CRESST REPORT 828

MEASUREMENT ERROR IN MULTILEVEL MODELS OF SCHOOL AND CLASSROOM ENVIRONMENTS: IMPLICATIONS FOR RELIABILITY, PRECISION, AND PREDICTION

MAY, 2013

Jonathan Schweig

National Center for Research
on Evaluation, Standards, & Student Testing
UCLA | Graduate School of Education & Information Studies

Measurement Error in Multilevel Models of School and Classroom Environments: Implications for Reliability, Precision, and Prediction

CRESST Report 828

Jonathan Schweig
CRESST/University of California, Los Angeles

May 2013

National Center for Research on Evaluation,
Standards, and Student Testing (CRESST)
Center for the Study of Evaluation (CSE)
Graduate School of Education & Information Studies
University of California, Los Angeles
300 Charles E. Young Drive North
GSE&IS Bldg., Box 951522
Los Angeles, CA 90095-1522
(310) 206-1532

The work reported herein was supported by grant number 52306 from the Bill and Melinda Gates Foundation with funding to the National Center for Research on Evaluation, Standards, and Student Testing (CRESST). Part of this research is made possible by a pre-doctoral advanced quantitative methodology training grant (R305B080016) awarded to UCLA by the Institute of Education Sciences of the US Department of Education.

The findings and opinions expressed in this report are those of the authors and do not necessarily reflect the positions or policies of the Bill and Melinda Gates Foundation or the US Department of Education.

To cite from this report, please use the following as your APA reference: Schweig, J. (2013). *Measurement Error in Multilevel Models of School and Classroom Environments: Implications for Reliability, Precision, and Prediction* (CRESST Report 828). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

TABLE OF CONTENTS

MEASUREMENT ERROR IN MULTILEVEL MODELS OF SCHOOL AND CLASSROOM ENVIRONMENTS: IMPLICATIONS FOR RELIABILITY, PRECISION, AND PREDICTION [1]

Jonathan Schweig
CRESST/ University of California, Los Angeles

Abstract

Measuring school and classroom environments has become central in a nation-wide effort to develop comprehensive programs that measure teacher quality and teacher effectiveness. Formulating successful programs necessitates accurate and reliable methods for measuring these environmental variables. This paper uses a generalizability theory framework to compare and contrast four widely used approaches for accounting for measurement error in school and classroom level variables. Then, this paper uses two empirical examples to demonstrate how each of these approaches lead to different conclusions about measurement precision, and influences the conclusions about relationships between the environmental variables and policy-relevant outcomes. Additionally, this paper shows how one widely used model may misrepresent the structure of the data in many survey administration scenarios.

Introduction

Developing comprehensive programs that measure teacher effectiveness is one of the most pressing policy issues in education today. The development of these programs has been catalyzed by a growing research consensus (Nye, Konstantopolous, & Hedges, 2004; Rivkin, Hanushek, & Kain, 2005; Rowan & Correnti, 2009) that effective teachers can make meaningful differences in the learning trajectories of students; the injection of large amounts of federal funding through initiatives including Race to the Top; and the technological advances that have enabled districts to track and store large amounts of data on student achievement. Measures of school and classroom environments are frequently included as key features of programs to measure teacher effectiveness.

Data about school and classroom environments is often collected through surveys administered to teachers and students, who function as raters of the environments in which they work and study. In such a measurement scheme, data from a lower level (individuals) are aggregated to establish higher level constructs that allow for inferences to be made about group qualities (Kozlowski & Klein, 2000). This process for constructing group-level variables is

[1] I would like to thank Joan Herman, Jia Wang, and Noelle Griffin for their support of this study; and thank Felipe Martinez and Noreen Webb, for reviewing the earlier version of this report and for their thoughtful feedback. The author is grateful to the North Carolina Education Research Data Center for data. My special thanks go to Laquita Stewart and Fred Moss. Without their assistance, this project would not have been possible.

widely studied in organizational psychology. Chan (1998) describes group variables that emerge in this way as being elemental composition variables.

In 2009, the Bill and Melinda Gates initiated The Measuring Effective Teaching (MET) project, one of the largest efforts in the United States to assemble an empirical research base to describe effective teaching. MET is based largely on information collected through five different sources. These include the Tripod Survey, developed by Ron Ferguson, which measures the quality of classroom learning environments, and the Working Conditions Survey (WCS), developed by the New Teacher Center, which measures the quality of school working conditions (Bill & Melinda Gates Foundation, 2010).

Information about classroom and school environments can be used for a variety of purposes. It can be used as a direct measure of teacher or school quality. For example, Memphis, Tennessee bases 5% of a teacher evaluation on student surveys. By 2013, 10% of teacher evaluation in Chicago public schools will be based on student surveys. (Butrymowicz, 2012). In New York City, teacher and parent surveys about the school environment can account for up to 15% of a school's score on its annual Progress Report. ("NYC School Survey", n.d.)

Environmental data can also predict important outcomes, such as student achievement and teacher retention. Preliminary results from the MET project, for example, demonstrate significant relationships between student's perceptions of classroom environment and estimates of teacher's value added (VAM) scores. (Bill & Melinda Gates Foundation, 2010) Ladd (2011) discusses how links between teacher mobility and working conditions can be used to develop and test teacher retention policies. Better understanding how targeted improvements in working conditions may improve retention is particularly critical for schools serving high-poverty, low-achieving student populations, where teacher turnover rates may be as high as 50% (Ingersoll, 2001).

In each of these applications, understanding the validity and reliability of the measure of school or classroom environment is essential. Recent work (Lüdtke, Marsh, Robitzsch, Trautwein, Asparouhov, & Muthén, 2008; Lüdtke, Marsh, Robitzsch, & Trautwein, 2011; Raudenbush & Sadoff, 2008; Shin & Raudenbush, 2010) has demonstrated that quantifying errors in measures of school and classroom environments has immediate consequences for assessing how well groups can be distinguished based on individual perceptions (James, 1982). As one example, Wei and Haertel (2011) note that standard errors of measurement are often used to make "margin of error" adjustments in order to determine whether a school or classroom is meeting a performance standard. Differences in the perception of reliability can lead to different determinations of whether or not a school or classroom is meeting performance standards.

There are also consequences for assessing how strongly school and classroom climate relate to external variables. Raudenbush, Martinez, Bloom, Zhu, and Lin (2011) point out that without a reliable and valid measure of the environment, in a situation such as that described by Ladd (2011), it would be impossible to determine whether strategies that fail to improve teacher retention failed because they improved working conditions but did not improve retention, or because they did not improve working conditions in the first place.

One area that has gone relatively unexplored is an explicit consideration of the sources of measurement error that may emerge under different data collection designs, and how these sources of error impact judgments about reliability and validity.

This article investigates these issues for five commonly used designs for studying classroom and school environments. Generalizability theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972) is used to examine the designs within a unified framework for examining the sources of error variance and estimating the reliability of the resulting indicators. Finally, the article explores how different conceptions of error may influence the strength of the relationship between aggregate indicators and external variables.

Conceptual Framework

Table 1 describes five commonly used designs for quantifying error variance in measures of school and classroom environments. Design A is a two level hierarchical model, where people are nested in groups. In a situation where students are rating classroom environment, this would mean that each student assigns a single rating to describe the overall environment of their classroom. A particular classroom's score is determined by averaging those individual ratings together. In a situation where teachers are rating school working conditions, it would mean that each teacher provides only one rating of working conditions, and a school rating is obtained by averaging across those ratings. Design A is technically configured as a one-way random effects ANOVA (Raudenbush & Bryk, 2002; Shrout & Fleiss, 1979). This design is referred to in Marsh, Lüdtke, Robitzsch, Trautwein, Asparouhov, & Muthén, (2009) as a "manifest-manifest" design because it assumes that both the individual ratings and the average across those ratings are error-free.

Table 1

Descriptions of Five Commonly Used Designs for Quantifying Error Variance in Measures of School and Classroom Environments

Model	Description	Selected References	Measurement error	Sampling error
Design A	People nested in groups, each person provides one rating	• Raudenbush & Sadoff (2008) • Marsh et al. (2009)	No	No
Design B	People nested in groups, each person provides one rating	• Raudenbush & Sadoff (2008) • Marsh et al. (2009) • Lüdtke et al. (2008)	No	Yes
Design C	People nested in groups, each person answers several items. The same items are administered to all people.	• Marsh et al. (2009) • Lüdtke et al. (2008) • Preacher, Zyphur, & Zhang (2010) • Kane & Brennan (1977)	Yes	Yes
Design D	People nested in groups, each person answers several items. The same items are administered to all people. *Items are random.*	• Brennan (2001a) • Kane & Brennan (1977)	Yes	Yes
Design E	People nested in groups, each person answers several items. Different items are administered to all people.	• Bryk & Raudenbush (1988) • Raudenbush, Rowan, & Kang (1991) • Raudenbush & Bryk (2002) • Brennan (2001a)	Yes	Yes

Design B also describes a two-level hierarchical design, with people nested in groups. Unlike Design A, Design B does not assume that the sample mean is an error-free measure of the population mean. Had a different group of raters been randomly selected, or a larger number of them, a slightly different observed mean rating would have been obtained. Marsh et al. (2009) refer to this design as a "manifest-latent" model that captures sampling error.

In Design C, individuals do not provide a single rating of the environment, but rather, answer a set of survey items about their environment. For example, students answer 10 questions about classroom quality and the responses to these 10 items are averaged together to give a composite rating. Design C is a "doubly latent" model where the composite rating incorporates random measurement error as an additional source of variance (Marsh et al., 2009; Preacher et

al., 2010). Importantly, Design C treats the items as fixed and cross-classified; that is, it is assumed that the same survey items are administered to all individuals, regardless of group (classroom or school) membership. Kane and Brennan (1977) note that this design is technically configured as a mixed-effects split-plot ANOVA, where students or teachers are treated as random and items are treated as fixed.

Design D is also a design with cross-classified items. Unlike Design C, however, this design incorporates random item effects—the specific items included on the survey are assumed sampled from a larger pool of items. These samples of items are administered to individuals across groups (classrooms or schools). In Design D, items are explicitly considered as sources of error variance. Kane and Brennan (1977) conceptualize this as a split plot random-effects ANOVA, with students or teachers random and items random.

Finally, Design E is a three-level hierarchical model, with items nested in people, and people nested in groups. Notably, in this model, the items are not cross-classified but nested. Substantively speaking, this corresponds to a situation in which every individual receives a different set of items from every other individual.

Each of these five designs makes a different set of theoretical assumptions about the nature of the data, depending on whether they treat items as fixed or randomly sampled, and as cross-classified or nested. Designs A, B, and C treat the items as fixed. Certainly, there are many sensible reasons why items may be considered as fixed when interest centers in drawing inferences for a particular set of items, or where the pool of possible items is relatively small (Webb & Shavelson, 2005). In the case of a survey about classroom quality, it may be that there are only specific dimensions of classroom quality that are of interest, such as instructional clarity and organization, and the survey items thus constitute a "census" (Bollen & Lennox, 1991) of items, rather than a sample. In a working conditions survey, there may be a limited number of aspects of school leadership that are of interest—support for discipline, communication of vision, etc.

Designs D and E on the other hand treat the items as randomly sampled, thus assuming that there is an infinite (or large) number of items that could have been included, each measuring school or classroom climate in the same way. De Boeck (2008) points out that random items are of particular interest in situations where items are generated by cloning existing items, or where it is beneficial to have large pools of potential items available for inclusion on a particular test or survey form. In many surveys of the school or classroom environment, the items occupy a gray area, where there is some evidence for treating them as fixed, and some evidence for treating them as random. Kane and Brennan (1977) compare and contrast a variety of designs with

students nested within classrooms, crossed with items. They explore designs where both items and students are treated as random, where items are treated as random and students treated as fixed, where items are treated as fixed and students treated as random, and where both are treated as fixed. Kane and Brennan (1977) note that while it is usually advisable to treat both students and items as random, there are times where a design that treats items as fixed may be appropriate.

Designs C and D treat the items as crossed with people. All survey takers respond to the same set of items, regardless of school or classroom membership. On the other hand, Design E treats the items as nested within people. This means that it is assumed that each survey taker responds to a unique set of items. While the hierarchical model represented by Design E is common in education research, it represents a misspecification of the structure of the data in the case of survey administrations, where items are typically crossed with people (Kane & Brennan, 1977). Raudenbush, Rowan, and Kang (1991) explored the measurement of school climate using Design E, but did not specifically justify treating the items as nested, rather than crossed. Thus, a pressing question with Design E is, how robust are the inferences drawn to this misspecification of the data structure?

Though emerging from separate research traditions, the five designs can be understood under a common lens through the perspective of generalizability theory. First outlined by Cronbach et al. (1972), generalizability (G) theory is a flexible framework that can be used to assess the accuracy of both individual and group distinctions from a unified perspective. G-theory acknowledges that measurement error comes from a variety of sources allowing for a direct comparison of the relative magnitude of each source (Shavelson, Webb, & Rowley, 1989). This framework offers a consistent lens for examining the implications of the different conceptions of error variance in the five designs described above, specifically as they pertain to: 1) the reliability of school and classroom climate aggregates and 2) the relationships between these aggregates and external variables or indicators.

Reliability of Group Means

Two different types of reliability coefficient may be of interest, depending on whether the unit of analysis is the group or the individual. The reliability of group means would be useful for assessing how well classrooms can be distinguished based on individual student ratings, or how well schools can be distinguished based on individual teacher ratings. By comparison, when the unit of analysis is the individual the reliability of individual ratings is of interest. In that situation, it is important to understand how much error there is in student or teacher ratings of their classrooms or schools. In the current study, the primary units of analyses are respectively the

classroom and the school, and so interest centers on the reliability of the group means under these five scenarios. However, this study also includes estimates of Cronbach's alpha (Cronbach, 1951), an individual-level reliability coefficient, since that coefficient is frequently reported as an assessment of the reliability of group means. While it is often assumed that high values of alpha imply high group-mean reliability, this is not always true. Brennan (1995) explored the relationship between the reliability of group means and the reliability of individual scores and outlined a series of situations where group means can be less reliable than individual scores and, importantly, vice versa. Indeed, if the unit of measurement is the group, it may be theoretically desirable to have low variance within-groups. As Bliese (2000) points out, in the ideal scenario involving aggregated measures, there would be no within-group variance at all—and thus, the implied reliability of individual ratings would approach zero.

By quantifying the various sources of error that contribute to observed variance, G-theory also makes it possible to answer policy-relevant questions such as the minimum number of individuals per group that need to be surveyed in order to maintain acceptable group-mean reliability. For example, classroom and school sizes can vary widely, and understanding how that variation in size impacts the reliability of aggregates obtained for those clusters is imperative if these indices are to be interpreted for practical use. Likewise, it may be important to ask how many items need to be administered. Improved understanding of the components of measurement error can help reduce response burden and improve information quality.

Relationships With External Variables

Differing conceptions of error and reliability can also have implications for understanding the relationship of group means with external variables. For illustration, this study uses both ordinary least squares regression and multilevel contextual effect models when the relationship between an outcome and a predictor is hypothesized to differ across levels of aggregation.

Contextual effects have a long history in the social sciences (Iversen, 1991; Raudenbush & Bryk, 2002; Shin & Raudenbush, 2010) and are often of interest from a policy perspective, as they permit to address questions involving a comparison of the predictive power of indicators at the individual and group levels. For example, a contextual effects model might be used to assess the impact on teacher retention of school environment indicators as perceived by individual teachers and in the aggregate as reported by all teachers across the school. Similarly we might investigate how aggregate measures of classroom environment, (in contrast to or beyond individual students' perceptions of classroom environment) influence student achievement.

This article applies the models of measurement and sampling error of Designs A-E to two empirical examples involving a widely used student survey of the classroom environment, and a

teacher survey of working conditions. In doing so, this article addresses the following research questions:

- How do the differences in the treatment of error variance in each design impact the estimated reliability of aggregated and individual level variables?
- How does variation in the number of items and the number of people impact estimates of group-mean reliabilities in each design?
- How do the different models impact the determination of relationships between aggregated variables and external variables?

Methods

Sample and Data Sources

The Tripod Classroom Environment Survey. The Tripod Survey assessment (Ferguson, 2010) is designed to assess seven dimensions of teaching practice, often referred to as the "Seven C's": Caring, Captivating, Conferring, Clarifying, Challenging, Controlling, Consolidating. This version of the Tripod Survey contains 36 items, and was administered in an urban school district in California in 2010. All items have 5-point scales (1=*totally untrue* and 5=*totally true*). For illustration purposes, this analysis focuses only on the 8 eight items contained in the Challenging scale, and only on classrooms with more than 5 students. The Challenging scale is intended to measure the degree to which the classroom supports "academic rigor". An example item is, "My teacher wants us to use our thinking skills, not just memorize things." The sample used in this analysis contained 5,508 students nested in 285 classrooms. The average classroom size was approximately 17 students, and the range was from 5 to 33 students. The external outcome variable is an evaluation report of the teacher's ability to create and maintain and effective environment for student learning based on administrative observations. All teachers are evaluated on a 4-point scale, with scores of 3 or 4 indicating that they have met a particular standard. Observation data was available for 135 teachers.

The Working Conditions Survey. This survey was designed to assess teaching conditions at the school level (New Teacher Center, 2008). The sample data comes from the 2008 survey, administered to both teachers and principals at schools in K-12 public and charter schools across the state of North Carolina. For this analysis, only surveys completed by teachers were considered, resulting in a data set with 88,936 individual teacher cases in 2,423 schools. Though the average school size is approximately 37 teachers, schools in this analysis range from 5 teachers to 146 teachers. The survey measures five theoretical constructs: Time, Decision Making, Leadership, Professional Development, and Facilities & Resources. For illustration, this study focuses on the eight items included in the Decision Making scale. The Decision Making

items ask teachers to rate how large a role teachers play in a variety of decisions that impact classroom and school practices. For example, "Teachers are centrally involved in decision making about educational issues." All items are judged on a 5-point scale (1=*no role at all*, and 5=*the primary role*). In the absence of other external indicators in the dataset, the outcome variable is another survey item asking teachers to rate their level of agreement (1=*strongly disagree* to 5=*strongly agree*) with the statement, "Overall, my school is a good place to teach and learn."

Analytic Methods

Reliability of group means. In order to address the first research question, we first obtain variance component estimates for Designs B, C, D, and E.[2] Table 2 summarizes the variance components estimable for each design. In Design B the variance of the school or classroom score effects is denoted σ_s^2 and is considered true variance reflecting the extent to which schools or classrooms differ from one another, on average. The remaining variance denoted by $\sigma_{p,ps,e}^2$ represents residual variance from sources not systematically incorporated into the model—specifically, the confounded variance of people, a person-group interaction, and random error. Variance components for Design B were estimated using the nlme package in R (Pinheiro, Bates, Saikat, Sarkar, & the R Development Core Team, 2012; see Appendix A for additional details on the estimation of variance components for several designs).

[2]In Design A, all variance is treated as true variance, and the model is assumed to be measurement error and sampling error free.

Table 2

Estimable Variance Components and Their Associated Reliability Coefficients and Standard Errors for Designs B Through E

Design B		Design C		Design D		Design E	
Component	*Description*	*Component*	*Description*	*Component*	*Description*	*Component*	*Description*
σ_s^2	Variance of groups	σ_s^2	Variance of groups	σ_s^2	Variance of groups	σ_s^2	Variance of groups
$\sigma_{p,ps,e}^2$	Conflated variance of people, and the interaction of people and groups, and residual variance	$\sigma_{p,ps}^2$	Conflated variance of people, and the interaction of people and group	$\sigma_{p,ps}^2$	Conflated variance of people, and the interaction of people and group	$\sigma_{p,ps}^2$	Conflated variance of people, and the interaction of people and groups
		$\sigma_{pi,spi,e}^2$	Conflated variance of person-item interaction, person-item-group triple interaction, and residual variance	σ_i^2	Variance of items	$\sigma_{i,pi,spi,si,e}^2$	Conflated variance of items, person-item interaction, person-item-group triple interaction, group-item interaction, and residual variance
		σ_{si}^2	Variance of interaction of group and items	σ_{si}^2	Variance of interaction of group and items		
				$\sigma_{pi,spi,e}^2$	Conflated variance of person-item interaction, person-item-group triple interaction, and residual variance		
Reliability Coefficients and Standard Errors							
$\lambda = \frac{\sigma_s^2}{\sigma_s^2 + \frac{\sigma_{p,ps,e}^2}{n}}$		$\lambda = \frac{\sigma_s^2}{\sigma_s^2 + \frac{\sigma_{p,ps}^2}{n_p} + \frac{\sigma_{si}^2}{n_i} + \frac{\sigma_{pi,psi,e}^2}{n_p n_i}}$		$\lambda = \frac{\sigma_s^2}{\sigma_s^2 + \frac{\sigma_{p,ps}^2}{n_p} + \frac{\sigma_{si}^2}{n_i} + \frac{\sigma_i^2}{n_i} + \frac{\sigma_{i,pi,psi,e}^2}{n_p n_i}}$		$\lambda = \frac{\sigma_s^2}{\sigma_s^2 + \frac{\sigma_{p,ps}^2}{n_p} + \frac{\sigma_{i,pi,psi,si,e}^2}{n_p n_i}}$	
$SE = \sqrt{\frac{\sigma_{p,ps,e}^2}{n_p}}$		$SE = \sqrt{\frac{\sigma_{p,ps}^2}{n_p} + \frac{\sigma_{si}^2}{n_i} + \frac{\sigma_{pi,psi,e}^2}{n_p n_i}}$		$SE = \sqrt{\frac{\sigma_{p,ps}^2}{n_p} + \frac{\sigma_{si}^2}{n_i} + \frac{\sigma_i^2}{n_i} + \frac{\sigma_{i,pi,psi,e}^2}{n_p n_i}}$		$SE = \sqrt{\frac{\sigma_{p,ps}^2}{n_p} + \frac{\sigma_{i,pi,psi,si,e}^2}{n_p n_i}}$	

In Design C items are recognized as contributing to score variance allowing to tease apart some of the confounded error in Design B. Score variance here can be decomposed into four main sources: Variance between school or classrooms is denoted σ_s^2. Second, $\sigma_{p,ps}^2$ is the confounded variance attributed to people, and the interaction of people and groups. Thus, some people may rate their schools higher or lower than others in the same school, but because each person rates only one school or classroom, it is not possible to determine whether this represents true differences in people's opinions, or if they reflect the quality of a match between people and their schools/classrooms. Third, σ_{si}^2 represents the group-item interaction. Substantively, this describes the extent to which schools or classrooms differ in their relative standing across items, averaged over people (i.e. classrooms or schools may have higher ratings on some items than others). Lastly, $\sigma_{pi,psi,e}^2$ represents confounded person-item and person-group-item interactions, and residual variance. Variance components for Design C were estimated in Mplus version 6.11 (Muthén & Muthén, 2010).

Design D yields the four variance components described in Design C and a fifth component, σ_i^2, representing the main effect of items. This describes the extent to which some items are rated more highly than others, averaging across people and groups. For Design D, variance components were estimated using urGENOVA (Brennan, 2001b).

Design E yields only three variance components. Although items are included in this model, they are nested within people and so their variance is confounded with other sources. Design E yields a main effect variance of schools/classrooms, denoted σ_s^2. $\sigma_{p,ps}^2$ represents the confounded variance attributed to people, and the interaction between people and groups. $\sigma_{i,pi,psi,si,e}^2$ represents the confounded variance of items, person-item interaction, person-group-item triple interaction, group-item interaction, and residual variance. It is worth noting that several of the variance sources that are separable in Design C and D are confounded into a single residual term in Design E. Variance components were estimated using the nlme package in R.

Once variance components have been estimated, it is possible to estimate group-mean reliability coefficients for each design. All of these coefficients are of a general form that represents a ratio of true score variance to true score plus error variance (Shavelson & Webb, 1991).

The formulas for the reliability coefficients that are implied by Designs B through E are presented in Table 2, along with the associated standard errors of measurement (Crocker & Algina, 1986). Design B results in a reliability index where the denominator contains the total variance of the mean—which is composed of the main effect variance of schools/classrooms and the residual variance averaged over the number of individuals per group. Design C averages the

person (within school/classroom) variance over the number of individuals per group, the variance of the school-item interaction over the number of items, and the residual variance over both the number of individuals and the number of items.

Design D contains all of the variance components of Design C with one additional component, $\frac{\sigma_i^2}{n_i}$, in the denominator. Design E yields an index with the variance of items incorporated into the denominator, in the confounded variance term $\sigma_{i,pi,spi,si,e}^2$, which is averaged over both the number of items and the number of people.

As can be seen in the formulas, group-mean reliability is a function of the estimated variance components, the number of items (n_i), and the number of individuals per group (n_p), and thus, the reliability coefficients will vary depending on the size of the group and the number of items included. As such, the reliability coefficients presented in Table 2 were estimated for a range of group sizes and item counts. Standard errors will also vary depending on group size and number of items, and so all else being equal larger groups have smaller standard errors. The final reliability coefficient estimated was Cronbach's alpha (Cronbach, 1951). Cronbach (2004) shows how alpha can be expressed:

$$\alpha = \frac{\sigma_p^2}{\sigma_p^2 + \frac{\sigma_{pi,e}^2}{n_i}}$$

σ_p^2 and $\sigma_{pi,e}^2$ represent variance between individuals and residual variance respectively, averaged over items (n_i). This formulation makes apparent that the true-score variance included in α, σ_p^2, does not correspond to our unit of analysis if surveys are intended to measure properties of schools or classrooms. Moreover, alpha is dependent only on the number of items—and, thus, if the number of items is held constant, alpha will be constant, regardless of the size of the school or classroom. An infrequently acknowledged aspect of alpha is that it does not consider the clustered data structure. There are no variance sources here that are attributable to schools or classrooms. As such, even for within-group reliability, alpha is an inappropriate coefficient (Raykov & Penev, 2009). Nevertheless, alpha is presented here for reference because it is still frequently reported (inappropriately) instead of coefficients appropriate for aggregate indicators.

Relationships with external variables. In order to address the second research question concerning the relationship between aggregated variables and outcomes of policy interest, two different types of regression models were used. For the Working Conditions Survey, a contextual-effects model was estimated using the nlme package for Designs A, B, D, and E, and using MPlus 6.11 for Design C. The contextual-effects model can be expressed

$$y_{ij} = \gamma_{00} + \gamma_{10}(x_{ij} - x_{.j}) + \gamma_{01}x_{.j} + u_j + e_{ij} \text{ (1)}$$

Where γ_{00} is a grand mean, $x_{ij} - x_{.j}$is a group mean centered predictor, $x_{.j}$is the group mean value of the predictor, and u_j and e_{ij} are normally distributed error terms with mean 0. An important feature of the contextual effects model is that the two predictors, $x_{ij} - x_{.j}$ and $x_{.j}$, are orthogonal by design. The group-mean centered predictor is included in order to unconfound the between and within effects of the predictor on the outcome. γ_{01} describes the fixed between-effect of $x_{.j}$on y_{ij}, and γ_{10} describes a fixed within-group effect. The difference between these effects, $\gamma_{01} - \gamma_{10}$, is referred to as a contextual effect; the effect of the group beyond the effect of the individuals within groups. The outcome variable used in this analysis is another survey item asking teachers to rate their overall opinion of the school.

Frequently, the contextual effects model is used on the observed data x_{ij}, and observed means, $x_{.j}$. Such analysis assumes that there is no sampling error, and no measurement error, a model consistent with Design A (Table 1). Because reliability is assumed to be perfect, Design A introduces bias into the estimation of γ_{01} and γ_{10} if measurement error exists. (Lüdtke et al., 2008; Preacher et al., 2010; Raudenbush & Sadoff, 2008). One way of adjusting for measurement error in contextual effects models is to use the means of the Empirical Bayes posterior distributions for the independent variables as predictors (Shin & Raudenbush, 2010), These Empirical Bayes estimates for each group are essentially a reliability weighted average of an observed mean and a grand mean. Schools with low reliability borrow more heavily from the grand mean, resulting in what is commonly referred to as "shrinkage." For Design B, adjusted means are estimated only for the schools, since the only source of error variance considered in that model is sampling variability. For Designs C-E, adjusted means are estimated for each teacher, as well as for each school. This is because these four designs assume there is individual level measurement error as well as sampling error (see Appendix B for more details on estimating adjusted contextual effects for several designs).

For the Tripod survey, a regular ordinary least squares (OLS) regression was used with observation ratings of a teacher's ability to create and maintain an effective learning environment as the outcome variable. OLS is used here because the outcome variable is measured at the classroom level, and so the relationship between the predictor and the outcome can only be impacted by group-level components (Preacher et al., 2010). Four different predictors were used to predict observation ratings: the observed classroom means (Design A), and bias-adjusted classroom means consistent with Designs B through E (Raudenbush & Sadoff, 2008). It is possible to get a rough sense of how different conceptions of error variance will influence regression parameter estimates in both the contextual effects model and the OLS regression. In

its simplest form (for one predictor, or for predictors uncorrelated), the slope parameter can be expressed:

$$\gamma_{01} = \frac{cov(X,Y)}{var(X)}$$

And the reliability-adjusted slope can be expressed

$$\gamma_{01}^{(T)} = \frac{cov\ (T,Y)}{var\ (Y)} = \frac{1}{\lambda}\gamma_{01}$$

Since the reliability, λ, ranges from 0 to 1, this means that the minimum value of $\gamma_{01}^{(T)}$is the observed fixed effect estimate, γ_{01}. The more that reliability decreases, the greater the magnitude of the adjusted slope coefficient. Technically, these equations only apply to situations where all groups are the same size, and, thus, have equivalent reliabilities. However, in the general case, the reliability estimated at the harmonic-mean group size (Brennan, 2001a) can give a rough estimate of the extent of the disattenuation.

In the contextual effects model, it is anticipated that both the within-group slopes and the between-group slopes will be disattenuated when adjusted values are used as predictors. In the OLS model, it is anticipated that the slopes will also be disattenuated. The extent of that disattenuation will be a function of reliability.

Results

How do the Differences in the Treatment of Error Variance in Each Design Impact the Estimated Reliability of Aggregated and Individual Level Variables?

Table 3 shows the variance components for Designs B through E. For the WCS, σ_s^2 is substantial for all of the designs. In Design B, for example, σ_s^2 accounts for about 18% of the total variance. In Design C, σ_s^2 is still significant, but the percentage of variance due to schools in this model is much smaller—approximately 8%. Design C attributes approximately 30% of the variance to teachers within schools, and approximately 7% to the interaction of schools and items. In Design D, approximately 4% of the variance is due to schools. About 23% is due to teachers-within-schools, and approximately 30% of the total variance is due to items. The σ_{si}^2component accounts for approximately 4% of the variance. In Design E, approximately 6% of the variance is between schools, and approximately 18% of the variance is between teachers-within-schools.

Table 3

Estimated Variance Components for Designs B Through E

	Design B		Design C		Design D		Design E	
Survey	Component	Estimate	Component	Estimate	Component	Estimate	Component	Estimate
	σ_s^2	0.101	σ_s^2	0.096	σ_s^2	0.070	σ_s^2	0.101
	$\sigma_{p,ps,e}^2$	0.446	$\sigma_{p,ps}^2$	0.363	$\sigma_{p,ps}^2$	0.377	$\sigma_{p,ps}^2$	0.297
WCS			σ_{si}^2	0.077	σ_i^2	0.499	$\sigma_{i,pi,psi,si,e}^2$	1.178
			$\sigma_{pi,psi,e}^2$	0.607	σ_{si}^2	0.065		
					$\sigma_{pi,psi,e}^2$	0.616		
	σ_s^2	0.099	σ_s^2	0.097	σ_s^2	0.095	σ_s^2	0.099
	$\sigma_{p,ps,e}^2$	0.459	$\sigma_{p,ps}^2$	0.392	$\sigma_{p,ps}^2$	0.390	$\sigma_{p,ps}^2$	0.386
Tripod			σ_{si}^2	0.021	σ_i^2	0.010	$\sigma_{i,pi,psi,si,e}^2$	0.581
			$\sigma_{pi,psi,e}^2$	0.549	σ_{si}^2	0.021		
					$\sigma_{pi,psi,e}^2$	0.549		

For the Tripod Survey, σ_s^2 is also substantial for all of the designs. In Design B, σ_s^2 accounts for approximately 18% of the variance. In Design C, approximately 10% of the variance is attributed to classrooms. Approximately 40% is attributed to students-within-classrooms, and approximately 2% is due to the classroom-item interaction, σ_{si}^2. In Design D, approximately 9% of the variance is between classrooms, and approximately 36% is between students within classrooms ($\sigma_{p,ps}^2$). Only approximately 1% of the variance is due to items. The classroom-item interaction, accounts for approximately 2% of the variance. In Design E, approximately 9% of the variance is between classrooms, and approximately 33% is between students-within-classrooms.

Because it decomposes variance into the largest number of distinct components, Design D in particular makes for an important point of comparison between the WCS and the Tripod Survey. In this design, it is clear that there is a large difference between these two surveys in the amount of variance that is attributable to items (σ_i^2). Additionally, the σ_{si}^2 effect is larger in the WCS than in the Tripod Survey.

Table 3 also shows that—for both surveys—the σ_s^2 component is equal for Designs B and E, but that the σ_s^2 components of Design C and D is slightly smaller than the σ_s^2component from Designs B and E. The magnitude of that difference is equal to

$$\frac{\sigma_{is}^2}{n_i}$$

Because the σ_{si}^2 effect is larger in the WCS, there is a larger difference between the σ_s^2 estimates in the WCS than there is in the Tripod Survey (see Appendix C for more details).

For both surveys, the estimates of $\sigma_{p,ps}^2$ differ between Design C and Design E. In fact, the difference in the estimates of teacher variance between Design C and Design E is equal to

$$\frac{\sigma_{i,si}^2}{n_i}$$

Because the σ_i^2 and σ_{si}^2 components are larger for the WCS than for the Tripod survey, there is larger of a discrepancy in the estimate $\sigma_{p,ps}^2$ between these designs for the WCS than there is for the Tripod (see Appendix C for more details).

These differences in the magnitude of the σ_i^2 and σ_{si}^2variance components have immediate implications for the estimated reliabilities under the different designs. Table 4 shows reliability estimates across a range of classroom and school sizes. The ranges of reliabilities presented in Table 4 show that, while larger groups can result in precise measurement, smaller groups can be highly unreliable.

Table 4

Group-Mean Reliability Coefficients for Designs B Through E Across a Range of Group Sizes

		Group mean reliability			
Survey	Design	Min group size	Mean group size	Max group size	Cronbach's alpha
WCS	Design B	0.531	0.854	0.970	0.840
	Design C	0.496	0.779	0.883	
	Design D	0.303	0.442	0.487	
	Design E	0.532	0.854	0.970	
Tripod	Design B	0.684	0.787	0.877	0.872
	Design C	0.666	0.765	0.854	
	Design D	0.656	0.754	0.842	
	Design E	0.683	0.787	0.877	

Note: Group sizes for WCS are 5, 25 and 143, respectively. Group sizes for Tripod are 10, 17, and 33 respectively.

For the WCS, the group-mean reliability estimated in Design C (λ =.779 at the mean group size) is about 10% lower than the reliabilities in Designs B and E. The reliability coefficient for Design D, when items are treated as crossed and random, is nearly 50% lower than the other estimates, because the σ_i^2 and σ_{si}^2variance components are relatively large in the WCS.

For the Tripod, the reliability estimates are fairly consistent across all four designs. While the group-mean reliability estimated in Design C (λ =.765 at the mean group size) is lower than the reliabilities in Designs B and E, it is only approximately 3% lower. The reliability coefficient for Design D, when items are treated as crossed and random, is only about 5% lower than the other estimates. This is because the σ_i^2 and σ_{si}^2variance components are relatively small in the Tripod.

Cronbach's alpha, presented in the last column, is constant across all group sizes. This means that, for the small groups, alpha highly over-estimates reliability. In the WCS, alpha overestimates group-mean reliability in Designs B and E (at the mean group size). In the Tripod Survey, alpha over-estimates group-mean reliability for all designs in all but the largest groups.

The group-mean reliabilities for Designs B and E are equal in both the WCS and Tripod examples for all group sizes (approximately .85 in the WCS, and .79 in the Tripod at the mean group size). This has the strong implication that accounting for error variance among the items has no impact on the estimate of group level reliability if the items are treated as nested (see Appendix D for more details).

There are important policy implications for these differences in reliability estimates, particularly in a performance evaluation context, such as those described in Memphis, Chicago or New York City. If teachers or schools whose aggregate ratings place them below a certain threshold are to be placed on a list for potential intervention, an informed sense of measurement reliability can be highly impactful. This can be shown by examining how the magnitude of the standard errors varies across designs.

Figure 1 shows the standard errors of measurement implied by Designs B through E for both the WCS and Tripod surveys, based on the formulas presented in Table 2. Because the standard errors are influenced by group size, they are presented at the minimum, harmonic mean, and maximum group sizes for both surveys.

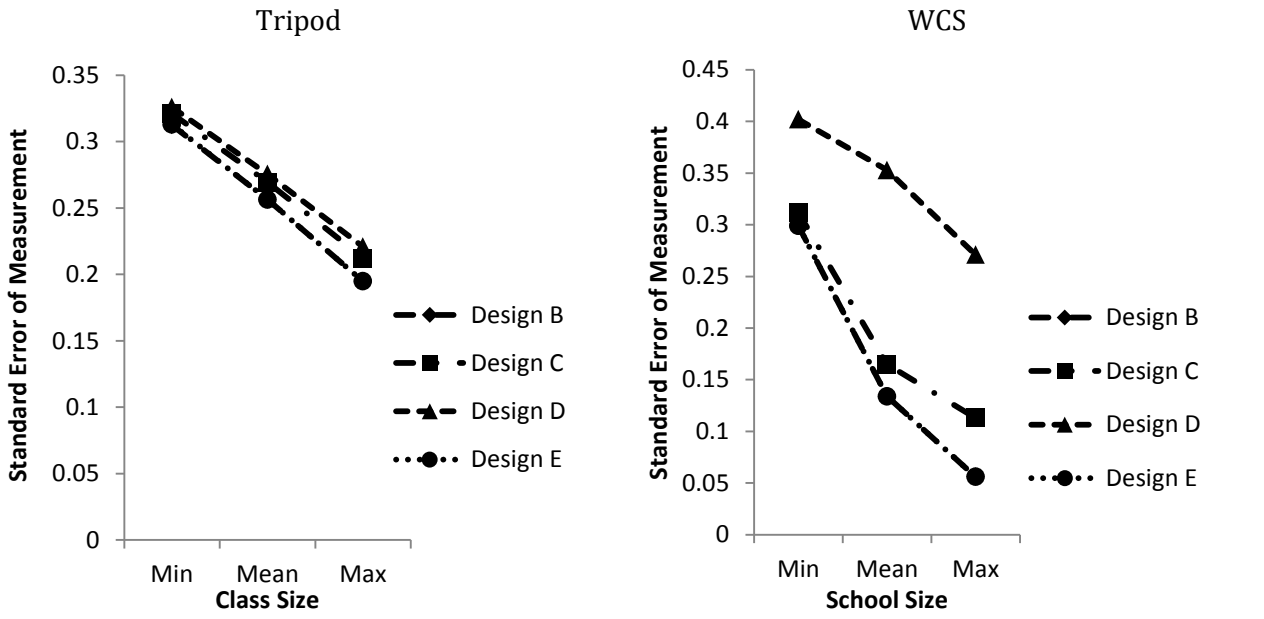

Figure 1. Standard errors of measurement based on reliability estimates for a range of group sizes.

For the WCS, the standard errors get smaller as the group sizes increase for all four designs, and the standard errors for Designs B and E are identical (hence the plots overlap). However, for schools that have only 5 raters (the minimum school size), the standard errors are uniformly large. If we were to consider a school with a median rating on the Decision Making scale, the 95% confidence interval implied by this standard error of measurement would range from approximately the 2nd percentile to the 96th percentile.

While standard errors decrease as a function of group size for all designs for the WCS, when items are treated as crossed and random, as in Design D, the standard errors are systematically larger than for all other designs, which is expected given that the estimated reliability is so much lower for this design. In this case, even for a school of average size, the 95% confidence interval spans from the 2nd percentile to the 96th percentile.

For the Tripod, some similar patterns are observed. Standard errors also decrease as group sizes increase, and the standard errors for Design B and E are identical. For all four designs, for small classrooms (10 students) with a median rating on the Challenging scale, the standard errors are large enough that the 95% confidence interval for a classroom's score ranges from approximately the 7th percentile to the 99th percentile. There is less of a difference between Design D and the other three designs for the Tripod survey, because there is less item variance in the Tripod survey, and the estimated reliability for Design D is more similar to the estimated reliabilities for Designs B, C, and E.

This highlights two important considerations in regard to standard errors of measurement. First, the sources of error variance that are quantified can have large impacts on the estimated size of the standard errors, particularly if item variance is large and is considered as a source of error variance. Second, standard errors vary as a function of group size, and while standard errors estimated at larger group sizes may be satisfactory for making determinations about whether a school or classroom has met the cut score for a particular standard, standard errors estimated at smaller group sizes may be unacceptably large.

How Does Variation in the Number of Items and the Number of People Impact Reliability in Each Design?

Based on the variance components in Table 3, it is possible to explore ways in which reliability can be improved by asking questions like, "with the current number of items, how many teachers need to be surveyed per school in order to get reliable estimates of the school means?" Or, "will adding additional items, or surveying more people result in greater gains in measurement precision?" Figure 2 shows how changes in group size and the number of items impact reliability estimates for Designs B through E. For investigation into the impact of sample size, items are held constant at 8. For investigation into the impact of number of items, group sizes are held constant at the respective harmonic mean group sizes.

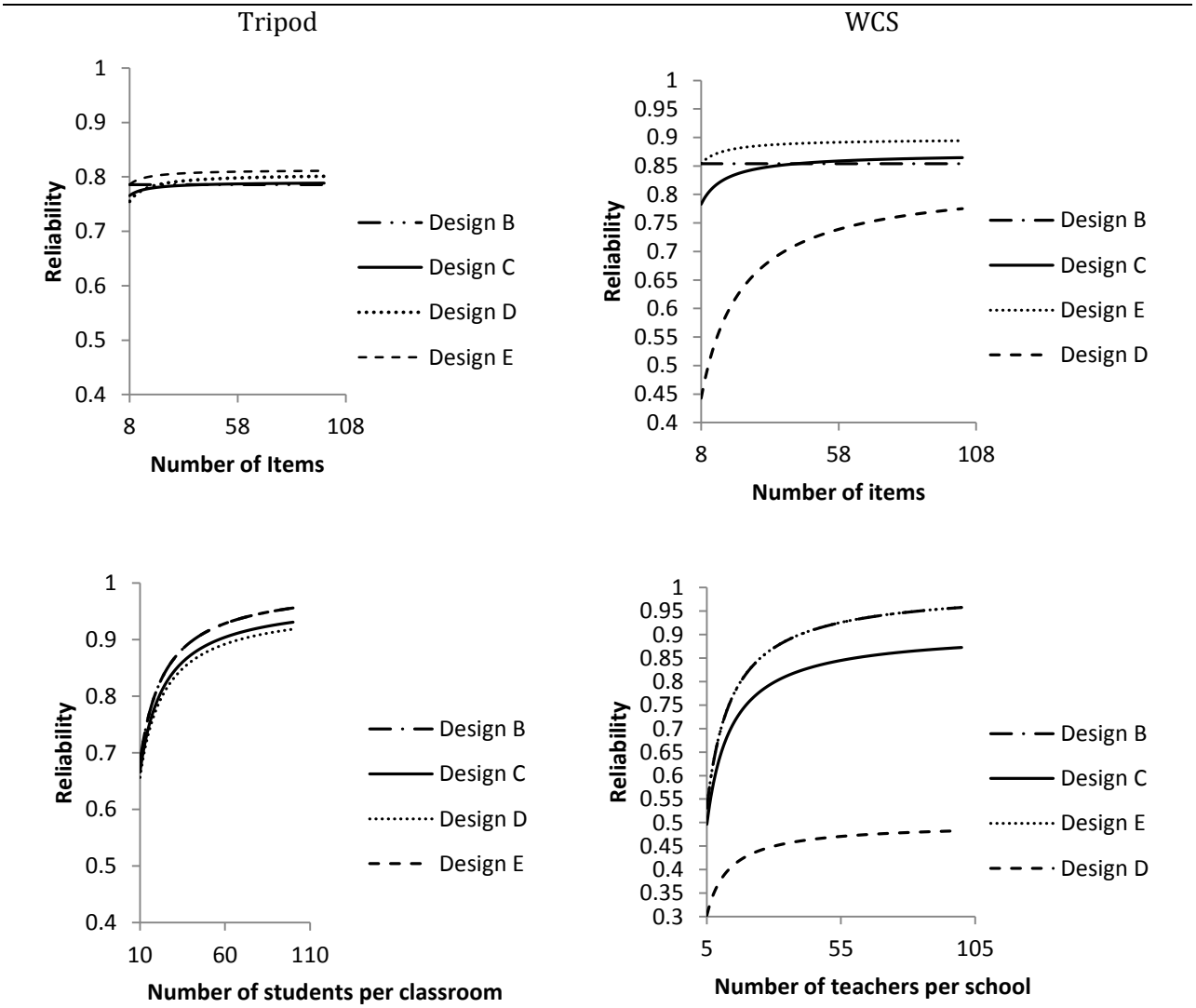

Figure 2. Changes in group-mean reliabilities for a variety of group size and item configurations.

There is an interesting contrast between the graphs for these two survey instruments. With regards to the WCS, for Designs B, C, and E, increasing the number of teachers per group has a larger impact on reliability than raising the number of items. Even with 100 items, the reliabilities for those three designs are less than .9. For Design D, however, increasing the number of items has a much larger impact on absolute reliability than increasing the number of raters. With 8 items, reliability does not reach .5 even with 100 raters per group.

With regards to the Tripod, a similar conclusion would be reached. Increasing the number of students per classroom has a larger impact in general than adding items. However, this is true across all designs—including Design D—and the profiles of all of the plots for the Tripod survey are essentially flat.

How Do the Different Models Impact the Relationships Between Aggregated Variables and External Variables?

Table 5 present results regarding the differences in correction to regression parameters that occur when different models are used to describe error variance. Design D, which had the lowest estimates of group mean reliability, has the highest adjustment to the between-group regression parameters. In fact, the between schools slope parameter is nearly 1.5 times as large in Design D as in the other designs. This is all consistent with expectations based on the estimated reliabilities for each design.

Table 5

Differences in Correction to Regression Parameters That Occur When Different Models Are Used to Describe Error Variance

Survey	Model	Between	Within	Compositional effect	Percent change
WCS	Design A	0.90	0.46	0.44	--
	Design B	0.99	0.46	0.53	20.45%
	Design C	1.13	0.55	0.58	31.82%
	Design D	1.65	0.62	1.03	134.09%
	Design E	1.21	0.68	0.53	20.45%

	Model	Parameter	Percent change
Tripod	Design A	0.64	--
	Design B	0.86	34.4%
	Design C	0.87	35.9%
	Design D	0.89	39.1%
	Design E	0.86	34.4%

Note. All parameters are statistically significant at the .01 level.

While it is possible to look only on the changes to the regression coefficients on their own, an interesting pattern emerges in the contextual effects. Design D results in the largest contextual effect for the WCS. The size of the contextual effect is nearly 1.3 times larger than the observed contextual effect. The contextual effect estimated in Design A can be interpreted as meaning, "for two teachers who rate their school environment equally, but work in schools that differ in school quality by one scale point, there is an expected difference in an individual teacher's feelings that their school is a good place to work and learn of .44 units. On this scale, that is than a change from "Neither disagree nor agree" to "Somewhat agree." However, in Design D, the

conclusion would be that there would be an anticipated change of almost one scale point—corresponding to a move from a feeling of ambivalence ("Neither disagree nor agree") to a feeling of agreement.

Because the σ_i^2 and σ_{si}^2 components are large in the WCS, the estimate of $\sigma_{p,ps}^2$ is smaller in Design E. This results in lower estimates of within-group reliability, and, thus, Design E results in the largest disattenuation of the within-group slope. In fact, this disattenuation is large enough that Design E results in a contextual effect equal to that of Design B. This suggests that, even though it accounts for item variance, using Design E can result in a downward bias in the estimate of the contextual effect. For the Tripod, Designs B and E give the same adjusted slope estimates. Design D, which has the lowest estimated reliability, has the largest disattenuated regression parameter. In the Tripod survey, because the σ_i^2 and σ_{si}^2 components are so much smaller (as compared to those in the WCS), the group-mean reliability estimates are much more similar across all designs, and the differences between the parameter estimates is much less pronounced.

Summary and Discussion

This study examined the reliability of group indicators created when surveys are administered, and individual responses aggregated in order to make classroom or school—level inferences. Consistent with previous work (Marsh et al., 2009; Raudenbush & Sadoff, 2008), this study suggests that how we account for measurement error in group-level variables has implications for the estimation of reliability, measurement precision, and biases in relationships with external variables. This study goes beyond prior research by demonstrating that it is not only the presence or absence of measurement error in the model that impacts precision and biases relationships with external variables. The specific sources of error variance that are quantified play a substantial role, as well. Specifically, it was shown that different models of error variance lead to a) different estimates of the reliability of aggregate measures, b) different conclusions about measurement precision, c) different sense of the effect of additional individuals or items for improving measurement precision, and d) different inferences about the strength of adjusted relationship with external variables. In addition, it was shown that Cronbach's alpha is an inappropriate coefficient for estimation of group-mean reliability.

A thorough consideration of the sources of error variance can have direct implications for policy. This raises the issues addressed in this paper from ones of psychometric interest to ones of practical importance. For example, differing conceptions of precision may impact whether or not it is justifiable to include a measure as part of a teacher or school evaluation. Or, it may impact how much weight is given to a measure in an evaluation composite. Careful consideration

of error variance can lead to a more nuanced understanding of how to think about how predictive working conditions are of teacher quit decisions, or how much variance in student achievement is attributable to variance in classroom climates. Moreover, the range of parameter estimates that can result depending on how error is quantified raises important questions. What does it mean for policy if the potential magnitude of the effect of interest has a range that is fairly large?

Of all of the sources of error variance that were considered in this study, two design considerations consistently had the largest impact on estimates reliability, precision, and regression parameters. Specifically, the decision of whether it is appropriate to treat items as crossed or nested, and whether it is appropriate to treat items as random or fixed. As is suggested in Kane and Brennan (1977), there is no "universally best" (p. 289) approach to treating items as random or fixed in scenarios where surveys are administered to teachers or students and the school or classroom is the unit of analysis. This decision, ultimately, can only be made in the context of a particular study (Kane & Brennan, 1977). However, it is rarely the case that a fully nested design appropriately describes the context of a particular survey administration. In the vast majority of cases, the fully nested design, Design E, misspecifies the structure of the error.

The results of this study show that this misspecification can have important practical consequences. First, and perhaps most surprisingly, the group-mean reliability implied by Design E is equal to the group-mean reliability implied by Design B. This means that, when groups are the unit of analysis, it is inconsequential whether you treat items as fixed or fully nested. And the reliability of class or school means is likely to be over-stated in Design E. The extent of this over-statement will be a function of the amount of item variance, and the size of the group-item interaction. What's more, the reliability of individual scores is likely to be understated in Design E, also as a function of item variance. In the case of a contextual effects model, this can impact inference about the magnitude of a contextual effect. Overall, the results show that when aggregated survey measures are going to be used as indicators of school and classroom environments, careful attention should be paid to the sources of error that are relevant to the data collection design.

Additional Questions and Limitations of the Current Study

This study described how several competing models of error variance can impact reliability, and can also impact parameter bias. Thoughtful consideration of the relevant sources of bias for the design and use of measures of classroom and school variables can be highly consequential. There are, however, several limitations of this study, and these present areas for future research and additional questions.

Hidden Facets

It is important to note that, due to the cross-sectional nature of the data set, only three main sources of variance are considered in this study. Variance due to groups (schools or classrooms), variance due to raters (teachers or students) and variance due to items. It may be that, in fact, there are other sources of error variance that are salient that were not included in this study. In G-Theory, these are referred to as "hidden" facets (Shavelson & Webb, 1991). Just as these models demonstrated that three competing models of error could result in different conclusions, incorporating other sources of variance into the design would also potentially impact results. For example, there is no variance attributed to occasion in any of the models investigated in this study—but it may be that if this same survey were administered at the beginning and end of the school year, there would be some variance attributed to the occasion on which the survey was administered.

In addition, with both the student surveys of classroom climate and the teacher surveys of schools, only one level of nesting is accounted for. For the Tripod, the assumption was that students were nested within classrooms. For the WCS, the assumption was that teachers were nested within schools. However, it is possible to imagine nested facets that are excluded here. For example, students nested within classrooms nested within teachers (nested within schools). Or teachers nested within departments nested within schools. Wei and Haertel (2011) suggest that omitted levels of variance can bias the estimation of variance components, and so this merits further consideration.

Assumption of Reflective Measurement

In the two examples used in this study, it is assumed that schools and classrooms have true scores on the climate variables in question, but that individuals do not. In other words, variance between schools represents true variance in the quality of working conditions, or variance between teachers represents true variance in the fairness of teachers, but variance between teachers in the same school (or students assigned to the same teacher) can be attributed to sampling variability and represents "noise."

This has important ramifications for the appropriateness of the measurement models used in this article. As Lüdtke et al. (2011) point out, the use of reflective, rather than formative indicators at level 2 is contingent upon the supposition that individuals do not have meaningful true scores on the construct of interest. In other words, the latent variable model is built on the supposition that variance between people is caused by school or classroom features, and not by true differences between people.

This is a nuanced issue in organizational climate research. Sirotnik's (1980) affective-descriptive continuum attempts to delineate items that are intended to measure individual, psychological constructs from items that are intended to measure organizational constructs. Many items defy categorization and fall somewhere between the two extremes. Take the following example (Sirotnik, p. 261) of five potential items about the construct of trust.

- I am generally a trusting type of person.
- I trust the staff members at this school.
- We trust one another at this school.
- You trust one another at this school.
- Staff members trust one another at this school.

Sirotnik suggests that the first item listed above is on the affective end of the continuum. This is signaled both by the item's "I-form" and because the item content relates to a psychological construct, an individual's trustworthiness. The fifth item is on the descriptive end of the continuum. It is a "they-form" item, and positions individuals as raters of a single organizational quality. The most interesting items are the middle three, which gradually progress from items that are clearly describable as psychological, into a space that is less clearly circumscribed. This raises important questions about what is being measured. Are the items measuring qualities of the classroom or school? Qualities of the teachers or students? Or something else?

The two surveys used in this article are based primarily on "they-form" items from the descriptive end of the continuum. However, that does not mean they do not reflect a certain amount of true variation in the psychological standing of the respondents. If school climate is to be measured by aggregating lower level responses to questions about climate, attention should be paid to whether items refer primarily to the psychological characteristics of individuals, or primarily to characteristics of organizations.

Tau-Equivalent Measures and Cross-Level Measurement Invariance

Several other assumptions were made in the course of this study. It is assumed that all of the items in the Decision Making and Challenging scales are tau-equivalent (Marcoulides, 1996) and that they all measure the constructs equally well. However, it may be the case that the items differentially tap on the Decision Making or Challenging constructs. This may make the "averaging over" inherent in G- Theory inappropriate.

It is also assumed that the concepts of Decision Making and Challenge can be constructed by averaging over individual responses. Much recent work in multilevel factor analysis has revealed that it is often the case that the structure of constructs at the group level is not

isomorphic (or even partially isomorphic) to the structure of constructs at the individual level (Hox, 2002; Holfve-Sabel & Gustaffsen, 2005; Zyphur, Kaplan, & Christian, 2008). One possible approach is to use a Structural Equation Modeling (SEM) framework, as Marsh et al. (2009) shows. This allows for the isomorphism assumption to be tested empirically (Lüdtke et al., 2011).

References

Bill & Melinda Gates Foundation. (2010). *Learning about teaching: Initial findings from the Measures of Effective Teaching project.* Retrieved from http://www.metproject.org/downloads/Preliminary_Findings-Research_Paper.pdf

Bliese, P. D. (2000). Within-group agreement, non-independence, and reliability: Implications for data aggregation and analyses. In K. J. Klein & S. W. J. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations: Foundations, extensions, and new directions* (pp. 349–381). San Francisco: Jossey-Bass.

Bollen, K. A., & Lennox, R. (1991). Conventional Wisdom on Measurement: A Structural Equation Perspective. *Psychological Bulletin*, 110, 305-14.

Brennan, R. L. (1995). The conventional wisdom about group mean scores. *Journal of Educational Measurement*, *32*(4), 385-396.

Brennan, R. L. (2001a). *Generalizability theory*. NY: Springer-Verlag.

Brennan, R. L. (2001b). *urGENOVA* (Version 2.1) [Computer software and manual]. Iowa City, IA: University of Iowa (Available on http://www.education.uiowa.edu/casma/).

Bryk, A. S., & Raudenbush, S. W. (1988). Toward a more appropriate conceptualization of research on school effects: A three-level hierarchical linear model. *American Journal of Education*, 65-108.

Butrymowicz, S. A. (2012, May 13) Student surveys for children as young as 5 years old may help rate teachers. *The Washington Post*. Retrieved from http://www.washingtonpost.com/local/education/student-surveys-may-help-rate-teachers/2012/05/11/gIQAN78uMU_story.html

Chan, D. (1998). Functional relations among constructs in the same content domain at different levels of analysis: A typology of composition models. *Journal of Applied Psychology*, *83,* 234-246.

Crocker, L., & Algina, J. (1986). *Introduction to Classical and Modern Test Theory*. Wadsworth Group; Belmont, CA.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16(3)*, 297–334.

Cronbach, L. J. (2004) .My current thoughts on coefficient alpha and successor procedures. *Educational Psychological Measurement*. *64*: 391-418.

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: John Wiley.

De Boeck, P. (2008). Random item IRT models. *Psychometrika, 73*, 533–559.

Ferguson, R. (2010, October 14). *Student perceptions of teaching effectiveness.* Retrieved from http://www.gse.harvard.edu/ncte/news/Using_Student_Perceptions_Ferguson.pdf

Holfve-Sabel, M., & Gustaffsson, J. (2005). Attitudes towards school, teacher, and classmates at classroom and individual levels: An application of two-level confirmatory factor analysis. *Scandinavian Journal of Educational Research, 49*(2): 187-202.

Hox, J. J. (2002). *Multilevel analysis: Techniques and applications*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Ingersoll, R. (2001). Teacher turnover and teacher shortages: An organizational analysis. *American Educational Research Journal, Vol. 38,* No. 3, pp. 499-534.

Iversen, G. R. (1991). *Contextual analysis* (Vol. 81). Newbury Park, CA: Sage.

James, L. R. (1982). Aggregation bias in estimates of perceptual agreement. *Journal of Applied Psychology, 67,* 219-229.

Kane, M. T., & Brennan, R. L. (1977). The generalizability of class means. *Review of Educational Research, 47,* 267-292.

Kozlowski, S. W. J., & Klein, K. J. (2000). A multilevel approach to theory and research in organizations: Contextual, temporal, and emergent processes. In K. J. Klein & S. W. J. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations.* San Francisco: Jossey-Bass. 3–90.

Ladd, H. (2011). Teachers' perceptions of their working conditions: How predictive of planned and actual teacher movement? *Educational Evaluation and Policy Analysis, 33*(2), 235-261.

Lüdtke, O., Marsh, H. W., Robitzsch, A., & Trautwein, U. (2011). A 2×2 taxonomy of multilevel latent contextual models: Accuracy and bias trade-offs in full and partial error-correction models. *Psychological Methods, 16* (4) 444-467.

Lüdtke, O., Marsh, H. W., Robitzsch, A., Trautwein, U., Asparouhov, T., & Muthén, B. (2008). The multilevel latent covariate model: A new, more reliable approach to group-level effects in contextual studies. *Psychological Methods, 13,* 203–229.

Marcoulides, G. A. (1996). Estimating variance components in generalizabiilty theory: The covariance structure analysis approach. *Structural Equation Modeling, 3,* 290-299.

Marsh, H. W., Lüdtke, O., Robitzsch, A., Trautwein, U., Asparouhov, T., & Muthén, B. (2009). Doubly-latent models of school contextual effects: integrating multilevel and structural equation approaches to control measurement and sampling error. Multivariate Behavioral Research, 44, 764–802.

Muthen, B. O. (1994). Multilevel covariance structure analysis. *Sociological methods & research*, 22(3), 376-398.

Muthén, B. O., & Muthén, L. K. (2010). Mplus (version 6.11) [computer software] Los Angeles: Muthen & Muthen.

New Teacher Center. (2008). *Validity and Reliability of the North Carolina Teacher Working Conditions Survey*. UC Santa Cruz. Retrieved from ncteachingconditions.org/research2008

"NYC School Surveys". (n.d.). Retrieved from http://schools.nyc.gov/Accountability/tools/survey/default.htm Nye, B., Konstantopoulos, S., & Hedges, L. V. (2004). How large are teacher effects? *Educational Evaluation and Policy Analysis*, *26*(3): 237-257.

Nye, B., Konstantopoulos, S., & Hedges, L. V. (2004). How large are teacher effects? *Educational evaluation and policy analysis*, *26*(3), 237-257.

Pinheiro, J., Bates, D., Saikat, D., Sarkar, D., & the R Development Core Team (2012). nlme: Linear and Nonlinear Mixed Effects Models. R package version 3.1-104.

Preacher, K. J., Zyphur, M. J., & Zhang, Z. (2010). A general multilevel SEM framework for assessing multilevel mediation. *Psychological Methods, 15*, 209–233.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models (2nd ed.).* Newbury Park, CA: Sage.

Raudenbush, S. W., Martinez, A., Bloom, H., Zhu, P., & Lin, F. (2010). *Studying the reliability of group-level measures with implications for statistical power: A six-step paradigm*. University of Chicago Working Paper.

Raudenbush, S. W., Rowan, B., & Kang, S. J. (1991). A multilevel, multivariate model for studying school climate with estimation via the EM algorithm and application to U.S. high-school data. *Journal of Educational Statistics, 16,* 295–330.

Raudenbush, S., & Sadoff, S. (2008). Statistical inference when classroom quality is measured with error. *Journal of Research on Educational Effectiveness, 1*, 138–154.

Raykov, T., & Marcoulides, G. A. (2011). *Introduction to psychometric theory*. New York, NY: Routledge.

Raykov, T., & Penev, S. (2009). Estimation of maximal reliability for multiple-component instruments in multilevel designs. *British Journal of Mathematical and Statistical Psychology*, *62*, 129–142.

Rivkin, S. G., Hanushek, E., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica*, *73*, 417–458.

Rowan, B., & Correnti, R. (2009). Studying reading instruction with teacher logs: Lessons from the Study of Instructional Improvement. *Educational Researcher*, *38*, 120–131.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer* (Vol. 1). Newbury Park, CA: Sage.

Shavelson, R. J., Webb, N. M., & Rowley, G. (1989). Generalizability theory. *American Psychologist*, 922-932.

Shin, Y., & Raudenbush, S. W. (2010). A latent cluster-mean approach to the contextual effects model with missing data. *Journal of Educational and Behavioral Statistics, 35*, 26–53.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass Correlations: Uses in Assessing Rater Reliability. *Psychological Bulletin, 2,* 420-428.

Sirotnik, K. A. (1980). Psychometric implications of the unit-of- analysis problem (with examples from the measurement of organizational climates). *Journal of Educational Measurement, 17*, 245–282.

Webb, N. M., & Shavelson, R. J. (2005). Generalizability theory: Overview. In B. S. Everitt & D. C. Howell (Eds.), *Encyclopedia of statistics in behavioral science* (Vol. 2, pp. 717–719). Chichester, UK: John Wiley & Sons Ltd.

Wei, X., & Haertel, E. (2011). The effect of ignoring classroom-level variance in estimating the generalizability of school mean scores. *Educational Measurement: Issues and Practice*. *30* (1) 13-22.

Zyphur, M., Kaplan, S., & Christian, M. (2008). Assumptions of cross-level measurement and structural invariance in the analysis of multilevel data: problems and solutions. *Group Dynamics: Theory, Research, and Practice*, *12*, 127–140.

Appendix A

The following appendix provides more detail on the models used to estimate variance components. In addition, this appendix provides syntax for estimating these models in MPlus and in the nlme package in R.

Design B

For Design B, the one-way random effects ANOVA, variance components were estimated using an empty model (Raudenbush & Bryk, 2002). The mixed equation can be expressed:

$$y_{ij} = \gamma_{00} + u_j + e_{ij}$$

y_{ij} here is taken as an individual level composite score. For the analyses in this study, y_{ij} was taken as either the average individual score across the 8 survey items for the WCS, or the average individual score across the 8 survey items for the Tripod. u_j is a school specific deviaition from the grand mean, γ_{00}, and e_{ij} is a person specific deviation. The variance of $u_j$represents the estimated school variance, σ_s^2, and the variance of $e_{ij}$is $\sigma_{p,ps,e}^2$, a residual variance that is composed of the variance attributable to teachers, a teacher-school interaction, and random error.

This design is implemented in R using the following syntax:

```
Library(nlme)
fit<-lme(y~1,random=~1|Group,data=data)
VarCorr(fit)
```

Alternately, this design can be implemented in MPlus:

```
TITLE: One-way Random-effects ANOVA in MPlus
DATA: File is data.dat;
VARIABLE : Names are yij;
ANALYSIS : Type is twolevel ; Estimator is ML;
MODEL:
%within%
yij;
%between%
yij;
OUTPUT: sampstat;
```

Design C

Design C incorporates variance components that result from a mixed-effects split-plot ANOVA. Estimation can be conducted using a multilevel factor analytic (MFA) framework. MFA relies on the covariance structure for parameter estimation. Muthen (1994) shows that for a two level model, an observed variable y_{ij} can be expressed as a linear function of a within level latent trait, η_w, and a between-level latent trait, η_B. ϵ_w and ϵ_B are residual variances that are independent and normally distributed. Given the factor model:

$$y_{ij} = \lambda_w \eta_w + \lambda_B \eta_B + \epsilon_w + \epsilon_B$$

The covariance structure of the observed variables can be expressed

$$\Sigma = \Lambda_{\mathrm{w}} \Phi_{\mathrm{w}} \Lambda'_{\mathrm{w}} + \Lambda_{\mathrm{B}} \Phi_{\mathrm{B}} \Lambda'_{\mathrm{B}} + \Theta_{\mathrm{w}} + \Theta_{\mathrm{B}}$$

In this study, y_{ij} is an $8 \times n_p$ matrix of individual item responses (for either the WCS or Tripod survey), Λ_{w} is an 8×1 vector of factor loadings for the individual level (student or teacher) latent variable (in this study, because of the assumption of tau-equivalence, all of the factor loadings are set to 1), and Λ_{B} is an 8×1 vector of between-level factor loadings (also set to 1) for the school or classroom level latent variable. Φ_{w} is a 1×1 matrix, containing the variance of the individual level latent trait, and Φ_{B} is a 1×1 matrix containing the variance of the between level latent trait.

Θ_{w} is an 8×8 diagonal matrix, containing individual level residual variances, and Θ_{B} is an 8×8 diagonal matrix, containing group level residual variances.

The variances contained in Φ_{w}, Φ_{B}, Θ_{w} and Θ_{B} can be used to estimate the variance components from Design C. Let:

$$\Phi_{\mathrm{w}} = \left|\sigma^2_{p,ps}\right|$$

$$\Phi_{\mathrm{B}} = |\sigma^2_{s}|$$

$$\Theta_{\mathrm{w}} = \begin{vmatrix} \sigma^2_{pi,psi,e} & & \\ & \ddots & \\ & & \sigma^2_{pi,psi,e} \end{vmatrix}$$

$$\Theta_{\mathrm{B}} = \begin{vmatrix} \sigma^2_{si} & & \\ & \ddots & \\ & & \sigma^2_{si} \end{vmatrix}$$

Following Raykov and Marcoulides (2011). σ_s^2 is the element of Φ_B, $\sigma_{p,ps}^2$ is the element of Φ_w, $\sigma_{pi,psi,e}^2$ is the average across the 8 elements of Θ_w, and σ_{si}^2 is the average across the 8 elements of Θ_B.

This design can be implemented in MPlus using the following syntax:

```
TITLE: Split-plot Mixed-effects ANOVA in MPlus
DATA: File is data.dat;
VARIABLE : Names are y1 y2 y3 y4 y5 y6 y7 y8;
CLUSTER : Group;
ANALYSIS : Type is twolevel ; Estimator is ML;
MODEL:
%within%
Factor_w by y1-y8@1;
%between%
Factor_b by y1-y8@1;
OUTPUT: sampstat;
```

Design E

For Design E, the variance components can be found using a three-level hierarchical linear model, expressed:

$$y_{ijk} = \gamma_{000} + r_k + u_{jk} + e_{ijk}$$

y_{ij} here is taken as an individual level item score. r_k is a school specific deviation from the grand mean, γ_{000}, u_{jk} is a person specific deviation from the grand mean, , and e_{ijk} is an item specific deviation. The variance of $r_k$represents the estimated school variance, σ_s^2, and the variance of $u_{jk}$is $\sigma_{p,ps}^2$ a conflated variance of teachers and a teacher school interaction. The variance of e_{ijk} is $\sigma_{i,pi,psi,si,e}^2$, a residual variance that is composed of the variance attributable to items, an item teacher interaction, a triple interaction of items, schools and teachers, an item-school interaction, and random error.

It is important to note that in order for this model to be implemented in R, the data file must have the appropriate structure. The file must be in so-called "long form", with each row representing a single item score, and a new variable for each Item ID, person ID, and group ID. This design can be implemented in R using the following syntax.

```
Library(nlme)
fit<- lme(y ~ 1, random = ~1 | group/person, data)
```

VarCorr(fit)

Appendix B

This appendix provides syntax (R and Mplus) for estimating contextual effects models for Designs A,B, and C. For Designs B and C, the contextual effects are corrected for measurement error.

Design A

R code:

```
library(nlme)
Null.Model<-lme(xij~1,random=~1|Group,data=Data)
VarCorr(Null.Model)
fit<-lme(yij~xij+xbar,random=~1|Group,data=Data)
```

In this code, the Null.Model is run to show estimated within and between variance components. VarCorr() outputs those estimates. The lme() command is used to estimate a contextual effects model, with an individual level predictor xij and a group mean predictor xbar. Please note that the coefficient on the xbar term here is the contextual effect, not the between-groups effect.

MPlus code:

```
TITLE: Contextual Effects model (Design A)
DATA: File is data.dat;
VARIABLE: Names are xij xbar yij;
cluster = Group;
within = xij ;
between = xbar;
centering = groupmean(xij);
ANALYSIS: Type is twolevel;
MODEL:
%within%
yij on xij
%between%
yij on xbar;
```

This Mplus code estimates two coefficients – the within-group slope and the between-group slope. Because the centering option is used, the contextual effect can be obtained as the difference of the two slopes.

Design B

R code:

```
library(nlme)
Null.Model<-lme(xij~1,random=~1|Group,data=Data)
VarCorr(Null.Model)
Data$EB<-predict(Null.Model)
fit<-lme(yij~xij+EB,random=~1|Group,data=Data)
```

In this code, the Null.Model is run to obtain Empirical Bayes means for each group. These means are saved into the data set and used as group level predictors. The lme() command is used to estimate a contextual effects model, with an individual level predictor xij and an adjusted group mean predictor EB. Please note that the coefficient on the EB term here is the contextual effect, not the between-groups effect.

MPlus code:

```
TITLE: Contextual Effects model (Design B)
DATA: File is data.dat;
VARIABLE: Names are xij yij;
cluster = Group;
ANALYSIS: Type is twolevel;
MODEL:
%within%
yij on xij
%between%
yij on xij;
```

This Mplus code estimates two coefficients – the within-group slope and the between-group slope. Because the centering option is used, the contextual effect can be obtained as the difference of the two slopes.

Design C

MPlus code:

```
TITLE: Contextual Effects model (Design C)
DATA: File is data.dat;
VARIABLE: Names are x1 x2 x3 x4 x5 x6 x7 x8 yij;
```

```
cluster = Group;
ANALYSIS: Type is twolevel;
MODEL:
%within%
fw by x1-x8@1;
yij on fw;
%between%
fb by x1-x8@1;
yij on fb;
```

This Mplus code estimates two coefficients – the within-group slope and the between-group slope. Because the centering option is used, the contextual effect can be obtained as the difference of the two slopes.

Appendix C

The following provides more detail on the Expected Mean Squares (EMS) basis for the variance components in Designs B, C, and E. It demonstrates why the estimated school variance, σ_s^2, and teacher within school variance, $\sigma_{p,ps}^2$, differ across designs. These EMS equations are found in Shavelson and Webb (1991).

Differences in $\boldsymbol{\sigma_{p,ps}^2}$ Across Designs

For Design C, the EMS equations for the variance components are

$$E(MS_s)^{(C)} = n_p n_i \sigma_s^{2(C)} + n_i \sigma_{p,sp}^{2(C)} + n_p \sigma_{si}^{2(C)} + \sigma_{pi,psi,e}^{2(C)} \quad (1)$$

$$E\left(MS_{p,ps}\right)^{(C)} = n_i \sigma_{p,ps}^{2(C)} + \sigma_{pi,psi,e}^{2(C)} \quad (2)$$

$$E(MS_{si})^{(C)} = n_i \sigma_{si}^{2(C)} + \sigma_{pi,psi,e}^{2(C)} \quad (3)$$

$$E\left(MS_{pi,psi,e}\right)^{(C)} = \sigma_{pi,psi,e}^{2(C)} \quad (4)$$

For Design E, the EMS equations for the variance components are

$$E(MS_s)^{(E)} = n_p n_i \sigma_s^{2(E)} + n_i \sigma_{p,ps}^{2(E)} + \sigma_{i,pi,si,psi,e}^{2(E)} \quad (5)$$

$$E\left(MS_{p,ps}\right)^{(E)} = n_i \sigma_{p,ps}^{2(E)} + \sigma_{i,pi,si,psi,e}^{2(E)} \quad (6)$$

$$E\left(MS_{i,pi,si,psi,e}\right)^{(E)} = \sigma_{i,pi,si,psi,e}^{2(E)} \quad (7)$$

Based on these EMS equations, it can be shown that the teacher variance, $\sigma_{p,ps}^2$ estimated in Design C will be smaller than the teacher variances estimated in Design E. Given that Equation (2) equals Equation (6)

$$n_i \sigma_{p,sp}^{2(C)} + \sigma_{pi,psi,e}^{2(C)} = n_i \sigma_{p,sp}^{2(E)} + \sigma_{i,pi,si,psi,e}^{2(E)} \quad (8)$$

So

$$\sigma_{p,ps}^{2(C)} - \frac{\sigma_{i,si}^2}{n_i} = \sigma_{p,ps}^{2(E)} \quad (9)$$

Which implies:

$$\sigma_{p,ps}^{2(E)} > \sigma_{p,ps}^{2(C)} \quad (10)$$

by the quantity $\frac{\sigma_{i,si}^2}{n_i}$.

Differences in σ_s^2 Across Designs

Given that Equation (1) equals Equation (5), we have:

$$n_p n_i \sigma_s^{2(C)} + n_i \sigma_{p,ps}^{2(C)} + n_p \sigma_{si}^{2(C)} + \sigma_{pi,psi,e}^{2(C)} = n_p n_i \sigma_s^{2(E)} + n_i \sigma_{p,ps}^{2(E)} + \sigma_{i,pi,si,psi,e}^{2(E)} \quad (11)$$

Which can be re-expressed:

$$n_p n_i \sigma_s^{2(3)} + n_i \sigma_{p,ps}^{2(3)} + n_p \sigma_{si}^{2(3)} = n_p n_i \sigma_s^{2(5)} + n_i \left(\sigma_{p,ps}^{2(5)} + \frac{\sigma_{i,si}^2}{n_i} \right) \quad (12)$$

And applying the equivalence in Equation (9), we have:

$$n_p n_i \sigma_s^{2(C)} + n_i \sigma_{p,ps}^{2(C)} + n_p \sigma_{si}^{2(C)} = n_p n_i \sigma_s^{2(E)} + n_i \sigma_{p,sp}^{2(C)} \quad (13)$$

Which implies:

$$\sigma_s^{2(C)} + \frac{\sigma_{si}^{2(C)}}{n_i} = \sigma_s^{2(E)} \quad (14)$$

By the quantity $\frac{\sigma_{si}^2}{n_i}$

Appendix D

The following provides more detail on the equivalence of the group-mean reliabilities for Design B and Design E. For Design B, the EMS equations for the variance components are :

$$E(MS_s)^{(B)} = n_p\sigma_s^{2(B)} + \sigma_{p,ps,e}^{2(B)} \quad (1)$$

$$E\left(MS_p\right)^{(B)} = \sigma_{p,ps,e}^{2(B)} \quad (2)$$

For Design E, the EMS equations for the variance components are :

$$E(MS_s)^{(E)} = n_p n_i\sigma_s^{2(E)} + n_i\sigma_{p,ps}^{2(E)} + \sigma_{i,pi,si,psi,e}^{2(E)} \quad (3)$$

$$E\left(MS_{p,ps}\right)^{(E)} = n_i\sigma_{p,ps}^{2(E)} + \sigma_{i,pi,si,psi,e}^{2(E)} \quad (4)$$

$$E\left(MS_{i,pi,si,psi,e}\right)^{(E)} = \sigma_{i,pi,si,psi,e}^{2(E)} \quad (5)$$

Results in Shavelson & Webb (1991) demonstrate that:

$$E(MS_s)^{(B)} = \frac{E(MS_s)^{(E)}}{n_i} \quad (6)$$

and:

$$E\left(MS_p\right)^{(B)} = \frac{E\left(MS_p\right)^{(E)}}{n_i} \quad (7)$$

Thus:

$$\sigma_{p,sp,e}^{2(B)} = \sigma_{p,sp}^{2(E)} + \frac{\sigma_{i,ps,si,psi,e}^{2(E)}}{n_i} \quad (8)$$

That is, the person-variance estimate in Design B is equal to the person variance estimate from Design E plus the residual variance averaged over the number of items.

By Equations (1), (3) and (6), we have:

$$n_p\sigma_s^{2(B)} + \sigma_{p,ps,e}^{2(B)} = n_p\sigma_s^{2(E)} + \sigma_{p,ps}^{2(E)} + \frac{\sigma_{i,pi,si,psi,e}^{2(E)}}{n_i} \tag{9}$$

By Equation (8) we can re-express Equation (9):

$$n_p\sigma_s^{2(B)} + \sigma_{p,sp,e}^{2(B)} = n_p\sigma_s^{2(E)} + \sigma_{p,sp,e}^{2(B)} \tag{10}$$

And so

$$\sigma_s^{2(B)} = \sigma_s^{2(E)} \tag{11}$$

Results from Brennan (2001a) shoe that the group-mean reliability coefficient for Design E can be expressed:

$$\lambda = \frac{\sigma_s^{2(E)}}{\sigma_s^{2(E)} + \frac{\sigma_{p,ps}^{2(E)}}{n_p} + \frac{\sigma_{i,pi,psi,si,e}^{2(E)}}{n_p n_i}} \tag{12}$$

By Equations (8) and (11), the reliability coefficient given in Equation (12) can be re-expressed:

$$\lambda = \frac{\sigma_s^{2(B)}}{\sigma_s^{2(B)} + \frac{\sigma_{p,ps,e}^{2(B)}}{n_p}} \tag{13}$$

Which is the group-mean reliability coefficient given by Brennan (2001a) for thr one-way random-effects ANOVA design given by Design B.