International Education Journal, 2006, 7(2), 110-126.
ISSN 1443-1475 © 2006 *Shannon Research Press*.
http://iej.com.au

Some problems in the analysis of cross-national survey data

John P. Keeves
School of Education, Flinders University *john.keeves@flinders.edu.au*

Petra Lietz
International University Bremen, Germany *p.lietz@iu-bremen.de*

Kelvin Gregory
School of Education, Flinders University *kelvin.gregory@flinders.edu.au*

I Gusti Ngurah Darmawan
School of Education, Flinders University *ngurah.darmawan@flinders.edu.au*

In this lead article three emergent problems in the analysis of cross-national survey data are raised in a context of 40 years of research and development in a field where persistent problems have arisen and where scholars across the world have sought solutions. Anomalous results have been found from secondary data analyses that would appear to stem from the procedures that have been employed during the past 15 years for the estimation of educational achievement. These estimation procedures are briefly explained and their relationships to the observed anomalies are discussed. The article concludes with a challenge to the use of Bayesian estimation procedure, while possibly appropriate for the estimation of population parameters would appear to be inadequate for modelling scores that are used in secondary data analyses. Consequently, an alternative approach should be sought to provide data on the performance of individual students, if a clearer and more coherent understanding of educational processes is to be achieved through cross-national survey research.

Cross-national research, survey research, secondary data analysis,
Bayesian estimation procedures, educational achievement

INTRODUCTION

As the number of school-aged children has grown rapidly world-wide and the demand for the provision of both primary and secondary education has increased at an even greater rate, it has gradually become essential to monitor educational standards. A little over 40 years ago the International Association for the Evaluation of Educational Achievement (IEA) was established and it set a pattern for the undertaking of the monitoring of educational achievement. Subsequently new bodies have been formed including the Programme for International Student Assessment (PISA) by the Organisation for Economic Cooperation and Development (OECD), the Southern and Eastern Africa Consortium for Monitoring Educational Quality (SACMEQ) by the Ministers of 15 Sub-Saharan African countries, and there have been many independent studies conducted in single countries supported by the World Bank and other agencies. These different bodies have had similar objectives, but have gone about their work in different ways that have also changed over time. It would seem that these bodies have had four essential tasks to fulfil, although some studies might not have sought to undertake particular tasks.

1. The countries involved were to be ranked on educational performance of a particular kind, with appropriate estimates of standard errors.
2. Trends over time in educational performance of a particular kind were to be monitored and, where possible, factors influencing stability and change in performance were to be identified for each country.
3. Similarities and differences in both the factors and the patterns of factors influencing educational performance both within and between countries were to be identified, as well as the stability and change in the effects of these factors over time.
4. Research workers in each of the countries involved were to be trained in the conduct of studies concerned with assessment and the evaluation of educational achievement in order to plan for the raising of the standards of performance in each of the countries participating in the testing programs.

Each of the different bodies involved in the conduct of such testing programs have carried out these tasks to different extents in accordance with the financial resources available and the capacities of the research workers engaged in the programs to undertake the necessary analysis and training. However, the expansion of the combined efforts of the several bodies now involved has gone well beyond the initial expectations of the founders of IEA. As a consequence there has evolved gradually an understanding of the factors influencing both the provision of educational services and educational performance at the levels of students, classrooms and teachers, schools and school systems. Nevertheless, there is much more to be learnt and much more to be done in order to raise the standards of education in the schools of every country involved.

PERSISTENT PROBLEMS IN THE PRIMARY ANALYSIS OF DATA

Since the 1980s research workers have become very aware of certain persistent problems in the primary analysis of cross-national survey data. Several major problems have been encountered.

1. There was a partial failure of the models employed in the scaling and combining of test and questionnaire items to fit the data in particular countries, and under some circumstances there was a partial failure of tests of fit to detect a lack of fit because of circularity in some of the procedures used.
2. There was a need to conduct multilevel analyses at two or more levels (namely, students, classes, schools and strata or sub-systems) in order to model effectively the data recorded for both dependent and independent variables. There was also the need to calculate the appropriate errors of sampling in order to estimate accurately the statistical significance of the estimates of the parameters associated with such variables and the relationships between them (see Darmawan and Keeves, this issue, pp. 161-174, and pp. 175-190).
3. There were marked differences between countries in the best models that explained adequately the variability in the data associated with the variables under consideration. While there are sometimes strong similarities between groups of countries, there are commonly marked differences both within and between countries in the effects of certain independent variables on certain key dependent variables that lead to imposing serious limitations on the generalisations that can be drawn from the analyses (see Gregory, this issue, pp. 151-160) and (see Skuza, this issue, pp. 191-208).
4. Difficulties were encountered in the use of rotated test and questionnaire items when attempts were made to extend the coverage of different aspects of the school curriculum through

increasing the numbers of test and questionnaire items administered, without imposing too great a burden on individual students.

5. There were problems in the identification of appropriate models for combining data obtained from the administration of test and questionnaire items to form meaningful and consistent composite measures for the latent variables under consideration.

Since the 1960s there has been an ongoing debate about these issues in the universities and institutes engaged in educational research related to these assessment programs that have led to marked advances in the analysis of data in the field of education. These advances have gradually spread more widely to such fields as forestry, genetics, public health and the social and behavioural sciences. These developments include the techniques involved in structural equation modelling (eg. LISREL, PLSPATH, STREAMS, MPlus), multilevel analysis (e.g. HLM, MLwiN, MPlus) and measurement (e.g. Quest, RUMM, Bigsteps, ConQuest). New procedures to reduce the errors of measurement in population estimates have also emerged from the Educational Testing Service and Boston College in the United States and the Australian Council for Educational Research that have involved the use of conditioning and plausible values, which have remained obscured from a wider less technical audience until more general papers have been written recently by Adams (2005) and Wu (2005) from the Australian Council for Educational Research, and the University of Melbourne. These papers have made more readily accessible certain ideas associated with the procedures being widely employed in cross-national testing programs at a time when there is some concern about certain anomalous results that are encountered in the secondary data analysis of cross-national data.

SOME EMERGENT PROBLEMS IN THE SECONDARY ANALYSIS OF DATA

Ongoing efforts have been made over time to improve the quality of assessment and evaluation procedures. They have included: (a) the development of instruments that go beyond the administration of multiple choice test items to employ constructed response items with partial credit being given for less than complete responses; (b) the raising of the level of response rates both within and across schools; (c) the making of effective provision for the estimation of missing data, so that the designed samples may be adequately filled; (d) greatly improved methods of statistical analysis to estimate both direct and indirect effects of variables that influence educational outcomes at the between student, between classroom, between school, and between system levels; and (e) the use of meta-analytic and trend analysis procedures (see, Chiu and Khoo, 2005) to combine results from different countries, different studies and over time in order to develop a better understanding of stability and change in educational provision around the world.

Nevertheless, three highly anomalous findings have emerged from the secondary analysis of data that cast serious doubts on the strength and appropriateness of certain procedures that are currently being widely employed: (a) to provide for different tests being administered to different students, (b) to compensate for missing data, and (c) to remove or reduce measurement error in order to improve the accuracy of population estimates. These procedures have been developed to overcome the limitations of test and sample design and response measurement. These three anomalous findings are considered briefly and in turn.

1. Meta-analysis of gender differences in reading achievement

In order to examine the gender differences in reading achievement at the middle secondary school level across a wide range of countries, Lietz (this issue, pp. 127-150) carried out a meta-analysis study that involved 147 data sets from a large number of testing programs including the IEA Reading Comprehension Study in 1970/71, the IEA Reading Literacy Study in 1990/91, the National Assessment for the Evaluation of Educational Progress Studies (NAEP) in the United

States from 1971 to 2003, the Programme for International Student Assessment (PISA) in 2000 in 43 countries, as well as the Australian ASSP and LSAY studies and many other smaller investigations. The meta-analysis was carried out using hierarchical linear modelling (HLM) procedures.

In the analysis, the outcomes examined were effect sizes, with their estimated errors, using a procedure advanced by Raudenbush and Bryk (2002, p. 209). The striking and anomalous finding was that the estimated effect size was substantially higher for the PISA studies (ê = 0.24) with similarly high values for the NAEP studies for the most recent decade (1992-2003), but not before that period. In contrast, studies prior to 1992 showed considerably lower effect sizes as reflected, for example, in the estimated effect size for the Reading Literacy Study, conducted by IEA in 1990-91 that was not significant (ê = 0.02). In general, in these more recent studies the girls were outperforming the boys with estimated effects that were noticeably greater than would be expected by chance (see Lietz, this issue, pp. 127-150). It is possible that these findings reflected the influence of cultural change, not only in the United States and Australia, but also in the 40 and more other countries of the world that have participated in the PISA and IEA studies. However, it is also possible that these effects arise from the item selection procedures employed to avoid gender bias, or from the procedures used for scaling and compensating for missing data and improving the accuracy of the national estimates of performance.

2. Mathematics Proficiency of Secondary School Students in South Africa

Howie (2002) undertook a secondary analysis of data on mathematics proficiency conducted as part of the Third (Trends in) International Mathematics and Science Study-Repeat (TIMSS-R) in South Africa in 1998/1999 under the auspices of the International Association for the Evaluation of Educational Achievement (IEA). The striking finding was that the highest achieving students, coming largely from the Western Cape Province, scored approximately 100 score points or one student standard deviation below the international average of 487 score points. The Western Cape Province was the wealthiest and most urbanised province in South Africa, where English was widely used. While only about seven per cent of the total sample spoke English as the main language in the home, the students from Western Cape Province were in the main English speakers at home. The remarkably low level of proficiency of these students in Western Cape Province, indicated that they were probably about three years[1] behind the international average in their level of achievement in mathematics. This finding suggested that a highly anomalous result existed that demanded rather more thorough investigation, not merely of cultural effects, but also into how these scores were estimated.

3. Trends in Bulgarian Eighth Grade Mathematics Performance from 1995 to 1999

Gregory and Bankov (2005) undertook a secondary analysis of the performance in mathematics achievement of eighth grade students in Bulgaria between 1995 and 1999 in the Third (Trends in) International Mathematics and Science Studies (TIMSS and TIMSS-R). In 1995, the Bulgarian students had a mean achievement of 527 that was above the international mean of 500. However, in 1999 the level of Bulgarian mathematics achievement fell to 511, which was still significantly above the international average. This involved a decline in performance of about one-sixth of a student standard deviation or approximately half a year of schooling that not only was statistically significant but was also of considerable practical significance. The possibility of unknown problems in the sample design could not be ignored. Nevertheless, a decline in performance of this magnitude could well indicate some anomaly in the scaling procedures used, or a marked change in the structure of the tests employed that was associated with an incompatibility of the

[1] This estimate for Mathematics achievement is obtained from Afrassa and Keeves (1999).

items in the different test booklets with the Bulgarian mathematics curriculum, where much of the content tested had been taught to the Bulgarian students between one to four years before the eighth grade. However, it was also possible that changes made in 1997 to the Bulgarian school system contributed to this anomalous result, and consequently the findings of the study raised politically sensitive issues. Nevertheless, in view of the other two anomalies discussed above, this result might indicate that a problem existed in the procedures used in the scaling of achievement data that warranted further examination.

SOME GENERAL ISSUES IN THE ANALYSIS OF DATA

There are several issues in the analysis of data that emerged over the period during which cross-national studies of educational achievement had been carried out, that need to be recognised and understood before the three anomalous effects considered above can be discussed and possibly addressed.

1. The effects of bias due to missing data

The occurrence of substantial missing data at the student and school levels, in general, would have the effect of introducing bias to both the estimates of the mean level of achievement and the variance. Such bias would be likely to inflate the mean level of achievement and reduce the variance, because it would seem likely that some lower performing students and schools from the designed samples would fail to participate in the study. These sources of bias would serve not only to distort the estimated level of performance, but also to reduce the capacity of the analysis of variance procedures employed in subsequent analyses to detect effects.

2. The effects of non-normal multilevel generating distributions of data

For the achievement test outcome variables and the indicators of attitude and the contextual variables formed by the summation of scores or by principal component procedures the underlying generating distributions would be likely to involve approximately normal distributions. However, there would be many key variables that would have to be included in the analyses of the data, such as the sex of a student and school type that could not be considered to be normally distributed. The failure to have underlying normal distributions would not only be likely to influence the use of significance tests, but could also influence the use of certain maximum likelihood estimation procedures. Sometimes, however, appropriate transformations could be used. Nevertheless, the underlying normality of the generating distributions would require very careful consideration, if and when maximum likelihood estimation procedures were employed.

3. The level of analysis problem

Only in the period since 1985 have effective analytical procedures been made available for an effective consideration of the multilevel analysis problem that has existed in educational research studies, where data were collected from students nested within schools. While over the past 40 years procedures have been employed to make some allowance for this aspect of the study and sample design in significance testing, it has not been possible to provide for the clustered sample design in the estimation of effects at appropriate levels until very recently. Even within the more highly developed countries there would sometimes, but not always, be substantial problems arising from the design of the sample, where these effects differed markedly between variables. Moreover, in many developing countries that currently participate in the IEA and PISA studies there would be very substantial design effects not only associated with individual schools but also associated with clearly identifiable regions and types of schools, as for example, academic, comprehensive and technical schools. These would require the use of a third level of analysis for

the appropriate estimation of statistical significance and the unbiased estimation of effects, since these school effects would be fixed effects and systematic in nature, and not random effects.

4. Bivariate and multivariate analysis

In a major debate that occurred 40 years ago the analysis of data in cross-national achievement studies shifted from an examination of bivariate relationships using simple analysis of variance procedures to an examination of multivariate relationships using regression procedures. Subsequently, a further development in the use of regression procedures led to the estimation of not only the direct effects of variables, but also the indirect effects of variables on the outcomes under consideration. However, the simplicity of bivariate relationships would seem still to have its appeal, whereas the real world of schooling would appear to be built out of a complex network of direct and indirect effects that required careful modelling at different levels of analysis. All efforts involved in the development of carefully constructed and trialled questionnaires would ultimately be wasted, if only bivariate relationships were examined and if multilevel and multivariate path models were not constructed to represent and tease out the effects of factors that influenced the educational outcomes within a particular education system and between educational systems (see, Chiu and Khoo, 2005).

5. The specification of regression models

The development and testing of regression models clearly would demand a thorough and systematic multilevel and multivariate analysis of variance using regression or maximum likelihood estimation procedures, with full recognition that each education system was likely to be very different from its neighbouring systems, because of the historical and cultural factors that had led to the formation in each country of a unique education system. While the questionnaires employed in the cross-national achievement surveys have sought to obtain meaningful data from students, teachers, and school principals, the questionnaires have frequently been returned with substantial missing and inconsistent information. Consequently, appropriate regression based procedures would be required to provide estimated values of missing data in those questionnaires where such data were missing or were inconsistent.

With the increasing number of countries involved in the surveys it is becoming more and more difficult to develop questionnaires that obtain meaningful data from the wide range of countries involved. Moreover, while in some highly developed countries there is little variability between schools in both their characteristics and the levels of achievement of their students, for many developing countries there are frequently wide disparities both in characteristics and levels of achievement. As a consequence, there are major differences between countries in the structures of the models that are constructed to explain optimally the differences between schools and students in their levels of achievement. Furthermore, there are likely to be large differences between countries in the explanatory power, in terms of proportion of variance explained, in the optimal models developed to account for variation in achievement outcomes.

6. The construction of tests and the sampling of test items

A major problem in the conduct of a testing program is that there is a relatively small limit to the number of test and questionnaire items to which a student can be asked to respond. This demands that in any content domain each student is required to answer only a sample of the test items that are employed to cover the content domain with adequate content and construct validities. A balanced incomplete block (BIB) design is currently widely used and compensation is made in estimating test scores not only for missing data but also for the different tests answered by different students.

Weiss and Yoes (1992) stated that there were two major approaches for estimating student performance, namely, the maximum likelihood method and the Bayesian estimation method. The maximum likelihood approach gave rise to two commonly used estimation procedures, either the Rasch (one parameter) modelling of the data or the three-parameter modelling of the item data collected in the testing program. The former modelling procedure provided measures that were said to be independent of the items sampled and the persons involved in the calibration of the scale of measurement. The former procedure also demanded that both the items and the persons tested must satisfy strict requirements of uni-dimensionality. The latter procedure, while claiming to be more accurate, has generally been found to be less robust.

In order to improve estimation and to compensate both for missing data and the BIB spiralling of the tests, an additional step beyond the maximum likelihood method involving the Bayesian estimation procedure has since 1992 been widely employed. In order to improve further the estimation, instead of relying solely on one estimated value, five plausible estimates have commonly been generated for subsequent analysis. These plausible values have been provided through the use of a so-called 'conditioning' procedure not only to replace the missing data, but also to replace all achievement test data, in order to improve both the effects of BIB spiralling, as well as to reduce the errors of measurement. It is argued in this article, that from the employment of the Bayesian estimation procedure that involves the formation of a prior distribution of estimated performance, the anomalous findings considered above may well arise.

A DISCUSSION OF PROCEDURES FOR ESTIMATING STUDENT PERFORMANCE

Adams (2005) and Wu (2005) have, with considerable clarity, in their published articles addressed these problems, in ways that were complementary and very informative. It was clear that because of BIB spiralling simple procedures that involved raw scores could no longer be employed to provide precise national estimates of the mean level of performance. However, maximum likelihood estimation procedures, involving either Rasch measurement or the three-parameter model could be used. Several other scoring procedures could also be used that were generally grouped within the two categories that Weiss and Yoes (1992) specified. These are listed below.

Maximum Likelihood Estimates (MLE)

One of three alternative procedures could be used:

(a) the Rasch model using the ConQuest, Quest, Bigsteps or RUMM programs that employed the one parameter measurement model;

(b) the three parameter model using BILOG or SAS/ETS enhance programs that employed the three parameter model; or

(c) the Weighted Maximum Likelihood Estimates (WMLE) obtained using the Warm Likelihood Estimation (WLE) procedure in which the maximum likelihood estimates for each individual were weighted by the information function for the set of items to which each individual had responded (Warm, 1989).

Bayesian Estimates

Two procedures could be used:

(d) the Plausible Values (PV) procedure that involved the use of five values which were sampled from a posterior distribution of the score for each individual; or

(e) the Expected A-Posterior Estimate (EAP) that involved calculating the mean of the posterior distribution for each student.

Three important statements were made about the different estimation procedures by Adams (2005) and Wu (2005) that were associated with the use of the Rasch model.

1. *Bias involved in estimates*

   All estimation procedures provided unbiased estimates of the mean score for the group.

2. *The maximum likelihood estimates*

   The maximum likelihood estimation procedure provided an unreduced estimate of the variance of the scores of the group. This was simply because no provision had been made to reduce the variance of the group scores that arose from errors of measurement.

3. *The Warm likelihood estimates*

   The Warm (1989) or weighted likelihood estimation procedure reduced the variance of the scores of the group by weighting each individual maximum likelihood estimated score distribution by the information function for each point estimate on the score distribution. The information function was defined by Fisher (1922) as the reciprocal of the precision with which a parameter was estimated. This information function was related to the square of the slope of the item characteristic curve at different points on the curve, and was standardised by dividing by the conditional variance:

   $$I(\theta,u) = \left(\frac{dp}{d\theta}\right)^2 \Big/ \text{conditional variance,}$$

   where $\frac{dp}{d\theta}$ = the slope at different points on a test characteristic curve,
   and $I(\theta,\ u)$ = the information function.

   Combining the likelihood distribution function (MLE) with the information function to form their product yielded the Warm likelihood function (WLF).

   The maximum value of the Warm likelihood function is referred to as the 'Warm likelihood estimate (WLE)' or the 'weighted likelihood estimate'.

Wu's simulation study

Wu (2005) has reported the results of a simulation study that provided information on the characteristics of the different estimates for both 3-item tests and 20-item tests where the generating distribution was N(0, 1) for the 3-item tests and N(2, 1) for the 20-item tests. These results are given in Table 1.

Wu (2005) drew the following conclusions from her simulation study.

1. MLE values might show some bias in mean values and greatly over estimated the variance of the generating distribution that was not adequately adjusted by a reliability correction. Thus variance associated with measurement error was clearly present in MLE values.
2. WLE values showed little bias in the mean values and overestimated the variance of the generating distribution. However, the Warm estimating procedure removed some but not all of the variance associated with measurement error. The reliability correction reduced the variance well below the expected value.
3. The plausible values (PV_1 to PV_5) were constructed to have an unbiased mean and an appropriate variance. It was argued that the measurement error had been removed by the conditioning process.

4. The estimated posterior (EAP) values that were formed as the mean of the plausible values were unbiased, but had as might be expected, substantially reduced variance. This variance was well adjusted by the reliability correction.

Table 1. Comparison of estimates for simulated 3-item test and 20-item test.[a]

	WLE	MLE	EAP	PV1	PV2	PV3	PV4	PV5	GV[c]
3-item Test Estimated Mean	-0.002	0.002	-0.002	-0.000	-0.004	-0.003	-0.002	-0.003	0
Standard Error	(0.030)	(0.039)	(0.036)	(0.041)	(0.042)	(0.042)	(0.041)	(0.041)	
Estimated Population Variance	1.950	2.350	0.359	0.995	1.004	1.002	1.004	1.001	1
Standard error	(0.263)	(0.178)	(.061)	(0.113)	(0.108)	(0.112)	(0.113)	(0.109)	
Corrected Value [b]			0.99						
	WLE	MLE	EAP	PV1	PV2	PV3	PV4	PV5	GV[c]
20-item Test Estimated Mean	1.966	2.117	2.002	2.002	2.002	2.000	2.003	2.003	2
Standard Error	(0.031)	(0.031)	(0.032)	(0.035)	(0.033)	(0.033)	(0.032)	(0.135)	
Estimated Population Variance	1.332	1.657	0.683	1.003	1.005	1.007	1.003	1.003	1
Standard Error	(0.051)	(0.056)	(0.047)	(0.059)	(0.061)	(0.061)	(0.063)	(0.059	
Corrected Value [b]	0.72	0.89	1.01						

[a] adapted from Wu (2005); [b] correction made for unreliability of estimates; [c] GV – Generated Values

Bayesian estimation, conditioning, and plausible values

The Bayesian estimation procedure that involved the construction of the prior distribution and its use in modifying the likelihood score distribution to form the posterior distribution has been referred to as a 'conditioning' procedure. Conditioning not only provided estimates for any missing scores, but it also refined the maximum likelihood estimates for all individuals. In addition these estimates were also replaced by five plausible values as well as an EAP estimate that was the mean of the five plausible values and the mean of the posterior distribution.

Adams (2005) and Wu (2005) presented evidence to support the case both for the use of plausible values and conditioning that would appear to have a high degree of credibility. Nevertheless, it is contended that through their cursory treatment of the construction of the prior distribution they failed to emphasise a potential shortcoming associated with the use of Bayesian estimates. Wu (2005, p. 125) recognised that a degree of bias might be associated with the estimates of population regression coefficients in the following words.

> The degree of bias of the regression coefficients will depend on test length and the partial correlation between the variable of interest and the latent variable, after controlling for any conditioning variables that were used. When a regression analysis is run using plausible values generated with a model that did not include the regressors, it is said that *model unspecification* has occurred. (Wu, 2005, p.125)

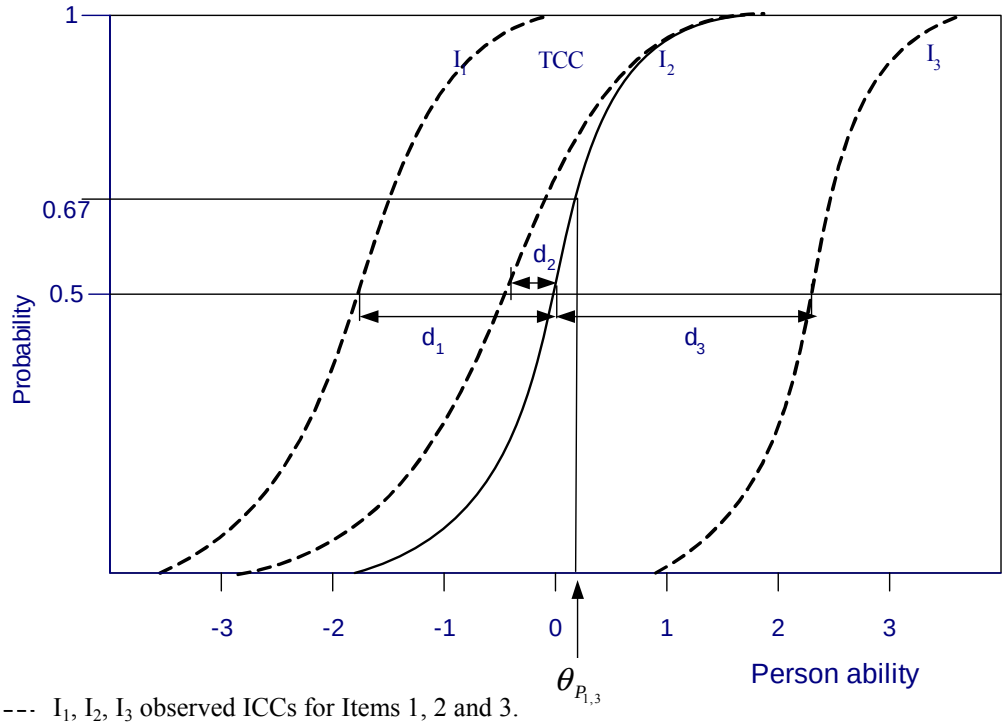
These qualifications are important but are clearly not enough and can be said to be both incomplete and inadequate. It is necessary to support the authors' assertions by a discussion of the

three anomalous cases that have been observed in secondary data analyses. It is also necessary to ask, with an understanding that has been developed from reading the papers by Adams (2005) and Wu (2005) and from experience in the primary and secondary analyses of cross-national data, whether suggestions can be advanced as to how the three observed anomalies might have arisen from the use of WLE, EAP or PV values. It is recognised that difficulties are encountered in the data analyses in such studies, and that the attempts made to provide for missing data and measurement error are necessary and desirable. Moreover, the authors apologise for failing to test fully their ideas by undertaking further analyses. However, before considering these anomalies, it is necessary to explain in greater detail the estimation procedures that are being employed in these studies.

A diagrammatic treatment of the estimation procedures

In the section that follows a diagrammatic explanation is presented of the estimation procedures without discussing these procedures using mathematical symbols. The figures are presented as illustrations of certain effects and are not derived from simulation or the use of particular measurements.

In Figure 1, three item characteristic curves are shown for Items 1, 2 and 3, the combined test characteristic curve for Items 1, 2 and 3 that were attempted by Person $P_{1,3}$, and the maximum likelihood estimate (MLE) for p the probability of a correct response for Person $P_{1,3}$, who responded correctly to Items 1 and 3, but not to Item 2.



------- I_1, I_2, I_3 observed ICCs for Items 1, 2 and 3.
——— TCC the observed test characteristic curve for Item 1, Item 2 and Item 3.
Where d_1, d_2 and d_3 are the item difficulty values and $d_1 + d_2 + d_3 = 0$.

Figure 1: Item characteristic curves for Items 1, 2, and 3 with I_1 and I_3 answered correctly

Since Person $P_{1,3}$ answered items I_1 and I_3 correctly with p = 0.67 a score of $\theta_{P_{1,3}}$ can be estimated. The zero for person ability is set at the average difficulty level of the three items when p = 0.5.

In Figure 2 it is shown how the response or likelihood distribution curve is trimmed to remove, in part, measurement error by weighting the likelihood function by the information function to provide a more precise estimate of the score of Person $P_{1,3}$, with reduced variance.

The likelihood distribution function is shown for Person $P_{1,3}$ who responded correctly to Item 1 and Item 3. The person's response or likelihood distribution function is combined with the information function to form the Warm likelihood function.

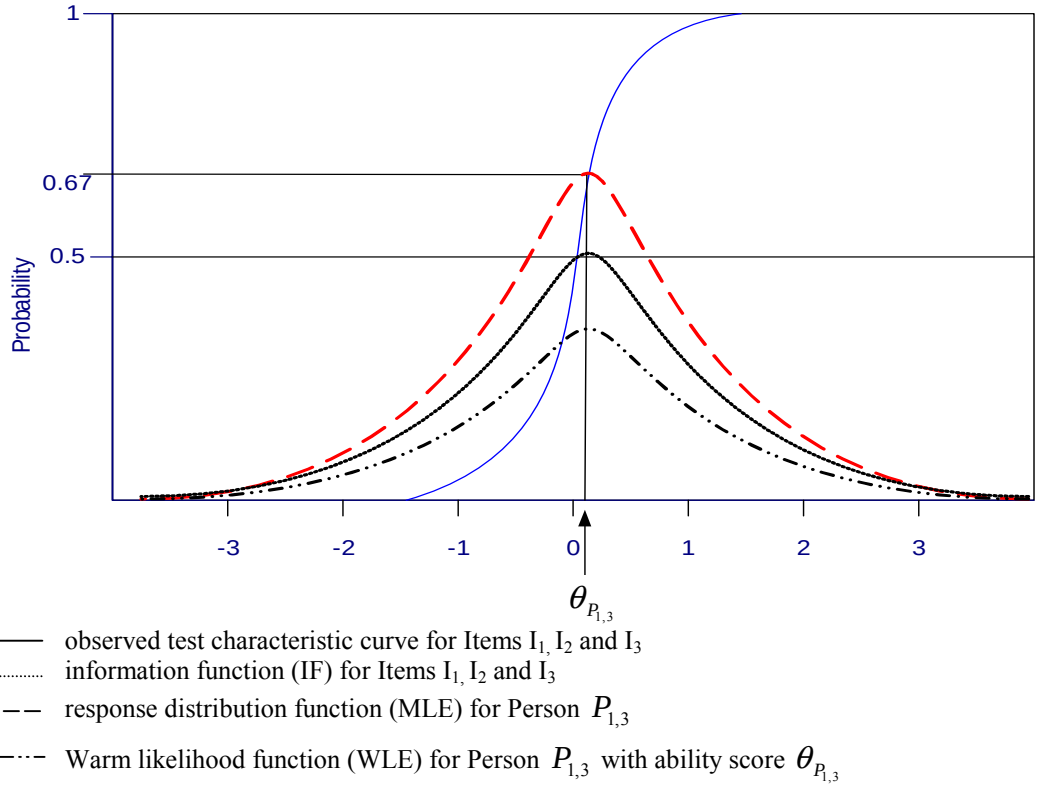


Figure 2: Response function curves for Items 1 and 3 answered correctly

It can be seen that the θ values for MLE and WLE remain close together showing little bias. However, the Warm likelihood function has less variance than the response distribution function for Person $P_{1,3}$, because it is formed by combining the information function with the response distribution function.

If missing data need to be imputed a prior distribution is clearly required, and consideration must be given, as to how best to produce this prior distribution.

A commonly used procedure is to employ a normal distribution N(0,1) as the prior distribution for each individual person and to impute the missing test score. However, it is also possible to construct a regression equation that best predicts the observed score for that individual using all known information about the group to which that individual belongs and to use this regression

equation as a prior distribution in a normalised or standardised[2] form in order to predict the score for the individual for whom a test score is missing. When some information is known about the individual the use of the prior distribution in this way to predict the missing score involves the use of Bayesian estimation procedures. If no information is known about the individual, scores are obtained at random using the posterior distribution for the group to which the individual belongs.

An extension of this principle is said to 'improve' or 'condition' the data by estimating the scores of all persons irrespective of whether or not their test scores are missing, and whether or not any other data are missing. In this estimation process the selection of a single best estimate proves inadequate, and the procedure currently adopted is to choose five estimates at random from the posterior distribution for the individual that is a combination of the likelihood or response distribution for the sub-group to which the individual belongs, and the normalised prior distribution, obtained by regression analysis procedures. If no specific information is known about the individual, the prior distribution represents the group and is based on the characteristics of the group to which the individual is said to belong, and scores can be estimated from the posterior distribution for the group. Clearly, at least five estimated scores are better than one. Moreover, because the posterior distribution is conditioned by the prior distribution to reduce measurement error and if all estimates are selected randomly from the posterior distribution, then the scores obtained follow the posterior distribution. Since the posterior distribution is a combination of two distributions, the scores that arise from the conditioning procedure form a distribution with reduced variance.

This procedure is presented in diagrammatic form in Figure 3.

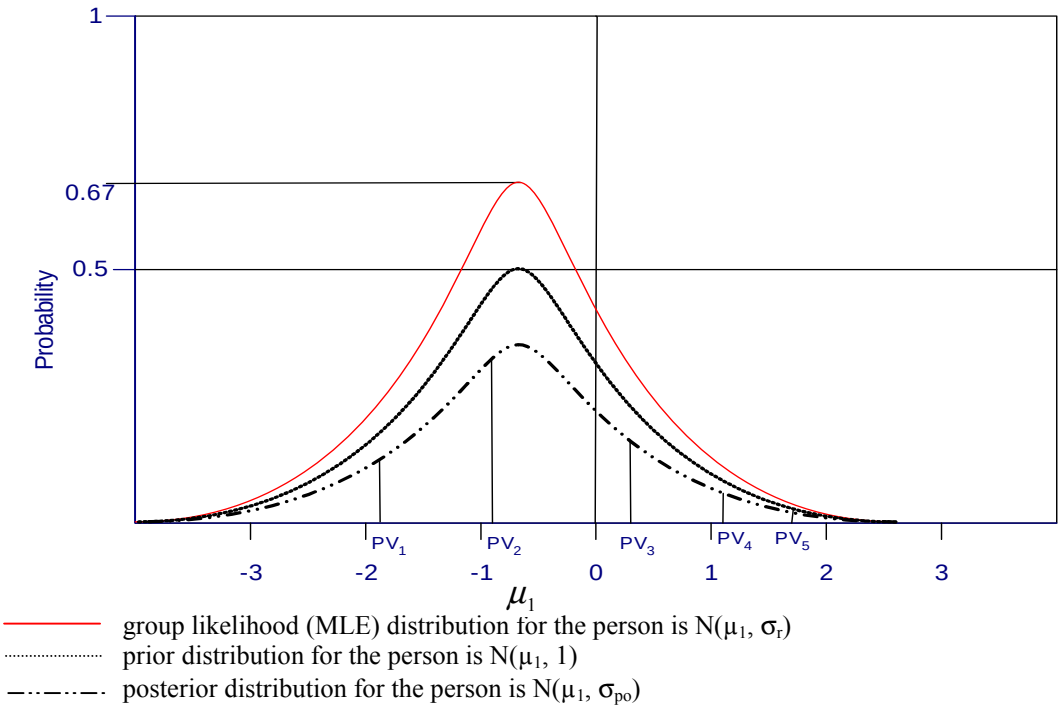


Figure 3: Bayesian Estimation in an ideal case for a specific person.

[2] A normalised or standardised distribution has a known mean of zero and a standard deviation of 1, and has a distribution that tends towards normality under the central limit theorem.

Consider the case of a person for whom a score is not available, but some information is known about that person and about the sub-group to which the individual person belongs.

The process of Bayesian estimation is shown for this specific person in which the sub-group likelihood response distribution is a $N(\mu_1,\sigma_r)$ distribution, and the prior distribution after regression analysis for that person is given by an approximately normal $N(\mu_1,1)$ score distribution. The likelihood response function is then weighted or multiplied at each level of θ by the corresponding value of the prior distribution to obtain a new likelihood function referred to as the posterior distribution. This posterior distribution for the individual persons is a $N(\mu_1,\sigma_{po})$ distribution. Five plausible values are then chosen at random from this posterior distribution for the missing data where some information about the individual person is available, and are shown as PV_1 to PV_5. Thus where information is known about the individual, that information is used to obtain the five plausible values. The expected posterior estimate (EAP) is the mean of the five plausible values that are obtained for each individual. Where no information is known about the individual, the prior distribution for the sub-group to which the individual belongs is used.

In Figure 4 the process is displayed of Bayesian estimation in the case of a low performing group of students who fit the regression model developed less than adequately and the group distribution exhibits positive skew. The prior distribution for the group is estimated from regression analyses and also exhibits positive skew. Consequently, the mean value of the posterior distribution for the group is likely to be seriously biased.

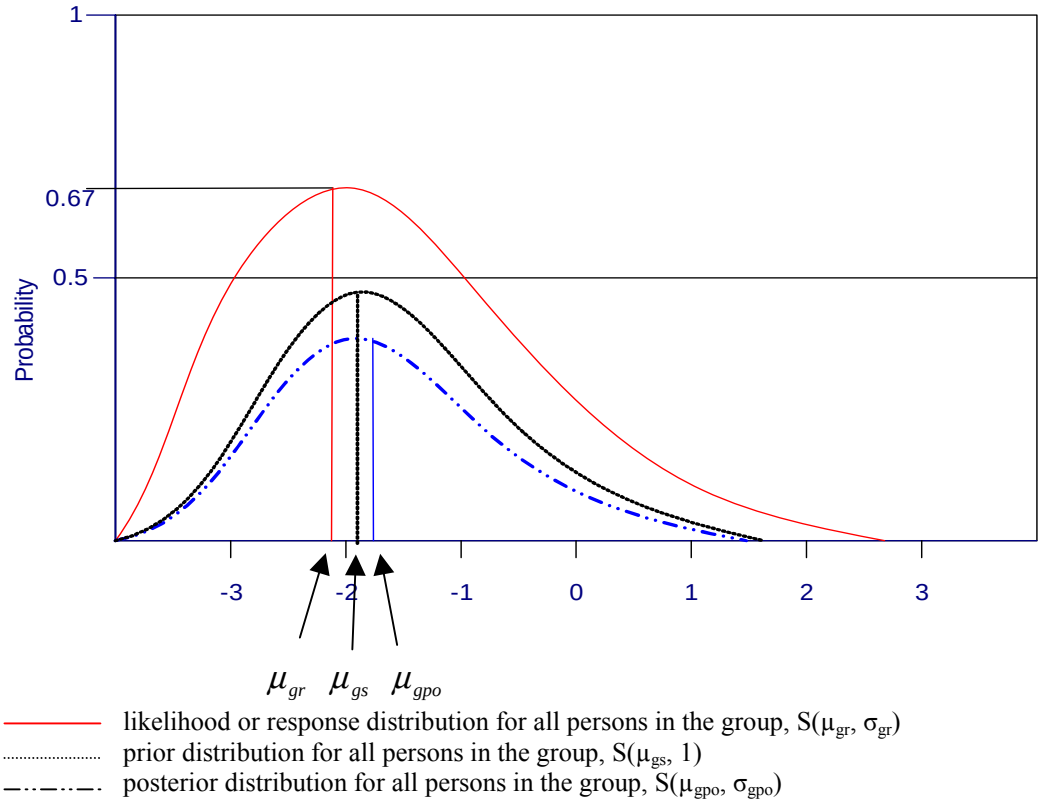


Figure 4: Bayesian estimation for a low performing group showing the response distribution for the group and the prior and posterior distributions

The major problem with the conditioning process and the use of the Bayesian estimation procedure is that if the prior distribution is not estimated well poor plausible values are obtained

for the individuals in the sample as well as for subgroups of individuals although the population estimates may be adequate.

OUR THOUGHTS ON THE ANOMALIES THAT WE HAVE OBSERVED

Meta analysis of gender differences and reading achievement

In the conduct of meta-analysis there was interest in the changes in important effects over time and in the differences that arose across different cultures and different education systems. It would seem likely that the gender effects that Lietz (this issue, pp. 127-150) compiled were derived before 1992 from raw score and Rasch scaled population estimates for male and female students. Since 1992, Bayesian procedures have been used in the estimation of both means and variances. These estimates would have reduced variance, as measurement error would have been removed. This reduction in variance would have given rise to larger effect sizes. Under these circumstances any attempt to undertake the meta-analysis of estimated effects might need to distinguish between, before the use of Bayesian procedures and after the use of Bayesian procedures. Other procedures such as raw scores and likelihood estimation procedures should yield similar mean values for male and female groups, where large groups were involved, but with much greater variances and consequently greatly reduced effect sizes.

Mathematics Proficiency of Students in South Africa

In the analysis of the data for the South African sample of students who were tested for proficiency in mathematics, there would be little doubt that any regression analyses of the data carried out would show that the characteristics of the South African sample were very different from other samples involved in the TIMSS studies. As a consequence the prior distributions used in the Bayesian estimation of the South African scores would differ markedly in variance from other national samples, probably casting serious doubts on the use of these procedures in the analyses of these data. What probably happened in the conditioning of the South African data is displayed in Figure 4 where the small group of higher performing English speaking students would be pulled back markedly towards the lower end of the scale in the Bayesian estimation process and conditioning operation.

Trends in Bulgarian Eighth Grade Mathematics Performance.

The possible explanation of the significant decline in mathematics achievement over the short time-span of four years probably lay in a shortcoming in the construction of the regression model that was used as the prior distribution in the Bayesian estimation of the posterior distribution from which the national mean value was estimated. It would seem possible that a variable that involved the changed structure of the strata employed for the sub-systems into which the schools were grouped was not examined in 1995 and 1999 in appropriate ways in the regression analysis to form the prior distribution for the Bayesian estimation procedure. However, it would also be possible that the different test booklets that were employed on the two occasions differed in important ways with respect to the mathematics curricula of Bulgarian schools at the eighth and lower grades. A consequence of this would be that the original score distributions were influenced differently on the two occasions by the lack of match between the curriculum and the different test booklets that were used to sample and estimate student achievement in mathematics.

Further comments

In Bayesian estimation the likelihood or response distribution is modified by the prior distribution to yield the posterior distribution of scores to differing extents for different countries, different sub-systems and different individuals. If the prior distribution reflected adequately the original

likelihood or response distribution it would seem that little distortion would be likely to occur. However, if the prior distribution did not reflect adequately the likelihood or response score distribution, as a consequence of important factors not being included in the prior distributional model, then the effects of those factors would most likely tend to disappear from the scores that were made available for secondary analysis.

Wu (2005), as stated above, warned against model mis-specification when regression analyses were undertaken to form the prior distribution. This suggested that any subsequent construction of explanatory models in secondary data analysis would largely be a waste of time and effort because of specification problems in the prior distribution and the magnitude of such effects would remain unknown.

Warm likelihood estimates (WLE) have also been provided in the data files to enable secondary data analyses to be undertaken with scores that were not modified by the effects of the prior regression-based distribution. However, little appears to be known about the effects of trimming the variance of the scores by the procedure proposed by Warm (1989). The use of this procedure must be expected to have consequences for the estimation of the effect sizes.

Further possibilities associated with the Bulgarian analysis could have arisen in two different ways. The regression analyses that were carried out in the forming of the prior distribution were most likely undertaken only at the student level through the use of the general linear model. If a two level model were used it might be possible to provide for effects at the student and school levels. However, many national school systems had strikingly large differences between regions or provinces and states, as well as school types, and the use of at least a three level model would seem to be required. In the formation of the posterior distribution it should be recognised that 'what you get out is strongly related to what you put in'. Unfortunately, little information has been made available on the nature of the variables employed in constructing the prior distribution in different countries and for different groups of students or on the amounts of variance involved at the different levels of the data. Furthermore, there has been little information provided on the nature and extent of differences between the different national education systems with respect to the strongest factors that were associated with the development and construction of the prior distribution within each system that had such a pivotal role in the conditioning process.

The BIB spiralling procedures that are built on the use of eight different test booklets and that serve to increase the range of content which can be assessed, employ items that are frequently clustered under a common stem. Thus the range of content assessed by each booklet is very limited. Our secondary analyses have shown that the different booklets operated very differently across countries in sampling student performance probably because of differences across countries in the structure of the curricula under survey. The relationships between the content of the items in the test booklets and the opportunity that the students in different countries had to learn that content and the performance of students in different countries has been a controversial issue over the past 40 years during which cross-national assessment programs have been operating. Unfortunately, little progress would appear to have been made over the years in the examination of curriculum design and time allocated to learning the content assessed by the test items and their effects on learning outcomes. These aspects are possibly involved in the anomalous effects recorded over time in the Bulgarian analyses of the TIMSS data.

CHALLENGING THE USE OF THE BAYESIAN PROCEDURE

Michell (1986, 2000) has raised questions about the nature of measurement in the behavioural and social sciences identifying the three theories of representational, operational and classical measurement. It would seem that, in practice, elements of all three theories are generally involved.

Moreover, measurement provides among its other functions, a structure for the use of mathematical symbols, ideas and relationships. Not only must the measures bear a representational relationship to qualities, but the measures must also express the relationships between such qualities that involve operational purposes.

Nevertheless, in the social and behavioural sciences abstract qualities are involved and measures of these abstract qualities are sought. Furthermore, information on how much of an abstract quality is involved, namely its measure is required. In addition, information about an abstract quality can only be obtained through the interaction of people with tasks that are associated with the quality under consideration. Consequently, in order to estimate the ability of a person all that can be observed is performance on a task. The difficulty of the task must both be sampled on multiple occasions and in multiple situations that involve probabilistic or stochastic relationships. Thus, several different sources of error must be taken into consideration. The discussion in this article is primarily concerned with those sources of error that are associated with performance errors that arise from:

(a) variability in the performance of the person involved,
(b) variability in the tasks being performed, and
(c) variability in the observation of performance on the tasks.

In general, tasks, observations and persons or cases are sampled, and thus sampling errors are also involved. The procedures adopted by Adams and Wu in their work seem to be directed towards certain operational aspects of measurement to the exclusion of other representational aspects, on the assumption that a so-called ‘true’ value is capable of being estimated. Such a ‘true’ value is unknowable in the social and behavioural sciences.

It is argued in this article that the work of Adams and Wu fails to satisfy the requirements of both representation and operation as they move beyond classical approaches. The use of plausible values is not appropriate for estimating the scores of individual students and certain subgroups of students. The plausible values and the EAP values are better suited for estimating the performance of a population. Consequently, it is also argued that other ways must be sought to allow for uncertainty and the use Bayesian estimation procedures should be rejected. Other error estimation procedures are available. For example, bootstrapping or jackknifing of items and persons with respect to their primary sampling units can be used to provide estimates of measurement error in the same way as bootstrapping and jackknifing are used to provide estimates of sampling error. This, however, seems to require a major rethinking of the strategy of data analysis that has evolved around the use of Bayesian estimation methods. These estimation procedures although apparently effective for the better estimation of population parameters, are made at the expense of individual and sub-group estimates, which are essential for the examination of multivariate and multilevel models. While information on trends in population mean values over time is of importance in the monitoring of educational outcomes, the development of a clearer and more coherent understanding of educational processes and how these change over time was not only the goal set by the founders of IEA, but remains today the most challenging task for those who believe in the importance of increasing the effectiveness of education and its contribution to human development.

REFERENCES

Adams, R. J. (2005) Reliability as a measurement design effect, *Studies in Educational Evaluation,* 31, (2/3), 162-172.

Affrassa, T.M. and Keeves, J.P. (1999) Changes in students' mathematics achievement in Australian lower secondary schools over time, *International Education Journal,* 1(1), 1-21.

Chiu, M.M. and Khoo, L. (2005) Effects of resources, distribution inequality, and privileged bias on achievement: Country, school, and student level analyses. *American Educational Research Journal*. 42(4), 575-604.

Fisher, R. A (1922) On the mathematical foundations of theoretical statistics, *Philosophical Transactions of the Royal Society of London* A, 222, 309-368.

Gregory, K.D. and Bankov, K. (2005) Exploring the Change in Bulgarian Eighth Grade Mathematics Performance from TIMSS 1995 to TIMSS 1999. In T. Plomp and S. Howie (Eds.), *Contexts of Learning Mathematics and Science Lessons Learned from TIMSS*. London: Routledge.

Howie, S. (2002) *English Language Proficiency and Contextual Factors Influencing Mathematics Achievement of Secondary School Pupils in South Africa.* The Hague: CIP-Gegerones Kononklyke Bibliotheck.

Michell, J. (1986) Measurement scales and statistics: A clash of paradigms. *Psychological Bulletin* 100(3), 398-407.

Michell, J. (2000) Normal science, pathological science and psychometrics. *Theory and Psychology*. 10(5), 639-667.

Raudenbush, S.W. and Bryk, A.S. (2002) *Hierarchical Linear Models: Applications and Data Analysis Methods* (2nd edition). Thousand Oaks, CA: Sage Publications.

Warm, T.A. (1989) Weighted likelihood estimation of ability in item response theory. *Psychometrika* 54(3), 427-450.

Weiss, D.J. and Yoes, M.E. (1992) Item Response Theory. In R.K. Hambleton and J.N. Zaal (1992) *Advances in Educational and Psychological Testing*: *Theory and Applications*. (pp. 69-95), Dordrecht, The Netherlands: Kluwer.

Wu, M. (2005) The role of plausible values in large-scale surveys. *Studies in Educational Evaluation.* 31, (2/3), 114-128.