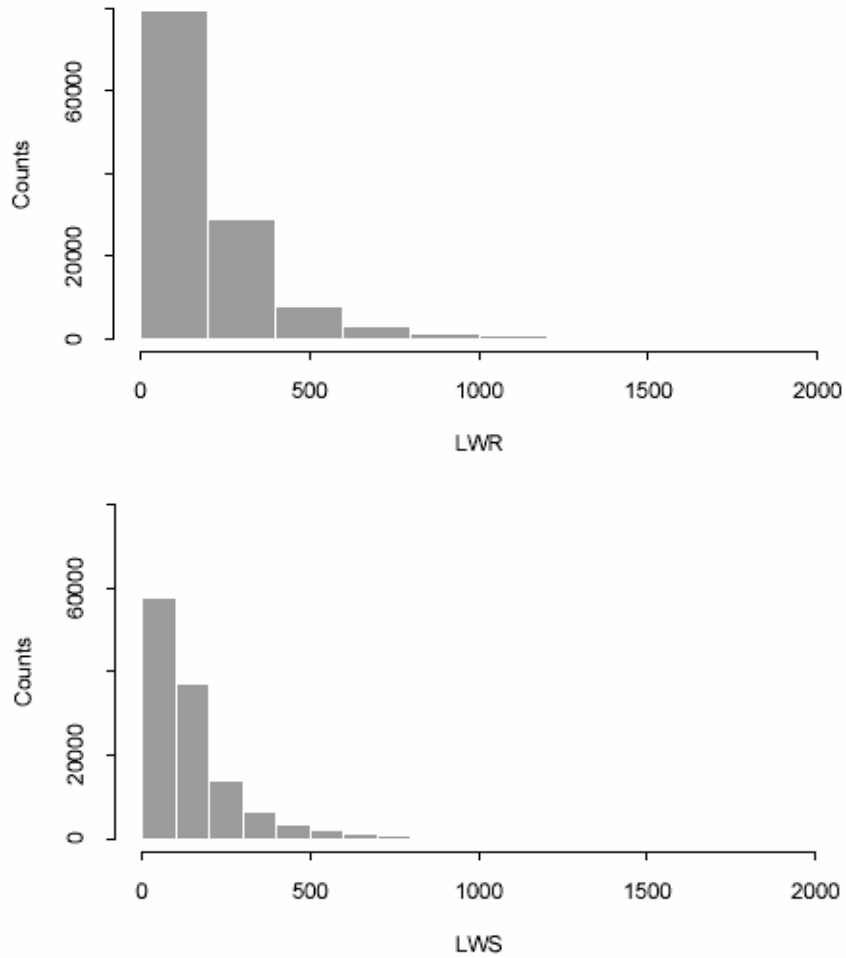


Figure 2: Histograms for standard errors of the estimated gene expression indexes.

Figure 3 presents the plot of residues of the fitted model versus predicted values for Gene 1007 (chosen randomly) from the two models, respectively. The horizontal line is the reference with the residue being zero. It is clear that the scatter plot from LWS gives a more random and symmetric pattern around the reference line, while LWR has more points further deviated away from the zero-line.

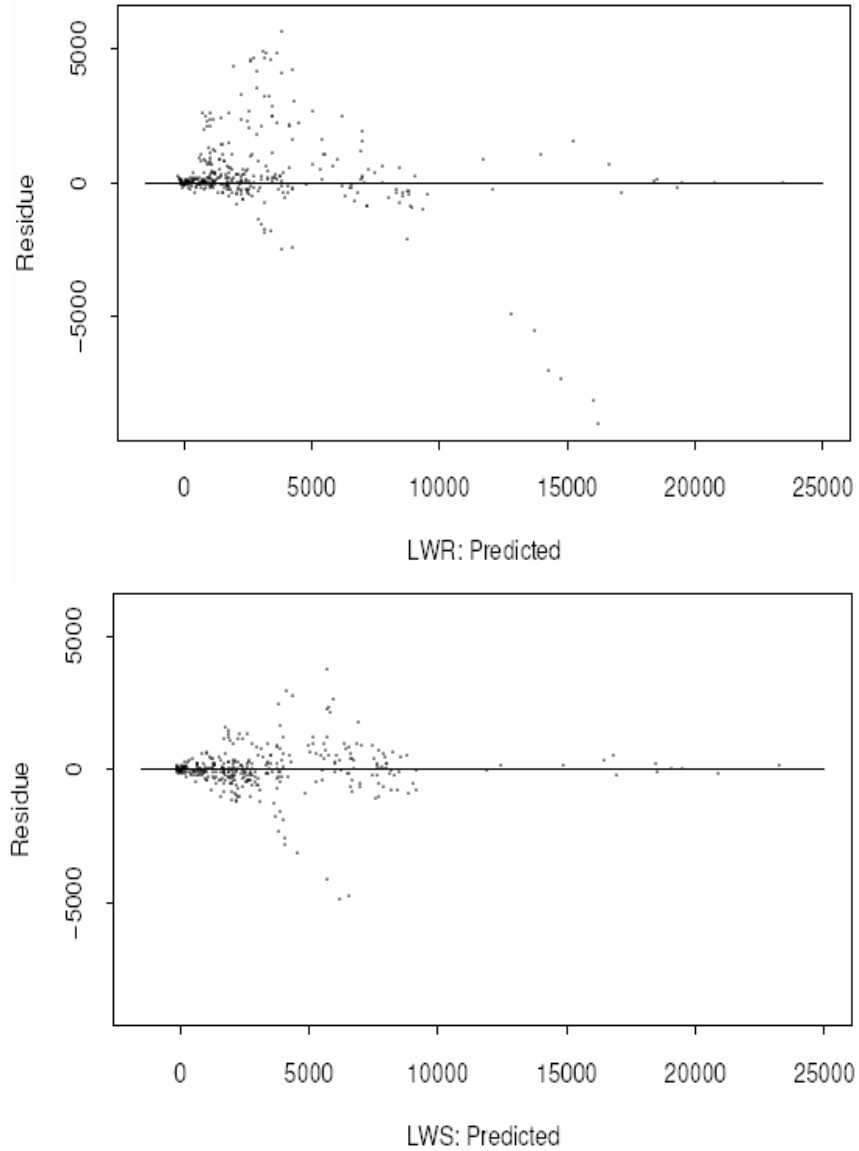


Figure 3: Residuals versus predicted values for gene 1007.

The better the model fits, the higher correlation of the predicted and observed PM values is supposed to be. Thus, correlations for all the probe sets are calculated for LWR and LWS. The histograms of the correlations obtained from the two models are shown in Figure 4, respectively. Note that most of the correlations obtained from LWS concentrate within 0.92 to 1, while the correlations from LWR even go below 0.90.

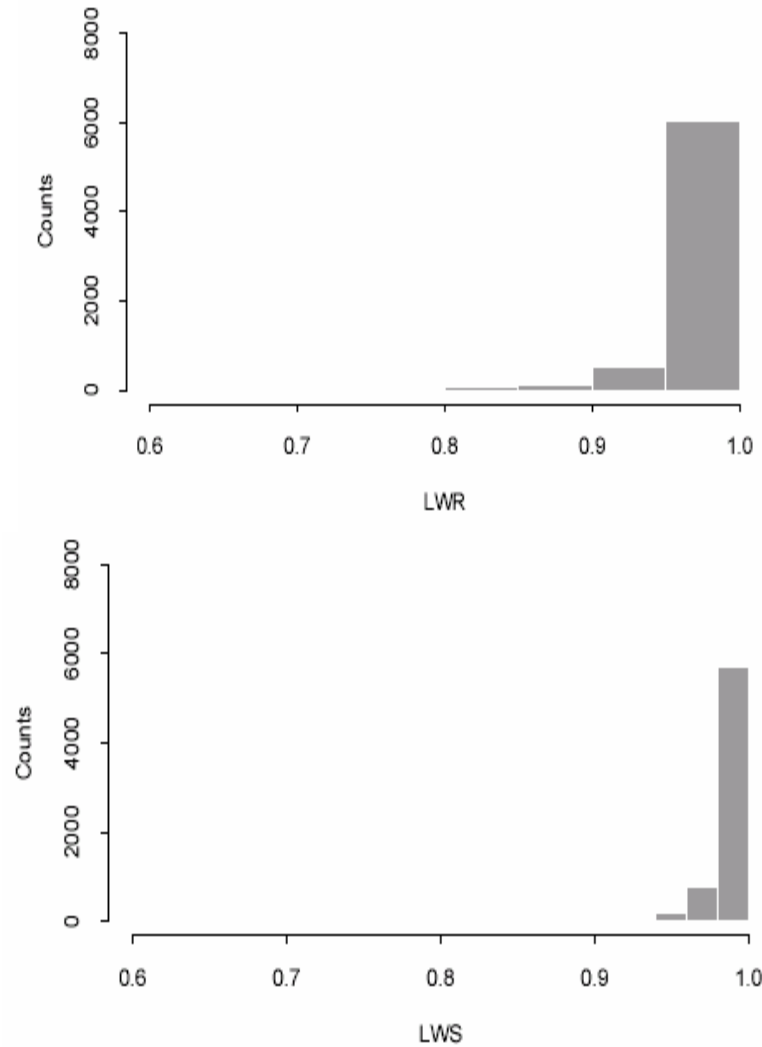


Figure 4: Histograms for correlations between observed and predicted PM intensities.

During the three consecutive days of the experiment, six replicated arrays for each group (Stimulated, 50:50, Starved) were produced. The manufacturing process and analytical methods, including normalization, assure the biological variation among the six independent arrays as low as possible. The variation of the gene expression indexes across the six replicated arrays may serve as a good statistic for comparing the two different regression models. A better model should be able to produce a smaller variation of the gene expression indexes among the six duplicates. In Table 1, the simple descriptive statistics of the sample variances of the expression indexes among

the six arrays in each condition (Stimulated, 50:50, Starved) are given to compare LWR with LWS. The result shows that the relationship generally holds that  $\text{Var}(\hat{q}_{LWS}) < \text{Var}(\hat{q}_{LWR})$ . In the Stimulated and 50:50 conditions, LWS yields much smaller variation among the six replicated arrays than LWR, while LWR and LWS perform roughly the same at the Starved condition. In other words, LWS gives more stable results such that the expression indexes from the six arrays in each condition (Stimulated, 50:50 and Starved) have a smaller variation than LWR.

Table 1: Descriptive statistics of sample variances among six arrays at each condition.

	Stimulated		50:50		Starved	
	LWR	LWS	LWR	LWS	LWR	LWS
Minimum	143.620	37.663	39.357	77.091	29.797	8.655
Maximum	3.029e7	3.653e7	1.210e8	8.310e7	6.217e7	4.129e7
Median	89881.3	76892.1	138872.6	118526.1	98829.2	88985.3
Mean	384779.9	369334.5	414866.1	395693.4	327800.8	329822.8

Assessing Gene Expression Measurements

In the experiment, the genes *Lys* and *Phe* were not spiked in starved samples, while *Dap* and *Thr* were not in stimulated sample. Therefore, 12 probe sets and 18 samples of the four spiked-in genes are known to be expressed or not in advance. Totally 144 probe sets should be expressed and 72 should be unexpressed. We obtained the number of expressed and unexpressed genes using the criterion of  $\hat{q}/S.E.(\hat{q}) > 6.0$ . The two methods (LWR and LWS) can detect the same number of expressed probe sets (132) and unexpressed probe sets (66). However, regarding the median standard error of the control probe sets, LWS gives a much smaller variation (S.E. of 177.2) of the estimated expression indexes than LWR (S.E. of 307.9). Hence, LWS is more reliable and stable for the estimation of the gene expression indexes.

Focusing on the four spiked-in genes, each gene known to be unexpressed should have a rank as low as possible among all the control genes. One probe set of *Thr* in a Stimulated condition that should be unexpressed has a unexpectedly high expression level. It is considered as an outlier and left out from our analysis. After averaging the expression indexes of each spiked-in probe set over their own six replicated arrays and calculating their ranks, the results are shown in Table 2. The ranks of the 11 unexpressed probe sets are listed with respect to the two models. The comparison between LWR and LWS based on the ranks is summarized with descriptive statistics as follows: LWS has the smaller median rank (6) and the smaller sum of ranks (68) with the smaller variance (13) while LWR has the median rank (8) and the sum of ranks (82) with the variance (17), respectively.

Table 2: Ranks of unexpressed genes among the control genes.

	<i>Dap1</i>	<i>Dap2</i>	<i>Dap3</i>	<i>Lys1</i>	<i>Lys2</i>	<i>Lys3</i>	<i>Phe1</i>	<i>Phe2</i>	<i>Phe3</i>	<i>Thr1</i>	<i>Thr2</i>
LWR	2	1	6	13	12	11	10	9	7	8	3
LWS	2	3	10	9	6	1	8	13	4	7	5

Moreover, we examined the ranks of the 11 probe sets of unexpressed control genes among all the genes in our study. Because we put no RNAs for these 11 probe sets, their measured expression levels should be close to zero and their ranks among all the genes should be among the lowest. As shown in Table 3, the ranks of the 11 probe sets detected from LWS are much lower

than those from LWR. In summary, LWS has the median rank (19) and the sum of ranks (312) with the variance (979) while LWR has the median rank (99) and the sum of ranks (2482) with the variance (79754), respectively. Based on the ranks of expression levels of the unexpressed control genes, LWS gives much better results than LWR.

Table 3: Ranks of unexpressed genes among all the genes in the study.

	<i>Dap1</i>	<i>Dap2</i>	<i>Dap3</i>	<i>Lys1</i>	<i>Lys2</i>	<i>Lys3</i>	<i>Phe1</i>	<i>Phe2</i>	<i>Phe3</i>	<i>Thr1</i>	<i>Thr2</i>
LWR	14	6	33	821	616	472	213	152	36	99	20
LWS	15	16	28	22	19	14	21	122	17	20	18

Among the spiked-in genes, *Dap* and *Thr* in 50:50 samples obtained 0.04 ng/8 $\mu$ g total RNA, 0 in stimulated and 0.08 for starved samples, while *Lys* and *Phe* in 50:50 samples obtained 0.04, 0.8 in stimulated and 0 for starved samples. Better gene expression index estimates should have the ability of differentiating between samples in which the underlying true gene expression levels vary. Hence, a sensible criterion is to assess an estimated expression index according to its correlation with the underlying true expression.

Intuitively, the true expression index should be proportional to the mRNA concentration. Thus higher correlation between the estimated expression indexes and mRNA concentrations is expected if the indexes are closer to the true expression levels. The correlation from LWR is 0.608 and from LWS is 0.609 where LWS is slightly higher than LWR. Similar results are obtained from the study of the correlations among the hybridization genes (*BioB*, *BioC*, *BioD*, *Cre*) and quantities of mRNA (2.5, 5, 25, 100).

To this end, we have made comparisons between the two regression models from several different perspective. LWR is a parametric regression model while LWS is a semiparametric model that is more robust in terms of model mis-specification.

Meanwhile, we notice that LWS gives slightly lower estimation of the expression indexes than LWR does generally. To compare LWR and LWS by combining the mean and variance of the expression indexes, we order all the measures and divide them into 50 quantile groups, then compute the median coefficient of variation (C.V.) for each group. Based on this criterion, LWS gives the average of all the median C.V.'s (0.088), which is smaller than that from LWR (0.094). Figure 5 shows a global and clear picture of the comparison. The median C.V. for each of the 50 groups from LWS is plotted against those from LWR. The straight line is the reference line with unit slope through the origin. It can be seen that most points in the square are above the reference line which indicates that the C.V.'s from LWS are smaller than those from LWR in general.

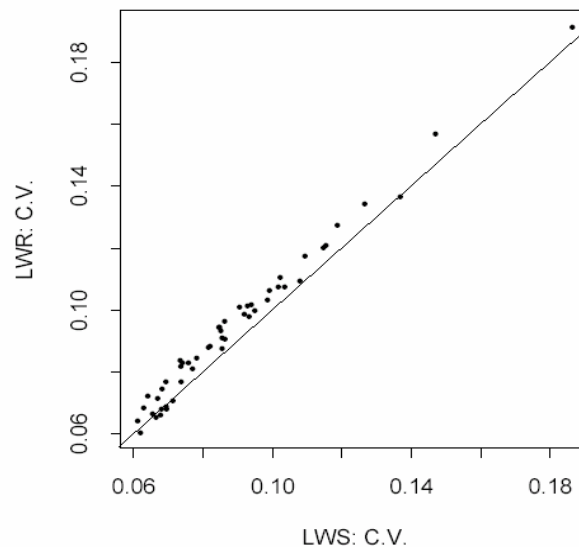


Figure 5: Comparison of coefficients of variation (C. V.) between LWR and LWS.

### Conclusion

Recently, much effort has been devoted to obtaining good estimates of the gene expression indexes, where Li-Wong's reduced model is widely used in applications. In this paper, we have proposed a semiparametric model based on Li-Wong's reduced model. The cubic spline smoothing technique allows a flexible functional form for MM expression intensities. Hence, it offers a better model-fitting procedure and captures the important gene expression patterns that might be missed by Li-Wong's reduced model.

From several aspects of comparison, our proposed model outperforms Li-Wong's reduced model. Practically and statistically, our new model is meaningful and easy to implement as well. The reason that we compare the proposed model with Li-Wong's reduced model is that the latter is very popular in practice and proved to perform better than the average expression index, the log-transformed average expression and others.

It is of interest to compare the proposed model with the new Affymetrix MAS 5.0 algorithm and other approaches. The variation of expression indexes changes positively with the intensity, which suggests a certain correlation or a linear trend between them. From the biological point of view, the genes are not independent, especially those that co-regulate. However, so far almost all model-based methods assume the variation has an independent structure. Therefore, a new methodology to incorporate the correlation structures needs to be developed.

For the comparison of measurements, we have extensively utilized the control genes which provided important and helpful information to our study. Control genes can also be used for normalization (Lemon et al., 2002). Hence if possible, we suggest that more control genes, especially those with more replicates should be used under reasonable biological consideration.

As to the model goodness-of-fit, there is no standard criteria available to justify and compare models with regard to the gene expression data where further research is needed. In the proposed model, the cubic spline smoothing is used, while the kernel smoothing (Speckman, 1988) and other nonparametric techniques may be applied to fit MM intensities as well. The proposed method can be improved in an adaptive

way as follows. We first test the goodness of fit of LWR based on the likelihood ratios. If there is no enough evidence to reject LWR, we would accept the estimates ( $\hat{q}$  and  $\hat{f}$ ) from LWR, otherwise we would proceed to LWS (spline).

### References

- Eubank, R. L. (1999). *Nonparametric regression and spline smoothing*. (2nd Ed). New York: Marcel Dekker.
- Heckman, N. E. (1986). Spline smoothing in a partly linear model. *Journal of the Royal Statistical Society, B* 48, 244-248.
- Hastie, T. J. & Tibshirani, R. J. (1990). *Generalized additive models*. London: Chapman and Hall.
- Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis K. J., Scherf, U. & Speed, T. P. (2002). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. (Accepted for publication in *Biostatistics*).
- Lipshutz R., Fodor S., Gingeras T., & Lockhart D. (1999). High density synthetic oligonucleotide arrays. *Nature Genetics*, 21, 20-24.
- Lemon W. J., Palatini J., Krahe R., & Wright F. A. (2002). Theoretical and experimental comparisons of gene expression indexes for oligonucleotide arrays. *Bioinformatics*, 18, 1470-1476.
- Li, C., & Wong W. H. (2001). *Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection*. Proceedings of the National Academy of Science, USA, 98, 31-36.
- Li, C., & Wong W. H. (2001). Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. *Genome Biology*, 2, 1-11.
- Lockhart, D. J., Dong, H. L., Byrne, M. C., Follettie, M. T., Gallo, M. V., Chee, M. S., Mittmann, M., Wang, C., Kobayashi, M., & Horton, H. (1996). Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnology*, 14, 1675-1680.

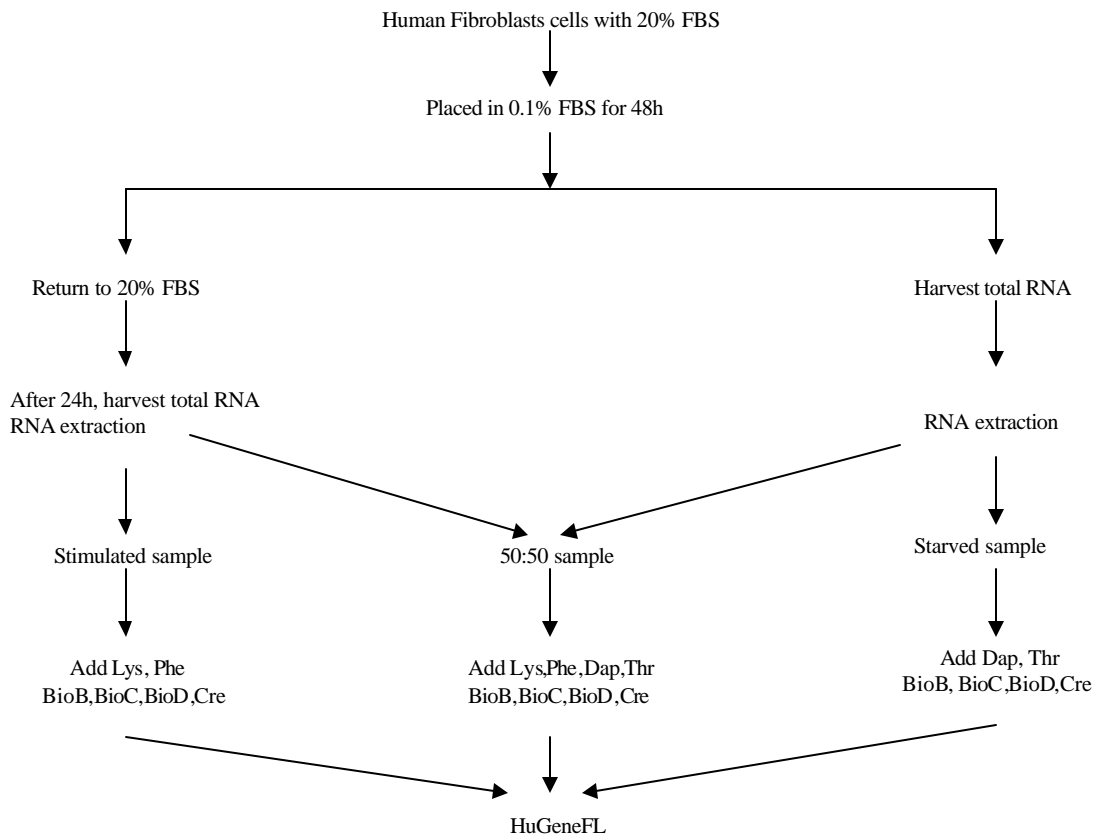
Schadt E., Li C., Ellis B., & Wong W. H. (2002). Feature extraction and normalization algorithms for high-density oligonucleotide gene expression array data. *Journal of Cellular Biochemistry*, 84, S37, 120-125.

Schadt E., Li C., Su C. & Wong W. H. (2001). Analyzing high-density oligonucleotide gene expression array data. *Journal of Cellular Biochemistry*, 80, 192-202.

Silverman, B. W. (1985). Some aspects of the spline smoothing approach to non-parametric regression curve fitting. *Journal of the Royal Statistical Society, B* 47, 1-52.

Speckman, P. (1988). Kernel smoothing in partial linear models. *Journal of the Royal Statistical Society, B* 50, 413-436.

#### Appendix: Experiment design float chart





*JMASM Algorithms and Code*  
**JMASM6: An Algorithm For Generating Exact Critical  
Values For the Kruskal-Wallis One-Way ANOVA**

Todd C. Headrick  
Southern Illinois University, Carbondale

---

A Fortran 77 subroutine is provided for computing exact critical values for the Kruskal-Wallis test on  $k$  independent groups with equal or unequal samples sizes. The subroutine requires the user to provide sorting and ranking routines and a uniform pseudo-random number generator. The program is available from the author on request.

Key words: Kruskal-Wallis test, nonparametric statistics, one-way ANOVA

---

### Introduction

Kruskal and Wallis (1952) derived a rank-based nonparametric test as an alternative to the one-way analysis of variance (ANOVA) on  $k$  independent groups. It has been demonstrated that the Kruskal-Wallis (K-W) test can have considerable power advantages over the ANOVA  $F$  test when the assumption of normality is violated (e.g., Aman & Headrick, 2003).

The null distribution of the K-W statistic is derived under the assumption that all  $N$  observations are from the same population. Because the number of ways  $N$  ranks may be divided into groups of  $n_1, \dots, n_k$  grows rapidly, most statistics textbooks (e.g., Conover, 1999; Gibbons, 1992; Siegel & Castellan, 1989) limit the reporting of exact critical values for the K-W statistic to no more than  $k=3$  groups and with  $n_j \leq 5$  observations per group.

The asymptotic null distribution of the K-W statistic is chi-squared with  $k-1$  degrees of freedom. As such, for  $k > 3$  and  $n_j > 5$  observations per group, the K-W asymptotic null distribution is recommended as the reference for making the decision to reject or fail to reject the null hypothesis that all  $k$  population distribution functions are identical (Conover, 1999, p. 289).

Most commonly used statistical software packages (e.g., Minitab; SPSS) that compute the K-W statistic only provide the asymptotic p-value. This may present a problem to an applied researcher because this p-value can be conservative relative to the exact p-value when both  $k$  and  $n_j$  are small. For example, for  $k = 5$  and  $n_j = 5$  for all  $j = 1, \dots, 5$ , the chi-squared critical value associated with  $\alpha = .05$  is 9.4876 whereas the exact critical value is 8.8985. Thus, using the asymptotic critical value for this design has the effect of lowering the Type I error rate from .050 to approximately .0363.

In view of the above, the purpose of this paper is to provide a subroutine that computes the exact critical values for the K-W statistic. The subroutine will compute critical values for any number of  $k$  populations with equal or unequal sample sizes.

---

Todd C. Headrick is Assistant Professor of Statistics, 222-J Wham Building, Mail Code 4618, Southern Illinois University-Carbondale, IL, 62901. His areas of research interest are statistical computing, nonparametric statistics, and optimization. Email him at headrick@siu.edu.

## Methodology

The subroutine initially generates  $N$  uniform pseudo-random numbers on the interval (0,1). It is assumed that the probability of obtaining any tied scores is zero. The uniform deviates are then ranked to form a permutation of the numbers from 1, ...,  $N$ . The algorithm then sequentially splits the permutation of ranks into  $k$  groups in accordance to the user's specified sample sizes of  $n_1, \dots, n_k$ . The K-W statistic is then computed as

$$H = \left( \frac{12}{N(N+1)} \sum_{j=1}^k \frac{R_j^2}{n_j} \right) - 3(N+1),$$

where  $R_j$  is the sum of the ranks in the  $j$ th group. This process is repeated until a sufficient number of  $H$  statistics are generated to adequately model the null distribution.

The algorithm then selects the critical values associated with the alpha levels of .01, .05, and .10. In general, the critical values returned by the subroutine are associated with a range of p-values. For example, for  $k=3$  and  $n_j=5$  for all  $j=1,2,3$ , the subroutine will return the exact critical value of 5.659997 for  $\alpha = .05$ . However, this critical value is associated with p-values ranging from approximately .0483 to .0537. As such, the program informs the user that the null hypothesis may be rejected if the computed K-W statistic is *strictly greater* than the critical value of 5.659997.

The method used in the subroutine for selecting critical values yields the same values reported in Conover (1999, Table A8, p. 539). It should be noted that this method is different from the method that was used for selecting the critical values reported in Gibbons (1992, Table K, p. 503) and Siegel & Castellan (1989, Table O, p. 356). Specifically, for the example above, these texts report a critical value of 5.78. This method indicates to the reader that the null hypothesis may

be rejected if the computed K-W statistic is *greater than or equal to* the critical value of 5.78.

## Conclusion

The program leaves it to the user to specify the number of K-W statistics to generate. The larger the value of  $N$  requires a larger number of K-W statistics to be generated to adequately model the null distribution. Thus, it is recommended that trials be repeated with an increasing number (e.g., 100,000; 500,000; 1,000,000, etc.) of K-W statistics generated in each trial run. It is suggested that the user terminate this process when changes in the critical values are less than  $10^{-4}$ .

## References

- Aman, S. Y., & Headrick, T. C. (2003). An empirical investigation of Type I error and power between data transformations to normality and nonparametric analysis in the context of one-way ANOVA. Annual meeting of the American Educational Research Association, SIG Educational Statisticians, Chicago, IL.
- Conover, W. J. (1999). *Practical nonparametric statistics* (3rd ed.). New York: Wiley.
- Gibbons, J. D. (1992). *Nonparametric statistical inference* (3rd ed.). New York: Marcel Dekker.
- Kruskal, W. H., & Wallis, W. A. (1952). Use of ranks in one-criterion analysis of variance. *Journal of the American Statistical Association*, 47, 583-621; errata, *ibid.*, 48, 907-911.
- Minitab (2000). *Minitab for Windows, release 13.3*, Minitab, Inc., State College, PA.
- Siegel, S., & Castellan, N. J. (1989). *Nonparametric statistics for the behavioral sciences* (2nd ed.). New York: McGraw-Hill.
- SPSS (2002). *SPSS for Windows, version 11.0*, SPSS, Inc., Chicago, IL.

```

      SUBROUTINE KW(K, N, M, ISAMP, CRIT)
C*****
C K is the specified number of populations in the one-way ANOVA.
C N is the specified total sample size.
C M is an array with K specified sample sizes  $n_1, \dots, n_k$ .
C ISAMP is the specified number of K-W
C sample statistics to be generated. CRIT is an array of critical
C values for the K-W test to be returned.
C*****
      REAL X(N), RKX(N), SMRKX(K), STAT(ISAMP), CRIT(3)
      INTEGER M(K)
      DOUBLE PRECISION DSEED
C*****
C Generate the specified number of K-W sample statistics.
C*****
      DO 10 I = 1, ISAMP
C*****
C Call the uniform random number generator routine. Generate an array X
C with N uniform deviates. Call the ranking routine. Rank the N uniform
C deviates in X and place these ranks into the array RKX.
C*****
      CALL UNIFORM (DSEED, N, X)
      CALL RANK (X, N, RKX)
C*****
C Compute the K-W statistic denoted below as H.
C*****
      IE = 0
      SUM1 = 0
      DO 20 J = 1, K
      IB = IE + 1
      IE = IE + M(J)
      SUM2 = 0
      DO 30 L = IB, IE
      SUM2 = SUM2 + RKX(L)
      SMRKX(J) = SUM2**2 / FLOAT(M(J))
30 CONTINUE
      SUM1 = SUM1 + SMRKX(J)
20 CONTINUE
      H = (12 / (FLOAT(N)*(FLOAT(N) + 1)))*SUM1 - 3*(FLOAT(N) + 1)
C*****
C Store the H statistic in the array STAT.
C*****
      STAT(I) = H
10 CONTINUE
C*****
C Sort the array STAT.
C*****
      CALL SORT (STAT, ISAMP)

```

```
C*****
C Obtain the critical values associated with the alpha levels
C .01, .05, and .10.
C*****
  DO 40 I = ISAMP, 1, -1
    PR = (FLOAT(ISAMP) - FLOAT(I)) / FLOAT(ISAMP)
        IF ( PR .LE. 0.01) THEN
          CRIT(1) = STAT(I)
        ELSEIF (PR .LE. 0.05) THEN
          CRIT(2) = STAT(I)
        ELSEIF (PR .LE. 0.10) THEN
          CRIT(3) = STAT(I)
        ELSEIF (PR .GT. 0.10) THEN
          GOTO 50
        END IF
  40 CONTINUE
  50 CONTINUE
    RETURN
  END
```

## A Recursive Algorithm For Fractionally Differencing Long Data Series

Joseph McCarthy  
Finance Department

Robert DiSario  
Department of Mathematics  
Bryant College

Hakan Saraoglu  
Finance Department

---

We propose a recursive algorithm to fractionally difference time series data. The algorithm eliminates the need to evaluate the gamma function directly, and hence avoids the overflow problem that arises when fractionally differencing a long data series. The proposed algorithm can be implemented using any general matrix programming language. An implementation using SAS is presented. The algorithm and the code provide a practical approach to including fractional differencing as part of a time series data analysis.

Key words: Fractionally differencing, time series

---

### Introduction

The process of differencing is widely used in time series data analysis. First differencing is often adequate to deal with nonstationary data for an ARIMA model. A useful generalization of integer differencing is fractional differencing. The resulting FARIMA models, or fractional ARIMA models, are often used for time series exhibiting long-range dependence (Beran (1994); Geweke and Porter-Hudak (1983); Granger and Joyeux (1980); Mandelbrot and Van Ness (1968)). Long-range dependent series have hyperbolically decaying autocorrelation functions, unlike the exponential decay found in autocorrelation functions for time series modeled by ARIMA.

Algorithms to do fractional differencing can be used in simulating FARIMA data, in fractionally differencing an empirical time series to obtain a series suitable for ARIMA modeling, and in testing for white noise of residuals after fitting a FARIMA model. Because long-range dependence is found in financial time series and in some geophysical time series, practical algorithms to accomplish fractional differencing are needed.

Statistical packages are beginning to incorporate modules to do fractional differencing. However, some of these modules are limited to very small data sets. For example, the SAS function FDIF can only handle approximately 171 observations (SAS release 8.2 Proc IML; SAS Institute, Inc. 2001). This limit is apparently due to use of the gamma function. Our proposed algorithm uses a recursive approach to eliminate the need to compute gamma directly. Thus it provides a practical way to fractionally difference a time series of much more than 171 observations. As discussed in the results section, we have tested this procedure for a time series as large as 10,000 observations. The algorithm that we describe could be implemented in any general matrix programming language. We provide an implementation using the matrix language SAS IML (SAS Institute, Inc. 1990).

---

Joseph McCarthy is a Professor in the Finance Department at Bryant College. He received a DBA in finance from the University of Colorado in 1983. His academic interests include fixed income instruments, non-linear modeling and wavelets. Robert DiSario is an Assistant Professor in the Department of Mathematics at Bryant College. He received a Ph.D. in statistics from Boston University in 1996. His academic interests include applied statistics and combinatorics. Hakan Saraoglu is an Associate Professor in the Finance Department at Bryant College. He received a Ph.D. in finance from Michigan State University in 1996. His academic interests include investments, international finance, and corporate finance.

### Method

Let  $y_t$  be obtained by taking the  $d^{\text{th}}$  difference of a time series  $x_t$ ;  $t = 0, 1, \dots, n - 1$ :

$$y_t = (1 - B)^d x_t, \tag{1}$$

where  $B$  is the backshift operator defined by

$$Bx_t = x_{t-1}.$$

If  $d=1$ , then  $y_t$  is the first difference:

$$y_t = (1 - B)x_t = x_t - Bx_t = x_t - x_{t-1}. \tag{2}$$

If  $d=2$ , then  $y_t$  is the second difference:

$$\begin{aligned} y_t &= (1 - B)^2 x_t = (1 - B)(x_t - x_{t-1}) \\ &= x_t - 2x_{t-1} + x_{t-2}. \end{aligned} \tag{3}$$

We could also obtain this second difference by expanding  $(1 - B)^2$  and applying the resulting second degree polynomial in  $B$  to  $x_t$ .

$$\begin{aligned} y_t &= (1 - B)^2 x_t = (1 - 2B + B^2)x_t \\ &= x_t - 2x_{t-1} + x_{t-2}. \end{aligned} \tag{4}$$

In general, for any integer  $d$ , the  $d^{th}$  difference can be found by expanding  $(1 - B)^d$  and applying the resulting polynomial in  $B$  to  $x_t$ . Fractional differencing ( $-0.5 < d < 0.5$ ) is defined in an analogous way. Expanding  $(1 - B)^d$  in a Taylor series (see Kaplan, 1984, p431):

$$\begin{aligned} (1 - B)^d &= 1 + \frac{d}{1!}(-B)^1 + \frac{d(d-1)}{2!}(-B)^2 \\ &\quad + \frac{d(d-1)(d-2)}{3!}(-B)^3 + \dots \\ &= \sum_{j=0}^{\infty} \frac{d(d-1)(d-2)\dots(d-(j-1))}{j!} (-1)^j B^j \end{aligned} \tag{5}$$

where the numerator in the above expression has  $j$  factors except when  $j=0$  where it is unity. Now by multiplying each factor in the numerator by  $-1$  we change the sign of each:

$$(1 - B)^d = \sum_{j=0}^{\infty} \frac{(-d)(1-d)(2-d)\dots((j-1)-d)}{j!} B^j \tag{6}$$

Next, multiplying by  $1 = \frac{\Gamma(j-j-d)}{\Gamma(-d)}$  and reversing the order of the factors in the product we obtain:

$$(1 - B)^d = \sum_{j=0}^{\infty} \frac{(j-1-d)(j-2-d)\dots(j-j-d)\Gamma(j-j-d)}{j!\Gamma(-d)} B^j \tag{7}$$

Finally, by repeatedly using the recurrence property of the gamma function:  $\Gamma(x) = (x-1)\Gamma(x-1)$  we can re-express the numerator as  $\Gamma(j-d)$ . Thus, we obtain

$$(1 - B)^d = \sum_{j=0}^{\infty} \frac{\Gamma(j-d)}{\Gamma(j+1)\Gamma(-d)} B^j,$$

which is a commonly used representation for the fractional differencing operator (Jensen, 1999).

To implement a fractional differencing algorithm it necessary to compute the coefficients in the above series:

$$C_j = \frac{\Gamma(j-d)}{\Gamma(j+1)\Gamma(-d)} \quad j=0,1,2,\dots \tag{8}.$$

Because these coefficients are used to multiply observations in the time series, this infinite sequence of coefficients can be truncated to the length of the data series.

A problem arises when calculating these coefficients because for large values of  $j$  the numerator and denominator become very large and exceed the computational capacity of the computer. For example, the gamma function evaluated at 171 is approximately 7.257E306. Our approach uses the recursive property of the gamma function,  $\Gamma(x) = (x-1)\Gamma(x-1)$ , to obtain a recursive property for the  $C_j$  as follows:

$$\begin{aligned}
 C_0 &= \frac{\Gamma(0-d)}{\Gamma(1)\Gamma(-d)} = 1 \\
 C_j &= \frac{\Gamma(j-d)}{\Gamma(j+1)\Gamma(-d)} \\
 &= \frac{(j-d-1)\Gamma(j-d-1)}{j\Gamma(j)\Gamma(-d)} \\
 &= \frac{(j-d-1)}{j} C_{j-1}
 \end{aligned} \quad (9).$$

Because the above recursive formula does not involve use of the gamma function, it is possible to calculate  $C_j$  for large values of  $j$ . It is only necessary to multiply  $C_{j-1}$  by  $\frac{(j-d-1)}{j}$  which is computationally trivial. Our SAS program which implements this appears in Appendix I. The key lines of code which recursively calculate  $C_j$  follow. Note that in SAS the array  $C_j [ ]$  must be indexed from 1 to  $n$ , rather than from 0 to  $n-1$ .

```

jj=0;
do i=1 to n;
  if i=1 then Cj[i] = 1;
  else Cj[i]= Cj[i-1]*((jj-d-
    1)/jj) ;
  jj=jj+1;
end;

```

The fractionally differenced time series,  $y_t$ , is obtained by convolving the input time series,  $x_t$ , with the vector of coefficients  $C_j$ . That is

$$y_t = (1-B)^d x_t = \sum_{j=0}^t C_j x_{t-j} \quad (10).$$

The lines of SAS code that implement the convolution appear below.

```

do i=1 to n;
  yt[i]=Cj[1:i]*xt[i:1];
end;

```

Using our approach we have been able to fractionally difference long data series. In the

results section we give an example using a series of 10,000 observations.

## Results

In the first example, we fractionally difference a small integer data series using  $d=.5$ , then fractionally difference the result again using  $d=.5$ . For this example, fractional differencing was done in two ways: first using the SAS function FDIF (SAS release 8.2 Proc IML); then using the code described above.

One reason for performing this test was to confirm that both approaches to fractional differencing produce the same result for a small time series. A second reason was to check that the  $d$  values are additive: fractional differencing twice with  $d=.5$  is the same as first differencing.

The data series and the two fractionally differenced series are presented in Table 1. The column labeled **XT** is the integer data series. **YJ** is the fractional difference of **XT** using 'Call FDIF' with  $d=.5$ . **ZJ** is the fractional difference of **YJ** using Call FDIF with  $d=.5$ . Next, **YT** is the fractional difference of **XT** using our recursive procedure with  $d=.5$ . Finally, **ZT** is the fractional difference of **YT** using the procedure with  $d=.5$ . Clearly, **YT** = **YJ** and **ZT** = **ZJ**, showing that the two procedures produce the same results for this small data series. Also, the reader can check that **ZT** and **ZJ** are the same as would be obtained by doing first differencing. The program that produced all four series appears in Appendix II.

In the second example we use our recursive method to fractionally difference a random series of 10,000 observations. Note that the method using the SAS FDIF function will not run on a time series that is longer than approximately 171 observations (using a Pentium IV, running at 1.7 GHz) and therefore was not included in this example. The SAS LOG in Appendix III shows that the program using our method successfully ran. Thus this method provides a practical way to fractionally difference long time series. Implementing this algorithm in SAS provides a convenient way to include fractional differencing as part of a complete analysis of a long memory time series.

## Conclusion

FARIMA models are commonly used to model long range dependent time series. In such cases, fractional differencing is often a useful part of the analysis. The practical way to fractionally difference a long time series is to use an algorithm that avoids calculating  $\gamma(n)$  directly. (Although not discussed in the results section, we also ran our program on a series of 100,000 observations using 5 minutes of CPU time). Our implementation in SAS is a convenient way to incorporate fractional differencing into time series data analysis.

## References

Beran, J. (1994). *Statistics for long memory processes*. New York: Chapman & Hall.

Geweke, J. & Porter-Hudak, S. (1983). The estimation and application of long memory time series models. *Journal of Time Series Analysis*, 4, 221-238.

Granger, C. W. J. & Joyeux, R. (1980). An introduction to long memory time series models and fractional differencing. *Journal of Time Series Analysis* 1(1), 15-29.

Jensen, M. (1999). Using wavelets to obtain a consistent ordinary least squares estimator of the fractional differencing parameter. *Journal of Forecasting*, 18, 17-32.

Kaplan, W. (1984). *Advanced calculus*. (3<sup>rd</sup> ed.). Reading, MA: Addison-Wesley.

Mandelbrot, B. B. & Van Ness, J.W. (1968). Fractional Brownian motions, fractional noises and applications. *SIAM Review*, 10, 422-437.

SAS Institute, Inc., (1990). *SAS/IML Software: Usage and reference*, Version 6 Cary, NC: SAS.

SAS Institute, Inc., (2001). *SAS/IML Software: Changes and enhancements*, Release 8.2. Cary, NC: SAS.

Table 1. Fractional differencing using SAS Call Fdif and using the recursive procedure.

XT	YT	ZT	YJ	ZJ
582	582	582	582	582
227	-64	-355	-64	-355
410	223.75	183	223.75	183
109	-160.75	-301	-160.75	-301
686	543.3281	577	543.3281	577
753	345.9688	67	345.9688	67
903	399.7793	150	399.7793	150
996	377.9981	93	377.9981	93
60	-647.4	-936	-647.4	-936
76	-201.273	16	-201.273	16
716	523.3205	640	523.3205	640
202	-272.01	-514	-272.01	-514
637	361.6509	435	361.6509	435
60	-394.921	-577	-394.921	-577
314	109.65	254	109.65	254
969	691.8636	655	691.8636	655
87	-524.382	-882	-524.382	-882
660	406.5947	573	406.5947	573
719	248.2841	59	248.2841	59
784	241.7671	65	241.7671	65

## Appendix I - SAS Program FRACDIFF.SAS

```
*****;
* fracdiff.sas *;
* *;
*****;
* create random data for fractional differencing algorithm *;
data one ;
* for n=171 both methods of fractional differencing work *;
* for n=172 call to fdif fails, but convolution works *;
do i=1 to 10000;
```



```

x=rand('NORMAL');
output;
end;
* fractional differencing algorithm implemented below *;
proc iml ;
  use one;
  read all into xx;
  index=xx[,1];
  xt=xx[,2];
  n=nrow(xt);
* d = fractional differencing parameter *;
  d=.5;
* initialization;
  yt=j(n,1,0);
  Cj=j(n,1,0);
* do loop calculates coefficients using recursive method *;
  jj=0 ;
do i=1 to n;
  if i=1 then Cj[i] = 1;
  else Cj[i]= Cj[i-1]*((jj-d-1)/jj) ;
  jj=jj+1;
end;
* Convolution follows. The arrays are indexed in reverse order to
implement the *;
* convolution. Also, the symbol for transpose in SAS IML is '      *;
do i=1 to n;
  yt[i]=Cj[1:i]^*xt[i:1];
end;
quit;

```

#### Appendix II - SAS Program TESTPROG4.SAS

```

*****;
* testprog4.sas *;
*          *;
*****;
data one ;
* for n=171 both methods of fractional differencing work *;
* for n=172 call to fdif fails, but convolution works *;
do i=1 to 20;
  x=int(rand('uniform')*1000);
  output;
end;
proc print data=one ;
run;
proc iml ;
  use one;
  read all into xx;
  index=xx[,1];
  xt=xx[,2];
  n=nrow(xt);d=.5;

```

```

* initialization;
yt=j(n,1,0);
zt=j(n,1,0);
yj=j(n,1,0);
zj=j(n,1,0);
Cj=j(n,1,0);
  call fdif(yj, xt, .5);
  call fdif(zj, yj, .5);
jj=0 ;
do i=1 to n;
  if i=1 then Cj[i] = 1;
  else Cj[i]= Cj[i-1]*((jj-d-1)/jj) ;
  jj=jj+1;
end;
do i=1 to n;
  * convolution follows *;
  yt[i]=Cj[1:i]^*xt[i:1];
end;
do i=1 to n;
  * convolution follows *;
  zt[i]=Cj[1:i]^*yt[i:1];
end;
print index xt yt zt yj zj;

```

#### Appendix III - SAS LOG for FRACDIFF.SAS

```

653 *****;
654 * fracdiff.sas *;
655 * *;
656 *****;
657
658
659 * create random data for fractional differencing algorithm *;
660
661 data one ;
662 * for n=171 both methods of fractional differencing work *;
663 * for n=172 call to fdif fails, but convolution works *;
664 do i=1 to 10000;
665 x=rand('NORMAL');
666 output;
667 end;
668
669
670 * fractional differencing algorithm implemented below *;
671

```

NOTE: The data set WORK.ONE has 10000 observations and 2 variables.

NOTE: DATA statement used:

real time 0.00 seconds

```

672 proc iml ;

```

```
NOTE: IML Ready
673 use one;
674 read all into xx;
675 index=xx[,1];
676 xt=xx[,2];
677 n=nrow(xt);
678 * d = fractional differencing parameter *;
679 d=.5;
680
681 * initialization;
682 yt=j(n,1,0);
683 Cj=j(n,1,0);
684
685 * do loop calculates coefficients using recursive method *;
686
687 jj=0 ;
688 do i=1 to n;
689   if i=1 then Cj[i] = 1;
690   else Cj[i]= Cj[i-1]*((jj-d-1)/jj) ;
691   jj=jj+1;
692 end;
693
694 * Convolution follows. Notice that the arrays are indexed in reverse
order to implement the
694! *;
695 * convolution. Also, the symbol for transpose in SAS IML is '
695! *;
696
697 do i=1 to n;
698   yt[i]=Cj[1:i]`*xt[i:1];
699 end;
700
701 quit;
NOTE: Exiting IML.
NOTE: 7659 workspace compresses.
NOTE: PROCEDURE IML used:
      real time      3.18 seconds
```

## Statistical Pronouncements

“I do not see that the sex of the candidate is an argument against her admission. After all, we are a university, not a bathing establishment” - David Hilbert, regarding Emmy Amalie Noether’s unsuccessful application to the faculty at Göttingen in 1915.

“As I understand De Moivre the ‘Original Design’ is the mean occurrence on an indefinite number of trials...The Deity fixed the ‘means’ and ‘chance’ provided the fluctuations...There is much value in the idea of the ultimate laws of the universe being statistical laws... [but] it is not an exactly dignified conception of the Deity to suppose him occupied solely with first moments and neglecting second and higher moments!” - Karl Pearson (1978, *The History of statistics in the 17<sup>th</sup> and 18<sup>th</sup> centuries against the changing background of intellectual, scientific and religious thought: Lectures given at University College London during the academic sessions 1921-1923*, p. 160.)

“We are passing from the scientific enthusiasm of the founders... to a period when men followed science as a profession, when the text-book writer appears seeking whom he may devour, and how his books will sell, rather than what new knowledge they may bring” - Karl Pearson (*ibid*, p. 176).

“You cannot too narrowly separate the history of statistics from the general history of science, still less from the history of philosophical and religious thought” - Karl Pearson (*ibid*, p. 213).

“Mathematicians have always been rather of a jealous nature...[and] there is some excuse..., for their reputation stands for posterity largely not on what they did, but on what their contemporaries attributed to them” - Karl Pearson, *ibid*, p. 226).

“It is idle to measure a man’s real value by the number of memoirs he writes, although that is very influential just now in academic appointments on both sides of the Atlantic - it is easier to count than to weigh” - Karl Pearson (*ibid*, p. 245).

“History is to no purpose unless you try to grasp the general character of a man and of the age in which he lived” - Karl Pearson (*ibid*, p. 248).

“Extreme mathematical power is not necessarily combined with an extremely logical mind” - Karl Pearson (*ibid*, p. 249).

“Mixed up with mathematics is the philosophy and the theology of the day” - Karl Pearson (*ibid*, p. 249).

“The advance of a science even like statistics is linked up with the general history of human ideas” - Karl Pearson (*ibid*, p. 303).

“The religious belief of men colors not only what they collect, but how they interpret it” - Karl Pearson (*ibid*, p. 319).

“A wise reformer, if he wishes practically to influence his generation, must know not only what is true, but how much of that truth his contemporaries can possibly digest” - Karl Pearson (*ibid*, p. 349).

“Extreme repugnance for computing [is] a sin of too many mathematical statisticians” - Karl Pearson (*ibid*, p. 426).

“However beautiful a mathematical theory, however completely it be worked out, its weaknesses or its successes can only be ascertained, when it has been submitted to the test of numerical evaluation” - Karl Pearson (*ibid*, p. 456).

“Experiments must be capable of being considered to be a random sample of the population to which the conclusions are to be applied. Neglect of this rule has led to the estimate of the value of statistics which is expressed in the crescendo ‘lies, damned lies, statistics’” - W. S. Gosset (“Student”), (1926, Mathematics and agronomy, *Journal of the American Society of Agronomy*, 18, p. 703.)

“Sampling is the central problem in statistics” - George W. Snedecor (1946, *Statistical methods*, p. 453).

“Modern statistical method is a science in itself” - S. S. Wilks (1948, *Elementary statistical analysis*, p. 1).

“Human progress is based on ‘permanencies’ or, rather, on our ability to detect permanencies both in the objects surrounding us and in changes in these objects” - Jerzey Neyman (1950, *First course in probability and statistics*, p. 1).

“In practical applications we seldom meet cases where the assumption of the existence of an a priori probability distribution seems to be justified; and even in those rare cases in which the latter assumption can be made, we usually do not know the shape of the a priori probability distribution and this makes the application of Bayes’ theorem impossible” - Abraham Wald (1950, *On the principles of statistical inference*, p. 26).

“An unfortunate publicity was given to discussions of the so-called foundations of probability, and thus the erroneous impression was created that essential disagreement can exist among mathematicians. Actually, these discussions concern only minor points which are of interest to but few specialists” - William Feller (1950, *An introduction to probability theory and its applications*, p. 6).

“The secret language of statistics, so appealing in a fact-minded culture, is employed to sensationalize, inflate, confuse, and oversimplify” - Darrell Huff, (1954, *How to lie with Statistics*, p. 8).

“I believe that the nonparametric techniques of hypothesis testing are uniquely suited to the data of the behavioral sciences” - Sidney Siegel (1956, *Nonparametric statistics for the behavioral sciences*, p. vii).

“Permutation tests are easy to define, but ...the numerical calculations required to carry them out are usually hopelessly tedious” - Henry Scheffé (1959, *The analysis of variance*, p. 313).

“A good (although debatable) case can be made for means and variances as indices of location and dispersion when the normality assumption holds; the argument loses much of its force, however, when the assumption fails” - James V. Bradley (1968, *Distribution-free statistical tests*, p. 12).

“The easiest way to abuse any statistical technique is to disregard and/or violate the assumptions necessary for the validity of the procedure” - Jean Dickinson Gibbons (1976, *Nonparametric methods for quantitative analysis*, p. 24).

“The rather naïve objection might be raised that educational data are rarely sufficiently non-normal to warrant concern. Perhaps the most effective means of dealing with such a notion on the part of an educational researcher is to suggest that he/she routinely construct relative frequency histograms of the data that they submit to statistical analysis. This time-honored but often neglected practice usually paints pictures of distributions that are unimagined by researchers who think of data in terms of the normal curve” - R. C. Blair (1981, A reaction to ‘Consequences of failure to meet assumptions underlying the fixed effects analysis of variance and covariance’, *Review of Educational Research*, 51, p. 503-504).

“Any reader who has penetrated this book to this point hardly needs convincing of the centrality of the concept of effect size... a moment’s thought suggests that it is, after all, what science is all about” - Jacob Cohen (1988, *Statistical power analysis for the behavioral sciences*, p. 531-532.)

“There is no physical entity that is the number 1. If there were, it would surely be in a place of honor in some great scientific museum, and past it would file a steady stream of mathematicians, gazing at 1 in wonder and awe” - John B. Fraleigh (1989, *A first course in abstract algebra*, p. 20.)

“The Monte Carlo method provides the experimental scientist with one of the most powerful tools available for planning experiments and analyzing data” - R. Bevington and D. Keith Robinson (1992, *Data reduction and error analysis for the physical sciences*, p. 76).

# Qualitative research has come a long way...

from this...



to this!



**qsr**

THE LATEST PRODUCTS  
HAVE ARRIVED

[www.qsrinternational.com](http://www.qsrinternational.com)

**NS**

**Nvivo**

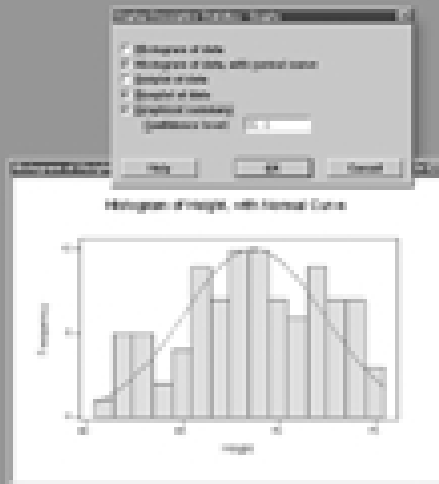
WORLD LEADING PRODUCTS FROM THE  
NUD\*IST LINE OF SOFTWARE

Read more about QSR software in this edition of JMASM.

# Take us for a test drive...

Download a **FREE demo\*** of **MINITAB Release 13 for Windows**

Visit [www.minitab.com/demo.htm](http://www.minitab.com/demo.htm) or call 1-800-488-3555



**M**INITAB is accurate, easy-to-use statistical software that integrates seamlessly into the statistics curriculum. With MINITAB, you can focus on teaching statistical concepts, not software, while your students gain valuable real-world computing skills. More than 4,000 colleges and universities worldwide, as well as thousands of distinguished companies, rely on MINITAB today.

- ▶ Basic and Advanced Statistics
- ▶ Regression and ANOVA
- ▶ Power and Sample Size
- ▶ Time Series and Forecasting
- ▶ Quality Statistics



**From start-ups to Fortune 500 companies, quality professionals worldwide rely on MINITAB**

- General Electric
- Ford Motor Company
- Honeywell International
- 3M
- DuPont
- British Telecom
- General Motors
- DaimlerChrysler
- Nokia
- Hewlett-Packard
- Motorola
- Lucent Technologies
- Eastman Kodak
- Polaroid
- Compaq
- Leading Six Sigma Consultants
- More than 4,000 colleges and universities worldwide

"Using MINITAB is like driving a car — all I need to do is put my key in the ignition, start the engine, and I'm on my way!"

**Shrikant Kulkarni**

Director, Engineering  
Master Blackbelt  
Carnet Company (a division of Precision Castparts, Inc.)



US Office, Minitab Inc., Tel: +1.800.488.3555 (US/CAN only), E-mail: [sales@minitab.com](mailto:sales@minitab.com)

\*Demo software is fully functional for 30 days.

© Minitab Inc., 2002. MINITAB and the MINITAB logo are registered trademarks of Minitab Inc. in the U.S. and other countries.

**Student rental options available – visit [www.e-academy.com/minitab](http://www.e-academy.com/minitab)**

# Announcing StatXact 5!

StatXact 5, with over 100 procedures and a 1500 page manual that is really a textbook on exact methods, provides the world's most comprehensive collection of exact procedures for significance tests and confidence intervals. Among its new features, StatXact 5 now gives you a host of new procedures for the commonly-encountered two-binomial situation. Based on recent research (Agresti and Min, *Biometrics* 2000; Chan and Zang, *Biometrics* 1999), these procedures will give you more powerful exact p-values, and shorter exact confidence intervals.

## New In StatXact 5

### Unconditional exact tests for 2 binomials:

- Superiority
- Non-inferiority
- Equivalence

More powerful exact unconditional tests and shorter exact confidence intervals for differences and ratios of proportions

Unconditional exact McNemar's test

Exact interaction tests in stratified 2xC tables

- comparison of C ordered binomials
- comparison two ordered multinomials

Exact test of trend for correlated binary data

Exact tests and confidence intervals for stratified Poisson data

While some standard software programs have a few exact tests, none has anywhere near the coverage of StatXact 5. StatXact 5, with over 100 tests and procedures, gives you exact p-values and confidence intervals for one-, two- and k-sample problems,  $R \times C$  contingency tables, stratified  $2 \times 2$  and  $2 \times C$  contingency tables, goodness-of-fit tests, measures of association, binomial data, multinomial data, and censored survival data. Plus, StatXact 5 gives you exact power and sample size capabilities.

CYTEL Software Corporation • 675 Massachusetts Ave., Cambridge, MA 02139 USA  
Tel (617) 661-2011 • Toll Free (US) 866-298-3511 • Fax (617) 661-4405  
<http://www.cytel.com> • E-mail: [sales@cytel.com](mailto:sales@cytel.com)

INTERNATIONAL DISTRIBUTORS: Ask Int'l (UK) e-mail: [cyteluk@asru.com](mailto:cyteluk@asru.com) • Tel: +44(0) 1227 795 240 • Fax: +44(0) 1227 795 201; ID2 (Belgium)  
• Tel: 32 2 6468918 • Fax: 32 2 4468662; Spadille Biostatistics ApS (Denmark) • [spadille@spadille.dk](mailto:spadille@spadille.dk) • Tel: 4548.484100 • Fax: 4248484200

**Cytel**  
STATISTICAL SOFTWARE



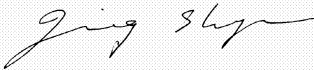
# Depth and flexibility for predicting numerical outcomes

As a clear leader in statistical software, SPSS has what you need for analysis — and the complete analytical process.

SPSS is a modular, tightly integrated, full-featured product line. It's available for Windows and Macintosh desktops. Alternatively, it's available for many high-performance server platforms. The SPSS product line covers the full analytical process. SPSS' offering includes products for database access, data cleaning and management, as well as a broad range of analytical capabilities, and high-quality tabular and graphical output. You can even publish your SPSS results to the Web. This enables people who don't have SPSS installed on their machines to interact with results using their Web browsers. SPSS products are available through a variety of pricing and licensing programs, including student, graduate student and campus-wide licenses.

Take a look at some highlights in SPSS' line-up for predicting numerical outcomes and learn about just one aspect of SPSS' many offerings for the analytical process.

Sincerely,



Jing Shyr, Ph.D.  
Vice President and Chief Statistician  
SPSS Inc.



Kyle A. Weeks, Ph.D.  
Senior Product Manager  
SPSS Inc.

## Linear Mixed Models procedure

Do you have data that display correlation and non-constant variability, such as data that represent students nested within classrooms or consumers nested within families? You can model not only means but also variances and covariances in your data using the powerful Linear Mixed Models procedure. Its flexibility means you can formulate a wide variety of models, such as multilevel models with fixed-effects covariances, hierarchical-linear models, random-effects models, random-coefficient models and linear-growth models. In addition, you can work with repeated measure designs, including incomplete repeated measurements in which the number of observations varies across subjects.

## General Linear Models (GLM) procedure — multivariate

Do you need a flexible procedure that works simultaneously with related multiple dependent variables? SPSS' GLM multivariate procedure does just that — providing flexible design and contrast options to estimate means and variances and to test and predict means. Mix and match categorical and continuous data to build models. Because GLM multivariate doesn't limit you to one data type, you have options giving you a wealth of model-building possibilities. Also, you can easily visualize relationships using profile plots (interaction plots) resulting from estimated predicted mean values.

## General Linear Models (GLM) procedure — repeated measures

Do you need to measure the same people over time, for example, to measure how overall employee satisfaction increases or decreases? Using SPSS' GLM repeated measures procedure you can analyze variances when you make the same measurement a fixed number of times on

individual subjects or cases. Get the flexibility to mix and match categorical and continuous-level predictors — including interactions. As with the GLM multivariate procedure, you can see relationships in your data using profile plots.

## Nonlinear Regression (NLR) and Constrained Nonlinear Regression (CNLR) procedures

Are you working with models that have nonlinear relationships, such as predicting coupon redemption as a function of time and number of coupons distributed? Estimate nonlinear equations using one of two SPSS procedures: NLR for unconstrained problems and CNLR for both constrained and unconstrained problems. CNLR empowers you to write your own algorithms. CNLR also gives you the flexibility to:

- Use linear and nonlinear constraints on any combination of parameters
- Estimate parameters by minimizing any smooth loss function (objective function)
- Compute bootstrap estimates of parameter standard errors and correlations

## Everything you need for predicting numerical outcomes

SPSS' procedures for predicting numerical outcomes aren't limited to the ones we just described. The following procedures help give SPSS 11.0 what you need for prediction:

- Linear Regression
- Weighted Least Squares Regression
- Two-Stage Least Squares
- Survival Analysis procedures
  - Cox Regression with time-dependent covariates
  - Kaplan-Meier
  - Life Tables

Want to know what other statistics — including stats for identifying groups and time-series analysis — and software SPSS offers for the complete analytical process? Visit [www.spss.com/statisticalmethods](http://www.spss.com/statisticalmethods) to download a white paper, "Complete end-to-end analysis with SPSS 11.0." Do you like what you see? You can buy SPSS 11.0 online at [www.spss.com/store](http://www.spss.com/store) or call (800) 543-9247.

SPSS BI helps people solve business problems using statistics and data mining. This predictive technology enables our customers in the commercial and public sectors to make better decisions and improve results. SPSS BI software and services are used successfully in a wide range of applications, including customer attraction and retention, cross-selling, survey research, fraud detection, Web site performance, forecasting and scientific research. SPSS BI's market-leading products include SPSS® Clementine®, AnswerTree®, DecisionTime® and SigmaPlot®

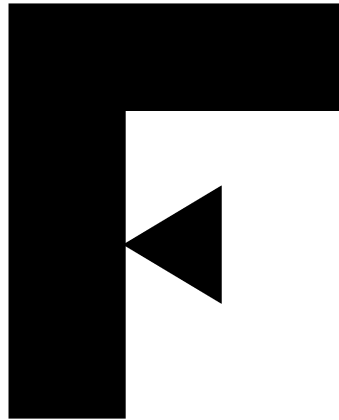


Call today  
for an SPSS  
product catalog  
(800) 543-9247

*“Perfection is achieved, not when there is nothing more to add, but when there is nothing left to take away.”*

- Antoine de Saint Exupery

F is a carefully crafted subset of the most recent version of Fortran, the world’s most powerful numeric language.



Using F has some very significant advantages:

- Programs written in F will compile with any Fortran compiler
- F is easier to use than other popular programming languages
- *F compilers are free* and available for Linux, Windows, and Solaris
- Several books on F are available
- F programs may be linked with C, Fortran 95, or older Fortran 77 programs

F retains the modern features of Fortran—modules and data abstraction, for example—but discards older error-prone facilities of Fortran.

It is a safe and portable programming language.

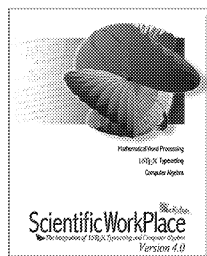
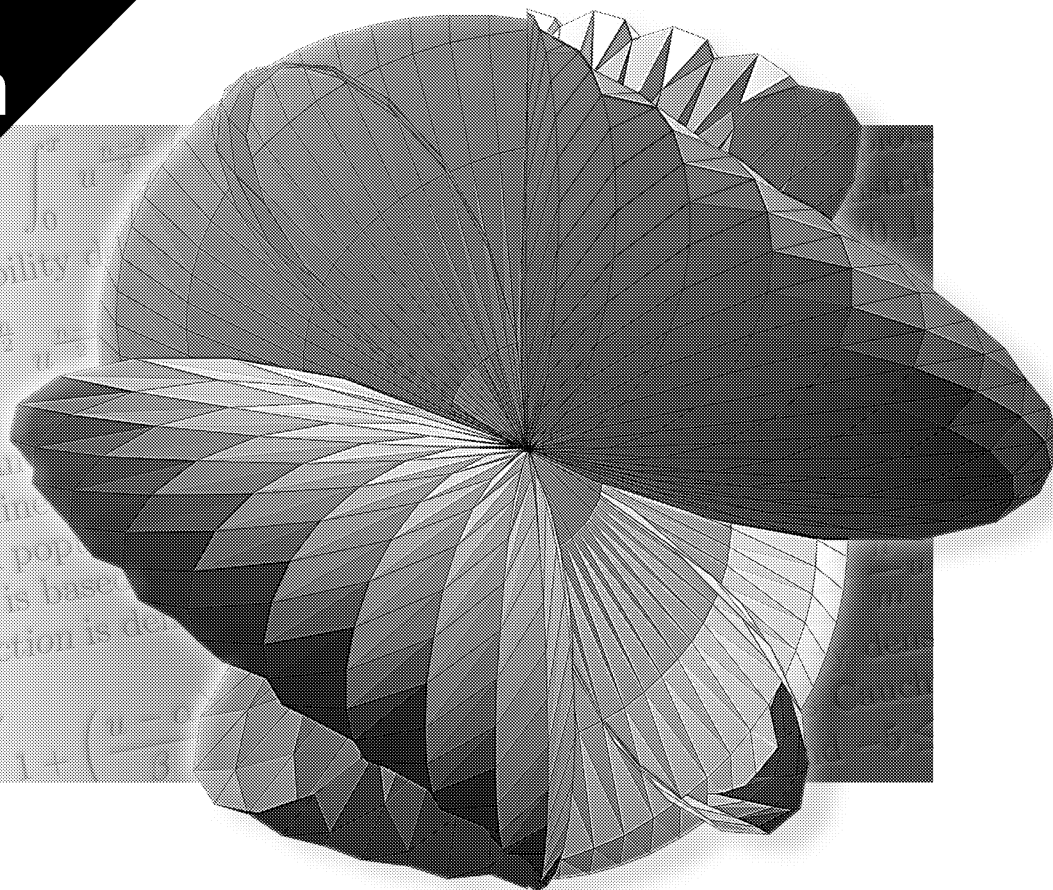
F encourages Module-Oriented Programming.

It is ideal for teaching a programming language in science, engineering, mathematics, and finance.

It is ideal for new numerically intensive programs.

The Fortran Company  
11155 E. Mountain Gate Place, Tucson, AZ 85749 USA  
+1-520-256-1455 +1-520-760-1397 (fax)  
<http://www.fortran.com> [info@fortran.com](mailto:info@fortran.com)

Introducing  
**Version  
4.0**



Math Word Processing  
**L<sup>A</sup>T<sub>E</sub>X** Typesetting  
Computer Algebra

The Gold Standard for Mathematical Publishing  
and the Easiest-to-Use Computer Algebra System

*Now in a new version!*

*Scientific WorkPlace* makes writing, publishing, and  
doing mathematics easier than you ever imagined  
possible.

- ◆ Enter text and mathematics naturally in the same paragraph
- ◆ Produce documents with or without L<sup>A</sup>T<sub>E</sub>X typesetting
- ◆ Produce portable L<sup>A</sup>T<sub>E</sub>X output
- ◆ Perform mathematical computations with both the *MuPAD*<sup>®</sup> and *Maple*<sup>®</sup> computer algebra engines
- ◆ Export documents as HTML, with mathematics exported as graphics or as MathML
- ◆ Use hyperlinking to create an entire web of *Scientific WorkPlace* documents
- ◆ And more

 **MacKichan**  
SOFTWARE, INC.

# Scientific WorkPlace<sup>®</sup>

Email: [info@mackichan.com](mailto:info@mackichan.com) ◆ Toll Free: 877-724-9673 ◆ Fax: 206-780-2857

Visit our website for free evaluation copies of all our software.

[www.mackichan.com/jmsm](http://www.mackichan.com/jmsm)



# Find Your Path With Us

**HIGHLY SPECIALIZED PERMANENT & CONTRACT  
OPPORTUNITIES AVAILABLE**

- BIostatISTICS
- SAS® PROGRAMMING
- DATA MANAGEMENT
- MARKETING SCIENCE
- RESEARCH
- MANUFACTURING

## **Permanent Placement**

Contact: Tracey Gmoser

800.989.5627

Fax: 212.818.9067

[perm@smithhanley.com](mailto:perm@smithhanley.com)

[www.smithhanley.com](http://www.smithhanley.com)

## **Contract Staffing**

Contact: Keith Shelly

800.684.9921

Fax: 407.805.3020

[contract@smithhanley.com](mailto:contract@smithhanley.com)

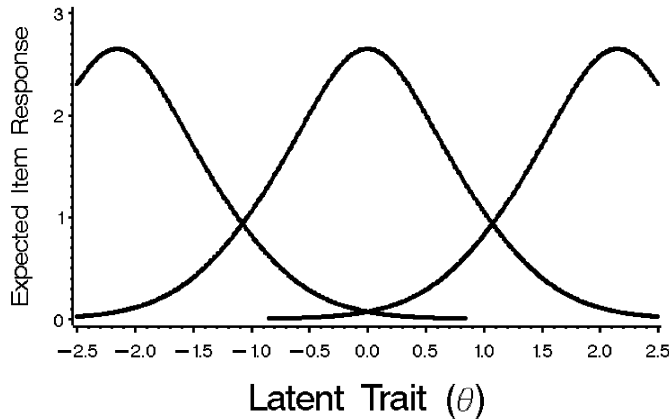
[www.smithhanley-consulting.com](http://www.smithhanley-consulting.com)

*Smith Hanley*

New York • Chicago • Houston • Southport • Orlando

# ***GGUM2000***

## *Item Response Theory Models for Unfolding*



The GGUM2000 software system estimates parameters in a family of item response theory (IRT) models that unfold polytomous responses to questionnaire items. These models assume that persons and items can be jointly represented as locations on a latent unidimensional continuum. A single-peaked, nonmonotonic response function is the key feature that distinguishes unfolding IRT models from traditional, "cumulative" IRT models. This response function suggests

that a higher item score is more likely to the extent that an individual is located close to a given item on the underlying continuum. Such single-peaked functions are appropriate in many situations including attitude measurement with Likert or Thurstone scales, and preference measurement with stimulus rating scales. This family of models can also be used to determine the locations of respondents in particular developmental processes that occur in stages.

The GGUM2000 system estimates item parameters using marginal maximum likelihood, and person parameters are estimated using an expected a posteriori (EAP) technique. The program allows for up to 100 items with 2-10 response categories per item, and up to 2000 respondents. The software is accompanied by a detailed user's manual. **GGUM2000 is free** and can be downloaded from:

<http://www.education.umd.edu/EDMS/tutorials>

Start putting the power of unfolding IRT models to work in your attitude and preference measurement endeavors. Download your free copy of GGUM2000 today!



## Are you involved in Data Modeling or Data Mining?

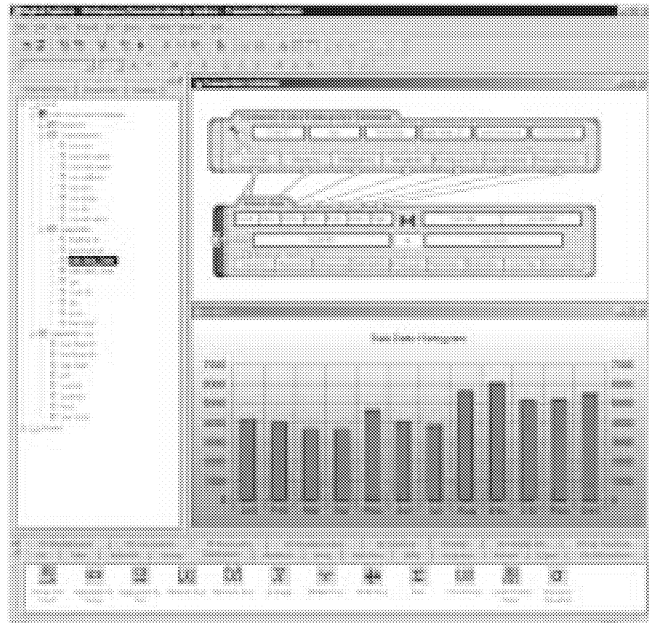
## Are you spending a large percentage of your time dealing with data issues?

If so, you will be happy to know that we have developed a tool that specifically addresses the data prep tasks associated with data modeling and data mining. The tool is called the Digital Excavator from Digital Archaeology ([www.digarch.com](http://www.digarch.com)). Data modelers are well aware of the time-consuming and sometimes frustrating nature of data set-up. In many cases data preparation can represent 60%-80% of the data mining project length. With Digital Archaeology's Digital Excavator, data preparation tasks are streamlined, results are more accurate, and the modeler has more time to focus on finding the appropriate mathematical solution--rather than wasting time with painful data issues. Digital Archaeology's software is intuitive, visual, self-documenting, and deploys what a number of analysts and customers have termed the "most elegant" user interface for data analysis and exploration ever conceived. It's the only tool specifically designed for the data prep tasks of data modeling.

**Visit our website and see for yourself! >>>> [www.digarch.com](http://www.digarch.com)**

Functions have been created which perform the following:

- Frequency Distributions
- Categorical Variable Profile
- Continuous Variable Profile
- Histograms
- De-duping
- Find and Replace Missing Values
- Find and Split Out Outliers
- Binning
- Correlation Matrix
- Cross-Tabs
- Panel Variables (Occupancy Map)
- Lag functions
- Decimal Scaling
- Rank and Sample Variables
- Recency, Frequency, Monetary Analysis
- N-Tile Distributions
- Gains Charts
- Many others



15721 COLLEGE BOULEVARD  
LENEXA, KS 66219  
1-877-DIGARCH (344-2724)  
[WWW.DIGARCH.COM](http://WWW.DIGARCH.COM)

# Ready to Take Your Next Step?

As your career climbs and  
each step requires careful  
planning, consider  
The Cambridge Group...

...Your Success is Our Business.

## **Business Statistics**

- Quantitative Analysis
- Marketing Sciences/Research
- Econometrics
- Quality Assurance

*stat@cambridgegroup.com*

## **Consultant & Contract Staffing**

- Biostatisticians
- SAS®/Statistical Programmers
- Clinical Data Managers
- Clinical Systems
- CRA's & Clinical Monitors
- Medical Writers
- Project Managers
- Bioinformatics

*contract@cambridgegroup.com*

## **Clinical Computing & Data Management**

- SAS Programming/Application
- Clinical Data Management
- Systems Design & Analysis

*QA@cambridgegroup.com*

## **Biostatistics**

- Clinical
- Preclinical/Nonclinical
- Health Outcomes
- PK/PD

*biostat@cambridgegroup.com*

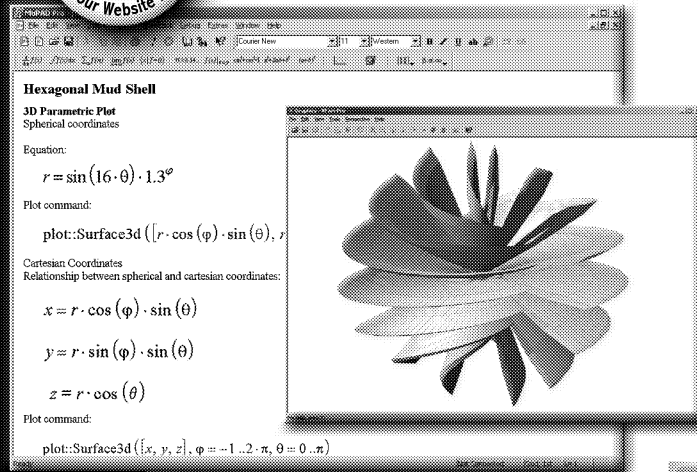


**THE CAMBRIDGE GROUP LTD**

**(800) 525-3396 fax (203) 226-3856**  
**www.cambridgegroup.com**

# MuPAD<sup>®</sup> Pro

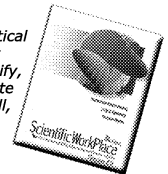
Version  
**2.0**  
See our Website for details



## The Open Computer Algebra System

MacKichan Software is proud to bring you MuPAD Pro, a full-featured computer algebra system in an integrated and open environment for symbolic and numeric computing. The MuPAD language has a Pascal-like syntax and allows imperative, functional, and object-oriented programming. Its domains and categories are like object-oriented classes that allow over-riding and overloading methods and operators, inheritance, and generic algorithms. A comfortable notebook interface includes a graphics tool for visualization, an integrated source-level debugger, a profiler, and hypertext help.

*Scientific WorkPlace, the proven solution for mathematical publishing, is an excellent companion to MuPAD Pro. It integrates with MuPAD so that you can evaluate, simplify, solve, and plot from inside your document, and evaluate functions that you have defined with MuPAD. Best of all, you typeset in LaTeX with just the click of a button.*



**MacKichan**  
SOFTWARE, INC.

Tools for Scientific Creativity Since 1981

Toll Free: 877-724-9673 • Email: [info@mackichan.com](mailto:info@mackichan.com)

Go to our homepage for free trial versions of all our products.

[www.mackichan.com/jmsm](http://www.mackichan.com/jmsm)

## Announcing the highly-anticipated new Numerical Recipes products

### Numerical Recipes in C++

The Art of Scientific Computing  
Second Edition

William H. Press, Saul A. Teukolsky,  
William T. Vetterling, and  
Brian P. Flannery

"This monumental and classic work is beautifully produced and of literary as well as mathematical quality. It is an essential component of any serious scientific or engineering library."

—*Computing Reviews*

This new version incorporates completely new C++ versions of the more than 300 *Numerical Recipes Second Edition* routines widely recognized as the most accessible and practical basis for scientific computing, in addition to including the full mathematical and explanatory contents of *Numerical Recipes in C*.

### Key Features:

- Includes linear algebra, interpolation, special functions, random numbers, nonlinear sets of equations, optimization, eigensystems, Fourier methods and wavelets, statistical tests, ODEs and PDEs, integral equations, and inverse theory.
- The routines, in ANSI/ISO C++ source code, can be used with almost any existing C++ vector/matrix class library, according to user preference

0-521-75033-4, Hardback, \$70.00

Visit [us.cambridge.org/numericalrecipes](http://us.cambridge.org/numericalrecipes) for more information on the complete line of Numerical Recipes products.

Available in bookstores or from



**CAMBRIDGE**  
UNIVERSITY PRESS

800-872-7423

[us.cambridge.org/mathematics](http://us.cambridge.org/mathematics)

### Other new Numerical Recipes products for your library...

#### Numerical Recipes Example Book [C++]

0-521-75034-2, Paperback, \$35.00

#### Numerical Recipes in C and C++ Source Code CDROM with Windows, DOS, or Macintosh Single Screen License

0-521-75037-7, CD-ROM, \$50.00

#### Numerical Recipes Multi-Language Code CDROM with LINUX or UNIX Single Screen License

Source Code for Numerical Recipes in C, C++, Fortran 77, Fortran 90, Pascal, BASIC, Lisp and Modula 2 plus many extras  
0-521-75036-9, CD-ROM, \$150.00

#### Numerical Recipes Multi-Language Code CDROM with Windows, DOS, or Macintosh Single Screen License

Source Code for Numerical Recipes in C, C++, Fortran 77, Fortran 90, Pascal, BASIC, Lisp and Modula 2 plus many extras  
0-521-75035-0, CD-ROM, \$90.00



# Numerical Recipes in Fortran from Cambridge University Press

## Numerical Recipes in Fortran 77

Volume 1 of Fortran Numerical Recipes  
Second Edition

*William H. Press, Saul A. Teukolsky,  
William T. Vetterling, and Brian P. Flannery*

"This reviewer knows of no other single source of  
so much material of this nature. Highly recommended."

—*Choice*

"...a valuable resource for those with a specific need for  
numerical software. The routines are prefaced with lucid, self-  
contained explanations...highly recommended for those who  
require the use and understanding of numerical software."

—*SIAM Review*

1992 992 pp. 0-521-43064-X Hardback \$70.00

### *Highlights include:*

- A chapter on integral equations and inverse methods
- Multigrid and other methods for solving partial differential equations
- Improved random number routines
- Wavelet transforms
- The statistical bootstrap method
- A chapter on "less-numerical" algorithms including compression coding and arbitrary precision arithmetic.

## Numerical Recipes in Fortran 77 Example Book

Second Edition

*William T. Vetterling, Saul A. Teukolsky, William H. Press, and Brian P. Flannery*

1992 256 pp. 0-521-43721-0 Paperback \$35.00

## Numerical Recipes in Fortran 90

The Art of Parallel Scientific Computing  
Volume 2 of Fortran Numerical Recipes  
Second Edition

*William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery*

"This present volume will contribute decisively to a significant breakthrough, as it provides models not only of the numerical algorithms for which previous editions are already famed, but also of an excellent Fortran 90 style."

—*From the Foreword by Michael Metcalf, one of Fortran 90's original designers and author of FORTRAN 90 Explained*

"This book is a classic and is essential reading for anyone concerned with the future of numerical calculation. It is beautifully produced, inexpensive for its content, and a must for any serious worker or student."

—*Computing Reviews*

Contains a detailed introduction to the Fortran 90 language and to the basic concepts of parallel programming, plus source code for all routines from the second edition of Numerical Recipes.

1996 576 pp. 0-521-57439-0 Hardback \$50.00

## Numerical Recipes Multi-Language Code CDROM with LINUX or UNIX Single Screen License

Source Code for Numerical Recipes in C, C++, Fortran 77, Fortran 90, Pascal, BASIC, Lisp and Modula 2 plus many extras

2002 0-521-75036-9 CD-ROM \$150.00

## Numerical Recipes Multi-Language Code CDROM with Windows, DOS, or Macintosh Single Screen License

Source Code for Numerical Recipes in C, C++, Fortran 77, Fortran 90, Pascal, BASIC, Lisp and Modula 2 plus many extras

2002 0-521-75035-0 CD-ROM \$90.00

Visit [us.cambridge.org/numericalrecipes](http://us.cambridge.org/numericalrecipes) for more information on the complete line of *Numerical Recipes* products.

Available in bookstores or from



**CAMBRIDGE**  
UNIVERSITY PRESS

800-872-7423

[us.cambridge.org/mathematics](http://us.cambridge.org/mathematics)

# NEW! XML PLUG-IN FOR sas®

**IMPORT XML-FORMATTED DATA INTO SAS  
EXPORT SAS DATA AS XML-FORMATTED DATA**

Now, an easy way to move CDISC and other XML-formatted data into or out of your SAS-based systems. You don't have to know perl, XSLT, Xpath, Java®, or exotic languages. The remarkable Tekoa™ XML plug-in does it all for you.

Provided free of charge to any SAS user currently wrestling with XML. Developed by Zurich Biostatistics, the pioneer in SAS/XML integration.

## **FREE. EASY. AND IT WORKS.**

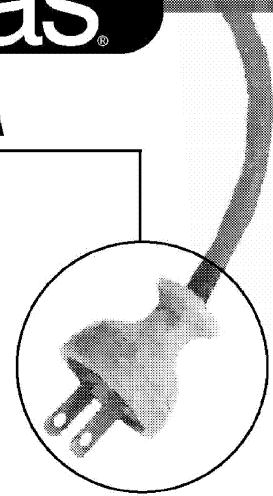
Just e-mail Michael Palmer (mcpalmer@zbi.net) and receive the fully-functional, proven Tekoa XML plug-in by e-mail.

*No charge. No obligation. No hassle. (We even support the tool. Imagine.)*

## **Zurich Biostatistics, Inc.**

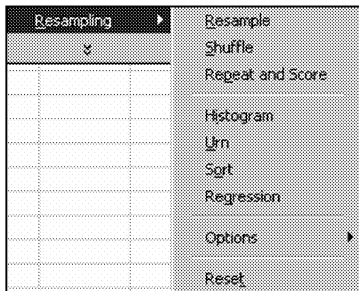
45 Park Place South, PMB 178, Morristown, NJ 07960 973 727-0025 [www.zbi.net](http://www.zbi.net)

***XML is easy if you know how. And we do.***



Tekoa XML Technology is a service mark of Zurich Biostatistics, Inc. SAS is a registered trademark of SAS Institute Inc. Java is a registered trademark of Sun Microsystems, Inc.

# Resampling Stats for Excel



Select the data you want to resample, select "resample" or "shuffle," then specify an output range for the resampled data. Calculate a statistic of interest, select "Repeat & Score," and the resampling operation will be repeated thousands (or tens of thousands) of times, and each time the value of your statistic of interest will be recorded. Does not use Excel's random number generator.

**View complete user guide and download free  
30-day trial at <[www.resample.com](http://www.resample.com)>**

\$249 commercial • \$149 personal/academic • \$89 student

612 N Jackson St., Arlington, VA 22201  
Tel 703-522-2713 • Fax 703-522-5846  
[stats@resample.com](mailto:stats@resample.com)



# ANNOUNCING LOGXACT 4

**Only LogXact 4 can fit a logistic regression model to the data on this page —even LogXact 2 cannot do it.**

To solve this problem, you need a very powerful exact logistic regression algorithm. LogXact 4 implements a ground-breaking network Monte Carlo algorithm (published in *JASA*, April 2000), extending the scope of LogXact to problems previously beyond its capacity.

PLUS! LogXact 4 also provides exact Poisson regression (used extensively for cohort studies in epidemiology).



## Take the Cytel Challenge

Data were gathered on 2,493 hospitalized patients, of whom 60 suffered from *clostridium difficile*, an acute form of diarrhea. Of interest was the relationship between the occurrence of diarrhea and age, length of hospital stage, sex, use of the antibiotic Clindomycin, and the use of the antibiotic Cephalexin.

When you have data like these (low response rates,

unbalanced covariates or small sample sizes), traditional logistic regression methods used by standard statistical packages often fail. But new LogXact 4 can fit models, test parameters and estimate coefficients even when the maximum likelihood method used by most statistical software can't. Plus, since LogXact gives you exact answers instead of approximations, it protects you from Type-I error.

### 60 cases of diarrhea among 2,493 hospitalized patients

	Group 1	Group 2	•	Group 18
Cephalexin	0	0	•	1
Clindomycin	0	0	•	0
Sex	1	1	•	0
Age	0	0	•	1
LOS	0	1	•	1
<b>Diarrhea/Total ( 60/2,493)</b>	<b>0/174</b>	<b>1/113</b>	<b>•</b>	<b>4/4</b>

**For the full data set, the solution and a free 30 day trial of LogXact 4, visit [www.cytel.com](http://www.cytel.com)**

**Call (617) 661-2011; fax (617) 661-4405; e-mail: [sales@cytel.com](mailto:sales@cytel.com)**

CYTEL Software Corporation • 675 Massachusetts Ave., Cambridge, MA 02139 USA  
Tel (617) 661-2011 • Toll Free (US) 866-298-3511 • Fax (617) 661-4405  
<http://www.cytel.com> • E-mail: [sales@cytel.com](mailto:sales@cytel.com)

INTERNATIONAL DISTRIBUTORS: Ask Int'l (UK) e-mail: [cyteluk@asru.com](mailto:cyteluk@asru.com) • Tel: +44(0) 1227 795 240 • Fax: +44(0) 1227 795 201; ID2 (Belgium)  
• Tel: 32 2 6468918 • Fax: 32 2 4468662; Spadille Biostatistics ApS (Denmark) • [spadille@spadille.dk](mailto:spadille@spadille.dk) • Tel: 4548.484100 • Fax: 4248484200

**Cytel**  
STATISTICAL SOFTWARE

## JOIN DIVISION 5 OF APA!

The Division of Evaluation, Measurement, and Statistics of the American Psychological Association draws together individuals whose professional activities and/or interests include assessment, evaluation, measurement, and statistics. The disciplinary affiliation of division membership reaches well beyond psychology, includes both members and non-members of APA, and welcomes graduate students.

Benefits of membership include:

- subscription to *Psychological Methods* or *Psychological Assessment* (student members, who pay a reduced fee, do not automatically receive a journal, but may do so for an additional \$18)
- *The Score* – the division's quarterly newsletter
- Division's Listservs, which provide an opportunity for substantive discussions as well as the dissemination of important information (e.g., job openings, grant information, workshops)

Cost of membership: \$38 (**APA membership not required**); student membership is only \$8

For further information, please contact the Division's Membership Chair, Yossef Ben-Porath ([ybenpora@kent.edu](mailto:ybenpora@kent.edu)) or check out the Division's website:

<http://www.apa.org/divisions/div5/>

---

## ARE YOU INTERESTED IN AN ORGANIZATION DEVOTED TO EDUCATIONAL AND BEHAVIORAL STATISTICS?

Become a member of the **Special Interest Group - Educational Statisticians** of the American Educational Research Association (SIG-ES of AERA)!

The mission of SIG-ES is to increase the interaction among educational researchers interested in the theory, applications, and teaching of statistics in the social sciences.

Each Spring, as part of the overall AERA annual meeting, there are seven sessions sponsored by SIG-ES devoted to educational statistics and statistics education.

We also publish a twice-yearly electronic newsletter.

Past issues of the SIG-ES newsletter and other information regarding SIG-ES can be found at <http://orme.uark.edu/edstatsig.htm>

To join SIG-ES you must be a member of AERA. Dues are \$5.00 per year.

For more information, contact Joan Garfield, President of the SIG-ES, at [jbg@umn.edu](mailto:jbg@umn.edu).

## Instructions For Authors

Follow these guidelines when submitting a manuscript:

1. The most recent American Psychological Association style guidelines are preferred.
2. Submissions are accepted via e-mail only. Send them to the Editorial Assistant at [ea@edstat.coe.wayne.edu](mailto:ea@edstat.coe.wayne.edu). Provide name, affiliation, address, e-mail address, and 30 word biographical statements for all authors in the body of the email message.
3. There should be no material identifying authorship except on the title page. A statement should be included in the body of the e-mail that, where applicable, indicating proper human subjects protocols were followed, including informed consent. A statement should be included in the body of the e-mail indicating the manuscript is not under consideration at another journal.
4. Provide the manuscript as an external e-mail attachment in MS Word for the PC format only. (Wordperfect and .rtf formats may be acceptable - please inquire.) Please note that Tex (in its various versions), Exp, and Adobe .pdf formats are designed to produce the final presentation of text. They are not amenable to the editing process, and are not acceptable for manuscript submission.
5. The text maximum is 20 pages double spaced, not including tables, figures, graphs, and references. Use 11 point Times Roman font. If the technical expertise is available, submit the manuscript in two column format.
6. Create tables without boxes or vertical lines. Place tables, figures, and graphs “in-line”, not at the end of the manuscript. Figures may be in .jpg, .tif, .png, and other formats readable by Adobe Illustrator or Photoshop.
7. The manuscript should contain an Abstract with a 50 word maximum, following by a list of key words or phrases. Major headings are Introduction, Methodology, Results, Conclusion, and References. Center headings. Subheadings are left justified; capitalize only the first letter of each word. Sub-subheadings are left-justified with indent.
8. Do not use underlining in the manuscript. Do not use bold, except for (a) matrices, or (b) emphasis within a table, figure, or graph. Do not number sections. Number all formulas, tables, figures, and graphs, but do not use italics, bold, or underline. Do not number references. Do not use footnotes or endnotes.
9. In the References section, do not put quotation marks around titles of articles or books. Capitalize only the first letter of books. Italicize journal or book titles, and volume numbers. Use “&” instead of “and” in multiple author listings.
10. *Suggestions for style*: Instead of “I drew a sample of 40” write “A sample of 40 was selected”. Use “although” instead of “while”, unless the meaning is “at the same time”. Use “because” instead of “since”, unless the meaning is “after”. Instead of “Smith (1990) notes” write “Smith (1990) noted”.

### Print Subscriptions

Print subscriptions including postage for professions is US \$60 per year; graduate students is US \$30 per year; and libraries, universities, and corporations is US \$195 per year. Subscribers outside of the US and Canada pay a US \$10 surcharge for additional postage. Online access is currently free at <http://tbf.coe.wayne.edu/jmasm>. Mail subscription requests with remittances to JMASM, P. O. Box 48023, Oak Park, MI, 48237. Email journal correspondence, other than manuscript submissions, to [jmasm@edstat.coe.wayne.edu](mailto:jmasm@edstat.coe.wayne.edu).

### Notice To Advertisers

Send requests for advertising information to [jmasm@edstat.coe.wayne.edu](mailto:jmasm@edstat.coe.wayne.edu).



# Lahey/Fujitsu Fortran

The standard for Fortran programming  
from the leader in Fortran language systems

SOFTWARE SOLUTIONS  
for Science & Engineering

## LF95 Fortran for Linux and Windows

Full Fortran 95/90/77 support  
Unsurpassed diagnostics  
Intel and AMD optimizations

IMSL compatible  
Fujitsu SSL2 math library  
Wisk graphics package

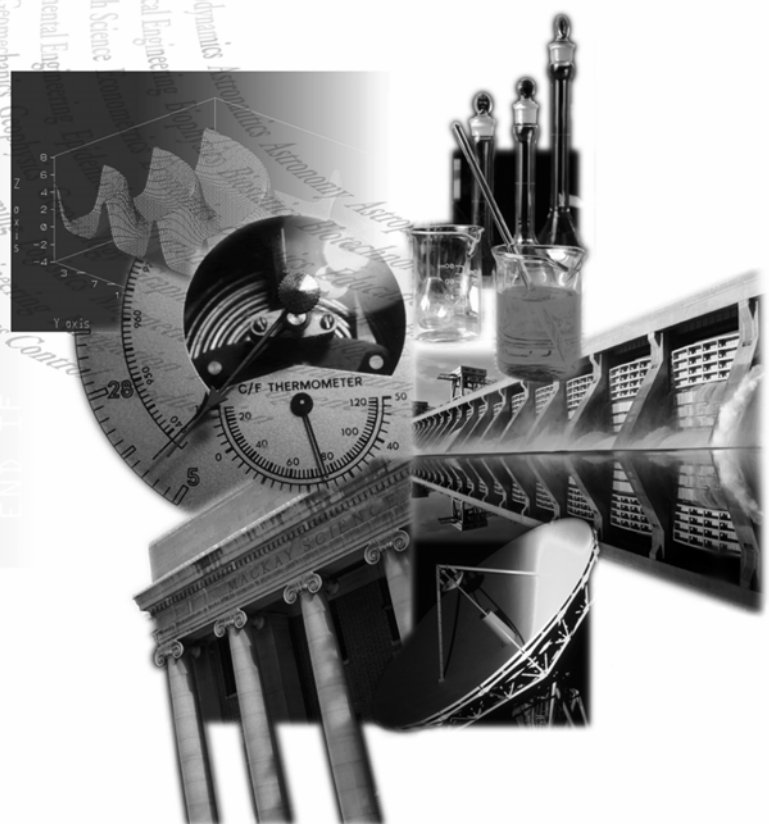
## LF Fortran for the Microsoft® .NET Framework - Coming Soon !

Visual Studio integration  
Windows / Web Forms designer  
Project and code templates

On-line integrated help  
XML Web services  
ADO.NET support

Visit [www.lahey.com](http://www.lahey.com) for more information

```
ELSE
  poly_coef
END IF
ELSE
  poly_coef
END IF
END FUNCTION poly_c
SUBROUTINE poly_ini
TYPE(poly), INTENT
REAL(fpkind), INTE
IF ( .NOT. PRESENT
  NULLIFY ( p%coef
ELSE
  m = UBOUND(v,i)
  IF ( max_degree
  ALLOCATE ( p%
  p%coeffs
ELSE
  ALLOC
  p%coeffs
END IF
END IF
```



Lahey Computer Systems, Inc.  
865 Tahoe Blvd - P.O. Box 6091  
Incline Village, NV 89450 USA  
1-775-831-2500  
[www.lahey.com](http://www.lahey.com)

# DataMinelt<sup>SM</sup>

announces

# Permutelt<sup>TM</sup> v2.0

## The fastest, most comprehensive and robust permutation test software on the market today.

Permutation tests increasingly are the statistical method of choice for addressing business questions and research hypotheses across a broad range of industries. Their distribution-free nature maintains test validity where many parametric tests (and even other nonparametric tests), encumbered by restrictive and often inappropriate data assumptions, fail miserably. The computational demands of permutation tests, however, have severely limited other vendors' attempts at providing useable permutation test software for anything but highly stylized situations or small datasets and few tests. Permutelt<sup>TM</sup> addresses this unmet need by utilizing a combination of algorithms to perform non-parametric permutation tests very quickly – often more than an order of magnitude faster than widely available commercial alternatives when one sample is large and many tests and/or multiple comparisons are being performed (which is when runtimes matter most). Permutelt<sup>TM</sup> can make the difference between making deadlines, or missing them, since data inputs often need to be revised, resent, or recleaned, and one hour of runtime quickly can become 10, 20, or 30 hours.

In addition to its speed even when one sample is large, some of the unique and powerful features of Permutelt<sup>TM</sup> include:

- the availability to the user of a wide range of test statistics for performing permutation tests on continuous, count, & binary data, including: pooled-variance t-test; separate-variance Behrens-Fisher t-test, scale test, and joint tests for scale and location coefficients using nonparametric combination methodology; Brownie et al. "modified" t-test; skew-adjusted "modified" t-test; Cochran-Armitage test; exact inference; Poisson normal-approximate test; Fisher's exact test; Freeman-Tukey Double Arcsine test
- extremely fast exact inference (no confidence intervals – just exact p-values) for most count data and high-frequency continuous data, often several orders of magnitude faster than the most widely available commercial alternative
- the availability to the user of a wide range of multiple testing procedures, including: Bonferroni, Sidak, Stepdown Bonferroni, Stepdown Sidak, Stepdown Bonferroni and Stepdown Sidak for discrete distributions, Hochberg Stepup, FDR, Dunnett's one-step (for MCC under ANOVA assumptions), Single-step Permutation, Stepdown Permutation, Single-step and Stepdown Permutation for discrete distributions, Permutation-style adjustment of permutation p-values
- fast, efficient, and automatic generation of all pairwise comparisons
- efficient variance-reduction under conventional Monte Carlo via self-adjusting permutation sampling when confidence intervals contain the user-specified critical value of the test
- maximum power, and the shortest confidence intervals, under conventional Monte Carlo via a new sampling optimization technique (see Opdyke, JMASM, Vol. 2, No. 1, May, 2003)
- fast permutation-style p-value adjustments for multiple comparisons (the code is designed to provide an additional speed premium for many of these resampling-based multiple testing procedures)
- simultaneous permutation testing and permutation-style p-value adjustment, although for relatively few tests at a time (this capability is not even provided as a preprogrammed option with any other software currently on the market)

For Telecommunications, Pharmaceuticals, fMRI data, Financial Services, Clinical Trials, Insurance, Bioinformatics, and just about any data rich industry where large numbers of distributional null hypotheses need to be tested on samples that are not extremely small and parametric assumptions are either uncertain or inappropriate, Permutelt<sup>TM</sup> is the optimal, and only, solution.

To learn more about how Permutelt<sup>TM</sup> can be used for your enterprise, and to obtain a demo version, please contact its author, J.D. Opdyke, President, DataMinelt<sup>SM</sup>, at [JDOpdyke@DataMinelt.com](mailto:JDOpdyke@DataMinelt.com) or [www.DataMinelt.com](http://www.DataMinelt.com).

DataMinelt<sup>SM</sup> is a technical consultancy providing statistical data mining, econometric analysis, and data warehousing services and expertise to the industry, consulting, and research sectors. Permutelt<sup>TM</sup> is its flagship product.