Practical Assessment, Research, and Evaluation

Volume 15 *Volume 15, 2010* Article 15

2010

Yule-Simpson's Paradox in Research

Heather Honoré Goltz

Matthew Lee Smith

Recommended Citation
Goltz, Heather Honoré and Smith, Matthew Lee (2010) "Yule-Simpson's Paradox in Research," *Practical Assessment, Research, and Evaluation*: Vol. 15 , Article 15.
DOI: https://doi.org/10.7275/dgcc-jv81
Available at: https://scholarworks.umass.edu/pare/vol15/iss1/15

Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 15, Number 15, October, 2010 ISSN 1531-7714

Yule-Simpson's Paradox in Research

Heather Honoré Goltz, *Houston VA HSR&D Center of Excellence and Baylor College of Medicine*
Matthew Lee Smith, *School of Rural Public Health, Texas A&M Health Science Center*

Yule (1903) and Simpson (1951) described a statistical paradox that occurs when data is aggregated. In such situations, aggregated data may reveal a trend that directly contrasts those of sub-groups trends. In fact, the aggregate data trends may even be opposite in direction of sub-group trends. To reveal Yule-Simpson's paradox (YSP)-type occurrences, researchers must simultaneously consider the effect of an intervention at specific levels and on the overall model to ensure datasets are accurately analyzed and research findings are appropriately interpreted. The primary objectives of this manuscript are to: (1) examine the history of YSP; (2) describe necessary and sufficient causes for YSP occurrences; (3) provide examples of YSP in research and explain YSP's relationship to multi-level modeling including Hierarchical Linear Modeling (HLM); and (4) discuss YSP's implications for researchers.

Yule-Simpson's paradox (YSP) is a statistical phenomenon that may occur when data is aggregated (Malinas, 2001; Simpson, 1951; Thompson, 2006; wa-Kivilu, 2003; Yule, 1903) or statistical test assumptions are violated (i.e., particularly assumptions of independence). When YSP occurs, aggregated data may reveal a trend that directly contrasts with that of its sub-groups. In fact, the aggregate data may be opposite in direction of these sub-groups (Thompson, 2006; i.e., trends or correlations existing within independent groups become inversed when groups are combined, which yields counter-intuitive statistical findings). A frequent cause of this statistical occurrence is when group data of unequal sizes are merged. This aggregation may create unequally weighted results, which produce misleading relationships and cause situations where findings may be interpreted inappropriately. This paradox is not limited to aggregate data from multiple groups; it may also occur when data is aggregated from multiple locations or (time) waves.

YSP is a relatively easy phenomenon to understand; however, Malinas (2001) describes it as a "logically benign, empirically treacherous hydra" (p. 265). Although YSP occurrences are often recognizable, without careful planning for data analyses, YSP-type effects may go unnoticed and lead to misinterpreted or erroneous study findings. The primary objectives of this manuscript are to: (1) examine the history of YSP; (2) describe necessary and sufficient causes for YSP occurrences; (3) provide examples of YSP in research and explain YSP's relationship to multi-level modeling including Hierarchical Linear Modeling (HLM); and (4) discuss YSP's implications for researchers.

HISTORY AND INTRODUCTION OF THE PARADOX

George Udny Yule (1871-1951) was a student and colleague of the statistician Karl Pearson (of *Pearson's r correlation*) and an eminent statistician in his own right (Kendall, 1952, p. 156). His contributions to the field included fourteen editions of *Introduction to the Theory of Statistics* and numerous texts on correlation and regression (Kendall, pp. 156-158). Yule was credited with inventing the *correlogram*, a method for plotting correlation coefficients, and developing the foundation of the *theory of autoregressive series* (Kendall, p.157).

In 1903, Yule wrote a paper titled "Notes on the Theory of Association of Attributes in Statistics." He began the paper by explaining that the most basic method for *statistical classification* is "division by dichotomy" (Yule, p. 121). This process involves

placing units of observation (e.g., patients, animals) into one of two *mutually exclusive* groups based on the presence or absence of an *attribute* (Yule, p. 121). These *attributes* are said to be *independent* for any given *universe* or *sub-universe* when the "chance of finding them together is the product of the chances of finding either of them separately" (p. 125).

When two (independent) attributes (e.g., the association of A and B) or their *contraries* are considered jointly, they may be *positively or negatively associated* (Yule, 1903, p. 126). The direction of this association is determined by whether the related attributes are *greater or less than the value* placed on the attribute when it is independent (Yule, p. 126). When three *attributes* (i.e., A, B, C) are considered, the relationship becomes more complex. In this situation, the pair-wise relationship between A and B or their *contraries* must be considered in terms of the third variable, C, or its *contrary* (i.e., A and B given that C is also present; Yule, 1903).

Despite the temptation of researchers to assume these attributes are independent, Yule wrote that one must actually determine whether these "attributes are independent, wholly or in part" (p. 132). These *partial associations* (interactions) are essential for testing the accuracy of the *total associations observed* (p. 132). Just because a pair of attributes is independent within a *sub-universe*, this does not permit statisticians to infer this relationship remains true for the *universe* or vice versa (p. 132). The association of A and B equal to zero does not always mean the association of A and B, given C, is also equal to zero. Further, one should not infer the inverse. Yule additionally explained that if A and B have a positive or negative value, the same might not hold true for A and B given C is present.

Yule (1903, pp. 133-134) illustrated these points using an example from population genetics. This example considers the inheritance of an attribute among same-sex parent/child pairs. Table 1 represents Yule's example in tabular form.

Table 1. Yule's Example of the Paradoxical Phenomenon

	Fathers (n=100)		Mothers (n=100)		Parents (n=100 couples)	
	With attribute	Without Attribute	With attribute	Without Attribute	With attribute	Without Attribute
Sons w/ attribute	25%	25%	-	-	-	-
Sons w/o attribute	25%	25%	-	-	-	-
Daughters w/ attribute	-	-	1%	9%	-	-
Daughters w/o attribute	-	-	9%	81%	-	-
Offspring w/ attribute	-	-	-	-	13%	17%
Offspring w/o attribute	-	-	-	-	17%	53%

Affected sons with an affected father (n=25) accounted for 25% of cases (p. 133). In contrast, affected offspring whose parents both had the attribute (i.e., 13 out of 30) accounted for 43.3% of cases (p. 133). This pooled data might cause researchers to incorrectly assume this attribute is more dominant or easily passed between generations. However, in reality the attribute is not particularly heritable through the mother or father's genetic line. Yule described this phenomenon as "quite a large but illusory inheritance created simply by the mixture of the two distinct records" (i.e., maternal and paternal inheritance; p. 133). Exploring this attribute inheritance by gender might help to explain the origins of this erroneous conclusion and others in similar situations.

Yule explained "there will be an apparent association between A and B in the universe…unless either A or B is independent of C" (p. 134). However, variables A, B, and C fail to meet this requirement in Yule's example. When the data are pooled, a positive association exists between both parents with the attribute (A) and offspring with the attribute (B), and male gender (C). A larger proportion of males have the attribute when compared to the proportion of females who have the attribute. The unequal proportion of

males with the attribute contributes a greater weighted average when the data are pooled. The pooled data then exhibit an association that runs counter to that found in at least one of the sub-groups. Yule referred to this phenomenon as the "fallacy of mixing distinct records" (p. 132). Although more commonly referred to as Simpson's paradox, we will refer to it as the Yule-Simpson's Paradox (YSP) throughout this text in deference to the contributions that Yule and Simpson made to our understanding of this paradox.

CONCEPTUAL MATURATION OF THE PARADOX

In 1951, E.H. Simpson published a paper titled "The Interpretation of Interaction in Contingency Tables." Simpson began this paper by considering a 2x2x2 (attributes A, B, and C) contingency table (p. 238). As in Yule's paper, Simpson wrote that this relationship contains partial associations (i.e., *first order interaction,* 2x2), as well as an interaction between all three variables (i.e., second order; p. 238). He demarcated the boundary between first and second order interactions by stating if A and B are associated (first order interaction), a *second order interaction* between AB and C will not exist if the *degree of association* for AB given C is the same as AB given C's contrary (p. 239). Simpson wrote that when there is no apparent "second order interaction, there is considerable scope for paradox and error…if A and B are associated positively in C and negatively in… (its contrary, γ) they may appear independent in the whole population" (p. 240).

Simpson (1951) illustrated this concept using a heuristic example of clinic patients (p. 241). In the example, patients received treatment or no treatment and were monitored for survival over time. Table 2 is a reproduction of the fourth table in the 1951 Simpson paper (p. 240).

Table 2. Simpson's Example of the Paradoxical Phenomenon (n=52)

	Male (n=20)		Female (n=32)	
	Untreated	Treated	Untreated	Treated
Alive	7.69%	15.38%	3.85%	23.08%
Dead	5.77%	9.62%	5.77%	28.85%

When the data were examined by gender, one might believe that both males and females responded favorably to treatment and survived (i.e., 8/52 and 12/52, respectively), compared to those who did not receive treatment (i.e., 4/52 and 2/52, respectively; Simpson, 1951, p. 241). However, when the data were aggregated, these positive associations vanished and there appeared to be no association between treatment and survival (p. 241).

YSP: NECESSARY AND SUFFICIENT CAUSES

Based on this historical overview of YSP, one might inquire "what is actually happening when this paradox occurs in a research study?" YSP can occur in the presence or absence of second or higher order interactions. In the former case, YSP occurs when aggregating data from multiple 2x2 contingency tables instead of using 2x2x2, or stratified tables. In the latter case, YSP occurs when there is no second interaction. However, in this case, B and C are correlated and either A or B (or their contraries) fails to be independent of C (i.e., the independence assumption is violated).

When a second order interaction occurs and/or the independence assumption is violated, the "moderator is a qualitative (e.g., sex, race, class) or quantitative (e.g., level of reward) variable that affects the direction and/or strength of the relation" between the independent (IV) and dependent (DV) variables (Baron & Kenny, 1986, p. 1174). The effect of the IV on the DV changes based upon the level of the moderator (i.e., C or its contrary). While a *linear* type moderator-interaction is "generally assumed," it is not the only type (p. 1175). Interactions may also be *quadratic* (curvilinear) or *step functions* (pp. 1175-1176).

Regardless of the type of moderator-interaction, when a moderator is considered, the relationship between the IV and DV may be altered in such a way that it changes direction. This change in the direction or strength of the IV and DV relationship has important implications for data analysis and interpretation. In such instances, potential moderator variables may include interactions between individual- and ecological-level factors (Kraemer et al., 2006, p. 605; Lièvre et al., 2002, p. 3) or unequal sample sizes unaccounted for within aggregate data (Cates, 2002, p. 2; Kunisaki, 2005, p. 1674). When researchers use data containing unaccounted-for moderators, calculated results often yield incorrect proportions, odds ratios, and relative risk

values (Cates, p.2; Sistrom & Garvan, 2004, p. 12). Interpretations based on these values may lead to erroneous conclusions concerning statistical, clinical, and practical significance. Further, these errors may compromise study validity and limit researchers' ability to compare results across studies (Lièvre et al., p. 2; Sistrom & Garvan, p. 18).

CONTEMPORARY EXAMPLES OF SIMPSON'S PARADOX

Examples of Yule-Simpson's paradox (YSP) occur in studies across vastly different fields. The paradox may occur in any instance where researchers aggregate data without accounting for potential moderators or independence of observations. According to Cohen (1986, p. 33), "any comparisons of probabilities, rates, or measurements that are weighted averages of component probabilities, rates, or measurements from subgroups" may be affected. Examples of YSP have been published in many fields including literature related to education, business and economics (Cohen, p. 34), cognitive psychology (Howe, Rabinowitz, & Grant, 1993), sports (Wardrop, 1995), and medical school admissions (Wainer & Brown, 2004). The following are examples from the fields of education, population sciences, and public safety:

Educational Testing and Measurement: SAT

Between 1981 and 2002, the *national average* for the verbal Scholastic Aptitude Test (SAT) score appeared to remain relatively stable at 504 points (Bracey, 2004, p. 32). However, during that same time period, the average verbal scores for all racial and ethnic subgroups increased by between eight and twenty-seven points. Bracey attributed this example of YSP to the changing demographics of SAT test-takers. Over this time period, the number of white students taking the SAT fell while the number of minority students rose. Performance improved across all racial and ethnic groups, but minority (excluding Asian Americans) students' average verbal scores remained below the national average. Higher numbers of increasing but below average scores resulted in a national average that not only failed to reflect subgroups' improved verbal scores; they failed to reflect any change at all (pp. 32-33).

Higher Education Admissions

University of California, Berkeley graduate school admissions data from the 1970's is one of the more well-known examples of YSP (Spellman, Price, & Logan, 2001). The university wanted to insure that female applicants were being treated fairly. Despite a concentrated effort to reduce potential for discrimination in graduate school admissions, university-wide data indicated that a higher number of females than males were being denied admission. One might have concluded that Berkeley was highly discriminatory against female graduate school applicants. Examining data by university department directly contradicted this conclusion. Further examination revealed females were applying in much greater numbers to departments with fewer available slots. When the data was aggregated, the entire university artificially appeared to have higher rejection rates for female applicants.

Demography

Demography is the study of human populations using characteristics such as birth and death rates. Cohen (1986) described a real-life example of Simpson's paradox involving a comparison of death rates in Costa Rica and Sweden. Historically, the Swedish have a reputation for being among the longest living people on the planet. In his example, the Costa Rican age-specific female death rates in the year 1960 were higher than the corresponding Swedish rates for the previous five-year period (p. 33). By contrast, its female crude death rate was substantially lower than the Swedish rate. This trend was also observed in Costa Rican males compared with Swedish males.

Intuitive assessment of this example would lead many to question, "How could a country with higher age-specific death rates have a lower crude death rate?" The answer is YSP. In 1960, Costa Rica had a much younger population than Sweden. There were more young Costa Ricans inhabiting age groups where age-specific death rates would be lower than that of the Swedish. At the same time, more Swedish were of the age where age-specific death rates would be higher. Comparison of each country's aggregated data produced the paradoxical conclusion that Costa Ricans lived longer (or died off less often) than the Swedish.

Public Safety

Engineers and risk analysts often use statistical models to aid in decision-making related to road safety. Davis (2004, pp. 1124) demonstrated YSP in research on the relationship between changes in neighborhood speed limits and occurrence of vehicle-pedestrian accidents. He accomplished this by simulating situations in which pedestrians ran into the street and stopped in

the path of oncoming traffic. Davis calculated the probability that pedestrians might be hit (*collision probability*) while considering the average traffic speed and volume at various residential sites (p. 1124). The model was then used to predict how changes in speed limit might affect pedestrian safety. Study results indicated reducing speed limits from 30mph to 25mph would increase frequency of collisions.

In this example, YSP occurs as a result of aggregating multi-site data and failure to consider *site* as a moderating variable (pp. 1124-1125). Hypothetical residential site number one had a much lower frequency of vehicle-pedestrian accidents than site number two, which were attributed to differences in traffic speed and volume. Each site exhibited a decline in frequency of collisions when speed limits were lowered in the statistical model, yet the aggregated multi-site data did not reflect this trend. In this case, results indicating that reducing speed limits from 30mph to 25mph would increase frequency of collisions were misleading and inaccurate. Although the study did not involve actual data from vehicle-pedestrian accidents, it did provide a solid example for the reader and researchers in the field of road safety research.

VISUAL REPRESENTATION OF YSP: A HEURISTIC DATASET

Issues of school violence and safety have been widely studied in the Post-Columbine era. Bullying, drug use, and firearms constitute major threats to student well-being and ability to succeed academically. For illustrative purposes, the authors adapted charts from Kocik (2001) and Paik (1985), and created a plausible YSP narrative pertaining to scores on a standardized achievement test and ratings of perceived school safety (i.e., responses ranging from 0% = "not at all safe" to 100% "completely safe") representing students from four high schools. Figure 1 contains a plot of the correlations (i.e., correlogram) between student academic achievement scores and their perceived safety.

When results were viewed for students at the four schools independently, each had a positive correlation between achievement and safety perceptions (i.e., higher achievement scores are related to higher perceived safety). Yet, when looking at the overall correlation between these scores in aggregate (i.e., for all students at all schools), a negative correlation is seen.

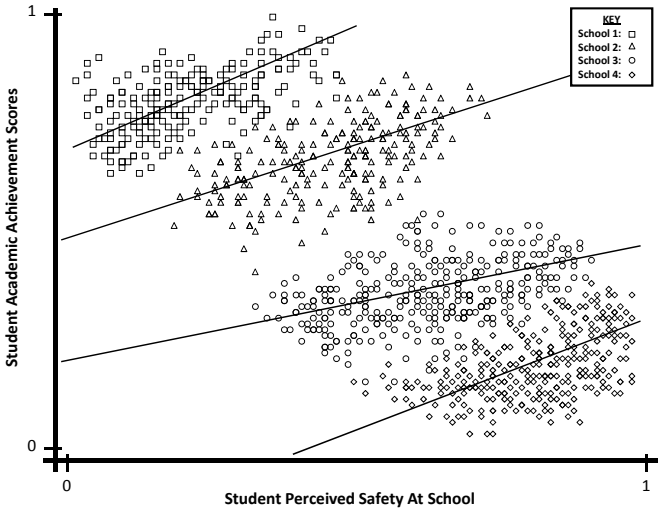

Figure 1. Scatterplot of Student Academic Achievement & Perceived Safety at School

Figure 2 uses the same heuristic dataset, but illustrates YSP originating from uneven subgroup sizes based on the distribution of student scores by school type (i.e., private schools are compared to public schools).

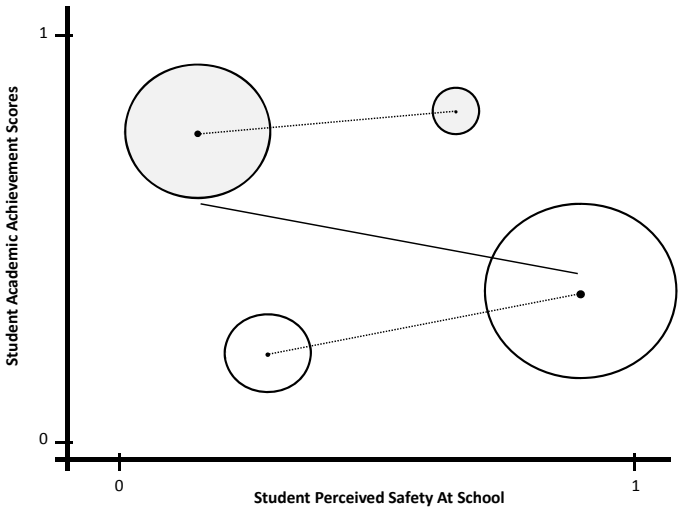

Figure 2. Student Academic Achievement & Perceived Safety at School: Comparison by School Type

Private schools are represented by the shaded circles and contain scores of students enrolled in Schools 1 and 2. Public schools are represented by non-shaded circles and contain scores of students enrolled in Schools 3 and 4. Circle sizes are indicative to the number of students within each subgroup (i.e., the larger the circle, the more students represented), and the black dot in the center of each circle represents the mean score for that subgroup. The broken line connecting circle centroids are show the correlation between scores based on school type. The solid line shows the overall correlation for all

students. As seen when the data is examined, a larger overall number of private school students scored higher on the achievement test and a larger number of public students reported higher levels of perceived safety ratings. The relationships between private school subgroups and between public school subgroups are positive; however, when scores are considered from both school types the overall relationship (solid black line) between achievement and perceived safety scores between subgroups is negative.

MULTI-LEVEL MODELING

Researchers may assume that what is "true in general is true for all individuals in a population" (Grimm & Yarnold, 2000, p. 342). This is not always the case. Most statistical methods have assumptions associated with characteristics of variables used in data analyses (e.g., shape, dispersion, level of measurement). Failure to consider the inherent assumptions in these methods may cause researchers to select inappropriate statistical tests, particularly with datasets containing higher order interactions (See Table 3). Violations of statistical test assumptions often occur when researchers attempt to use *conventional statistical* methods to explore hierarchical data (Roberts, 2004, p. 31; wa-Kivilu, pp. 249-250).

Hierarchical or nested data structure is a by-product of the human condition. An individual's interactions with others and his or her environment occur within hierarchical structures (i.e., successively larger clusters such as families, schools, and communities; wa-Kivilu, 2003, p. 249). Each of these contexts has an effect on the individual. According to Roberts (2004), "neglecting the fact that individuals…may be nested inside other larger clusters will often lead researchers to erroneous conclusions about their data" (p. 30). Thus, outcomes are examined across all levels of nested data. Violating test assumptions by using them with hierarchical data can cause invalid results "resulting in a Type I error or Type II error, or over- or under-estimation of significance or effect size(s)" (wa-Kivilu, 2003, p. 250).

Hierarchical data is particularly problematic for researchers. wa-Kivilu (2003) states that individuals within hierarchies are more *homogenous* "than people randomly sampled from the entire population" (p. 250). *Homogeneity* may increase over time with increased exposure to other individuals within the hierarchy (wa-Kivilu, p. 250). These individuals will have a higher degree of *intra-class correlation* (i.e., *ICC* or "amount of variance explained by the grouping structure") than individuals from another group (Roberts, 2004, p. 32; wa-Kivilu, p. 250). When ICC exists within a dataset, "the assumption of independent observations has been violated" (Roberts, p. 32; wa-Kivilu, p. 250). Conventional statistical tests that rely on this assumption cannot be used for most hierarchical data analyses (Roberts, p. 32; wa-Kivilu, p. 250).

Hierarchical Linear Modeling

After considering the historical and theoretical origins of YSP, one might inquire "How can knowledge of multi-level modeling techniques help researchers to resolve YSP?" Multi-level analyses allow researchers to track how individual members of a group change over time, as well as how this change relates to other variables (Grimm & Yarnold, 2000, p. 343). They are *robust* to violations of "assumptions of independence, linearity, reliability of measurement and normality" (wa-Kivilu, 2003, p. 250). These assumptions combined with the implied hierarchical levels of data are vital to efficient and effective statistical analyses (Grimm & Yarnold, p. 343; Roberts, pp. 30-31; wa-Kivilu, p. 250).

Researchers and statisticians have created an array of programs appropriate for analyzing hierarchical datasets over the past two decades (O'Connell & McCoach, 2004; Roberts, 2004. Multi-level modeling may be performed using general statistical analysis programs such as SAS and SPSS, or specialized programs such as Hierarchical Linear Modeling (HLM), MLwiN, and Mplus. For the purposes of this paper, we will focus on HLM as we have had basic exposure to this software package.

The HLM program allows researchers to build an initial hierarchical model (the null model) and to test it against hypothesized regression models (Roberts, 2004). The null model must contain at least two levels (e.g., students within schools); level-1 refers to the lowest unit within the model. In addition to information on nesting, levels provide information concerning fixed and random coefficients, variance, and covariates within the regression model. A basic two-level HLM model contains at least one dependent/outcome variable derived from individual scores or measures (level-1) and independent/predictor (parameter) variables, those related to membership within a group, in Level-2. Error terms are also added to the model to address variance due to nesting within the dataset.

The HLM 2-level model building process generally involves creating a null or benchmark model (Roberts,

2004). This null model contains fixed estimates for the intercept and random error variance estimates for the individual (level-1) and group (level-2), but no level-2 predictor variables. HLM is similar to an ANOVA in that the null model tests groups' deviations from the *overall grand mean* of the dependent variable (i.e., between group differences in the dependent variable) and uses the initial parameter estimates as a "yardstick" for successive models. However, HLM parameter estimates will differ somewhat from ANOVA because they are created using Empirical Bayes estimation, resulting in shrinkage of outliers towards the mean. The next step in model building involves creating a random intercept or random coefficient model (Roberts, 2004). The model is run by adding level-2 predictor variables one at a time. Each new model also contains estimates for new slopes, intercepts, and related variances; outputs are interpreted to determine the fixed or random effect for each new parameter and the unique variance it contributes to the model (i.e., variation in individual scores based on a facet of group membership). A Chi-Square statistic is generated by testing the new model against the null and determining whether adding the new parameter generates significant variance in the slopes and intercepts among groups. Model fit statistics consider "parsimony" within the model and can be used in conjunction with the Chi-Square statistic to determine whether the "overall model fit" increases or decreases with the addition of a specific parameter. If the parameter fails to explain a significant amount of variance within the dependent variable, it may be removed from the model. The process continues until the full model is constructed and deemed to have a "good fit."

On the surface, HLM is similar to simple linear regression (one-level) in that it uses regression equations to model parameters. However, HLM differs from simple regression models in that it: 1) incorporates an error term for each model level rather than one (assumed random) error term for the entire regression equation and 2) involves a step-wise model building process rather than simultaneously considering all parameters. Thus, HLM helps explain variance in individual scores due to group membership, offers opportunities to test within and between-level interactions and main effects, and honors data structure and statistical test assumptions. It has additional utility in that the program can also be used to perform other General Linear Model-based analyses including ANCOVA, regression, and bootstrapping (Roberts, 2004).

The 2003 National Assessment of Educational Progress (NAEP) test provides an illustrative example of HLM application relevant to YSP (Braun, Jenkins, & Grigg, 2006). In this pilot study, 4th grade charter and non-charter public school students were compared in terms of NAEP performance using HLM modeling. Researchers determined that charter school students scored significantly lower than their non-charter school counterparts, even after adjusting for student-level characteristics in the models (i.e., charter students averaged 4.7 points lower on the 4th grade NAEP mathematics test and 4.2 points on the reading test). The use of HLM in this example aided researchers to account for potential confounding due to sampling (e.g., lower numbers of charter schools sampled versus non-charter public schools, uneven distribution of charter schools across states), student-level characteristics (e.g., potential differences between public school students and those enrolled in charter schools affecting academic performance such as parental support), and school-level characteristics (e.g., schools located in states with lower mean achievement scores versus those in higher scoring states). Thus, their final HLM models were more reflective of actual academic achievement differences in 4th grade charter and non-charter students than analyses that aggregated data and failed to account for data structure, potential confounders, and unique error terms.

CONCLUSIONS

The relative ease with which unaccounted for moderators, hierarchical data, and data aggregation introduce YSP-type effects in data analyses is an important research issue. Spurious and confounding variables (i.e., moderators) have potential to distort statistical findings (Baker & Kramer, 2002, p. 2; Kunisaki, 2005, p. 1674). For this reason it is imperative for researchers to select research study plans that honor statistical assumptions and are capable of uncovering moderators that may *mask* or *alter* the relationship between predictor and outcome variables (Clark et al., 2005, p.1463; Sistrom & Garvan, 2004, p. 18). Several researchers suggest altering study designs to include *randomized multi-arm* trials to isolate the effects of unobserved variables (Baker & Kramer, 2002, p. 2; Sistrom & Garvan, 2004, p. 18). Researchers also recommend using large cross-sectional studies with stratified samples as a means of preserving data's hierarchical structure (Kraemer et al., 2006, p. 606).

Additionally, researchers should facilitate on-going collaboration with a statistician whenever feasible. Failing to involve statisticians in the research design and planning stages represents an often "missed opportunity." Statisticians are trained in appropriate application of statistical tests and techniques that would limit the chance of improperly analyzing data (e.g., using techniques that ignore nested data) and using techniques such as data aggregation that would lead to unnoticed YSP-type effects. Beyond data analysis, incorporating statisticians into research teams insures that data will be interpreted and disseminated in a manner that is both accurate and effective.

Until researchers are able to consistently detect the counterintuitive and contradictory statistical findings resulting from YSP, these effects will continue to lurk unnoticed within datasets. When YSP is undetected in datasets, reported study results and interpretations of study findings may be erroneous and have potential to mislead readers. Careful study design and thorough analysis may prevent researchers from becoming victim to this paradox. Researchers must cautiously consider the *big picture* while remaining mindful of its parts.

References

Baker, S. G., & Kramer, B. S. (2002). The transitive fallacy for randomized trials: If A bests B and B bests C in separate trials, is A better than C? *BMC Medical Research Methodology 2*(13), available at http://www.biomedcentral.com/1471-2288/2/13.

Baron, R., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, *51*(6), 1173-1182.

Bracey, G.W. (2004). Simpson's paradox and other statistical mysteries. *American School Board Journal*, 191(2), 32-34.

Braun, H., Jenkins, F., & Grigg, W. (2006). *A Closer Look at Charter Schools Using*

Hierarchical Linear Modeling(NCES 2006-460). U.S. Department of Education, National Center for Education Statistics, Institute of Education Sciences. Washington, DC: U.S. Government Printing Office.

Cates, C. J. (2002). Simpson's paradox and calculation of number needed to treat from meta-analysis. *BioMed Central Medical Research Methodology*, *2*(1), available at http://www.biomedcentra.com/1471-2288/2/1.

Clark, A. G., Boerwinkle, E., Hixson, J., & Sing, C. F. (2005). Determinants of the success of whole-genome association testing. *Genome Research*, *15*, 1463-1467.

Cohen, J.E. (1986). An uncertainty principle in demography and the unisex issue. *The American Statistician*, *40*(1), 32-39.

Davis, G.A. (2004). Possible aggregation biases in road safety research and a mechanism approach to accident modeling. *Accident Analysis and Prevention*, *36*, 1119-1127.

Grimm, L. G., & Yarnold, P. R. (2000). *Reading and understanding more multivariate statistics*. Washington, DC: American Psychological Association.

Howe, M.L., Rabinowitz, F.M. & Grant, M.J. (1993). On measuring (in)dependence of cognitive processes. *Psychological Review*, *100*(4), 737-747.

Kendall, M. G. (1952). George Udny Yule, C.B.E., F.R.S. *Journal of the Royal Statistical Society 115*(1), 156-161.

Kocik, J. (2001). Proof without words: Simpson's Paradox. *Mathematics Magazine*, *74*(5), 399.

Kraemer, H. C., Wilson, K. A., & Hayward, C. (2006). Lifetime prevalence in pseudocomorbidity in psychiatric research. *Archives of General Psychiatry*, *63*(6), 604-608.

Kunisaki, K. (2005). Simpson's paradox [Letter to the editor]. *Critical Care Medicine*, *33*(7), 1673-1674.

Malinas, G. (2001). Simpson's paradox: A logically benign, empirically treacherous hydra. *The Monist*, *84*(2), 265-283.

O'Connell, A. A., & McCoach, D. B. (2004). Applications of Hierarchical Linear Models for evaluations of health interventions. *Evaluation & The Health Professions*, *27*(2), 119-151.

Paik, M. (1985). A graphic representation of a three-way contingency table: Simpson's Paradox and correlation. *The American Statistician*, *39*(1), 53-54.

Roberts, J. K. (2004). An introductory primer on Multilevel and Hierarchical Linear Modeling. *Learning Disabilities*, *2*(1), 30-38.

Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society*, *13*(2), 238-241.

Sistrom, C. L., & Garvan, C. W. (2004). Proportions, odds, and risk. *Radiology*, *230*(1), 12-19.

Spellman, B. A., Price, C. M., & Logan, J. (2001). How two causes are different from one: The use of (un)conditional information in Simpson's paradox. *Memory & Cognition*, *29*(2), 193-208.

Thompson, B. (2006). *Foundations of behavioral statistics: An insight-based approach*. New York: Guilford.

Wainer, H. & Brown, L.M. (2004). Two statistical paradoxes in the interpretation of group differences: Illustrated with medical school admission and licensing data. *The American Statistician*, *58*(2), 117-123.

wa-Kivilu, M. w. (2003). Understanding the structure of data when planning for analysis: application of Hierarchical Linear Models. *South African Journal of Education*, *23*(4), 249-253.

Wardrop, R.L. (1995). Simpson's paradox and the hot hand in basketball. *The American Statistician*, *49*(1), 24-28.

Yule, G. U. (1903). Notes on the theory of association of attributes in statistics. *Biometrika*, *2*(2), 121-134.

Citation:

Goltz, Heather Honoré & Matthew Lee Smith (2010). Yule-Simpson's Paradox in Research. *Practical Assessment, Research & Evaluation*, 15(15). Available online: http://pareonline.net/getvn.asp?v=15&n=15.

Acknowledgements:

This work was supported in part by the Houston VA HSR&D Center of Excellence (HFP90-020). The views expressed in this article are those of the author(s) and do not necessarily represent the views of the Department of Veterans Affairs.

The authors would like to thank Dr. Bruce Thompson for stimulating our interest in Yule-Simpson's Paradox and providing an opportunity to present an early draft at the 2006 Southwestern Educational and Research Association (SERA) Meeting.

Corresponding Author:

Heather Honoré Goltz
Houston VA HSR&D Center of Excellence
2002 Holcombe Blvd. (Mail Stop 152)
Houston, TX 77030
heather.honore [at] va.gov