Studying the Reliability of Group-Level Measures with Implications for Statistical Power: A Six-Step Paradigm[1]

THIS DRAFT: March 25, 2011

Stephen W. Raudenbush[a], Andres Martinez[b], Howard Bloom[c], Pei Zhu[c], and Fen Lin[a]

[a]University of Chicago

[b]University of Michigan

[c]MDRC

[1] The work reported here has been supported by the William T. Grant Foundation under the grant "Building Capacity for Evaluating Group-Level Interventions". The authors are especially grateful to Bob Granger, Ed Seidman and Vivian Tseng, from the William T. Grant Foundation, for their continued advice and encouragement; to Robert C. Pianta and Andrew J. Mashburn, for generously lending data from the Multi-State Study of Pre-Kindergarten for the case study presented here; and to the participants of the joint meetings between the William T. Grant Foundation and the Forum for Youth Investment held in Washington, D.C., on December of 2006 and July of 2007, for their interest in discussing these ideas.

Abstract

Many youth development programs aim to improve youth outcomes by raising the quality of social interactions occurring in groups such as classrooms, athletic teams, therapy groups, after-school programs, or recreation centers. As a result, evaluators are increasingly interested in determining whether such programs significantly improve "group quality." We consider methods for studying the reliability of measures of group quality, with implications for the design of evaluation studies, and we illustrate these methods using a large-scale data set on classroom observations. Our approach enables the analyst to compare options for improving reliability, including increasing the number of raters per classroom, increasing the number or length of occasions of measurement, or improving the training of raters. These inferences depend on model assumptions, and we develop and illustrate a method for testing the sensitivity of these inferences to errors of model misspecification. We then consider the implications of such investments for the statistical power of experiments that assess the impact of intervention on group quality. Our six-step approach extends generalizability theory and uses it to improve research on environments in which youth develop.

I. Introduction

Many programs aim to improve youth outcomes by improving the quality of social processes occurring in groups within which youth develop. This is the case, for example, for after-school programs (e.g., Hirsch and Wong, 2005), teacher professional development programs (e.g., Kinzie, Whitaker, Neesen, Kelley, Matera, and Pianta, 2005), comprehensive school reform programs (e.g., Borman, Slavin, Cheung, Chamberlain, Madden, and Chambers, 2005), and training programs for coaches (e.g., Smith, 2006). Whether the programs are athletic, educational, social, or therapeutic, they commonly operate on groups such as teams, schools, classrooms, or therapy groups, and seek to promote high-quality social interactions among the members of the group to ultimately improve the skill, knowledge or emotional regulation of the participating youths.

The programs we focus on are thus based on a theory that has two parts: 1) a theory of how the program changes group quality; and 2) a theory of how group quality affects youth outcomes. Knowing whether a program affects group quality is essential to understanding when, why, and how it affects the outcomes of participating individuals.

To see why, consider first a study in which groups were randomly assigned to one of two conditions (program versus control) and evaluators found no impact of the program on personal outcomes. Assume that the study had adequate statistical power and that the personal outcomes were well measured. Two possibilities emerge.

First, it may be that the program improved group quality in all the ways intended but that the persons in the program group failed to benefit from the changes in group functioning. For example, suppose that a teacher-training intervention caused teachers to

change their methods of classroom instruction in all the ways intended, but the instructional changes produced no change in student achievement. Clearly, the underlying theory about the link between group quality (in this case, instructional quality) and person outcomes (in this case, student achievement) was incorrect. On the other hand, it may be that the program had little or no impact on group quality. In this case, the theory about how improved group quality affects youth outcomes was never tested!

Suppose that the evaluators did not administer a reliable and valid measure of group quality. Then it would not be possible to distinguish between these two contradictory explanations for the finding of no program effect on youth outcomes, and it would be difficult to draw any conclusions for the study that might guide future program development.

Even in a happier scenario in which the evaluation did produce convincing evidence of a positive effect on youth outcomes, interpretation depends on knowing whether the intervention changed group quality in the ways intended. If the intended group changes did not occur, then the program must have had its effect on persons through channels that the program designers failed to anticipate. Knowing this result is particularly important in attempts to reproduce the positive effect in a new setting.

In sum, assessing the impact of a program on group quality is essential in building a science of intervention to improve youth outcomes. In this article we focus on how to assess the reliability of group quality measures, and on how to design studies of interventions that aim to improve group quality.

The Interrelated Problems of Measurement and Power

If the advancement of intervention science requires an understanding of how the intervention affects group quality, the reliable and valid measurement of group quality is essential. Yet, the measurement of neighborhoods, schools, classrooms, day care centers, community centers, after-school programs, and other group settings is comparatively under-developed despite more than a century of intense interest in measuring personal attributes such as cognitive skill and personality. Raudenbush and Sampson (1999) used the term “ecometrics” to refer to the science of validly measuring ecological settings, as distinct from the much more thoroughly developed science of measuring psychological attributes, widely known as “psychometrics.”

Although valid measurement is essential, our focus here is in one aspect of valid measurement, namely, *reliability*. Reliability is the consistency of results of applying a measurement procedure under conditions that vary in ways deemed irrelevant to the true value of what is being measured. One cannot make valid inferences about group quality based on unreliable data. Reliability is thus a necessary if not sufficient condition for valid inference about the quality of a group’s functioning. Motivating our focus on reliability is its influence on a study’s statistical power. Low reliability will weaken statistical power, thus diminishing the chance of discovering true intervention effects on group quality.

A cost-benefit trade-off arises in designing studies to have high power. The resources available can be used to increase the number of groups in the study or, given a fixed number of groups, to obtain more reliable measures. For example, an experimenter using observations to assess classroom quality can typically enhance reliability by

improving rater training, dispatching multiple raters to the same classroom, or observing each classroom on multiple occasions. Such enhancements cost money – money that could instead be used to increase sample sizes. So a smart analysis of the costs and benefits associated with increasing reliability is in order. This paper presents a conceptual framework for investigating these trade-offs so that researchers can make efficient uses of available resources to enhance power. In addition, the paper illustrates how such analyses can be done using a large data set of classroom observations.

We draw heavily on generalizability theory (Cronbach and Gleser, 1965; Brennan, 2001), a widely used approach among psychometricians for quantifying multiple sources of measurement error. We expand this approach in several ways. First, because rater effects tend to be large in observational studies of classrooms, we re-formulate the model to illuminate the potential benefit of increasing the number of raters per classroom to investing more in rater training. Second, we suggest methods for sensitivity analysis, enabling the investigator to explore the likely practical significance of omitting relevant sources of measurement error from the model used to analyze the data. Third, we explicitly link improvements in measurement to gains in statistical power for detecting the impact of interventions on group quality. These augmentations to generalizability theory lead us then to recommend a “six-step” paradigm for studying errors of measurement.

Background

Generalizability theory (GT), introduced by Cronbach, Gleser, Nanda, and Rajaratnam (1972)[2], provides a framework for conceptualizing and investigating the sources of error in a measure. Brennan (1977, 1992a, 1992b, 2001), Shavelson, Webb and Rowley (1989), and Shavelson and Webb (1991) have developed, extended, and applied GT. The theory links a methodological study to the design of a penultimate substantive study.

Two Kinds of Studies

Our ultimate interest is the design of a study to test the effectiveness of the program in improving group quality. Following the language of GT, we call this the "decision study" (D-study), because it will inform decisions about the value of the interventions. To inform the design of this study requires preliminary information from a "generalizability study" (G-study), the purpose of which is to quantify the sources of error that arise in the measurement process. Subject to cost constraints, the G-study enables the researcher to minimize errors of measurement in the D-study. A well-designed G-study can produce information about:

- the relative importance of various sources of error, including, for example, rater inconsistency, temporal instability, and item inconsistency;

[2] An earlier presentation to the key ideas in GT can be found in Lindquist (1953). Also, Cronbach, Rajaratnam and Gleser (1963) and Gleser, Cronbach and Rajaratnam (1965) present GT explicitly though more succinctly than Cronbach, Gleser, Nanda, and Rajaratnam in their 1972 seminal book. Gleser et al. (1965), in particular, apply GT to study the adequacy of generalizations made from a set of observations that are classifiable according to two aspects of the measuring process and show how to use estimates of the variance components to design more efficient studies.

- the likely benefits of measurement re-design (e.g., recruiting more raters, training them better, observing more often, using more items) in improving reliability and therefore power;
- the impact of measurement reliability on the statistical power of the primary study;
- the minimum detectable treatment effect size that the primary study can detect with a pre-set level of power.

Application of G-Theory to the Measurement of Social Settings

GT has been used extensively to analyze person-level measures. Its use for studying classroom-level measures is less common, though there are some examples (e.g., Gillmore, Kane, & Naccarato, 1978; Kane & Brennan, 1977; Kane, Gillmore & Crooks, 1976; McGaw, Wardrop, & Bunda, 1972; Smith, 1979). In related prior work, Raudenbush, Rowan, and Kang (1991) developed a multivariate, three-level hierarchical linear model to investigate sources of error in studies of school climate measured by interviewing teachers. They showed how adding more items can reduce the effect of item inconsistency on school-level reliability. More importantly, sampling more teachers per school is important when the reports of teachers attending the same school are somewhat inconsistent. Raudenbush and Sampson (1999) combined GT and item response theory to assess the reliability and validity of measures of social and physical disorder in urban neighborhoods.

In this article, we consider observations that have been obtained by trained raters on multiple occasions. None of the earlier work that we are aware of, however, discusses the framework we propose for investigating the trade-offs associated with the

measurement of group quality, the sensitivity of results to key model assumptions, or the statistical power of evaluation studies.

In the G-study, we explicitly represent the potential sources of error in a *theoretical model*. In practice, however, the data actually collected may not pick up every source of error represented in this model. We therefore also define an *observational model*. The relationship between the theoretical and the observational models will inform about the quality of the design of the reliability study. A well-designed G-study can be extremely useful to insure that the resources available for the primary study are efficiently used to maximize statistical power. In contrast, a poorly designed G- study may provide little new information or—worse—actually *mis-inform* research designers about the reliability they can anticipate in the primary study, leading to poor choice of sample size and incorrect calculations of power for the primary study.

Given the possibility that the observational model under-identifies the theoretical model, it is essential to explicitly state the assumptions under which the observational model gives correct results. These "identifying assumptions" will typically not be testable. However it is possible to test the sensitivity of the results to the failure of these assumptions. The idea is to vary the assumptions and watch for variation in the key conclusions. We therefore recommend and illustrate how such a sensitivity analysis can be conducted.

This reasoning leads us to build on GT to recommend a "six-step" paradigm for studying errors of measurement in group processes. Step 1 is to hypothesize the salient sources of measurement error and to combine these in a theoretical model. Step 2 is to design a study of these sources of error. No matter how well-designed such a study might

be, it will often not enable the researcher to isolate every source of error specified in the theoretical model. So this step requires one to write down a model for the observed data (the "observational model") and to explicate the relationship between the theoretical model and the observational model. This step is crucial because it enables the investigator to explicate the assumptions under which the analytic results will apply. Step 3 defines these "identifying assumptions" by equating the sources of variance in the observational model to those in the theoretical model. In Step 4, one analyzes the data under the identifying assumptions and considers alternative options for maximizing reliability in light of the results. A key question then arises: how sensitive are the decisions based on such an analysis to the identifying assumptions? So, in Step 5, we undertake sensitivity analyses that explore the extent to which key conclusions vary under alternative assumptions. The result of this process is a range of plausible values for the reliability with which a specific study design can measure the outcome. Step 6 explores the implications of this range of plausible reliabilities for the statistical power of an evaluation study.

III. An Empirical Case Study

To illustrate how the proposed six-step framework can be applied in a realistically complex situation, we now examine the variance components of the Classroom Assessment Scoring System (CLASS) as used in the Multi-State Study of Pre-Kindergarten (MSSPK) conducted by the National Center for Early Development (Pianta, Howes, Burchinal, Bryant, Clifford, Early & Barbarin, 2005). The MSSPK was not designed as a G-study and thus any limitations for examining reliability reflect the fact that doing so was not its original purpose. Nonetheless, the MSSPK is remarkable in

allowing estimation of sources of variation we believe to be theoretically important—sources of variation typically ignored in past studies of classrooms. So we regard this study as particularly useful for our purposes and thank the authors of MSSPK for permission to use it.

Measuring Instrument and Setting

Researchers at the University of Virginia developed the CLASS to assess classroom quality in preschool through third-grade classrooms (Pianta, La Paro and Hamre, 2006).[3] Its focus is the social and emotional environment within classrooms as manifested by interactions between teachers and students. Among the most extensively used classroom-observational instruments, CLASS has been important in the State-wide Early Education Programs Study (Clifford, et al., 2005) and the Early Child Care and Youth Development Study.[4]

Using the CLASS, raters observe multiple "segments" per class per day. Each segment consists of 20 minutes of observation followed by 10 minutes of coding. Each segment yields a series of numeric assessments that are later aggregated into higher-order domains (Hamre, Mashburn, Pianta, Locasle-Crouch & La Paro, 2006).

Step 1: Hypothesize sources of measurement error and specify a theoretical model

Sources of error. For the purposes of this case study, we consider classrooms, raters, segments, and days as the main sources of variance. The day variance arises because mean levels of classroom quality may vary from day to day. This variance, also referred to as the between-day variance, is distinct from the within-day variance, which is represented by the segments. We assume both of these sources of variation to be random

[3] See http://www.classobservation.com/ for details.
[4] See http://www.nichd.nih.gov/research/supported/seccyd.cfm for details.

and focus only on "short-term" instabilities, regarded here as part of the error variance, alongside the rater variance. We deliberately omit other sources of variation, including item inconsistency, as plausible sources of error variance, to keep the example manageable.

The following interactions are possible and all contribute, in theory, to the error variance: *classroom-by-rater variance,* reflecting the fact that differences between classrooms may vary across raters; *classroom-by-day variance,* the variation that arises when the difference between the ratings of two classrooms changes from day to day independent of the rater; *rater-by-day variance,* the variation that arises when the "toughness" or "leniency" of raters changes from day to day in a non-systematic way; *rater-by-segment variance,* the variation that arises when the "toughness" or "leniency" of raters changes from one segment to the next, within a day, in a non-systematic way; and *classroom-by-day-by-rater variance,* the variation that arises when the difference between the ratings of two classrooms changes from day to day and the changes in those differences vary from one rater to the next.

Theoretical model. All these sources of variation are represented in the following theoretical model:

$$y_{rs(cd)} = \mu + \alpha_c + \beta_r + \gamma_d + \pi_{s(cd)} + (\alpha\beta)_{cr} + (\alpha\gamma)_{cd} + (\beta\gamma)_{rd} + (\beta\pi)_{rs(cd)} + (\alpha\beta\gamma)_{crd} \quad (1)$$

Here, $y_{rs(cd)}$ represents the rating for classroom *c* obtained by rater *r* on segment *s* of day *d*; μ is the mean across all observations; α_c, β_r, γ_d, and $\pi_{s(cd)}$ are random effects associated with classrooms, raters, days and segments, respectively; $(\alpha\beta)_{cr}$ is the classroom-by-rater interaction effect; $(\alpha\gamma)_{cd}$ is the classroom-by-day interaction effect;

$(\beta\gamma)_{rd}$ is the rater-by-day interaction effect; $(\beta\pi)_{rs(cd)}$ is the rater-by-segment interaction effect; and $(\alpha\beta\gamma)_{crd}$ is the classroom-by-rater-by-day interaction effect.[5]

A positive value of α_c indicates that classroom *c* has an above-average "true quality" and a positive value of β_r indicates that rater *r* gives, on average, more favorable ratings than does the average rater. Also, a positive value of $(\alpha\beta)_{cr}$ indicates that rater *r* rated classroom *c* higher than expected given the true quality of classroom *c* and the overall tendency of rater *r* to be lenient versus strict. The interaction terms $(\alpha\beta)_{cr}$, $(\alpha\gamma)_{cd}$ and $(\beta\gamma)_{rd}$ can be estimated if the measurement study crosses at least some classrooms with raters, at least some classrooms with days, and at least some raters with days, respectively. Similarly, the classroom-by-rater-by-day interaction term, $(\alpha\beta\gamma)_{crd}$, can be estimated if the design crosses at least some classrooms, raters and days.

The subscript "*rs(cd)*" indicates that a segment is, by definition, nested within a given day for a given classroom. A segment is by definition a unique interval of life in a particular classroom in a particular day. This nesting precludes the interaction of the segment effect with the classroom effect and of the segment effect with the day effect. Thus the theoretical model does not contain a segment-by-classroom, a segment-by-day, a segment-by-classroom-by-day or a segment-by-classroom-by-day-by-rater interaction effect. However it is possible to have more than one rater per segment so the only interaction term for segments is the rater-by-segment interaction term, $(\beta\pi)_{rs(cd)}$.

---

[5] Notice this model is equivalent to a three-way cross-classified model with random effects α_c, β_r and γ_d in which the error term has been replaced by $\pi_{s(cd)} + (\beta\pi)_{rs(cd)}$.

In fully-crossed designs all the main effects and interactions presented in (1) are estimable. However, fully-crossed designs are not typically economically possible or even logistically feasible. Therefore, in reality, it is likely that some of the theoretical variance components turn out to be fully confounded.

Yet, whatever the limitations particular designs may impose, we regard Equation (1) as the theoretical measurement model in which $y_{rs(cd)}$ is a fallible measure (or "observed score") of classroom quality, where $\mu + \alpha_c$ is the "true score," and the expression $\beta_r + \gamma_d + \pi_{s(cd)} + (\alpha\beta)_{cr} + (\alpha\gamma)_{cd} + (\beta\gamma)_{rd} + (\beta\pi)_{rs(cd)} + (\alpha\beta\gamma)_{crd}$ is the measurement error. The total variation of $y_{rs(cd)}$ is

$$Var(y_{rs(cd)}) = \sigma^2_{class} + \sigma^2_{rater} + \sigma^2_{day} + \sigma^2_{segment} + \sigma^2_{class\ x\ rater} + \sigma^2_{class\ x\ day} + \sigma^2_{rater\ x\ day} + \sigma^2_{rater\ x\ segment} + \sigma^2_{class\ x\ rater\ x\ day} \quad (2)$$

where σ^2_{class} is "true-score variance" and the remaining terms constitute the "error variance."

Step 2: Design a study and specify the observational model

Table 1 defines the design requirements for estimation of each source of variation in our theoretical model (Equation 1). The available CLASS data allow estimating some—but not al—of these sources. Thus our observational model will not fully map onto its theoretical counterpart. Let us consider the design of the MSSPK and its implications for the observational model.

--- Insert Table 1 Here ---

The MSSPK is a descriptive study of $C = 240$ pre-school centers located in six states with one classroom per center and 40 centers from each state in the study.[6] Data for the present analysis are from the August 2001 through June 2002 collection wave in which $R = 26$ raters participated. On average, each rater visited 15.8 classrooms; a classroom-rater combination was repeated, on average, on 2.6 different days, and 6 segments were recorded for each classroom-rater-day combination, on average. In total, 6,473 ratings were generated.[7]

In the MSSPK, classrooms were never assessed at the same time by multiple raters. Therefore, the segment and the segment-by-rater variances are fully confounded. Actually no classroom was ever assessed by more than one rater on any given day, so the variance associated with the three-way interaction between classrooms cannot be estimated. We shall see that this three-way interaction is fully confounded with the classroom-by-day variance.

In addition, not all the two-way interactions between classrooms, raters and days are estimable despite classrooms, raters and days all being partially crossed. The amount of crossing is not sufficient to stably estimate separately all the variances that arise from the two-way interactions of these three main effects. In the classroom-by-rater cross-tabulation, 221 of the 240 classrooms are partially crossed with 25 of the 26 raters, so the

---

[6] As noted earlier, we use this descriptive study as a reliability study.

[7] These are arithmetic (not harmonic) means. The difference when the total sample size is calculated with these numbers is due to approximation and to some cases that were discarded. To be exact, $R = 26$ (number of raters), $\overline{C}_r = 411/26$ (classrooms per rater), $\overline{D}_{cr} = 1{,}067/411$ (days per classroom-rater combination), and $\overline{S}_{dcr} = 6{,}390/1{,}067$ (segments per day-classroom-rater combination). Thus the total sample size used is $N = R \cdot \overline{C}_r \cdot \overline{D}_{cr} \cdot \overline{S}_{dcr} = 6{,}390$. The extra 83 cases (1.28%) were discarded because of missing data in the outcomes explored.

classroom-by-rater interaction variance can be estimated (see the appendix for more details).

In principle, the classroom-by-day and the day-by-rater variances can also be estimated because classrooms and days were partially crossed and days and raters were also partially crossed. Yet these variances are confounded because the amount of crossing was insufficient to estimate them separately. In fact, no two raters ever observe a given classroom on the same day, and very rarely did a rater observe 2 classrooms on a given day (certainly never 3 or more classrooms on a given day). So the classroom-by-day variance cannot be distinguished from the day-by-rater variance. Since the variance of the three-way interaction between classrooms, raters and days was fully confounded with the classroom-by-day variance, this means that the classroom-by-day variance, the day-by-rater variance and the classroom-by-rater-by-day variance are not separable from each other. As we shall see, an identifying assumption about the confounding of these theoretical variance components will be required to draw conclusions about the measurement properties of CLASS from this study.

The variances associated with the main effects of classrooms, raters and days are all estimable and we assume are not confounded with the variances of their interactions. Of the 240 classrooms, 158 were observed by at least 2 raters. Of the remaining 82 classrooms, 63 were observed by a rater who also observed at least one of the 158 classrooms with multiple raters. Thus, in reality, 221 classrooms were partially crossed with raters. We thus see that the classroom-by-rater variation is not confounded with the classroom variation. In addition, all classrooms were observed on at least two days and at least two classrooms were observed in 158 of the 167 days. All classrooms and days were

then partially crossed and thus the classroom-by-day interaction is not confounded with the classroom variation.

The observational model. The data actually collected thus give place to the following observational model:

$$y_{s(crd)} = \mu + \alpha_c^* + \beta_r^* + \gamma_d^* + \pi_{s(crd)}^* + (\alpha\beta)_{cr}^* + (\alpha\gamma)_{rd}^* . \quad (3)$$

Notice first that the sub-index "*s(crd)*" is different from the one in the theoretical model (Equation 1). This indicates the nesting of segments within classrooms, days *and* raters. Notice also this model has 6 random effects, whereas the number of random effects in the theoretical model (Equation 1) is 9. The difference arises precisely because of the confounding of some of the theoretical variances. The superscript "*" is used here to annotate that the random effects (and their variances) in the observational model are not necessarily equal to their corresponding terms in the theoretical model.

The relationship between the variance components in the theoretical model and those in the observational model is in Table 2. As shown, the variances for the main effects of classrooms, raters and days, and for the classroom-by-rater interaction effect, are estimable and match exactly their theoretical counterparts. In contrast, for the reasons discussed before, the classroom-by-day, the day-by-rater and the classroom-by-day-by-rater variances are confounded in what is denoted $(\alpha\gamma)_{rd}^*$, while the segment and the segment-by-rater variances are confounded in what is denoted $\pi_{s(crd)}^*$.

--- Insert Table 2 Here ---

Because the estimate of classroom variance matches that of the theoretical model, model misspecification will not inflate or deflate the estimate of the "true-score" variance. In that sense, our data cannot *mis-inform* about the reliability of the classroom

measure because there is no confounding with the true score. Rather, the confounding is in the error variance. Yet, the confounding in $(\beta\gamma)^*_{rd}$ and in $\pi^*_{s(crd)}$ does not allow gathering full information on all the theoretical variance components and thus the design *somewhat under-informs* about the ways to improve the reliability: We cannot draw inferences about the separate sources of error variation in the theoretical model without making further assumptions.

Step 3: State identifying assumptions

When the observational model does not match the theoretical model, which will typically be the case, the investigator must make some assumptions in order to proceed. The idea then is to make such "identifying assumptions." In the current case, we make the following assumptions before checking sensitivity:

1. We see from Table 2 that $\sigma^{*2}_{segment} = \sigma^2_{segment} + \sigma^2_{rater\ x\ segment}$, meaning that the measurement study cannot separate the variance of the main effect of segments from the segment-by-rater interaction variance. We provisionally assume $\sigma^2_{rater\ x\ segment} = 0$, in which case $\sigma^{*2}_{segment} = \sigma^2_{segment}$.

2. Next, we see from Table 2 that $\sigma^{*2}_{class\ x\ day} = \sigma^2_{class\ x\ day} + \sigma^2_{rater\ x\ day} + \sigma^2_{class\ x\ rater\ x\ day}$. We provisionally assume $\sigma^2_{rater\ x\ day} = \sigma^2_{class\ x\ rater\ x\ day} = 0$, in which case $\sigma^{*2}_{class\ x\ day} = \sigma^2_{class\ x\ day}$. Our rationale is based on the assumption that what happens in classrooms does vary significantly from one day to the next while there is less a priori reason to believe that raters vary from day to day. Of course this assumption cannot be checked with the available data, and that fact leads us to rely on a sensitivity analysis to test the credibility of our results.

Under assumptions (1) and (2), the theoretical model and the observational model match, and the variance and reliability of the classroom means will become

$$Var\left(\bar{y}_{\bullet(\bullet c\bullet)}\right)=\sigma^{*2}_{class}+\frac{\sigma^{*2}_{rater}+\sigma^{*2}_{class\,x\,rater}}{R_c}+\frac{\sigma^{*2}_{day}+\sigma^{*2}_{class\,x\,day}}{D_c}+\frac{\sigma^{*2}_{segment}}{D_c S_{cd}} \qquad (4)$$

and

$$\text{Rel}\left(\bar{y}_{\bullet\bullet c\bullet}\right)=\frac{\sigma^{*2}_{class}}{\sigma^{*2}_{class}+\left(\frac{\sigma^{*2}_{rater}+\sigma^{*2}_{class\,x\,rater}}{R_c}+\frac{\sigma^{*2}_{day}+\sigma^{*2}_{class\,x\,day}}{D_c}+\frac{\sigma^{*2}_{segment}}{D_c S_{cd}}\right)} \qquad (5)$$

where R_c is the number of raters who observe each class, D_c is the number of days each class is observed, and S_{cd} is the number of segments per class per day. Our sensitivity analyses (Step 5) will re-do the analysis below under a range of alternative assumptions.

Step 4: Analyze the data and draw tentative conclusions under the identifying assumptions

We obtained restricted maximum likelihood estimates of the variance components from the observational model for a measure of *instructional climate*. This measure is a weighted composite of two factors measured by the CLASS: concept development and quality of feedback.[8]

--- Insert Table 3 Here ---

Fitting the observational model (Equation 3) yields variance estimates shown in Table 3, providing some insight about how adding raters, day or segments will affect the *noise* of the classroom mean measure under our identifying assumptions. This in turn makes it possible to anticipate how the reliability of the measure would change in

[8] Factor weights are 0.99 for concept development and 0.77 for quality of feedback. The outcome was standardized so that the sum of the error variances is unity.

response to varying the deployment of observational resources. The largest source of error variation is $\sigma^{*2}_{segment}$, the confounded segment and segment-by-rater variance. Given this large variation, and given our identifying assumptions, we conclude tentatively that adding segments will be very helpful in increasing the reliability. The next largest source of error variance is the rater variance. The noise associated with this variation can be reduced by adding raters. In contrast, day noise is relatively small so adding days will not boost the reliability as much. Yet the variation is non-trivial so a modest improvement in reliability is achievable by adding days. Moreover, adding days will also reduce the impact of the classroom-by-day variance. However, the benefit of adding raters will be greater given the comparatively large values of rater variance and rater-by-classroom variance relative to the day variance and the day-by-classroom variance. There are two caveats: a) the benefit of adding raters or days or segments must be weighed against costs; and b) these conclusions may hinge on the validity of our identifying assumptions.

How can we examine the impact of more training for raters? Our results indicate that there is non-negligible variation between raters $(\hat{\sigma}^2_{rater} = .29)$ and that the rater-by-classroom variance is also non-negligible $(\hat{\sigma}^2_{class\ x\ rater} = .11)$. This suggests perhaps that increased training of raters might be a good investment. How can our data inform such an option?

The standard representation of reliability will not give us an answer. We therefore modify the expression for reliability by noting that, according to our model, the correlation between the ratings of two raters (rater *r* and rater *r'*) during the same segment is

$$Corr(y_{crs(d),}y_{cr's(d)}) = \rho_{raters} = \frac{\sigma^2_{class} + \sigma^2_{day} + \sigma^2_{class\,x\,day} + \sigma^2_{segment}}{\sigma^2_{class} + \sigma^2_{day} + \sigma^2_{class\,x\,day} + \sigma^2_{segment} + \sigma^2_{rater} + \sigma^2_{class\,x\,rater}}. \quad (6)$$

Presumably, training raters would increase this correlation, and it is possible for trainers to monitor the increase in this correlation as training proceeds. To discern the effect of such an increase, we define also two other correlations: the correlation ρ_d between the true quality on two days (day *d* and day *d'*), and the correlation ρ_s between the true quality on two segments (segment *s* and segment *s'*):

$$\rho_d = Corr[\mu + \alpha_c + \gamma_d + (\alpha\gamma)_{cd}, \mu + \alpha_c + \gamma_{d'} + (\alpha\gamma)_{cd'}] = \frac{\sigma^2_{class}}{\sigma^2_{class} + \sigma^2_{day} + \sigma^2_{class\,x\,day}},$$

$$\rho_s = Corr[\mu + \alpha_c + \gamma_d + (\alpha\gamma)_{cd} + \pi_{s(d)}, \mu + \alpha_c + \gamma_d + (\alpha\gamma)_{cd} + \pi_{s'(d)}] = \frac{\sigma^2_{class} + \sigma^2_{day} + \sigma^2_{class\,x\,day}}{\sigma^2_{class} + \sigma^2_{day} + \sigma^2_{class\,x\,day} + \sigma^2_{segment}}. \quad (7)$$

We can now re-express our reliability formula (Equation 5) in terms of these correlations:

$$Rel(\bar{y}_{\bullet\bullet c\bullet}) = \frac{\rho_r \rho_d \rho_s}{\rho_r \rho_d \rho_s + \left(\frac{1-\rho_r}{R_c} + \rho_r \rho_s \frac{1-\rho_d}{D_c} + \frac{\rho_r (1-\rho_s)}{D_c S_{cd}} \right)}. \quad (8)$$

This form of the expression for the reliability helps us assess the impact of increasing inter-rater consistency, that is, ρ_r.

Increasing the number of raters. The number of raters can significantly affect reliability. We see from Table 4 that increasing R_c from 1 to 10 increases the reliability from .55 to .88 when $\rho_d = .70$ and $D_c = 7$. However, the impact of increasing the number of raters greatly diminishes when the inter-rater correlation ρ_d is very high. For example, increasing the number of raters from 1 to 10 increases the reliability only from .80 to .93 when $\rho_d = .90$ and $D_c = 7$.

--- Insert Table 4 Here ---

Increasing the number of days. Increasing the number of days of observation per classroom has a comparatively modest impact on reliability. For example, increasing the number of days from 4 to 10 increases reliability from .82 to .87 when $R_c = 4$ and $\rho_d = .80$ under our identifying assumptions.

Increasing inter-rater correlation. Improved training can presumably improve inter-rater correlation and hence boost reliability. Table 4 shows that this effect is large when the number of raters per classroom is small. However, the impact of increasing inter-rater correlation diminishes as the number of raters per classroom increases. Thus we see that when we increase the inter-rater correlation from .60 to .90, the reliability increases dramatically, from .45 to .85 when $R_c = 1$ (given $D_c = 7$). However, when $R_c = 10$, the same increase in inter-rater correlation boosts the reliability modestly, from .85 to .93 (again given $D_c = 7$).

In summary, there is a clear trade-off between increasing the number of raters per classroom and improving training to boost the inter-rater correlation. Hiring more raters may be a simpler and surer way of increasing reliability. But if improving the rater training is comparatively inexpensive, researchers may consider increasing the intensity of rater training as an option.

In Table 4, entries where reliability exceeds .80 are in bold, highlighting there are a number of options for achieving high reliability. Each entry represents a combination of the number of days, the number of raters, and the amount of training each rater receives as reflected in the inter-rater correlation. Associated with each option, in principle, is an

overall cost. In principle, such a table can be used to select the most cost effective strategy for achieving a given level of reliability.

Step 5: Sensitivity analysis

All of the results under Step 4 were contingent on the validity of our identifying assumptions. To assess the sensitivity of our results to departures from these assumptions, we consider eight alternative scenarios. The scenarios arise from varying the assumptions 1) about the confounding of the segment variation with the segment-by-rater variation and 2) about the confounding of the classroom-by-day variation with the rater-by-day variation and the classroom-by-rater-by-day variation shown in Table 2. A summary of the scenarios considered in our sensitivity analysis is in Table 5. For the confounding of segment variation with the segment-by-rater variation ($\sigma^{*2}_{segment} = \sigma^{2}_{segment} + \sigma^{2}_{rater\,x\,segment}$), in scenarios 1 through 4 we assume $\sigma^{2}_{rater\,x\,segment} = 0$, thus $\sigma^{*2}_{rater} = \sigma^{2}_{rater}$, while in scenarios 5 through 8 we assume $\sigma^{2}_{rater\,x\,segment} = \sigma^{2}_{segment} = \sigma^{*2}_{segment} / 2$. On the confounding of the classroom-by-day, the rater-by-day and the classroom-by-rater-by-day variances ($\sigma^{*2}_{class\,x\,day} = \sigma^{2}_{class\,x\,day} + \sigma^{2}_{rater\,x\,day} + \sigma^{2}_{class\,x\,rater\,x\,day}$), in scenarios 1 and 5, we assume $\sigma^{2}_{rater\,x\,day} = \sigma^{2}_{class\,x\,rater\,x\,day} = 0$, thus $\sigma^{*2}_{class\ x\,day} = \sigma^{2}_{class\ x\,day}$; in scenarios 2 and 6, we assume $\sigma^{2}_{class\,x\,day} = \sigma^{2}_{class\,x\,rater\,x\ day} = 0$, thus $\sigma^{*2}_{class\,x\,day} = \sigma^{2}_{rater\,x\,day}$; in scenarios 3 and 7, we assume $\sigma^{2}_{class\,x\,xrater\,x\,xday} = 0$ and $\sigma^{*2}_{class\,x\,day} / 2 = \sigma^{2}_{class\,x\,day} = \sigma^{2}_{rater\,x\,day}$; finally, in scenarios 4 and 8, we assume $\sigma^{*2}_{class\,x\,day} / 3 = \sigma^{2}_{class\,x\,day} = \sigma^{2}_{rater\,x\,day} = \sigma^{2}_{class\,x\,rater\,x\,day}$.

--- Insert Table 5 Here ---

Next we estimate the reliability under varying sample sizes for each of these scenarios.[9] The results are in Table 6. As shown, if R_c and D_c are the same, the differences in reliability are small. That is, results are insensitive to varying S_{cd} only. What is more, results are generally insensitive to assumptions about the variance components as longs as R_c and D_c are not too different. When R_c is much bigger than D_c, or vice versa (e.g., in a 4 to 1 or 7 to 1 ratio), the reliability starts to vary modestly between scenarios. Even then, the differences would have little impact on the power analysis, the subject to which we now turn. The sensitivity analysis allows us to estimate power under the range of reliabilities generated by alternative identifying assumptions.

--- Insert Table 6 Here ---

Step 6: Reliability, sample size, and power to detect treatment effects

We now investigate the implications of our G-study for the design of a primary study when the outcome measure is group quality. To illustrate, we consider the case in which the primary study will have a "multi-site" experimental design. The sites will be schools; within schools, classrooms will be randomly assigned to treatments.

Multi-site cluster-randomized trial with group-level outcomes. Suppose a new teacher training program aims to improve the quality of the student-teacher interactions in classrooms. Assume eligible teachers are assigned at random to treatment and control conditions within each school participating in the study.

Program evaluation model. Let t_{ck} denote the *true latent quality score* for classroom $c \in \{1,...,C\}$ in school $k \in \{1,...,K\}$ and assume for simplicity that all schools

[9] Note the expressions for the variance and the reliability (Equations 4 and 5) need to be revised whenever the identifying assumptions change.

in the study have the same number of classrooms. Under a fully balanced design with *C/2* classrooms assigned to the treatment condition and *C/2* to the control condition at each school, the model for evaluating the program effects might then be

$$t_{ck} = \gamma_0 + r_{0k} + (\gamma_1 + r_{1k})W_{ck} + u_{ck} \qquad (9)$$

where γ_0 is the overall mean; γ_1 is the average treatment effect; W_{ck} is a contrast indicator (½ for treatment classrooms and -½ for control classrooms); $u_{ck} \sim N(0, \sigma_c^2)$ are classroom-specific random effects; $r_{0k} \sim N(0, \tau_{00})$ are random effects associated with the school means; and $r_{1k} \sim N(0, \tau_{11})$ are random effects associated with the school-specific treatment effects. Thus τ_{11} quantifies the heterogeneity in treatment impact across sites. Here, σ_c^2 is the between-classroom variance.[10]

Let $\bar{y}_{\bullet ck}$ be the *mean observed outcome* for classroom $c \in \{1,...,C\}$ in school $k \in \{1,...,K\}$. That is, $\bar{y}_{\bullet ck}$ is a measure of t_{ck}, and the notation of $\bar{y}_{\bullet ck}$ reflects that it is a classroom-level measure obtained by multiple raters on multiple days and segments, as in the previous section. Thus, if $\bar{e}_{\bullet ck}$ denotes the measurement error in $\bar{y}_{\bullet ck}$ as a measure of t_{ck}, we have $\bar{y}_{\bullet ck} = t_{ck} + \bar{e}_{\bullet ck}$, where the measurement errors $\bar{e}_{\bullet jk}$ are independently distributed with a mean of 0 and variance of σ_e^2. The reliability of $\bar{y}_{\bullet jk}$ is therefore $\lambda = \sigma_c^2 / (\sigma_c^2 + \sigma_e^2)$.

Paralleling Raudenbush & Liu (2000), the variance of the treatment effect is

---

[10] Notice the model assumes random site effects. In other cases, sites may be regarded as fixed and study results would be limited to the study sample. If the number of sites is small, a fixed-effects model is often more appropriate since it is difficult to generalize to a large population from a small number of sites.

$$Var(\hat{\gamma}_1) = \frac{\tau_{11} + 4\left(\sigma_c^2 + \sigma_e^2\right)/C}{K} \tag{10}$$

and the power of a test of $H_0 : \gamma_{01} = 0$ against $H_1 : \gamma_{01} \neq 0$ is determined by the non-central $F(\alpha; 1, K-1; \phi)$ statistic with non-centrality-parameter

$$\phi = \frac{\gamma_1^2}{Var(\hat{\gamma}_1^2)} = \frac{K\gamma_1^2}{\tau_{11} + 4(\sigma_c^2 + \sigma_e^2)/C}. \tag{11}$$

Standardization. Define the standardized effect size $\delta = \gamma_1 / \sigma_c$ and the variance of the site-specific standardized effect sizes $\sigma_\delta^2 = \tau_{11} / \sigma_c^2$. Dividing the numerator and denominator of (11) by σ_c^2 yields a useful and mathematically equivalent re-expression:

$$\phi = \frac{K\gamma_1^2 / \sigma_c^2}{\tau_{11} / \sigma_c^2 + 4(\sigma_c^2 + \sigma_e^2)/(C\sigma_c^2)} = \frac{K\delta^2}{\sigma_\delta^2 + 4(C\lambda)^{-1}}. \tag{12}$$

The non-centrality parameter shown in Equation (12) determines the power to detect a treatment effect. It is clear that increasing the number of schools or classrooms, or the reliability of the outcome measure, all result in increased power. Interestingly, Equation (12) shows that lack of reliability in essence reduces the effective *C*, that is, the number of classrooms sampled per school. If $\lambda = 1$, the second term in the denominator is 4/*C*. However, suppose $\lambda = .5$, then the second term in the denominator would be 4/(.5**C*). Power in this case would be the same as reducing the number of classes per school by half, given $\lambda = 1$.

Reliability and power. To study the extent to which the reliability of measurement of classroom quality affects the power in a multi-site group-randomized trial, consider a study with *C* = 6 classrooms in each of *J* = 15 schools. At each school, 3 classrooms are assigned to the treatment condition and 3 to the control condition. Let .15 of the total true

standardized outcome variation lie between schools and set at .05 the level for determining statistical significance. How much power will the study have to detect an effect size of .75, assuming the effect size variability is .05? [11]

Under the assumption that the outcome measure is fully reliable, the power to detect effect sizes .75 and larger is about .93. The study would then be "adequately powered" given common conventions for acceptable power. Power, however, rapidly decreases as the reliability of the outcome measure decreases. This relationship is shown in Figure 1. Notice how low reliabilities can hurt the power of the study. For example, a study using an outcome with reliability of 0.25 will have a statistical power of only about 0.39 and would now be regarded as badly underpowered. [12]

--- Insert Figure 1 Here ---

IV. Summary and Conclusions

Measuring group-level quality allows researchers to study the impact of an intervention on processes believed essential to the improvement of youth outcomes. These processes might include the instructional quality of classrooms, the social

---

[11] Although an effect size of 0.75 may seem exceedingly large, recall that the standardization for group-level outcomes is achieved by dividing the observed difference in mean outcomes for the treatment and control groups by the true between-classroom variation only, not by the total observed variation. This will tend to increase the magnitudes of the observed effect sizes. In addition, group-level outcomes tend to have smaller standard deviations than do individual-level outcomes, which also increase their observed effect sizes. Furthermore, interventions targeted to promote group quality will most likely have proportionally larger direct effects on the group processes than they will on individual outcomes. For all these reasons, the magnitudes of effect sizes for group-level outcomes that are estimated as described above may be much larger those typically observed for individual outcomes. However, much remains to be learned about how large effect sizes for interventions designed to improve group quality.

[12] The "Optimal Design for Longitudinal and Multilevel Research" software can be used for performing power analyses of impact studies taking into account the reliability of the outcome measure. This software includes options for power analysis of cluster-randomized trials and blocked (or multi-site) cluster-randomized trials. The software and its documentation are freely available from the website of the William T. Grant Foundation (http://www.wtgrantfoundation.org/).

environment of after-school programs, or the degree of teamwork on athletic teams. Most group-level outcomes of interest are, however, not directly observable. Rather, they are "latent" quantities and as such, will typically be measured by multiple observations or interviews. Measurement error will typically be an important concern.

Measurement error can drastically reduce the power of a seemingly well-powered study. Small measurement errors allow increased power to detect non-zero program effects on group quality. In studies using direct observation, measurement error can be reduced by increasing the number of raters per group, the number of observation days per group, the number of observation segments per day, or the number of raters in a given day or segment. How raters are assigned may matter as well. It will typically be quite useful to assign raters in such a way that over time, each group is observed by more than one rater. Good rater selection and training will also play a key role in increasing reliability and boosting power, as illustrated in our example.

When planning the primary study – that is, the study of the impact of the intervention on group quality—each of these improvements in reliability will generally cost money. Training raters, observing for more days or more segments per day, assigning multiple observers to each group, for instance, generally increase the cost of data collection given a fixed sample size. The money so invested could, in principle, have been invested in boosting the sample size—e.g., the number of classrooms per school or the number of schools. Ideally, one would know the likely impact of each investment on power and, based on this knowledge, optimally allocate the resources available for research.

A G-study is useful to inform such planning decisions. The first step is to define the plausible sources of measurement error and then to write down a statistical model that represents how those sources combine to generate errors of measurement. In Step 2, we design the reliability study in a way that hopefully will distinguish and quantify the most important sources of measurement error. We do this knowing that no matter how sophisticated the design of the reliability study, some theoretically plausible components of error are likely to be confounded. Assumptions must then be made about which sources of variation are large and which are negligible, and that is the focus of Step 3. Step 4 reaches tentative conclusions the validity of which depend on the identifying assumptions. These assumptions cannot be checked. However, using Step 5, the analyst can conduct a sensitivity analysis to determine whether key conclusions about the design of the primary study would change substantially under alternative assumptions.

Two issues emerge in planning the D-study, Step 6. The first involves possible sources of bias. If classrooms, for example, are nested within raters in the primary study, bias can easily creep in. This bias can be reduced by insuring that each rater observes an equal number of treatment and control classrooms. Ideally, classrooms within treatment conditions would also be randomly assigned to the raters if a nested design is selected for the primary study.

The second issue in designing the decision study, of course, is to maximize power. Equivalently, given a fixed power, one minimizes the minimum detectable treatment effect—the smallest treatment effect that can be discovered with adequate power. For any possible design decision (e.g., more raters per classroom, more days per classroom, more training of raters), one can use the results of the G-study to discern the

expected impact on measurement reliability. Next, one can study how those improvements in reliability are likely to affect the precision and power of the D-study. One might then assess the likely cost of each strategy for increasing reliability and hence power. These investments could then be weighed against the key alternative way to improve power: boosting the sample size. In principle, there is an optimal design that will maximize power with available resources by wisely allocating resources between improvements in measurement reliability and increases in sample size.

A limitation of this study is that we have not developed several possible alternative modeling strategies. For example, we might conceptualize “days” as having sequence effects: perhaps teachers and students become acclimated to the presence of the observer, and one may discover a systematic association between the sequence number of the observation and classroom quality. This would encourage us to specify fixed effects of days or of a polynomial function of days. However, the interaction between such a fixed effect and raters or classrooms would nonetheless be regarded as random. Similarly, segment variation may be explained by fixed effects of time of day. It may be that quality is greater in the morning than in the afternoon, for example. Again, the interaction between time of day and raters or time of day and classrooms would be regarded as random. Although we avoid exploring these possibilities in the current article, it may be wise in practice to explore a full range of alternative plausible models.

Table 1: Sources of variability and corresponding random effects in the theoretical model plus research design requirements to estimate the variance component.

Source	Random effect	Design requirements for the variance component to be estimable
Classrooms (measurement unit)	$\alpha_c \sim N\left(0, \sigma^2_{class}\right)$	Multiple classrooms with more than one observation for some of them
Raters	$\beta_r \sim N\left(0, \sigma^2_{rater}\right)$	Multiple raters with more than one observation for some of them
Days	$\gamma_d \sim N\left(0, \sigma^2_{day}\right)$	Multiple days with more than one observation for some of them
Segments	$\pi_{s(cd)} \sim N\left(\sigma^2_{segment}\right)$	Multiple segments per day per rater (this is the residual term).
Classroom-by-rater interaction	$(\alpha\beta)_{cr} \sim N\left(\sigma^2_{class\,x\,rater}\right)$	Some classrooms are observed by multiple raters who observe more than one of those classrooms.
Classroom-by-day interaction	$(\alpha\gamma)_{cd} \sim N\left(\sigma^2_{class\,x\,day}\right)$	Some classrooms are observed on multiple days during which more than one of those classrooms is observed.
Rater-by-day interaction	$(\beta\gamma)_{rd} \sim N\left(\sigma^2_{rater\,x\,day}\right)$	Some raters make observations on multiple day during which multiple observations are made.
Rater-by-segment interaction	$(\beta\pi)_{rs(cd)} \sim N\left(\sigma^2_{rater\,x\,segment}\right)$	Multiple raters per segment across classroom-day combinations.
Classroom-by-rater-by-day interaction	$(\alpha\beta\gamma)_{crd} \sim N\left(\sigma^2_{classroom\,x\,rater\,x\,day}\right)$	For some raters, classrooms and days are at least partially crossed. & For some days, classrooms and raters are at least partially crossed. & For some classrooms, raters and days are at least partially crossed.

Table 2: Random effects in the observational model and equivalence between estimators of the variance components in the observational model with those in the theoretical model for the empirical case study.[1]

Random effects in the observational model	Equivalence between variance components in the observational model with those in the theoretical model
$\alpha_c^* \sim N\left(0, \sigma_{class}^{*2}\right)$	$\sigma_{class}^{*2} = \sigma_{class}^2$
$\beta_r^* \sim N\left(0, \sigma_{rater}^{*2}\right)$	$\sigma_{rater}^{*2} = \sigma_{rater}^2$
$\gamma_d^* \sim N\left(0, \sigma_{day}^{*2}\right)$	$\sigma_{day}^{*2} = \sigma_{day}^2$
$\pi_{s(crd)}^* \sim N\left(0, \sigma_{segment}^{*2}\right)$	$\sigma_{segment}^{*2} = \sigma_{segment}^2 + \sigma_{rater\,x\,segment}^2$
$(\alpha\beta)_{cr}^* \sim N\left(0, \sigma_{class\,x\,rater}^{*2}\right)$	$\sigma_{class\,x\,rater}^{*2} = \sigma_{class\,x\,rater}^2$
$(\alpha\gamma)_{rd}^* \sim N\left(0, \sigma_{class\,x\,day}^{*2}\right)$	$\sigma_{class\,x\,day}^{*2} = \sigma_{class\,x\,day}^2 + \sigma_{rater\,x\,day}^2 + \sigma_{class\;x\,rater\,x\,day}^2$

[1] All and only random effects and variance components from the observational model are denoted with a *.

Table 3: Variance estimates for the instructional climate composite.

	Variance [1,2]					
	σ^{*2}_{class}	σ^{*2}_{rater}	σ^{*2}_{day}	$\sigma^{*2}_{class\,x\,rater}$	$\sigma^{*2}_{class\,x\,day}$	$\sigma^{*2}_{segment}$
Instructional Climate Composite	.11	.29	.08	.11	.13	.39

[1] Restricted maximum likelihood estimates.

[2] Standardized estimates of the error variances (not including σ^{*2}_{class}) may not add up to 1 due to rounding.

Table 4: Reliability under varying R_c, D_c and ρ_r.[1]

R_c	ρ_r	D_c 1	4	7	10
1	.6	.39	.44	.45	.45
	.7	.46	.54	.55	.55
	.8	.54	.65	.67	.67
	.9	.62	.77	.80	.81
4	.6	.59	.72	.74	.75
	.7	.63	.77	**.80**	**.81**
	.8	.66	**.82**	**.86**	**.87**
	.9	.69	**.87**	**.90**	**.92**
7	.6	.63	.79	**.81**	**.83**
	.7	.66	**.83**	**.86**	**.87**
	.8	.68	**.86**	**.89**	**.91**
	.9	.70	**.88**	**.92**	**.94**
10	.6	.66	**.82**	**.85**	**.86**
	.7	.67	**.85**	**.88**	**.90**
	.8	.69	**.87**	**.91**	**.92**
	.9	.70	**.89**	**.93**	**.94**

[1] Here $\rho_d = \rho_s = .75$ and $S_{dc} = 6$.

Table 5: Sensitivity analysis: Estimated Variance Components Under Alternative Identifying Assumptions

Source [1]	Scenario [2, 3]							
	1	2	3	4	5	6	7	8
σ^2_{class}	.11	.11	.11	.11	.11	.11	.11	.11
σ^2_{rater}	.29	.29	.29	.29	.29	.29	.29	.29
σ^2_{day}	.07	.07	.07	.07	.07	.07	.07	.07
$\sigma^2_{class\ x\ rater}$	.12	.12	.12	.12	.12	.12	.12	.12
$\sigma^2_{class\ x\ day}$	.14	0	.07	.05	.14	0	.07	.05
$\sigma^2_{class\ x\ rater}$	0	.14	.07	.05	0	.14	.07	.05
$\sigma^2_{class\ x\ rater\ x\ day}$	0	0	0	.05	0	0	0	.05
$\sigma^2_{segment}$	.38	.38	.38	.38	.19	.19	.19	.19
$\sigma^2_{rater\ x\ segment}$	0	0	0	0	.19	.19	.19	.19

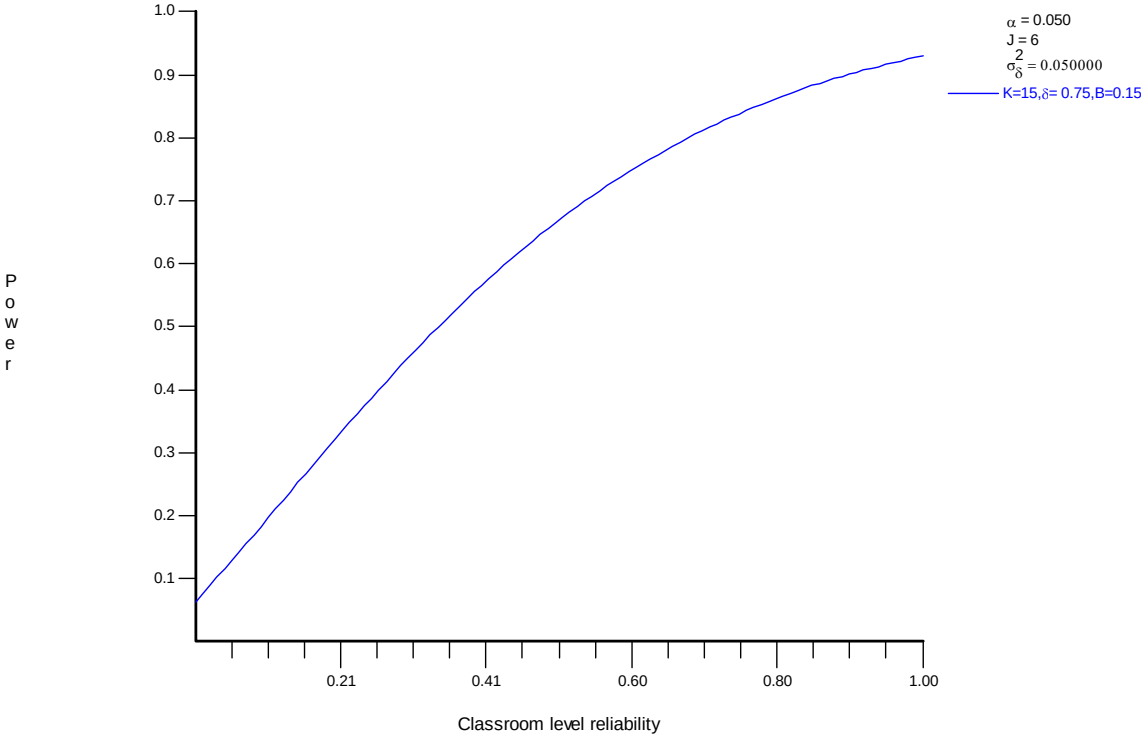
[1] Source of variance from the theoretical model.
[2] Assumptions about the estimates obtained from the observational model.
[3] Any differences in total variation are due to rounding.

Table 6: Reliability under different scenarios for varying R_c, D_c and S_{cd}.[1]

R_c	D_c	S_{cd}	Scenario							
			1	2	3	4	5	6	7	8
1	1	2	.12	.12	.12	.12	.12	.12	.12	.12
1	1	4	.13	.13	.13	.13	.13	.13	.13	.13
1	1	6	.14	.14	.14	.14	.14	.14	.14	.14
1	1	8	.14	.14	.14	.14	.14	.14	.14	.14
1	4	2	.18	.15	.16	.16	.16	.14	.15	.14
1	4	4	.18	.16	.17	.16	.17	.15	.16	.16
1	4	6	.19	.16	.17	.17	.18	.15	.17	.16
1	4	8	.19	.16	.17	.17	.18	.16	.17	.16
1	7	2	.19	.16	.17	.17	.17	.14	.15	.15
1	7	4	.20	.16	.18	.17	.18	.15	.17	.16
1	7	6	.20	.16	.18	.17	.19	.16	.17	.16
1	7	8	.20	.16	.18	.17	.19	.16	.17	.17
4	1	2	.18	.22	.20	.20	.20	.25	.23	.23
4	1	4	.21	.27	.24	.24	.23	.29	.26	.26
4	1	6	.23	.29	.25	.26	.24	.31	.27	.28
4	1	8	.23	.30	.26	.27	.24	.32	.28	.28
4	4	2	.35	.35	.35	.35	.35	.35	.35	.35
4	4	4	.38	.38	.38	.38	.38	.38	.38	.38
4	4	6	.39	.39	.39	.39	.39	.39	.39	.39
4	4	8	.40	.40	.40	.39	.40	.40	.40	.39
4	7	2	.41	.39	.40	.39	.39	.37	.38	.38
4	7	4	.43	.41	.42	.41	.42	.40	.41	.40
4	7	6	.44	.41	.42	.42	.43	.41	.42	.41
4	7	8	.44	.42	.43	.42	.44	.41	.42	.42
7	1	2	.19	.25	.22	.22	.23	.30	.26	.27
7	1	4	.23	.31	.27	.28	.25	.35	.30	.31
7	1	6	.25	.34	.29	.30	.27	.37	.31	.32
7	1	8	.26	.36	.30	.31	.27	.39	.32	.33
7	4	2	.41	.43	.42	.42	.43	.45	.44	.44
7	4	4	.45	.48	.46	.46	.46	.49	.47	.48
7	4	6	.46	.50	.48	.48	.47	.50	.49	.49
7	4	8	.47	.50	.49	.49	.48	.51	.49	.50
7	7	2	.49	.49	.49	.48	.49	.49	.49	.48
7	7	4	.52	.52	.52	.52	.52	.52	.52	.52
7	7	6	.53	.53	.53	.53	.53	.53	.53	.53
7	7	8	.54	.54	.54	.53	.54	.54	.54	.53

[1] D_{cr} = 1 throughout.

Figure 1: Reliability vs. power for a multi-site cluster-randomized trial with group-level outcomes.



Appendix: Notes on data cross-tabulations

To assess how many of the terms in the theoretical model can be estimated from the data at hand, we applied the principles listed in Table 3 to our illustrative example. The results were as follows.

On the classroom-by-rater cross-tabulation:

- Of the 240x26=6240 cells, 5829 have no observations and the remaining 411 cells have at least 4 observations (non-empty cells: 6.59%).
- 82 classrooms were observed only by 1 rater; 147 classrooms were observed by 2 raters; 10 classrooms were observed by 3 raters; and 1 classroom was observed by 5 raters (no classroom was observed by two or more raters simultaneously).
- All raters assessed multiple classrooms (max: 22 classrooms per rater; min: 3; average: 15.8; harmonic mean: 13.1).
- Some classrooms, however, are nested within raters who rate only the classrooms that are nested within them (19 classrooms are nested in one rater who only rated those 19 classrooms).

On the classroom-by-day cross-tabulation:

- Of the 240*167=40,080 cells, 39,013 have no observations and the remaining 1067 cells have at least 2 observations (non-empty cells: 2.66%).
- There were 2 or more classrooms assessed in 158 of the 167 days and only 1 classroom assessed in the remaining 9 days. The classrooms assessed in those 9 days were also assessed in other days.
- All classrooms are assessed on multiple days (max: 8 days per classroom; min: 2; average: 4.4; harmonic mean: 4.2).

On the day-by-rater cross-tabulation:

- Of the 167*26=4,342 cells, 3287 have no observations and the remaining 1055 cells have at least 2 observations (non-empty cells: 24.30%).
- Only in 9 days was only 1 rater employed, and the raters employed in those 9 days assessed in other days that were not nested. There were multiple raters employed in the remaining 158 days.
- All raters assess on multiple days (max: 88 days per rater; min: 3; average: 40.6; harmonic mean: 24.3).

On the three-way cross-tabulation of classrooms, raters and days:

- Of the 240*167*26=1042080 cells, 1041013 have no observations and the remaining 1067 cells have at least 2 observations. Notice these 1067 cells necessarily have to be the same non-blank cells in the classroom-by-day cross-tabulation.
- Of the 1067 non-blank cells, 1043 are unique rater-by-day combinations and only 12 rater-by-day combinations have 2 classrooms, meaning that only on 12 occasions did a rater visit two classrooms on the same day. Those 12 occasions are nested within 7 raters.
- All classrooms are assessed on multiple rater-by-day combinations (max: 8 rater-by-day combinations per classroom; min: 2; average: 4.4; harmonic mean: 4.2).

- On the classroom-day-by-rater (cd.r) cross-tabulation:
  - There is one rater for each classroom-day combination.
  - All raters assess on multiple classroom-day combinations (41.04 on average).
- On the classroom-rater-by-day (cr.d) cross-tabulation:
  - There are at least 2 days in 360 of the 411 classroom-rater combinations.
  - There are at least 2 classroom-rater combinations in 158 of the 167 days.
- On the rater-day-by-classroom (rd.c) cross-tabulation:
  - There is only 1 classroom in 1,043 of the 1,057 rater-day combinations. The remaining 12 have 2 classrooms.
  - There are at least 2 rater-day combinations for all classrooms.

References

Bloom, H. S. (2005). Randomizing Groups to Evaluate Place-Based Programs. In H. S. Bloom (Ed.), *Learning More from Social Experiments: Evolving Analytic Approaches* (pp. 115-172). New York: Russell Sage Foundation.

Bloom, H. S. (2007). Using Covariates to Improve Precision for Studies That Randomize Schools to Evaluate Educational Interventions. *Educational Evaluation and Policy Analysis*, 29(1), 30-59.

Borman, G. D., Slavin, R. E., Cheung, A., Chamberlain, A., Madden, N., & Chambers, B. (2005). Success for All: First-Year Results From the National Randomized Field Trial. *Educational Evaluation and Policy Analysis*, 27(1), 1-22.

Brennan, R. L. (1977). *Generalizability Analyses: Principles and Procedures.* ACT Technical Bulletin No. 26. American College Testing, Inc.

Brennan, R. L. (1992a). *Elements of generalizability theory* (revised edition). Iowa City, IA, American College Testing, Inc.

Brennan, R. L. (1992b). "Generalizability theory." *Educational Measurement: Issues and Practice* 11(4): 27-34.

Brennan, R. L. (2000). "(Mis) Conception About Generalizability Theory." *Educational Measurement: Issues and Practice* 19(1): 5-10.

Brennan, R. L. (2001). *Generalizability Theory*. New York, NY, Springer-Verlag, Inc.

Clifford, R. M., Barbarin, O., Chang, F., Early, D. M., Bryant, D., Howes, C., Burchinal, M., & Pianta, R. (2005). What is pre-kindergarten? Characteristics of public pre-kindergarten programs. Applied Developmental Science, 9(3), 126-143.

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

Cronbach, L. J., G. C. Gleser, et al. (1972). *The dependability of behavioral measures: Theory of generalizability for scores and profiles*. New York, Wiley.

Cronbach, L. J., N. Rajaratnam, et al. (1963). "Theory of Generalizability - A Liberalization of Reliability Theory." *British Journal of Statistical Psychology* 16(2): 137-163.

Donner, A., & Klar, N. (2000). *Design and Analysis of Cluster Randomization Trials in Health Research*. London: Arnold Publishers.

Gillmore, G. M., Kane, M. T., & Naccarato, R. W. (1978). The Generalizability of Student Ratings of Instruction: Estimation of the Teacher and Course Components. *Journal of Educational Measurement,* 15(1), 1-13.

Gleser, G. C., Cronbach, L. J., & Rajaratnam, N. (1965). *Generalizability of scores influenced by multiple sources of variance.* Psychometrika, 30(4), 395-418.

Hamre, B. K., Mashburn, A. J., Pianta, R. C., Locasle-Crouch, J., & La Paro, K. M. (2006). *Classroom Assessment Scoring System Technical Appendix*. Greensboro, NC.

Hedges, L. V. (2007). Intraclass Correlation Values for Planning Group-Randomized Trials in Education. *Educational Evaluation and Policy Analysis*, 29(1), 60-87.

Hirsch, B. J., & Wong, V. (2005). After-School Programs. In D. L. DuBois & M. J. Karcher (Eds.), *Handbook of Youth Mentoring*. Thousand Oaks, California: Sage Publications, Inc.

Kane, M. T., & Brennan, R. L. (1977). The Generalizability of Class Means. *Review of Educational Research,* 47(2), 267-292.

Kane, M. T., Gillmore, G. M., & Crooks, T. J. (1976). Student Evaluations of Teaching: The Generalizability of Class Means. *Journal of Educational Measurement*, 13(3), 171–183.

Kinzie, M., Whitaker, S., Neesen, K., Kelley, M., Matera, M., & Pianta, R. (2005). *State-wide Web-based Professional Development & Curricula for Early Childhood Educators: Design & Infrastructure*. Paper presented at the World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education 2005, E-Learn 2005.

Kirk, R. E. (1982). *Experimental Design: Procedures for the Behavioral Sciences* (Second ed.). Belmont, CA: Brooks/Cole.

Lindquist, E. F. (1953). *Design and Analysis of Experiments in Psychology and Education.* Boston: Houghton Mifflin.

McGaw, B., Wardrop, J. L., & Bunda, M. A. (1972). Classroom Observation Schemes: Where Are the Errors? *American Educational Research Journal*, 9(1), 13-27.

Medley, D. and H. Mitzel (1963). Measuring classroom behavior by systematic observation. *Handbook of research on teaching*. N. L. Gage. Chicago, Rand McNally.

Murray, D. M. (1998). *Design and Analysis of Group-Randomized Trials*. New York: Oxford University Press, Inc.

Pianta, R. C., La Paro, K. M., Payne, C., Cox, M. J., & Bradley, R. (2002). The relationship of kindergarten classroom environment to teacher, family and school characteristics and child outcomes. *The Elementary School Journal*, 102(3), 225-238.

Pianta, R., Howes, C., Burchinal, M., Bryant, D., Clifford, R. M., Early, D. M., & Barbarin, O. (2005). Features of pre-kindergarten programs, classrooms, and teachers: Prediction of observed classroom quality and teacher-child interactions. *Applied Developmental Science*, 9(3), 144-159

Pianta, Robert C., Karen M. La Paro and Bridget K. Hamre (2006) Classroom Assessment Scoring System Manual: Middle/Secondary Version (June: University of Virginia).

Raudenbush, S. W. (1997). Statistical Analysis and Optimal Design for Cluster Randomized Trials. *Psychological Methods*, 2(2), 173-185.

Raudenbush, S. W., & Liu, X. (2000). Statistical Power and Optimal Design for Multisite Randomized Trials. *Psychological Methods*, 5(2), 199-213.

Raudenbush, S. W., Rowan, B., & Kang, S. J. (1991). A Multilevel, Multivariate Model for Studying School Climate in Secondary Schools with Estimation via the EM Algorithm and Application to U.S. High-School Data. *Journal of Educational Statistics,* 16(4), 295-330.

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

Raudenbush, S. W., Martinez, A., & Spybrook, J. (2007). Strategies for Improving Precision in Group-Randomized Experiments. *Educational Evaluation and Policy Analysis*, 29(1), 5-29.

Sahai, H., & Ageel, M. I. (2000). *The Analysis of Variance: Fixed, Random, and Mixed Models*. Boston: Birkhauser.

Shavelson, R. J., Webb, N. M., & Rowley, G. L. (1989). *Generalizability Theory.* American Psychologist, 44(6), 922-932.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability Theory: A Primer.* Thousand Oaks, CA: Sage.

Shochet, P. A. (2005). *Statistical Power for Random Assignment Evaluations of Education Programs*. Princeton, NJ: Mathematica Policy Research.

Smith, R. E. (2006). Understanding Sport Behavior: A Cognitive-Affective Processing Systems Approach. *Journal of Applied Sport Psychology, 18*(1), 1-27.