Journal of Research on Educational Effectiveness, 1: 138–154, 2008

ISSN: 1934-5747 print / 1934-5739 online
DOI: 10.1080/19345740801982104

Routledge
Taylor & Francis Group

METHODOLOGICAL STUDIES

Statistical Inference When Classroom Quality is Measured With Error

Stephen W. Raudenbush and Sally Sadoff
University of Chicago, Chicago, Illinois, USA

Abstract: A dramatic shift in research priorities has recently produced a large number of ambitious randomized trials in K-12 education. In most cases, the aim is to improve student academic learning by improving classroom instruction. Embedded in these studies are theories about how the quality of classroom must improve if these interventions are to succeed. The problem of measuring classroom quality then emerges as a major concern. This article first considers how errors of measurement reduce statistical power in studies of the impact of interventions classroom quality. We show how to use information about reliability to compute power and plan new research. At the same time, errors of measurement introduce bias into estimates of the association between classroom quality and student outcomes. We show how to use knowledge about the magnitude of measurement error to eliminate or reduce this bias. We also briefly review research on the design of studies of the reliability of classroom measures. Such studies are essential to evaluate promising new classroom interventions.

Keywords: Group-randomized experiments, hierarchical linear models, errors in variables

THE CENTRALITY OF VALID CLASSROOM ASSESSMENT IN EDUCATIONAL RESEARCH

The premise of this article is that the valid measurement of classroom process is essential to the advance of large-scale quantitative social science in education. Quantitative assessment of social interactions in classrooms has a long and distinguished history (cf. Brophy & Good, 1970). However, this tradition has had comparatively little influence on the design and conduct of recent large-scale randomized experiments in school settings, and this past work has not

Address correspondence to Stephen W. Raudenbush, Department of Sociology, 1126 East 59th Street, Chicago, IL 60637, USA. E-mail: sraudenb@uchicago.edu

significantly affected recent large-scale surveys of schools, classrooms, and student outcomes. There are exceptions of seminal importance, including the Study of Instructional Improvement (cf. Rowan, Camburn, & Correnti, 2004), several applications of the "CLASS" framework (Pianta, Hamre, Mashburn, & Downer, in press), and in the Third International Science and Mathematics Study (see Stigler, Gallimore, & Hiebert, 2000), and this article draws heavily on these studies.

Classroom Assessment in Intervention Research

A seismic shift in national research priorities over the past 6 years has led to a dramatic increase in the number of large-scale randomized experiments designed to test the impact of educational interventions on student outcomes. Spybrook (2007) identified 55 such trials supported by the Institute for Education Sciences. Of these, the vast majority assigned groups, typically schools or classrooms rather than individuals, to interventions. The majority of the innovative interventions attempted to improve student learning by improving classroom teaching.

A major aim of these studies is to evaluate the impact on student learning of assignment to an innovative classroom intervention. This aim can be achieved, in principle, without measuring the quality of classroom instruction. However, the interpretation of findings from such a study will typically be ambiguous.

Consider a study in which the assignment of schools or classrooms to a novel instructional innovation is found to have no significant impact on student learning. Assume that the study design was unbiased and provided adequate statistical power to detect a nonnegligible effect. Two explanations immediately arise. Program evaluators refer to these as "theory failure" versus "implementation failure" (Rossi, Lipsey, & Freeman, 2004).

First, it may be that the innovation changed classroom instruction in the ways intended but that those classroom changes made no difference in student learning. The term *theory failure* describes this scenario because the theory that links intended changes in instruction to intended student outcomes will have proven incorrect.

Second, the innovation may never have been effectively implemented in classrooms. Perhaps the innovators lacked skill in working with teachers or perhaps the teachers lacked the skill, knowledge, or motivation to put the innovative ideas to work in their teaching. In any case, program theory about the relationship between the intended instruction and student outcomes was never tested, leading to "implementation failure."

Without valid assessments of instructional process, it would be impossible to distinguish between these two explanations, severely limiting the study's contribution to knowledge. One would never know whether the theory underlying the program had in fact been tested.

Suppose instead that assignment to the innovation did produce gains in student learning. One might then assume that the innovation "worked" by

improving instruction in the ways the program designers intended. But without valid measurements of instruction, this conclusion would be unwarranted. Perhaps the innovation "worked" in other ways, an assertion that could not be probed without studying the impact of the innovation on instruction. Once again, a failure to measure key aspects of classroom life yields major ambiguities in the findings.

The task of measuring of classroom processes challenges researchers to make precise their theory about how and when an innovation comes to be successful. More broadly, if we are to develop a science of the links between school organization, the work of teaching, and children's cognitive development, tools for measuring classroom qualities and processes become essential.

Classroom Assessment in Large-Scale Surveys

Let us now fast-forward into a world where much is known about which innovations improve classroom life and which aspects of classroom life are crucial to student development. A question of obvious importance then involves access: What is the distribution of classroom quality so defined? To what extent do students who vary by demography and geography have access to these educational opportunities? Answering these questions requires the assessment of classroom quality in large-scale surveys. To disregard such an aim would be equivalent in medical research to disregarding the importance of knowing which heart patients have access to bypass surgery or which kidney patients have access to dialysis.

However, we need not wait for such an accretion of knowledge about "what works" in classrooms to justify the large-scale assessment of classroom process. The accumulation of knowledge about classroom life and student learning is gradual, and experience shows that when investigators take care in assessing classroom process, great opportunities arise for the discovery of new ideas about how to improve education. Perhaps the Third International Science and Mathematics Study is the most convincing case in point: Mathematics instruction in countries whose students excel in math looks very different from the instruction observed in countries where children fare less well (Stigler et al., 2000). These findings are important not only for generating explanations for cross-national differences in mathematics achievement but also for the design of new instructional interventions within countries, and, in particular, within the United States, whose students score disappointingly low in math.

The Role of Classroom Measures in Studies of School Improvement

Consider a study of a new innovation designed to improve instruction and thereby to improve student learning. For simplicity, let us denote Z as indicating assignment to the innovation. That is, $Z = 1$ if a class participates in the innovation and $Z = 0$ if not. Of obvious interest is the association between

Z and student outcome variable Y. However, for reasons just described, the researchers are also interested in (a) whether Z affects classroom instruction, here denoted Q, and (b) the association between instruction received, Q, and student outcome Y. We are interested in three questions:

1. *The association between Z and Y*. This is a standard three-level analysis (students within classrooms within schools, classrooms being the unit of treatment) and is described in detail in books on hierarchical linear models (cf. Goldstein, 2003; Raudenbush & Bryk, 2002). Fully documented software for planning such studies is available on the Web site of the W.T. Grant Foundation (http://wtgrantfoundation.org). We do not consider these models further in this article.
2. *The association between Z and Q*. This appears to be a standard two-level model with Z and Q both defined on classrooms nested within schools. However, we postulate that Q is measured with error and investigate the implications of these errors of measurement for studies of the impact of an innovation Z on a classroom process, Q. An extra variance component is therefore added, making the model essentially a three-level model (errors of measurement within classrooms within schools).
3. *The association between Q and Y*. This is a three-level model with Y at Level 1 and Q at Level 2 (between classrooms). We are interested in the implications of the fact that Q is measured with error for inferences regarding association between classroom process and Y.

The discussion here is relevant not only to studies of the impact of innovations but also to surveys. Surveys enable study of the distribution of access to high-quality classrooms. In these studies, the explanatory variables are characteristics of schools and students and classroom quality, Q, is an outcome. Such surveys also enable specification of models in which Q is an explanatory variable with Y is a student outcome.

The remainder of this article is organized as follows. The next section considers the implications of measurement error of classroom process Q for the design of studies that use Z as an explanatory variable and Q as an outcome. The following section considers the implications of the measurement error for studies of the effect of Q on Y. The last section briefly reviews the design of studies that aim to quantify sources of error in measuring Q and considers implications for further research.

STUDIES OF THE IMPACT OF INNOVATIONS ON CLASSROOM PROCESS

We now consider the case in which the aim is to study the impact of receiving a classroom-level intervention on classroom quality, Q. However, Q is measured with error. The key result is that although these errors of measurement do not

cause bias, they do reduce precision and power. As a result, unreliability of classroom measurement requires sampling a larger number of classrooms or schools than would otherwise be required to obtain a desired level of statistical power.

Spybrook (2007) identified two experimental designs most commonly used in studies of interventions designed to improve learning. Design 1 is a school randomized trial: Whole schools are assigned at random; teachers and children in the experimental schools participate in the intervention. In contrast, in Design 2, schools are first sampled and then classrooms within schools are assigned at random to treatments. We consider these two designs separately because the reliability of classroom measurement affects statistical power differently in these two cases.

Statistical Inference for Design 1: Randomization at the School Level

Let $Z_k = 1$ if school k is assigned to receive the experimental innovation and $Z_k = 0$ if school k is assigned to the control condition, for schools $k = 1, \ldots, K$. Let Q_{jk} denote the "true" quality of classroom j in school k on a dimension of classroom quality for classrooms $j = 1, \ldots, J_k$. This is a two-level design with classrooms nested within schools at Level 1 and schools varying at Level 2.

Model. A simple Level 1 model represents variation between classrooms within schools:

$$Q_{jk} = \theta_{0k} + c_{jk}, \quad c_{jk} \sim N\left(0, \tau_c^2\right). \tag{1}$$

Here θ_{0k} is the mean classroom quality in school k, c_{jk} is a classroom random effect and τ_c^2 is the variance between classrooms within schools on quality. This variance is assumed for simplicity to be constant across schools, though this assumption can be relaxed. The classroom random effects are mutually independent.

At Level 2, the mean classroom quality varies across schools partly as a function of treatment assignment:

$$\theta_{0k} = \alpha_{00} + \alpha_{01} Z_k + s_{0k}, \quad s_{0k} \sim N\left(0, \omega_{s0}^2\right). \tag{2}$$

Here α_{00} is the average classroom quality in the control group; α_{01} is the average causal effect on quality of being assigned to the experimental innovation, and s_{0k} is a school-level random effect having between-school variance ω_{s0}^2. The school-level random effects are assumed independent of each other and of the classroom random effects. Substitution of Equation 2 into Equation 1 yields the combined model

$$Q_{jk} = \alpha_{00} + \alpha_{01} Z_k + s_{0k} + c_{jk}. \tag{3}$$

Estimation and Hypothesis Testing: No Measurement Error. Suppose now that Q_{jk} could be observed. Estimation of Equation 3 would provide an unbiased estimate of the treatment effect α_{01} if schools were randomly assigned to treatments and no attrition emerged. The analyst may wish to add covariates to Equation 1 or 2, but such an addition would have no consequences for the principles we are developing here. In particular, suppose that the researcher is planning a balanced design, with $K/2$ schools in each treatment condition and J classrooms per school. Then the minimum variance unbiased estimate would be the simple difference between the arithmetic means of the outcome in the experimental and control groups respectively, that is

$$\hat{\alpha}_{01} = \bar{Q}_E - \bar{Q}_C, \tag{4}$$

where

$$\bar{Q}_E = \sum_{k=1}^{K/2} \sum_{j=1}^{J} Q_{jk}/(KJ/2)$$

and

$$\bar{Q}_C = \sum_{k=K/2+1}^{K} \sum_{j=1}^{J} Q_{jk}/(KJ/2)$$

where the data have been organized such that the students are clustered within schools and then sorted so that the first $K/2$ schools are the experimental schools. The variance of this estimator (Raudenbush, 1997) is

$$Var(\hat{\alpha}_{01}) = 4\left[\omega_{s0}^2 + \tau_c^2/J\right]/K \tag{5}$$

The null hypothesis $H_0 : \alpha_{01} = 0$ can be tested using a central F statistic with degrees of freedom 1, K-2. Under the alternative hypothesis $H_a : \alpha_{01} = \alpha_{01}^* > 0$, the computed F ratio will be distributed as a noncentral F with degrees of freedom 2, K-2 and noncentrality parameter

$$\psi = \frac{\alpha_{01}^{*2}}{Var(\hat{\alpha}_{01})} = \frac{K\alpha_{01}^{*2}}{4[\omega_{s0}^2 + \tau_c^2/J]}. \tag{6}$$

Power increases with ψ: as the number of schools, K, and the squared effect magnitude α_{01}^{*2} increase, so does power. Increasing the number of classrooms per school, J, helps, but this benefit diminishes to zero as J increases unless the between-school variance, ω_{s0}^2, is null.

It is often convenient to plan research based on prior beliefs about standardized effect sizes. Past research may give some guidance on likely effect sizes in standardized units when no information is available about

scale-specific effects. Define the standardized effect size $\delta = \alpha^*_{01}/\sqrt{\omega^2_{0s} + \tau^2_c}$ and the intraschool correlation (the fraction of variation that lies between schools) as $\rho = \omega^2_{0s}/(\omega^2_{0s} + \tau^2_c)$. Therefore $\alpha^{*2}_{01} = \delta^2(\omega^2_{0s} + \tau^2_c)$ and $\omega^2_{0s} + \tau^2_c = \omega^2_{0s}/\rho$. Making these substitutions into Equation 6 yields the equivalent expression

$$\psi = \frac{K\delta^2}{4[\rho + (1-\rho)/J]}. \tag{7}$$

Measurement Error Model, Statistical Inference, and Power. We now consider the effect of measurement error on statistical inference and power. Suppose that we do not observe the true classroom quality but rather a fallible indicator W_{jk} following the model

$$W_{jk} = Q_{jk} + e_{jk}, \qquad e_{jk} \sim N\big(0, \sigma^2_e\big) \tag{8}$$

where the measurement errors $e_{jk}$are mutually independent of each other and of all other random effects. Now the minimum variance unbiased estimate is the mean difference

$$\hat{\alpha}_{01} = \bar{W}_E - \bar{W}_C, \tag{9}$$

with sampling variance

$$Var(\hat{\alpha}_{01}) = 4\big[\omega^2_{s0} + \big(\tau^2_c + \sigma^2_e\big)/J\big]/K. \tag{10}$$

As before, the null hypothesis $H_0 : \alpha_{01} = 0$ can be tested using a central F statistic with degrees of freedom 1, K-2. Under the alternative hypothesis $H_a : \alpha_{01} = \alpha^*_{01} > 0$, the computed F ratio will be distributed as a noncentral F with degrees of freedom 1, K-2 and noncentrality parameter

$$\psi = \frac{\alpha^{*2}_{01}}{Var(\hat{\alpha}_{01})} = \frac{K\alpha^{*2}_{01}}{4[\omega^2_{s0} + (\tau^2_c + \sigma^2_e)/J]}. \tag{11}$$

We now define the reliability of measurement as

$$\lambda = \tau^2_s/\big(\tau^2_c + \sigma^2_e\big), \tag{12}$$

the usual ratio of "true score variance" to observed covariance. This reliability is equivalent to the correlation between two realizations of W based on the same measurement procedures as before but using randomly different raters, items, occasions, and so forth, that generate random errors of measurement (see the last section in this article). Substituting $\tau^2_c + \sigma^2_e = \tau^2_c/\lambda$ in Equation 11 and standardizing as before so that $\delta = \alpha^*_{01}/\sqrt{\omega^2_{0s} + \tau^2_c}$ and $\rho = \omega^2_{0s}/(\omega^2_{0s} + \tau^2_c)$,

we now obtain the standardized noncentrality parameter

$$\psi = \frac{K\delta^2}{4[\rho + (1-\rho)/(J\lambda)]}. \tag{13}$$

Inspection of Equation 13 reveals that setting $\lambda = 1$ yields the noncentrality parameter we obtained when no measurement error was present (see Equation 7). In effect, Equation 13 tells us that unreliability reduces the effective sample size of classrooms per school. For example, Equation 2.13 with $\lambda = .5$ is equivalent to Equation 7 with J reduced by half.

Statistical Inference for Design 2 (Classrooms Randomized Within Schools)

We now consider the case where K schools are again sampled. However, rather than assigning these schools to the experimental innovation, we instead sample J classrooms within those schools and assign those classrooms at random to the experimental innovation or to a control. Let $Z_{jk} = 1$ is classroom j within school k is assigned to receive the experimental innovation and $Z_{jk} = 0$ if that classroom is assigned to the control condition. As before, let Q_{jk} denote the "true" quality of classroom j in school k on a dimension of classroom quality for classrooms $j = 1, \ldots, J_k$. This is again a two-level design with classrooms nested within schools at Level 1 and schools varying at Level 2. However, now the treatment indicator is at Level 1 rather than Level 2.

Model. A simple level-1 model is then

$$Q_{jk} = \theta_{0k} + \theta_{1k} Z_{jk} + c_{jk}, \quad c_{jk} \sim N\left(0, \tau_c^2\right). \tag{14}$$

Here θ_{0k} is the mean classroom quality for the control classrooms in school k, θ_{1k} is the average causal effect of the experimental innovation on classroom quality in school k, c_{jk} is a classroom random effect, and τ_c^2 is the variance between classrooms (within treatments) within schools on quality. As before, the classroom random effects are mutually independent. The key difference between Designs 1 and 2 is that, because the causal variable varies within schools in Design 2, it is now possible to identify a school-specific treatment effect, θ_{1k}. This effect may be modeled as varying randomly over schools.

Therefore, at Level 2, both θ_{0k} and θ_{1k} may vary from school to school:

$$\begin{array}{l} \theta_{0k} = \alpha_{00} + s_{0k}, \\ \theta_{1k} = \alpha_{10} + s_{1k}, \end{array} \begin{pmatrix} s_{0k} \\ s_{1k} \end{pmatrix} \sim N \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \omega_{s0}^2 & \omega_{s01} \\ \omega_{s10} & \omega_{s1}^2 \end{pmatrix} \right] \tag{15}$$

Here α_{00} is the average level of quality in the control group; α_{01} is the average causal effect of being assigned to the experimental innovation; and s_{0k}, s_{1k}

are school-level random effects having between-school variances ω_{s0}^2, ω_{s1}^2 respectively and covariance $\omega_{s01} = \omega_{s10}$. The pair of school-level random effects, while correlated with each other, are assumed independent of the pairs of random effects associated with other schools and also of the classroom random effects.

Substitution of Equation 15 into 14 yields the combined model

$$Q_{jk} = \alpha_{00} + \alpha_{01} Z_{jk} + s_{0k} + s_{1k} Z_{jk} + c_{jk}. \tag{16}$$

Estimation and Hypothesis Testing: No Measurement Error. Suppose for now that Q_{jk} could be observed. Estimation of Equation 16 would provide an unbiased estimate of the treatment effect α_{01} if schools were randomly assigned to treatments, no attrition emerged, and no spillover effects between classrooms were present. Suppose that the researcher is planning a balanced design, with $K/2$ schools in each treatment condition and J classrooms per school. Then the minimum variance unbiased estimate would be the simple difference between the arithmetic means of the outcome in the experimental and control groups respectively, that is,

$$\hat{\alpha}_{01} = \bar{Q}_E - \bar{Q}_C, \tag{17}$$

where

$$\bar{Q}_E = \sum_{k=1}^{K} \sum_{j=1}^{J/2} Q_{jk}/(KJ/2)$$

and

$$\bar{Q}_C = \sum_{k=1}^{K} \sum_{j=J/2+1}^{J} Q_{jk}/(KJ/2)$$

where the data have been sorted by school; within schools, the data are sorted such that the first $J/2$ classrooms are the experimental classrooms. The variance of this estimator (Raudenbush & Liu, 2000) is

$$Var(\hat{\alpha}_{01}) = \left[\omega_{s1}^2 + 4\tau_c^2/J\right]/K, \tag{18}$$

and the null hypothesis $H_0 : \alpha_{01} = 0$ can be tested using a central F statistic with degrees of freedom 1, K-1. Notice that the between-school variance component ω_{s0}^2 does not influence Equation 18: One of the key advantages of randomizing within schools is that school-level variance in the mean outcome is removed from the experimental error variance. Under the alternative hypothesis

$H_a : \alpha_{01} = \alpha_{01}^* > 0$, the computed F ratio will be distributed as a noncentral F with degrees of freedom 1, K-1 and noncentrality parameter

$$\psi = \frac{\alpha_{01}^{*2}}{Var(\hat{\alpha}_{01})} = \frac{K\alpha_{01}^{*2}}{\omega_{s1}^2 + 4\tau_c^2/J}. \tag{19}$$

A key factor that can undermine power is ω_{s1}^2, which measures the heterogeneity of the treatment effect across schools. If this heterogeneity is large, increasing the total number of schools, K, contributes greatly to power. If the heterogeneity is null, increasing the number of classrooms, J, is as helpful as is increasing K. If schools are regarded as fixed rather than random, the F ratio under the null hypothesis takes on a noncentral F distribution with 1, $K(J$-2) degrees of freedom and noncentrality parameter Equation 19 with $\omega_{s1}^2 = 0$. The main effect of treatment in the fixed effects model becomes difficult to interpret, however, if the treatment effect is heterogeneous across schools.

It may again be convenient to work with standardized effect sizes. We now define the standardized effect size somewhat differently than in the case of Design 1. Specifically, we set $\delta = \alpha_{01}^*/\tau_c$, the ratio of the true effect size to the true standard deviation of the outcome within schools. We also standardize the heterogeneity of the treatment effect variance such that $\sigma_\delta^2 = \omega_{s1}^2/\tau_c^2$. With these definitions in mind, the noncentrality parameter (Equation 19) is equivalent to

$$\psi = \frac{K\delta^2}{\sigma_\delta^2 + 4/J}. \tag{20}$$

Measurement Error Model, Statistical Inference, and Power. To assess how reliability of measurement affects power, we apply the same measurement model as in Design 1, namely, Equation 8, and for the balanced case ($J/2$ classrooms per treatment in each school k) apply the minimum variance unbiased estimate $\hat{\alpha}_{01} = \bar{W}_E - \bar{W}_C$, with sampling variance

$$Var(\hat{\alpha}_{01}) = \left[\omega_{s01}^2 + 4\left(\tau_c^2 + \sigma_e^2\right)/J\right]/K. \tag{21}$$

Using our previous definition of the reliability, that is, $\lambda = \tau_c^2/(\tau_c^2 + \sigma_e^2)$, our noncentrality parameter now becomes

$$\psi = \frac{K\delta^2}{\sigma_\delta^2 + 4/(J\lambda)}. \tag{22}$$

Inspection of Equation 22 again reveals that setting $\lambda = 1$ yields the noncentrality parameter we obtained when no measurement error was present (see

Equation 20). As in Design 1, Equation 22 tells us that unreliability reduces the effective sample size of classrooms per school.

Illustrative Example. The consequences of reliability for statistical power are graphed in Figure 1. The example is based on data collected by Mashburn et al. (in press). In that data set, 82 classrooms were nested within 18 schools. About 18% of the variance in outcomes lay between schools. We assume that the heterogeneity of the effect size is about .05. Power is plotted as a function of standardized effect sizes of 0.50, 0.75, and 1.00. Reliability makes a fairly substantial contribution to power. Thus, for an effect size of 0.75, the figure indicates that power would be about .60 if the reliability were $\lambda = .50$, whereas power would be about .80 for $\lambda = .80$.

STUDIES OF THE IMPACT OF CLASSROOM PROCESS ON STUDENT OUTCOMES

In this section, we consider the case in which the aim is to assess the contribution of classroom quality Q to student learning Y. As before, classroom quality is measured with error. The key result in this case is that measurement error creates bias. Specifically, if the effect of classroom quality on student

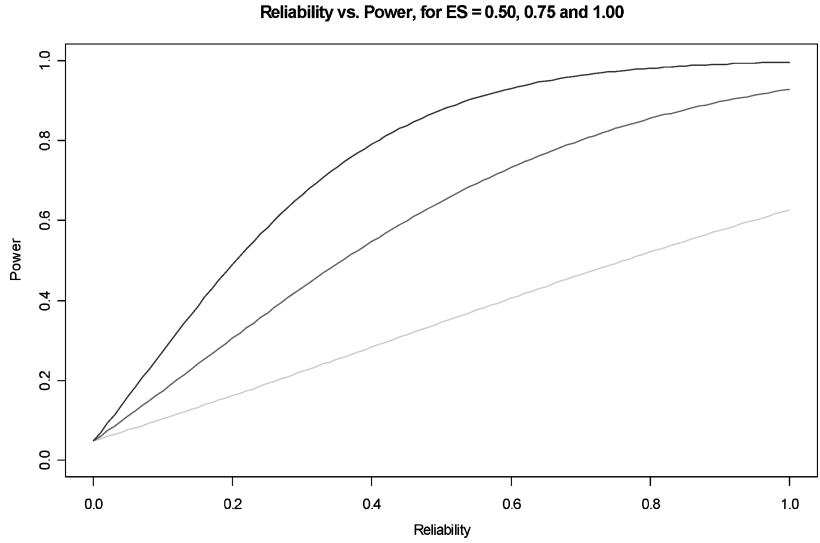

Figure 1. Random-effects model assumed; harmonic mean of classrooms per school (J) is 4.56; schools (K) are 18; effect size variability is 0.05, variance because of blocking is 0.18 of total variance across schools.

outcomes is positive, the estimate based on a fallible measure of classroom quality will be negatively biased. Although unreliability also reduces power, thereby increasing the need to sample more classrooms or schools, the problem of bias is now central. Increasing the sample size will not reduce the bias.

One of the benefits of conducting a study of the reliability of classroom measurement is that information from such a study can be used to correct the bias that arises from measurement error. We now consider how the bias can be so corrected.

Bias Arising From Measurement Error

As before, Q_{jk} denotes the "true" classroom quality on some dimension of interest. We are now interested in using Q_{jk} to predict student outcome Y_{ijk}, where students are indexed by $i = 1, \ldots, n_{jk}$. In this setting it will typically be required to control for one or more covariates X_{jk} because classrooms will not be randomly assigned to levels of quality Q_{jk}. It is common practice is to incorporate as covariates those prior characteristics of teachers, schools, and students that predict classroom quality and are related to Y_{ijk}. More sophisticated and useful versions of this strategy seek to compare "high" Q_{jk} classrooms in the experimental condition to observationally similar classrooms in the control condition (Peck, 2003). Causal inference proceeds cautiously under the assumption that the association between Q_{jk} and unobserved covariates is null conditional on observed covariates. This is the assumption of ignorable assignment of Q_{jk} given the observed covariates.

If Q_{jk} were observed, we could estimate the three-level model that would identify the impact of Q_{jk} on Y_{ijk} given the observed covariates. For simplicity, we consider a single covariate, X_{jk} . At Level 1 (between children within classrooms), we have

$$Y_{ijk} = \pi_{0jk} + \varepsilon_{ijk}, \qquad \varepsilon_{ijk} \sim N\left(0, \sigma_{\varepsilon}^2\right), \tag{23}$$

where $\pi_{0jk}$is the mean outcome in class j of school k and ε_{ijk} is a child-level random effect independently and identically distributed and independent of all other random effects. At Level 2 (between classrooms within schools) we model the classroom mean as a function of Q_{jk} and X_{jk}:

$$\pi_{0jk} = \beta_{00k} + \beta_{01k} Q_{jk} + \beta_{02k} X_{jk} + r_{0jk}, \qquad r_{0jk} \sim N\left(0, \tau_r^2\right). \tag{24}$$

Here β_{00k} is a school-specific random intercept and β_{01k} is a random coefficient representing the unique partial effect of Q_{jk} on π_{0jk}; β_{02k} represents the partial association between X_{jk} and π_{0jk} within school k. The regression

coefficients jointly vary over schools according to the Level 3 model

$$\begin{aligned}\beta_{00k} &= \gamma_{000} + u_{00k} \\ \beta_{01k} &= \gamma_{010} + u_{01k} \\ \beta_{02k} &= \gamma_{020}.\end{aligned} \quad \begin{pmatrix} u_{00k} \\ u_{01k} \end{pmatrix} \sim N\left[\begin{pmatrix} u_{00k} \\ u_{01k} \end{pmatrix}, \begin{pmatrix} \omega_{u00} & \omega_{u01} \\ \omega_{u10} & \omega_{u11} \end{pmatrix} \right] \tag{25}$$

In this model we have constrained $\beta_{02k}$to be invariant over schools. Substituting Equation 25 into Equation 24 and Equation 24 into Equation 23, we have the combined or "mixed" model

$$Y_{ijk} = \gamma_{000} + \gamma_{010} Q_{jk} + \gamma_{020} X_{jk} + r_{0jk} + u_{00k} + u_{01k} Q_{jk} + \varepsilon_{ijk}, \tag{26}$$

Inference would be straightforward if we could directly observe Q_{jk} . However, as in earlier sections, we do not observe Q_{jk} but rather the fallible variable W_{jk} where

$$W_{jk} = Q_{jk} + e_{jk}, \quad e_{jk} \sim N\left(0, \sigma_e^2\right). \tag{27}$$

Complicating matters a bit is that Q_{jk} is related to X_{jk} (recall that we included X_{jk} to control for confounding). We therefore have

$$Q_{jk} = \alpha_0 + \alpha_1 X_{jk} + s_{0k} + c_{jk}, \qquad s_{0k} \sim N\left(0, \omega_{s0}^2\right); \qquad c_{jk} \sim N\left(0, \tau_c^2\right). \tag{28}$$

If we were to estimate a regression of the same form as Equation 26 but now using the fallible W_{jk} as a predictor in place of the unknown Q_{jk}, the estimates of γ_{010}, γ_{020} would be biased.

Correcting for the Bias

Although the bias in estimating the effect of classroom quality on student outcomes can be severe, obtaining a good estimate of the reliability $\lambda = \tau_c^2/(\tau_c^2 + \sigma_\varepsilon^2)$ can be most helpful. To see why, let us consider conditional expectation of Q_{jk} given W_{jk}, X_{jk}, and s_{0k}, the school-level random component of Q_{jk} that is

$$\hat{Q}_{jk} = [E(Q_{jk}|W_{jk}, X_{jk}, s_{0k})] = \lambda W_{jk} + (1 - \lambda)(\alpha_0 + \alpha_1 X_{jk} + s_{0k}). \tag{29}$$

Taking a second expectation gives us

$$\begin{aligned} Q_{jk}^* &= E(Q_{jk}|W_{jk}, X_{jk}) = E(\hat{Q}_{jk}|W_{jk}, X_{jk}) \\ &= \lambda W_{jk} + (1 - \lambda)(\alpha_0 + \alpha_1 X_{jk} + s_{0k}^*) \end{aligned} \tag{30}$$

where

$$s_{0k}^* = E(s_{0k}|W_{jk}, X_{jk}) = \lambda_{2k}(\bar{W}_{.k} - \alpha_0 - \alpha_1 \bar{X}_k)$$
$$\text{with } \lambda_{2k} = \frac{\omega_{0s}^2}{\omega_{0s}^2 + \tau_c^2/(J_k\lambda)}. \tag{31}$$

Here

$$\bar{W}_{.k} = \sum_{j=1}^{J_k} W_{jk}/J_k,$$

and

$$\bar{X}_{.k} = \sum_{j=1}^{J_k} X_{jk}/J_k.$$

The conditional expectation Q_{jk}^* is the empirical Bayes posterior mean of the latent variable Q_{jk} given the observed data and is available in the empirical Bayes residual file output by now-standard software for hierarchical linear models (Raudenbush & Bryk, 2002, Equations 8.29, 8.30).

The important point is that substitution of a consistent estimate of Q_{jk}^* for the unknown Q_{jk} in Equation 26 will eliminate the large-sample bias in the estimation of the regression coefficients γ_{010}, γ_{020} under standard assumptions that we now make clear. Let us now examine the conditional expectation of Y given the observed covariate X and the fallible indicator W. The first step is to condition also on the random effects u_{0k}, u_{1k} :

$$\begin{aligned} &E(Y_{ijk}|X_{jk}, W_{jk}, u_{00k}, u_{01k}) \\ &= \gamma_{000} + \gamma_{010}Q_{jk}^* + \gamma_{020}X_{jk} + E(r_{0jk}|X_{jk}, W_{jk}, u_{00k}, u_{01k}) \\ &+u_{00k} + u_{01k}Q_{jk}^* + E(\varepsilon_{ijk}|X_{jk}, W_{jk}, u_{00k}, u_{01k}). \end{aligned} \tag{32}$$

Under our assumptions that the Level 1 and Level 2 random effects are conditionally independent of the level three random effects and of X and Q,

$$\begin{aligned} E(r_{0jk}|X_{jk}, W_{jk}, u_{00k}, u_{01k}) &= E(r_{0jk}) = E(\varepsilon_{ijk}|X_{jk}, W_{jk}, u_{00k}, u_{01k}) \\ &= E(\varepsilon_{ijk}) = 0. \end{aligned}$$

Therefore,

$$E(Y_{ijk}|X_{jk}, W_{jk}, u_{00k}, u_{01k}) = \gamma_{000} + \gamma_{010}Q_{jk}^* + \gamma_{020}X_{jk} + u_{00k} + u_{01k}Q_{jk}^*. \tag{33}$$

Our second step is to take the expectation over the distribution of the Level 3 random effects:

$$E(Y_{ijk}|X_{jk}, W_{jk}) = \gamma_{000} + \gamma_{010}Q^*_{jk} + \gamma_{020}X_{jk} + E(u_{00k}|X_{jk}, W_{jk}) + E(u_{01k}|X_{jk}, W_{jk})Q^*_{jk}.$$

Once again, we assume that the random effects are independent of the X and W, in which case $E(u_{00k}|X_{jk}, W_{jk}) = E(u_{00k}) = E(u_{01k}|X_{jk}, W_{jk}) = E(u_{01k}) = 0$. Therefore, we have

$$E(Y_{ijk}|X_{jk}, W_{jk}) = \gamma_{000} + \gamma_{010}Q^*_{jk} + \gamma_{020}X_{jk}. \quad (34)$$

Thus, substitution of a consistent estimate of Q^*_{jk}will eliminate the large-sample bias associated with the error with which W measures Q. The assumptions are standard ones: that treatment assignment is ignorable, meaning that Q is conditionally independent of unobservables (the random effects) given the observable X, that the measurement errors are also ignorable, and that the linearity of Y in X and Q holds.

DESIGNING STUDIES OF QUANTIFY ERRORS IN MEASURES OF CLASSROOM PROCESS

As shown in the previous section, knowing the reliability λ is extremely useful in removing bias associated with measurement error when classroom quality is a predictor of student outcomes. The second section had revealed the utility of knowing λ in studies of the impact of innovations on classroom quality. Specifically, λ is a crucial input in determining the power of a study to detect the impact of an innovation on classroom quality.

The question thus arises: How does one design a study to reveal the reliability of measurement of classroom quality? This is a topic taken up in detail in Raudenbush, Martinez, Bloom, Zhu, and Lin (2007). Their key idea is that one must first conceptualize the primary sources of error in measuring quality and then design a study in which the importance of these sources can be separately identified. Raudenbush and Sampson (1999) applied this approach to measuring neighborhoods and described it as the science of measuring ecological settings or "ecometrics" as distinct from psychometrics, though both use similar statistical tools. In particular, generalizability analysis as developed by Cronbach and Gleser (1965) is particularly useful.

To illustrate, Raudenbush, Martinez, Bloom, Zhu, and Lin (2007) described data collected at the National Center for Early Development and Learning (see Mashburn et al., in press). In this study, raters coded aspects of quality of classrooms on multiple days. Within each day, classrooms were observed on multiple 20-min segments. Random variation associated with raters, days, and

segments combined additively and in interaction to generate a complex error of measurement. The authors showed how increasing the number of raters or days or segments per day would plausibly affect reliability given estimation of the magnitude of the components of error variance generated by these sources. By increasing reliability, we have seen that we can increase the power of a study designed to assess the impact of a new intervention Z on classroom quality Q. We have also seen that a good estimate of reliability is extremely useful in reducing bias in studies that aim to assess the impact of classroom quality Q on student outcomes Y.

Studies of sources of measurement error can also reveal potential biases that can arise even in randomized experiments. Suppose that some raters are more likely than others to be assigned to observed classrooms randomly assigned to the experimental treatment. This unbalanced assignment might produce a substantial bias if rater effects are large. The same problem can arise in assigning days of the week or month to classrooms in the experimental versus control condition. Although it may not be possible to balance all sources of error by experimental condition, a measurement study can reveal the most important sources of error, enabling the researcher to focus on balancing these sources, thereby removing the most salient sources of bias.

This discussion leaves many questions unanswered. Consider, for example, studies of validity that postulate a confirmatory factor structure among items. Use of factor analysis generally entails the assumption that measurement errors are independent. However, the fact that multiple items are assessed contemporaneously by the same rater implies that the errors of measurement will in fact be correlated. So factor analyses must remove these rater and temporal effects if the model assumptions hold.

We have considered rather simple cross-sectional designs. In fact, the knowledge we seek to produce in children will result from *sequences* of instruction across grades. For example, instruction in word decoding and complex oral language in grades K-1 followed by instruction in reading comprehension and writing in Grades 2 and 3 may be essential to produce a high level of reading comprehension by the end of Grade 3. Such sequential treatments pose problems of causal inference that do not arise in cross-sectional settings. These difficulties combine with problems of measurement error to pose significant methodological challenges that we cannot discuss in this article but that are the subjects of ongoing research.

ACKNOWLEDGMENTS

The original draft of this article was prepared for the meeting "Approaches to Assessing Classroom Quality," jointly sponsored by the Spencer Foundation and the W.T. Grant Foundation on February 21, 2007, in Chicago. The work reported here was supported by funds from the W.T. Grant Foundation in support of the grant "Building Capacity for Group-Randomized Experiments"

and by funds from the Spencer Foundation for the project "Improving Research on Instruction: Models, Designs, and Analytic Methods."

REFERENCES

Brophy, J. E., & Good, T. L. (1970). Teachers' communications of differential expectations for children's classroom performance: Some behavioral data. *Journal of Educational Psychology, 61*, 365–374.

Cronbach, L. J., & Gleser, G. C. (1965). *Psychological tests and personnel decisions*. Urbana: University of Illinois Press.

Goldstein, H. (2003). *Multilevel Statistical Models (3rd ed.)*. Edward Arnold.

Mashburn, A. J., Pianta, R. C., Hamre, B. K., Downer, J. T., Barbarin, O., Bryant, D., et al. (in press). Pre-K program standards and children's development of academic, language and social skills. *Child Development*.

Peck, L. (2003). Subgroup analysis in social experiments: Measuring program impacts based on post-treatment choice. *American Journal of Evaluation, 24*(2), 157–187.

Raudenbush, S. W. (1997). Statistical analysis and optimal design for cluster randomized trials. *Psychological Methods, 2*(2), 173–185.

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

Raudenbush, S. W., & Liu, X. (2000). Statistical power and optimal design for multisite randomized trials. *Psychological Methods*, 5(2), 199–213.

Raudenbush, S. W., Martinez, A., Bloom, H., Zhu, P., & Lin, F., (2007). *The reliability of group-level measures and the power of group-randomized studies*. Unpublished manuscript, University of Chicago Department of Sociology.

Raudenbush, S. W., & Sampson, R. (1999). Ecometrics: Toward a science of assessing ecological settings, with application to the systematic social observations of neighborhoods. *Sociological Methodology*, 29, 1–41.

Rossi, A. P., Lipsey, M. W., & Freeman, H. E. (2004). *Evaluation: A systematic approach* (7th ed.). Thousand Oaks, CA: Sage.

Rowan, B. E., Camburn, E., & Correnti, R. (2004). Using teacher logs to measure the enacted curriculum in large-scale surveys: Insights from the Study of Instructional Improvement. *Elementary School Journal, 105,* 75–102.

Spybrook, J. H. (2007). *The statistical power of group randomized trials funded by the Institute of Education Sciences*. Unpublished doctoral dissertation, University of Michigan School of Education, Ann Arbor.

Stigler, J., Gallimore, R., & Hiebert, J. (2000). Using video surveys to compare classrooms and teaching across cultures: Examples and lessons from the TIMSS Video Studies. *Educational Psychologist, 35*(2), 87–100.