Confidence Bounds and Power for the Reliability of Observational Measures on the Quality of a Social Setting

Yongyun Shin

yshin@vcu.edu

Department of Biostatistics

Virginia Commonwealth University

P.O. Box 980032

830 East Main Street

Richmond, VA 23298-0032

and

Stephen W. Raudenbush

sraudenb@uchicago.edu

Department of Sociology

University of Chicago

1126 E. 59th Street

Chicago, IL 60637

Abstract

Social scientists are frequently interested in assessing the qualities of social settings such as classrooms, schools, neighborhoods, or day care centers. The most common procedure requires observers to rate social interactions within these settings on multiple items and then to combine the item responses to obtain a summary measure of setting quality. A key aspect of the quality of such a summary measure is its reliability. In this paper we derive a confidence interval for reliability, a test for the hypothesis that the reliability meets a minimum standard, and the power of this test against alternative hypotheses. Next, we consider the problem of using data from a preliminary field study of the measurement procedure to inform the design of a later study that will test substantive hypotheses about the correlates of setting quality. The preliminary study is typically called the "generalizability study" or "G-study" while the later, substantive study is called the "decision study" or "D-study." We show how to use data from the G study to estimate reliability, a confidence interval for the reliability, and the power of tests for the reliability of measurement produced under alternative designs for the D study. We conclude with a discussion of sample size requirements for G studies.

KEY WORDS: Confidence Interval; D Study; G Study; Power; Reliability; Teaching Quality.

1 Introduction

Social scientists often seek information about the quality of social processes occurring in groups, including for example, after-school programs (e.g., Hirsch and Wong 2005), teacher professional development programs (e.g., Kinzie et al. 2005), comprehensive school reform programs (e.g., Borman et al. 2005), training programs for coaches (e.g., Smith et al. 2007), and day care centers (e.g., Pianta et al. 2005). Ratings of setting quality may be regarded as predictors of youth outcomes or as outcomes in studies of interventions designed to improve setting quality. For concreteness in this paper, we focus on the classroom as the key setting of

interest, and our interest focuses on the quality of teaching that occurs within each classroom, though the methods we propose apply to many other settings.

Policy makers and educators seek objective measures of teaching quality for use in evaluating teachers, helping teachers improve instruction, and studying the impact of interventions that aim to improve teaching and learning. One of the most popular methods for obtaining such measures is to assign trained observers to visit a classroom, rating the teaching on a series of, say, n items, to be aggregated into an overall score, typically a mean.[1] Because raters vary in their skill, we can expect variability between their ratings of a given class at a given time. For this reason, it often makes sense to dispatch more than one rater to each classroom and then to aggregate over rater responses, thereby averaging over the random rater differences. The reliability of the summary measure of teaching on any occasion then depends on the variability in the item responses within ratings, the number of items, the heterogeneity among the raters, and the number of raters. One may also wish to aggregate over occasions within a teacher, though our interest in this paper is the reliability of the measure for capturing the quality of teaching defined on a single occasion.

The reliability of the measure is an important criterion in various studies of teaching quality. In intervention studies designed to detect the impact of a teacher training program on teaching quality, low reliability of the measure of teaching quality will constrain the statistical power of the study. In another study, the measure of teaching quality may serve as an explanatory variable where the outcome is student learning. In this case, low reliability will not only reduce power but also produce a biased estimate unless care is taken to adjust for measurement error (Raudenbush and Sadoff 2008; Shin and Raudenbush 2010).

1.1 Goals for the Paper

In this paper, we aim to achieve two goals. First, we derive confidence intervals for reliability, tests of the hypothesis that the reliability achieves a given minimum bound, and

[1] An alternative, increasingly popular approach is to obtain videotapes of classroom interactions and to rate the teaching observed on the videotape. This facilitates multiple ratings of each class at each occasion.

the power of those tests against alternative hypotheses. We show how the sample sizes, including the number of classrooms sampled, the number of raters to rate a classroom, and the number of items, affect the width of the confidence intervals and the power of the tests. These effects of sample sizes depend, of course on estimated variance components, and we discuss this relationship.

Second, we consider the problem of how to use data from a preliminary study of the measurement procedure to inform the design of a larger study that will test hypotheses about the associations between classroom quality and other variables. Specifically, we show how to use data from the preliminary study to estimate reliability, a confidence interval for the reliability, and the power of tests for the reliability of measurement produced under alternative designs for the later study. We simulate confidence intervals for the later study, and these intervals take into account the uncertainty about variance component estimates in the preliminary study. These results have implications for the design of the preliminary study itself.

1.2 Generalizability Studies

Because of the importance of obtaining reliable measures of teaching quality, educational researchers have recently carried out a number of field studies the aim of which is to assess reliability and to inform decisions about the optimal number of items per measure and raters per classroom. Such a field study is commonly known as a "generalizability study," or "G study" (c.f. Brennan 2001). The G study provides information about the relative importance of various sources of error of measurement, for example, errors arising from different items or raters. The idea is to use the results of the G study to plan the measurement protocols in what is called the "decision study" or "D study," the study that will generate conclusions about the impact of an intervention or the association between variables in a population. If the G study is useful, it will provide some assurance that the measurement procedures used in the D study are cost effective in insuring adequate reliability of measurement.

One problem with this line of work has been that the G studies rarely produce confidence intervals for the anticipated reliability in the D study. Instead, the G study typically produces point estimates of this reliability so that planners of the D study have no sense of how much uncertainty is associated with the conclusions based on the G study. The question then arises about how large the sample sizes in a G study must be to obtain reasonable estimates of the reliability in the D study. Robert Brennan, a pioneering leader in educational measurement, has labeled the unknown uncertainty about reliability measures as "the Achilles heel" of research on sources of error in educational measurement (Brennan 2001). Brennan (2001) gives a comprehensive overview on parametric and nonparametric estimation of standard errors and approximate confidence intervals for variance components and their ratios in analysis of variance. Burdick and Graybill (1992) provide details on estimation of confidence intervals for variance components and their ratios. However, the reliability itself, which is a function of these variance components, is typically an important numerical indicator of the quality of the measure and is directly related to the attenuation bias that will arise when this measure is correlated with other variables (Raudenbush and Sadoff 2008). We want to use the variance estimates and their standard errors from a G study to estimate or predict a confidence interval for the reliability that will be obtained in the D study, and we want to assess the power of tests of the minimum bound for the reliability obtained in the D study.

1.3 Approaches

One might imagine that a field study would insure that every rater would observe each and every classroom. However, such a design is generally too costly. An often-used alternative is an incomplete balanced block design: One assigns K raters and J classrooms to each of B blocks. Every rater within a block observes every classroom within that block, yielding K ratings for each of the J classrooms within the block, with KJ ratings per block and KJB ratings overall.

To clarify the logic of our approach, the next section begins with simple case of a field

study for which $B = 1$, that is, there is a single block. Section 3 then elaborates to the general case of B blocks. Section 4 shows how to use the results of a G study to plan a D study. The Discussion section follows at last.

2 One Complete Block

Before we consider a reasonably general study of B blocks, it is instructive to study a simple design where every rater rates every classroom. This creates one complete block. Let K raters rate each of J classrooms on n instruction items. Based on the model presented below, this section expresses reliability as a function of variance components and sample sizes, derive the estimators for the variance components and thus for the reliability, and then expresses power to detect a desired level of reliability and a confidence interval for reliability.

2.1 Model

The model of interest is expressed as

$$Y_{ijk} = \mu + \alpha_j + \beta_k + (\alpha\beta)_{jk} + \epsilon_{ijk} \quad (1)$$

where Y_{ijk} is a rating score, μ is the grand mean, $\alpha_j \sim N(0, \sigma^2_\alpha)$ is the main effect of classroom j, $\beta_k \sim N(0, \sigma^2_\beta)$ is the main effect of rater k, $(\alpha\beta)_{jk} \sim N(0, \sigma^2_{\alpha\beta})$ is the interaction effect between classroom j and rater k and ϵ_{ijk} is a random error involving instruction item i for $i = 1, 2, \cdots, n$, $j = 1, 2, \cdots, J$ and $k = 1, 2, \cdots, K$. A more elaborate model may specify main and interaction effects involving items. For example, we may express ϵ_{ijk} as an addition of orthogonally decomposed random components $\epsilon_{ijk} = \gamma_i + (\gamma\alpha)_{ij} + (\gamma\beta)_{ik} + (\gamma\alpha\beta)_{ijk}$ for the main effect $\gamma_i \sim N(0, \sigma^2_\gamma)$ of item i, the interaction effect $(\gamma\alpha)_{ij} \sim N(0, \sigma^2_{\gamma\alpha})$ between item i and classroom j, the interaction effect $(\gamma\beta)_{ik} \sim N(0, \sigma^2_{\gamma\beta})$ between item i and rater k and the three-way interaction effect $(\gamma\alpha\beta)_{ijk} \sim N(0, \sigma^2_{\gamma\alpha\beta})$ among item i, classroom j and rater k. In the generalizability theory framework (Brennan 2001), this model has a fully

crossed $item \times classroom \times rater$ design where classrooms are the objects of measurement and where raters and items are random facets. However, we want to focus our paper on the basic logic of reliability, confidence intervals and power and minimize mathematically complicated expressions. Therefore, we assume $\gamma_i = (\gamma\alpha)_{ij} = (\gamma\beta)_{ik} = 0$ which simplifies the fully crossed model to the model (1) where $\epsilon_{ijk} = (\gamma\alpha\beta)_{ijk}$. As we will show in the next section, this assumption does not restrict the fully crossed model as much as it seems to do. Having clarified the logical core of the problem in this comparatively simple case, the next steps in our future research will involve generalizing that logic to a broader range of designs and outcome types. We illustrate the developed methods via analysis of data from Classroom Assessment Scoring System (CLASS) that uses a 7 point scale for each item where the median responses tend to be around 3 or 4 with quite symmetric distributions (La Paro et al. 2004; Raudenbush et al. 2010).

2.2 Reliability

The measure of classroom quality is $\mu + \alpha_j$ in the model (1). The purpose of the studies we aim to inform is to compare classrooms on a measure of classroom quality. Therefore, we focus on the deviation score α_j whose observed measure is $\hat{\alpha}_j = \bar{Y}_{.j.} - \bar{Y}_{...}$ for $\bar{Y}_{.j.} = \sum_{k=1}^{K}\sum_{i=1}^{n} Y_{ijk}/(nK)$ and $\bar{Y}_{...} = \sum_{j=1}^{J} \bar{Y}_{.j.}/J$. We define reliability of the observed classroom effects given σ^2_α, $\sigma^2_{\alpha\beta}$ and σ^2 as

$$\lambda_\alpha(n, K) \quad = \quad \frac{\sigma^2_\alpha}{\sigma^2_\alpha + \sigma^2_{\alpha\beta}/K + \sigma^2/(nK)} \tag{2}$$

which is the correlation between the observed effects of a classroom over a pair of randomly parallel realizations of the measurement procedure. That is, $\lambda_\alpha(n, K) = cor(\hat{\alpha}^1_j, \hat{\alpha}^2_j)$ where $\hat{\alpha}^1_j = \bar{Y}^1_{.j.} - \bar{Y}^1_{...}$ is an estimator of α_j based on a random sample of K raters and n items from large populations of raters and items, respectively, and $\hat{\alpha}^2_j = \bar{Y}^2_{.j.} - \bar{Y}^2_{...}$ is a second estimator based on a second random sample of K raters and n items from the same populations. In

the generalizability theory framework, the observed mean and universe scores for classroom j are $\bar{Y}_{.j.}$ and $\mu + \alpha_j$ respectively. Then, the observed and universe deviation scores for the classroom are $\bar{Y}_{.j.} - E_j(\bar{Y}_{.j.}) = \alpha_j + \overline{(\alpha\beta)}_{j.} + \bar{\epsilon}_{.j.}$ and α_j respectively for the expectation E_j taken over the units of classrooms, $\overline{(\alpha\beta)}_{j.} = \sum_k (\alpha\beta)_{jk}/K$ and $\bar{\epsilon}_{.j.} = \sum_k \sum_n \epsilon_{ijk}/(nK)$. The variance of the difference in the two deviation scores yields relative error variance $var\left[\overline{(\alpha\beta)}_{j.} + \bar{\epsilon}_{.j.}\right] = \sigma^2_{\alpha\beta}/K + \sigma^2/(nK)$. Reliability equation (2) is a generalizability coefficient,[2] the ratio of $var(\mu + \alpha_j)$ to $var(\mu + \alpha_j) + var\left[\overline{(\alpha\beta)}_{j.} + \bar{\epsilon}_{.j.}\right]$ (Brennan 2001). Equation (2) is also the reliability of the classroom effect estimator $\hat{\alpha}_j$, $var(\alpha_j)/var(\hat{\alpha}_j) = \frac{J}{J-1}\lambda_\alpha(n, K) \propto \lambda_\alpha(n, K)$.

2.2.1 What Determines Reliability?

Equation (2) depends on variances and sample sizes. The three variance components of the reliability estimate work in the following ways:

1. The more heterogenous the classrooms are in quality, the larger will be the between-classroom variance σ^2_α and therefore the larger will be the reliability;

2. The more raters disagree when they observe a classroom, the larger will be the rater-by-classroom variability $\sigma^2_{\alpha\beta}$ and therefore the lower the reliability;

3. The more inconsistent the items are, the larger will be the item variance σ^2 and therefore the lower the reliability.

2.2.2 How Should Resources Be Allocated?

Because Equation (2) also depends on two sample sizes, it reveals how to allocate resources as follows:

[2]The $item \times classroom \times rater$ design has a generalizability coefficient $\frac{\sigma^2_\alpha}{\sigma^2_\alpha + \sigma^2_{\alpha\beta}/K + \sigma^2_{\gamma\alpha}/n + \sigma^2_{\gamma\alpha\beta}/(nK)}$ which simplifies to the equation (2) under a single assumption of no item-by-classroom interaction effect, i.e. $\sigma^2_{\gamma\alpha} = 0$. Therefore, the model (1) does not restrict the fully crossed design as much as it seems to do in terms of the reliability.

1. The larger the number K of raters who observe a given classroom, the higher the reliability, assuming either $\sigma^2_{\alpha\beta} > 0$ or $\sigma^2 > 0$ or both. Increasing K will be especially helpful in increasing reliability when $\sigma^2_{\alpha\beta}$ or σ^2 is large.

2. Adding items will increase the reliability whenever $\sigma^2 > 0$. Adding items will be especially helpful when σ^2 is large.

3. Increasing n or K or both will increase the reliability when $\sigma^2_{\alpha\beta}$ is small but σ^2 is large. However, there may be tradeoffs. Suppose, for example, it is very expensive to train raters and very cheap to increase the number of items. Then, increasing n rather than K will be much more cost effective in boosting reliability. In contrast, if increasing K is cheap but increasing n is expensive, for example, in generating and validating a new instrument, then increasing K will be more cost effective than will increasing n, and this assertion will be even more true when $\sigma^2_{\alpha\beta}$ is also appreciable.

2.3 Estimation

Reasonable effect estimators are $\hat{\mu} = \bar{Y}_{...}$, $\hat{\alpha}_j = \bar{Y}_{.j.} - \bar{Y}_{...}$, $\hat{\beta}_k = \bar{Y}_{..k} - \bar{Y}_{...}$, and $\widehat{(\alpha\beta)}_{jk} = \bar{Y}_{.jk} - \bar{Y}_{.j.} - \bar{Y}_{..k} + \bar{Y}_{...}$ for $\bar{Y}_{.jk} = \sum_{i=1}^{n} Y_{ijk}/n$ and $\bar{Y}_{..k} = \sum_{j=1}^{J} \bar{Y}_{.jk}/J$. The sums of squares are $SSA = nK \sum_j \hat{\alpha}_j^2$, $SSB = nJ \sum_k \hat{\beta}_k^2$, $SSAB = n \sum_k \sum_j \widehat{(\alpha\beta)}_{jk}^2$ and $SSE = \sum_k \sum_j \sum_i \hat{\epsilon}_{ijk}^2$ for $\hat{\epsilon}_{ijk} = Y_{ijk} - \hat{\mu} - \hat{\alpha}_j - \hat{\beta}_k - \widehat{(\alpha\beta)}_{jk}$ so that the expected mean squares are

$$\begin{aligned} E(MSA) &= E[SSA/(J-1)] = nK\sigma^2_{\alpha} + n\sigma^2_{\alpha\beta} + \sigma^2, \\ E(MSB) &= E[SSB/(K-1)] = nJ\sigma^2_{\beta} + n\sigma^2_{\alpha\beta} + \sigma^2, \\ E(MSAB) &= E\{SSAB/[(J-1)(K-1)]\} = n\sigma^2_{\alpha\beta} + \sigma^2, \\ E(MSE) &= E\{SSE/[JK(n-1)]\} = \sigma^2. \end{aligned} \tag{3}$$

Equating each mean square to its expectation and solving for the parameters yield

$$\begin{aligned} \hat{\sigma}^2 &= MSE \\ \hat{\sigma}^2_{\alpha\beta} &= (MSAB - MSE)/n \\ \hat{\sigma}^2_{\alpha} &= (MSA - MSAB)/(nK) \\ \hat{\sigma}^2_{\beta} &= (MSB - MSAB)/(nJ). \end{aligned} \tag{4}$$

A reasonable estimator for the reliability (2) is

$$\hat{\lambda}_\alpha = 1 - \frac{MSAB}{MSA}. \tag{5}$$

The mean rating for classroom j as the measure of teaching quality is of interest to educators. The estimator for the mean rating is $\hat{\mu} + \hat{\alpha}_j = \bar{Y}_{.j.} \sim N(\mu, \sigma^2_j)$ for $\sigma^2_j = \sigma^2_\alpha + \sigma^2_\beta/K + \sigma^2_{\alpha\beta}/K + \sigma^2/(nK)$. Then, a $(1-s) \times 100\%$ confidence interval for the mean rating $\mu + \alpha_j$ of classroom j is

$$\bar{Y}_{.j.} \pm t_{1-s/2,JK} \times \hat{\sigma}_j \tag{6}$$

for the $(1 - s/2) * 100$th percentile $t_{1-s/2,JK}$ from the t distribution with JK degrees of freedom and $\hat{\sigma}^2_j = \hat{\sigma}^2_\alpha + \hat{\sigma}^2_\beta/K + \hat{\sigma}^2_{\alpha\beta}/K + \hat{\sigma}^2/(nK)$.

2.4 Hypothesis Test and Power

Our interest in this section is in deriving the power with which we can reject a null hypothesis that the reliability equals a given minimum bound in favor of an alternative hypothesis that the reliability will exceed the given minimum bound. In the appendix, we

show that a ratio of unreliability to unreliability estimator for $B = 1$ is

$$F(\lambda_\alpha) = \frac{1-\lambda_\alpha}{1-\hat{\lambda}_\alpha} \sim F_{J-1,(J-1)(K-1)} \tag{7}$$

for $\frac{1}{1-\hat{\lambda}_\alpha} = \frac{MSA}{MSAB}$ and $F_{J-1,(J-1)(K-1)}$, the F distribution with $J-1$ numerator and $(J-1)(K-1)$ denominator degrees of freedom. Let $f_{J-1,(J-1)(K-1),\beta}$ denote the β-quantile from $F_{J-1,(J-1)(K-1)}$. We have the following:

Theorem 2.1 *Suppose that we want to test $H_0 : \lambda_\alpha = \lambda_{\alpha 0}$ against an alternative hypothesis $H_a : \lambda_\alpha > \lambda_{\alpha 0}$ at a significance level s and let $f_0^s = f_{J-1,(J-1)(K-1),1-s}$. Under the H_a, the power to detect $\lambda_\alpha = \lambda_{\alpha 1} > \lambda_{\alpha 0}$ is*

$$P\left[F(\lambda_{\alpha 1}) > f_0^s \frac{1-\lambda_{\alpha 1}}{1-\lambda_{\alpha 0}}\right]. \tag{8}$$

Proof The random variable (7) implies that the test statistic under $H_0 : \lambda_\alpha = \lambda_{\alpha 0}$ is $F(\lambda_{\alpha 0}) = \frac{1-\lambda_{\alpha 0}}{1-\hat{\lambda}_\alpha} \sim F_{J-1,(J-1)(K-1)}$ for $\lambda_{\alpha 0} \geq 0$ such that $P\left[F(\lambda_{\alpha 0}) > f_0^s\right] = s$. Then, under $H_a : \lambda_\alpha > \lambda_{\alpha 0}$, the power to detect $\lambda_\alpha = \lambda_{\alpha 1} > \lambda_{\alpha 0}$ is equal to $P\left[\frac{1-\lambda_{\alpha 0}}{1-\hat{\lambda}_\alpha} > f_0^s\right] = P\left[F(\lambda_{\alpha 1}) > f_0^s \frac{1-\lambda_{\alpha 1}}{1-\lambda_{\alpha 0}}\right]$ where $F(\lambda_{\alpha 1}) = \frac{1-\lambda_{\alpha 1}}{1-\hat{\lambda}_\alpha} \sim F_{J-1,(J-1)(K-1)}$. ∎

The power depends positively on $\lambda_{\alpha 1}$ but negatively on f_0^s and $\lambda_{\alpha 0}$. Consequently, given f_0^s and $\lambda_{\alpha 0}$, the higher the n or the K in the study given variance components, the higher the $\lambda_{\alpha 1}(n, K)$ in equation (2), and thus the higher the power. Moreover, the larger the number of classrooms J or raters K in the study, the lower the f_0^s and thus the higher the power although the impact on f_0^s of K is relatively weak to that of J.

Equation (8) may be reexpressed to find an effective reliability size

$$\lambda_{\alpha,1-\beta} \;=\; 1 - f_{J-1,(J-1)(K-1),\beta} \times (1-\lambda_{\alpha 0})/f_0^s \tag{9}$$

that achieves a desired power of $(1-\beta)$ given $\lambda_{\alpha 0}$ and f_0^s. In designing a study, the study planner may select a desired reliability size with adequate power in equation (9) and then

set it equal to the reliability in equation (2).

2.5 Confidence Interval for Reliability

In this section, we derive a confidence interval for the reliability of measurement whose width represents the level of uncertainty involved in a study design. The confidence interval facilitates selection of the design with the minimal uncertainty among feasible study designs. The following theorem expresses a confidence interval for λ_α:

Theorem 2.2 *Let* $(1-\hat{\lambda}_\alpha) = \frac{MSAB}{MSA}$. *A* $(1-s)\times 100$ *% confidence interval for* λ_α *is*

$$1-(1-\hat{\lambda}_\alpha)\times\left(f_{J-1,(J-1)(K-1),1-s/2}, f_{J-1,(J-1)(K-1),s/2}\right). \tag{10}$$

Proof The random variable (7) implies

$$\begin{aligned} 1-s &= P\left[(1-\hat{\lambda}_\alpha)f_{J-1,(J-1)(K-1),s/2} < 1-\lambda_\alpha < (1-\hat{\lambda}_\alpha)f_{J-1,(J-1)(K-1),1-s/2}\right] \\ &= P\left[1-(1-\hat{\lambda}_\alpha)f_{J-1,(J-1)(K-1),s/2} > \lambda_\alpha > 1-(1-\hat{\lambda}_\alpha)f_{J-1,(J-1)(K-1),1-s/2}\right]. \end{aligned}$$ ∎

The width (10) depends positively on two quantities:

$$1-\hat{\lambda}_\alpha = \frac{MSAB}{MSA} = 1-\frac{\hat{\sigma}^2_\alpha}{\hat{\sigma}^2_\alpha+\hat{\sigma}^2_{\alpha\beta}/K+\hat{\sigma}^2/(nK)}, \tag{11}$$

$$\left(f_{J-1,(J-1)(K-1),s/2}, f_{J-1,(J-1)(K-1),1-s/2}\right). \tag{12}$$

Therefore, given the variance component estimates, the larger the n or the K in the study, the lower the unreliability estimator (11), and thus the narrower the confidence interval (10). Moreover, the larger the number of classrooms J or raters K ceteris paribus in the study, the narrower the interval (12) and thus, the narrower the confidence interval (10) although the impact on the width (12) of K is relatively weak to that of J. Given variance component estimates and sample sizes, the model (1) may be simulated to generate the confidence interval (10). An illustrative simulation will be given for general $B \geq 1$ in this paper.

2.6 Expected Confidence Interval for Reliability

Simulating confidence interval (10) for reliability may take too long to generate useful results. It is helpful to have an alternative method that enables study planners to compare the uncertainty involved in multiple study designs within a reasonable amount of time. To do that, we express a $(1-s)\times 100\%$ *expected* confidence interval (10) as

$$1-\frac{(J-1)(1-\lambda_\alpha)}{J-3}\times\left(f_{J-1,(J-1)(K-1),1-s/2},f_{J-1,(J-1)(K-1),s/2}\right) \tag{13}$$

where $E\left(\frac{MSAB}{MSA}\right)=\frac{(J-1)(1-\lambda_\alpha)}{J-3}$ for $\frac{MSAB}{MSA(1-\lambda_\alpha)}\sim F_{(J-1)(K-1),J-1}$. Then, the study planner may explore multiple study designs to select one that minimizes the expected width (13), that is, the expected uncertainty.

A study design has its true population reliability (2) as a function of the sample sizes given the true variance components. If randomly parallel measurement procedures or replications of the study design were to repeat many times and produce as many estimated confidence intervals (10), then $(1-s)\times 100\%$ of the intervals would capture the true reliability. The expected interval (13) represents the average estimated interval. The smaller the s, the more likely this interval is to contain the true reliability.

3 B Incomplete Blocks

It is costly to have every rater rate every classroom. A study becomes cost effective if KB raters and JB classrooms are divided into B incomplete blocks. Each block has J classrooms and randomly assigned K raters where each rater rates n instruction items for each classroom. This creates a balanced incomplete randomized block design.

3.1 Model

A reasonably general model for such a design is

$$Y_{ijkb} = \mu + \gamma_b + \alpha_{jb} + \beta_{kb} + (\alpha\beta)_{jkb} + \epsilon_{ijkb} \quad (14)$$

where Y_{ijkb} is a rating score, μ is the grand mean, $\gamma_b \sim N(0, \sigma^2_\gamma)$ is the effect of block b, $\alpha_{jb} \sim N(0, \sigma^2_\alpha)$ is the main effect of classroom j, $\beta_{kb} \sim N(0, \sigma^2_\beta)$ is the main effect of rater k, $(\alpha\beta)_{jkb} \sim N(0, \sigma^2_{\alpha\beta})$ is the interaction effect between classroom j and rater k and $\epsilon_{ijkb} \sim N(0, \sigma^2)$ is a random error involving instruction item i for $i = 1, 2, \cdots, n$, $j = 1, \cdots, J$, $k = 1, 2, \cdots, K$ and $b = 1, 2, \cdots, B$. Equation (14) is the model (1) for $B = 1$. An effective blocking scheme yields classrooms more homogeneous within than across blocks. Schools or school districts, for example, may be such blocks. Such a blocking scheme controls for the block differences due to, for example, high-performing and low-performing schools such that the inferences on the classroom quality are precise and generalizable to the classrooms in all such schools. Because raters are randomly assigned to blocks, it is reasonable to assume that the rater effects are not different across blocks.

3.2 Reliability

The reliability of $\hat{\alpha}_{jb} = \bar{Y}_{.j.b} - \bar{Y}_{...b}$ for $\bar{Y}_{.j.b} = \sum_k \sum_i Y_{ijkb}/(nK)$ and $\bar{Y}_{...b} = \sum_j Y_{.j.b}/J$ is identical to the equation (2). Therefore, the statements about the equation (2) are also valid for the reliability.

3.3 Estimation

Reasonable estimators for μ, γ_b, α_{jb}, β_{kb} and $(\alpha\beta)_{jkb}$ are $\hat{\mu} = \bar{Y}_{....}$, $\hat{\gamma}_b = \bar{Y}_{...b} - \bar{Y}_{....}$, $\hat{\alpha}_{jb} = \bar{Y}_{.j.b} - \bar{Y}_{...b}$, $\hat{\beta}_{kb} = \bar{Y}_{..kb} - \bar{Y}_{...b}$, $\widehat{(\alpha\beta)}_{jkb} = \bar{Y}_{.jkb} - \bar{Y}_{.j.b} - \bar{Y}_{..kb} + \bar{Y}_{...b}$ and $\hat{\epsilon}_{ijkb} = Y_{ijkb} - \bar{Y}_{.jkb}$ for $\bar{Y}_{.jkb} = \sum_{i=1}^n Y_{ijkb}/n$, $\bar{Y}_{..kb} = \sum_{j=1}^J \bar{Y}_{.jkb}/J$ and $\bar{Y}_{....} = \sum_b \bar{Y}_{...b}/B$. The sums of squares are $SSG = nJK \sum_b \hat{\gamma}^2_b$, $SSA = nK \sum_b \sum_j \hat{\alpha}^2_{jb}$, $SSB = nJ \sum_b \sum_k \hat{\beta}^2_{kb}$, $SSAB = n \sum_b \sum_k \sum_j \widehat{(\alpha\beta)}^2_{jkb}$ and

$SSE = \sum_b \sum_k \sum_j \sum_i \hat{\epsilon}^2_{ijkb}$. Then, the expected mean squares are

$$\begin{aligned}
E(MSG) &= E[SSG/(B-1)] = nJK\sigma^2_\gamma + nK\sigma^2_\alpha + nJ\sigma^2_\beta + n\sigma^2_{\alpha\beta} + \sigma^2 \qquad (15)\\
E(MSA) &= E\{SSA/[(J-1)B]\} = nK\sigma^2_\alpha + n\sigma^2_{\alpha\beta} + \sigma^2\\
E(MSB) &= E\{SSB/[(K-1)B]\} = nJ\sigma^2_\beta + n\sigma^2_{\alpha\beta} + \sigma^2\\
E(MSAB) &= E\{SSAB/[(J-1)(K-1)B]\} = n\sigma^2_{\alpha\beta} + \sigma^2\\
E(MSE) &= E\{SSE/[JKB(n-1)]\} = \sigma^2.
\end{aligned}$$

By equating each mean square to its expectation and solving for the parameters, we obtain the variance estimators (4), the reliability estimator (5) and

$$\hat{\sigma}^2_\gamma = (MSG - MSA - MSB + MSAB)/(nJK). \qquad (16)$$

The mean rating for classroom j within block b as the classroom measure of teaching quality is $\mu+\gamma_b+\alpha_{jb}$. The estimator is $\hat{\mu}+\hat{\gamma}_b+\hat{\alpha}_{jb} = \bar{Y}_{.j.b} \sim N(\mu, \sigma^2_{jb})$ for $\sigma^2_{jb} = \sigma^2_\gamma + \sigma^2_\alpha/J + \sigma^2_\beta/K + \sigma^2_{\alpha\beta}/(JK) + \sigma^2/(nJK)$. A $(1-s)\times 100\%$ confidence interval for a classroom mean rating $\mu+\gamma_b+\alpha_{jb}$ is

$$\bar{Y}_{.j.b} \pm t_{1-s/2,JKB} \times \hat{\sigma}_{jb} \qquad (17)$$

for the $(1-s/2)*100$th percentile $t_{1-s/2,JKB}$ from the t distribution with JKB degrees of freedom and $\hat{\sigma}^2_{jb} = \hat{\sigma}^2_\gamma + \hat{\sigma}^2_\alpha/J + \hat{\sigma}^2_\beta/K + \hat{\sigma}^2_{\alpha\beta}/(JK) + \hat{\sigma}^2/(nJK)$.

3.4 Hypothesis Tests and Power

In this section, we express the power to reject $H_0 : \lambda_\alpha = \lambda_{\alpha 0}$ in favor of an alternative hypothesis that the reliability will exceed the given minimum bound $\lambda_{\alpha 0}$. In the appendix,

we show that the ratio of unreliability to unreliability estimator is

$$F(\lambda_\alpha) = \frac{1-\lambda_\alpha}{1-\hat{\lambda}_\alpha} \sim F_{(J-1)B,(J-1)(K-1)B} \tag{18}$$

for $1-\hat{\lambda}_\alpha = MSAB/MSA$. We have the following theorem:

Theorem 3.1 *Suppose that we want to test a null hypothesis $H_0 : \lambda_\alpha = \lambda_{\alpha 0}$ against an alternative hypothesis $H_a : \lambda_\alpha > \lambda_{\alpha 0}$ at a significance level s and let $f_0^s = f_{(J-1)B,(J-1)(K-1)B,1-s}$. Under the H_a, the power to detect $\lambda_\alpha = \lambda_{\alpha 1} > \lambda_{\alpha 0}$ is equation (8).*

Proof The random variable (18) implies the test statistic $F(\lambda_{\alpha 0}) = \frac{1-\lambda_{\alpha 0}}{1-\hat{\lambda}_\alpha} \sim F_{(J-1)B,(J-1)(K-1)B}$ under $H_0 : \lambda_\alpha = \lambda_{\alpha 0}$ for $\lambda_{\alpha 0} \geq 0$ such that $P\left[F(\lambda_{\alpha 0}) > f_0^s\right] = s$. Then, under $H_a : \lambda_\alpha > \lambda_{\alpha 0}$, the power to detect $\lambda_\alpha = \lambda_{\alpha 1} > \lambda_{\alpha 0}$ is $P\left[\frac{1-\lambda_{\alpha 0}}{1-\hat{\lambda}_\alpha} > f_0^s\right] = P\left[F(\lambda_{\alpha 1}) > f_0^s \frac{1-\lambda_{\alpha 1}}{1-\lambda_{\alpha 0}}\right]$ where $F(\lambda_{\alpha 1}) \sim F_{(J-1)B,(J-1)(K-1)B}$. ∎

Just as in the case where $B = 1$, we see that the power depends positively on $\lambda_{\alpha 1}$ but negatively on f_0^s and $\lambda_{\alpha 0}$ and the same observations about the power (8) above also applies here. Moreover, the larger the JB ceteris paribus, the lower the f_0^s and thus the higher the power. Given $\lambda_{\alpha 0}$ and f_0^s, the power (8) may be reexpressed as an effective reliability size

$$\lambda_{\alpha,1-\beta} \;=\; 1 - f_{(J-1)B,(J-1)(K-1)B,\beta} \times (1-\lambda_{\alpha 0})/f_0^s \tag{19}$$

that achieves a desired power of $(1-\beta)$.

3.5 Estimating a Confidence Interval for Reliability

Our aim in this section is to express a confidence interval for λ_α that will be achieved in the study and to illustrate how the width changes across multiple study designs on average.

Theorem 3.2 *Let $(1-\hat{\lambda}_\alpha) = \frac{MSAB}{MSA}$. A $(1-s)\times 100\%$ confidence interval for λ_α is*

$$1 - (1-\hat{\lambda}_\alpha) \times \left(f_{(J-1)B,(J-1)(K-1)B,1-s/2}, f_{(J-1)B,(J-1)(K-1)B,s/2}\right). \tag{20}$$

This result follows from the proof for Theorem 2.2 where we replace $f_{(J-1),(J-1)(K-1),s/2}$ and $f_{(J-1),(J-1)(K-1),1-s/2}$ with $f_{(J-1)B,(J-1)(K-1)B,s/2}$ and $f_{(J-1)B,(J-1)(K-1)B,1-s/2}$ respectively. The confidence interval (20) depends positively on two quantities: equation (11) and

$$\left(f_{(J-1)B,(J-1)(K-1)B,s/2}, f_{(J-1)B,(J-1)(K-1)B,1-s/2}\right). \tag{21}$$

Consequently, the larger the n or the K given variance estimates, the lower the unreliability estimator (11), the narrower the width (20) and thus the less uncertainty will be involved in the study. In addition, the larger the number of classrooms JB or raters K ceteris paribus in the study, the narrower the interval (21) and thus, the narrower the confidence interval (20). In the next section, we show how to simulate the confidence interval (20) based on the model (14) and to compare the uncertainty across multiple study designs.

The expected value of the confidence interval (20) is a useful alternative in comparing the amount of uncertainty involved across multiple study designs in a reasonable amount of time. A $(1-s)\times 100\%$ expected confidence interval may be expressed as

$$1-\frac{(J-1)B(1-\lambda_\alpha)}{(J-1)B-2}\times\left(f_{(J-1)B,(J-1)(K-1)B,1-s/2}, f_{(J-1)B,(J-1)(K-1)B,s/2}\right) \tag{22}$$

where $E\left(\frac{MSAB}{MSA}\right)=\frac{(J-1)B(1-\lambda_\alpha)}{(J-1)B-2}$ for $\frac{MSAB}{MSA(1-\lambda_\alpha)}\sim F_{(J-1)(K-1)B,(J-1)B}$.

3.6 Properties of Reliability Estimator (5)

From $\frac{MSAB}{MSA(1-\lambda_\alpha)}\sim F_{(J-1)(K-1)B,(J-1)B}$, we have $E(\hat{\lambda}_\alpha)=1-\frac{(J-1)B(1-\lambda_\alpha)}{(J-1)B-2}\to\lambda_\alpha$ as $JB\to\infty$. In addition, $var(\hat{\lambda}_\alpha)=var\left(\frac{MSAB}{MSA}\right)\approx\frac{2K(1-\lambda_\alpha)^2}{(K-1)(J-1)B}$ for JB large that can be made arbitrarily small as JB increases. Because $E(\hat{\lambda}_\alpha-\lambda_\alpha)^2<E(\hat{\lambda}_\alpha^2)-2E(\hat{\lambda}_\alpha)^2+\lambda_\alpha^2=var(\hat{\lambda}_\alpha)+[\lambda_\alpha^2-E(\hat{\lambda}_\alpha)^2]\to 0$ as $JB\to\infty$, for $\epsilon>0$, we have

$$\begin{aligned} P(|\hat{\lambda}_\alpha-\lambda_\alpha|\geq\epsilon) &= P\left[(\hat{\lambda}_\alpha-\lambda_\alpha)^2\geq\epsilon^2\right] \\ &\leq E(\hat{\lambda}_\alpha-\lambda_\alpha)^2/\epsilon^2\to 0 \text{ as } JB\to\infty \end{aligned}$$

by Markov's inequality. Therefore, $\hat{\lambda}_\alpha$ is asymptotically unbiased and consistent with its variance tending to zero as JB increases. The Theorems 3.1 and 3.2 depends on the exact distribution (18) based on the model (14).

3.7 Illustrative Example

In this section, we consider variance component estimates $(\sigma^2_\alpha, \sigma^2_\beta, \sigma^2_{\alpha\beta}, \sigma^2) = (0.11, 0.29, 0.11, 0.39)$ from the variance component analysis of the CLASS used in the Multi-State Study of Pre-Kindergarten (MSSPK) by the National Center for Early Development (Raudenbush et al. 2010) and show how to use the power (8) and the expected confidence interval (22) to be informative about the design for a study.

Given the variance components, Figure 1 shows the impact of alternative study designs on power (8) and reliability size (19) for testing $H_0 : \lambda_\alpha = 0.5$ VS $H_a : \lambda_\alpha > 0.5$ at a significance level $s = 0.05$. Graph (a) draws power (8) against the number of JB classrooms

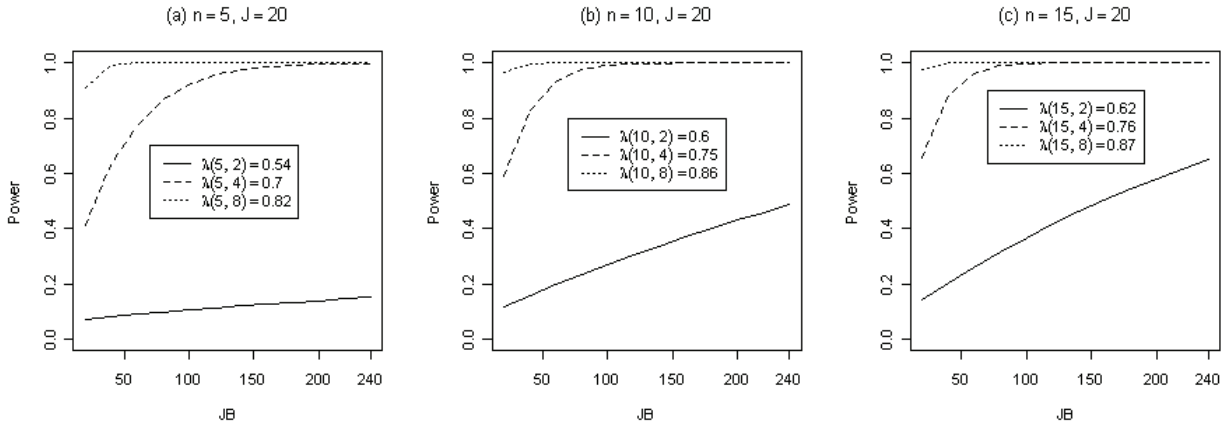

Figure 1: Each graph draws power (8) against JB for $K = 2, 4, 8$ given $J = 20$, $\lambda_{\alpha 0} = 0.5$ and $f_0^{0.05}$. Across graphs, n increases from 5 to 10 to 15. The legend shows $\lambda(n, K)$.

ranging 20 to 240 classrooms given $n = 5$ items and $J = 20$ classrooms per block. The legend displays $\lambda(n, K)$ and indicates that the solid, thick-dotted and thin-dotted lines are for the frequency of rating each classroom by 2, 4, or 8 raters respectively. Graphs (b) and (c) are drawn identically except for $n = 10$ and $n = 15$ items given respectively. Within the range of

the study design, it is hard to obtain adequate power and reliability size given $K = 2$ raters. As the number of K raters increases given other sample sizes, the power and the reliability size increase markedly. The impact of increasing n from 5 to 10 items is noticeable on the power as well as the reliability, but the impact grows weak when n changes from 10 to 15 items, ceteris paribus. Increasing the number of JB classrooms increases the power but does not affect the reliability size given variances and other sample sizes. Graph (a) also reveals that increasing the number of JB classrooms does not yield decent power and reliability size with low sample sizes $(n, K) = (5, 2)$.

To illustrate the general utility of the expected confidence interval (22), Figure 2 draws $\lambda_\alpha(n, K)$ and the interval on the vertical axis that will be achieved against alternative study designs (a) K given $(n, J, B) = (15, 20, 3)$; (b) n given $(K, J, B) = (2, 20, 3)$; (c) J given $(n, K, B) = (15, 2, 3)$; and (d) B given $(n, K, J) = (15, 2, 20)$ on the horizontal axis. A

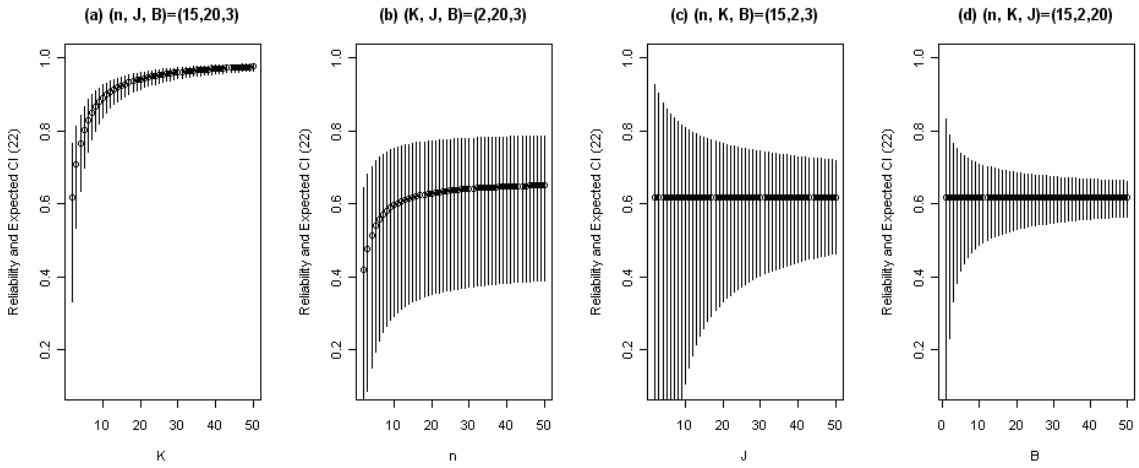

Figure 2: Graphs (a) to (d) draw the reliability and the expected 95% confidence interval (22) that will be achieved against study designs (a) K given $(n, J, B) = (15, 20, 3)$; (b) n given $(K, J, B) = (2, 20, 3)$; (c) J given $(n, K, B) = (15, 2, 3)$; and (d) B given $(n, K, J) = (15, 2, 20)$ on the horizontal axis.

reliability value is plotted as a dot. As the number of raters K increases, the reliability (2) rises and the expected confidence interval (22) shrinks markedly in the graph (a). The strong impact of increased K on the reliability and the expected width is most pronounced in the low range of $K \leq 5$. As the number of items n increases in the graph (b), the reliability

rises and the expected confidence width narrows at low values of n, but the impact becomes very weak from $n \approx 10$. The reliability, not a function of J and B given variances, stays constant in the graphs (c) and (d). As J or B grows in the graphs, the expected confidence intervals shrink quickly at low values, but slowly at high values of J or B. Therefore, the increased number of raters exhibits the most pronounced impact in increasing the reliability and narrowing the confidence interval (20) on average. Overall, the impact of sample sizes has a diminishing return in reducing the expected uncertainty in the study. This implies that it is important to have all sample sizes above the respective low ranges to capitalize on the strong impact.

Thus far, we have derived a confidence interval for reliability, a test for the hypothesis that the reliability meets a minimum bound, and the power of this test against alternative hypotheses that can apply to either G or D study. In the next section, we illustrate how to use a G study to compare the alternative designs for a D study.

4 Using a G study to Compare Alternative Designs for a D Study

We now show how to use the results from a G study to inform the design of a D study. The unknown true variance components are assumed to stay constant over time. If the G-study sample sizes were infinite, one would know the variance components. One could then simply substitute these known values into equation (2) to compute reliability for given sample sizes (n_D, K_D), to be used in the D study. One would then select the sample sizes for the D study that achieved acceptable reliability at minimum cost. In reality, of course, the G study will have a finite sample, so that the variance components will be estimated with error. The point estimates and their uncertainty estimates depend on the sample sizes of the G study. Consequently, the estimates as well as the G study design are informative about the design of a D study. In this section, we consider, as the G-study variance component estimates,

$\hat{\theta} = (\hat{\sigma}^2_\alpha, \hat{\sigma}^2_\beta, \hat{\sigma}^2_{\alpha\beta}, \hat{\sigma}^2) = (0.11, 0.29, 0.11, 0.39)$ from the MSSPK, and show how to use the G study to inform the design of a D study. We consider $H_0 : \lambda_\alpha = 0.5$ against $H_a : \lambda_\alpha > 0.5$ at a significance level s throughout this section.

4.1 Determining Sample Sizes in the D Study

The variance estimates produce a reliability estimate $\hat{\lambda}_\alpha = \hat{\lambda}_\alpha(n_D, K_D) = \lambda_\alpha(\hat{\theta}, n_D, K_D) = \frac{0.11}{0.11+0.11/K_D+0.39/(n_D K_D)}$ for a D-study design as an ordered pair (n_D, K_D). The test rejects H_0 if $P[F > (1-0.5)/(1-\hat{\lambda}_\alpha)] < s$ for $F \sim F_{(J_G-1)B_G,(J_G-1)(K_G-1)B_G}$ given a G-study design $(n_G, K_G, J_G B_G)$. Then, letting $\lambda_{\alpha 1} = \hat{\lambda}_\alpha$ under $H_a : \lambda_\alpha = \lambda_{\alpha 1} > 0.5$ produces the power of the test $P\left[F > f_0^s \frac{1-\lambda_{\alpha 1}}{1-0.5}\right]$ for $f_0^s = f_{(J_G-1)B_G,(J_G-1)(K_G-1)B_G,1-s}$ that will be achieved in the D study. A $(1-s)\times 100\%$ confidence interval for λ_α that is anticipated in the D study is

$$1 - (1-\hat{\lambda}_\alpha) \times \left(f_{(J_G-1)B_G,(J_G-1)(K_G-1)B_G,1-s/2}, f_{(J_G-1)B_G,(J_G-1)(K_G-1)B_G,s/2} \right). \tag{23}$$

Note that neither the confidence interval nor the power depends on n_G given the variance estimates from the G study. The MSSPK is a study of $J_G B_G = 240$ classrooms from preschool to third grade, 40 classrooms in each of 6 states ($B_G = 6$). Out of a total of 26 raters available from August 2001 to June 2002 (Raudenbush et al. 2010), we approximate $K_G \approx 4$ raters per block. The study planner may now select, among feasible D-study designs $\{(n_D, K_D)\}$ given the G-study results, the design (n_D^*, K_D^*) that produces the desired reliability, power and uncertainty represented as the width of the confidence interval (23) at minimal cost.

In the next two sections, we assume that $\hat{\theta}$ was obtained under hypothetical G-study designs $\{(n_G, K_G, J_G B_G)\}$, and illustrate, given the $\hat{\theta}$, the impacts of the G-study designs and feasible D-study designs (the sample sizes) on reliability, power and confidence interval that will be obtained in a D study.

4.2 Estimating Statistical Power in the D Study

To illustrate the impacts of D-study sample sizes (n_D, K_D) and G-study sample sizes $(n_G, K_G, J_G B_G)$ on reliability and power that will be obtained in the D study, we consider a small set of values of $K_D \in \{1, 2, 3, 4, 8\}$ for the D study. This simplification implies that, for any particular choice n_D, we consider 5 possible D-study designs, that is $(n_D, 1), (n_D, 2), (n_D, 3), (n_D, 4)$ and $(n_D, 8)$. Finally, we will consider a set of hypothetical G-study designs $\{(n_G, K_G, J_G B_G)\}$ with $K_G \in \{2, 4, 8\}$. Therefore, for every choice of $(n_G, J_G B_G)$, we have 3 possible G-study designs, that is $(n_G, 2, J_G B_G), (n_G, 4, J_G B_G)$ and $(n_G, 8, J_G B_G)$.

Figure 3 compares the impacts of D-study designs $\{(n_D, K_D)\}$ and G-study designs $\{(n_G, K_G, J_G B_G)\}$ on reliability and power given $J_G = 20$. The number of items n_D increases from 5 to 10 to 15 across the three rows given K_D while the number of raters K_D changes from 2 to 3 to 4 to 8 across the four columns given n_D. The graphs for $K_D = 1$ do not appear in the Figure because one rater produces the reliability sizes $\lambda_{\alpha 1}(n_D, 1)$ equal to 0.37, 0.42 and 0.45 for n_D =5, 10 and 15 respectively, below the acceptable minimum reliability $\lambda_{\alpha 0} = 0.5$. Therefore, $K_D = 1$ is not useful within the range of the G-study designs $\{(n_G, K_G, J_G B_G)\}$. In each graph, power on the vertical axis is drawn against the total number of classrooms $J_G B_G$ ranging 20 to 240 on the horizontal axis. The legend in graph (a) applies to all graphs and indicates that the solid, thick-dotted and thin-dotted lines draw power against $J_G B_G$ for $K_G = 2, 4$ and 8 given $J_G = 20$ respectively. The D-study design (n_D, K_D) and the $\lambda_{\alpha 1}(n_D, K_D)$ are displayed on top of each graph. As K_D increases across columns given n_D, both $\lambda_{\alpha 1}$ and power increase dramatically. As the number of n_D items increases across rows given K_D, both $\lambda_{\alpha 1}$ and power also increase. The reliability and the power increase appreciably from $n_D = 5$ to 10, but relatively weakly from $n_D = 10$ to 15, given K_D. Within each graph, power increases as the sample sizes $J_G B_G$ and K_G increase in the G study. Therefore, a D-study design (n_D, K_D) as well as a G-study design $(n_G, K_G, J_G B_G)$ are positively associated with the power.

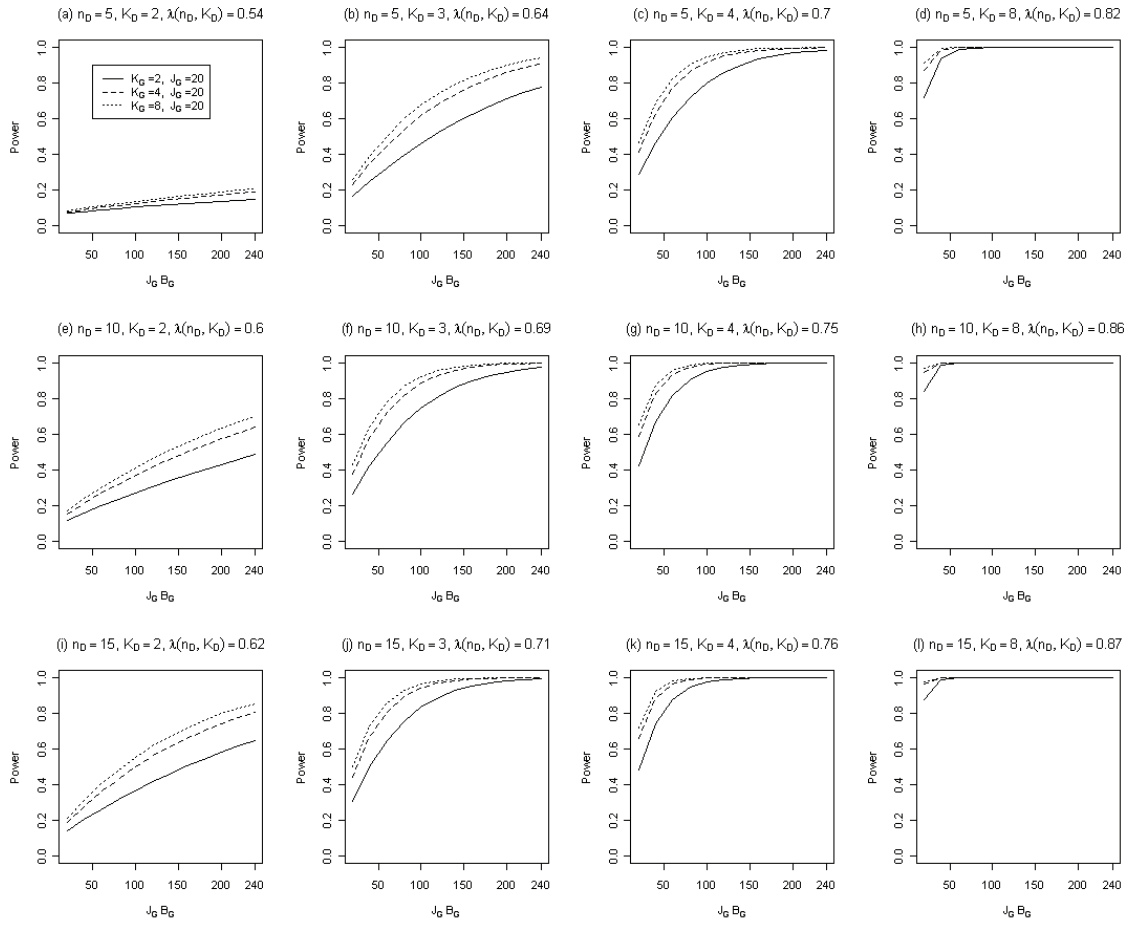

Figure 3: Each graph draws power (8) against $J_G B_G$ for $K_G = 2, 4, 8$ given $J_G = 20$, $\lambda_{\alpha 0} = 0.5$ and $f_0^{0.05} = f_{(J_G-1)B_G,(J_G-1)(K_G-1)B_G,0.95}$. Across rows, n_D increases from 5 to 10 to 15 while across columns, K_D changes from 2 to 3 to 4 to 8. The legend in graph (a) applies to all graphs.

4.3 Estimating Uncertainty in the D study

This section illustrates the impacts of D-study and G-study sample sizes on the confidence interval for reliability that will be obtained in the D study. To show the impacts, we may simulate a G study based on the model (14) given, say, $\mu = 1$, $\sigma_\gamma^2 = 0.2$ and θ equal to the G-study variance component estimates used above. As an illustrative example, we simulate a G-study design $(n_G, K_G, J_G B_G) = (5, 4, 10 \times 6)$ one thousand times to produce 1000 confidence intervals (23) for each of the feasible D-study designs $\{(n_D, K_D)\}$ considered in Figure 3. Then, the 1000 widths may be drawn in a boxplot to represent the uncertainty

that will be anticipated for a D study under the feasible D-study design (n_D, K_D). The narrower the widths overall, the lower the boxplot, and therefore the less uncertainty will be in the D study.

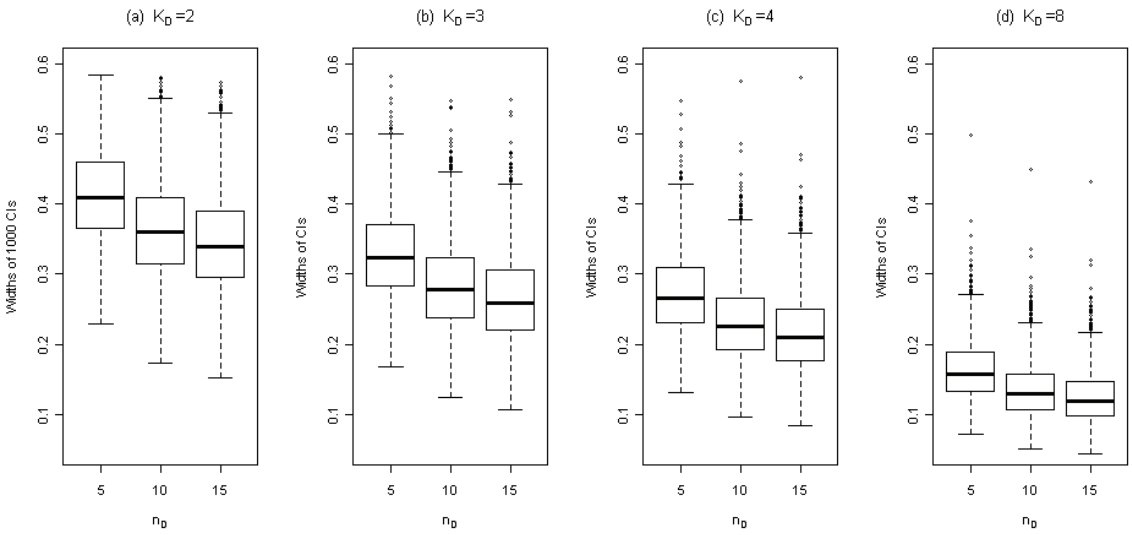

Figure 4: Graphs (a), (b), (c) and (d) are drawn given K_D =2,3,4 and 8 respectively. Each graph draws three boxplots for n_D =5,10 and 15 respectively given K_D. Each boxplot draws the widths of the 1000 confidence intervals (23) for λ_α given a specific D-study design (n_D, K_D). The horizontal axis labels the D-study sample size n_D.

Figure 4 displays the boxplots. The vertical axis for each boxplot represents the widths of the 1000 confidence intervals given a specific D-study design (n_D, K_D). Each of the four graphs (a) to (d) draws three boxplots for n_D =5,10 and 15 respectively given K_D. Graphs (a), (b), (c) and (d) are given K_D=2,3,4 and 8 respectively and, otherwise, drawn in the same way. Across graphs (a) to (d), we see the dramatic reduction in uncertainty that is anticipated in the D study as the number of K_D raters increases given n_D. Within each graph, we see the (relatively weak) decrease in the uncertainty as the number of n_D items increases given K_D.

Although the boxplots in Figure 4 are revealing, they are all drawn given a single G-study design $(n_G, K_G, J_G B_G) = (5, 4, 10 \times 6)$. Unlike the plots in Figure 3 on power, they do not show us the impacts of multiple G-study and D-study designs on the uncertainty anticipated in the D study. Moreover, it took six hours to simulate the G-study design to produce Figure

4 by a fully automated procedure in the statistical software package R on a laptop computer with a 2.53 GHz processor and 3 GB memory. Although it takes only minutes to produce the results with smaller designs, it takes much longer to produce the results with a larger design. Therefore, it is desirable to use an alternative method that compares the impacts of multiple G-study and D-study designs on the uncertainty within a reasonable amount of time without regard to the sample sizes. The expected confidence interval (22) for reliability provides one such method.

Under $H_a : \lambda_\alpha = \lambda_{\alpha 1} > 0.5$, the expected confidence interval that will be obtained in the D study may be expressed as

$$1 - \frac{(J_G - 1)B_G(1 - \lambda_{\alpha 1})}{(J_G - 1)B_G - 2} \times \left(f_{(J_G-1)B_G,(J_G-1)(K_G-1)B_G,1-s/2}, f_{(J_G-1)B_G,(J_G-1)(K_G-1)B_G,s/2} \right). \quad (24)$$

Figure 5 illustrates the impacts of D-study sample sizes and G-study sample sizes on reliability and the expected confidence interval (24). It has the main heading on top of each graph, G- and D-study sample sizes, horizontal axis labels and the legend that are identical to and set up identically to those of Figure 3. It only replaces each graph of power in Figure 3 with one of the expected confidence intervals (24) represented on the vertical axis. Each graph has solid, thick-dotted and thin-dotted confidence intervals for $K_G = 2, 4$ and 8 respectively given the same $J_G B_G$, which had to be jittered not to overlap. In each graph, the expected confidence interval and the reliability size $\lambda_{\alpha 1}$ (the dots) on the vertical axis are drawn against the total number of classrooms $J_G B_G$ ranging 20 to 240 given $J_G = 20$ on the horizontal axis. The D-study design (n_D, K_D) and the $\lambda_{\alpha 1}(n_D, K_D)$ are displayed on top of each graph in an identical way to the corresponding graph of Figure 3. As K_D increases across columns given n_D, the $\lambda_{\alpha 1}$ increases and the expected uncertainty decreases dramatically. As the number of n_D items increases across rows given K_D, $\lambda_{\alpha 1}$ increases and the expected interval shrinks, but relatively weakly. Within each graph, the expected interval reduces as the sample sizes $J_G B_G$ and K_G increase in the G study. Therefore, both

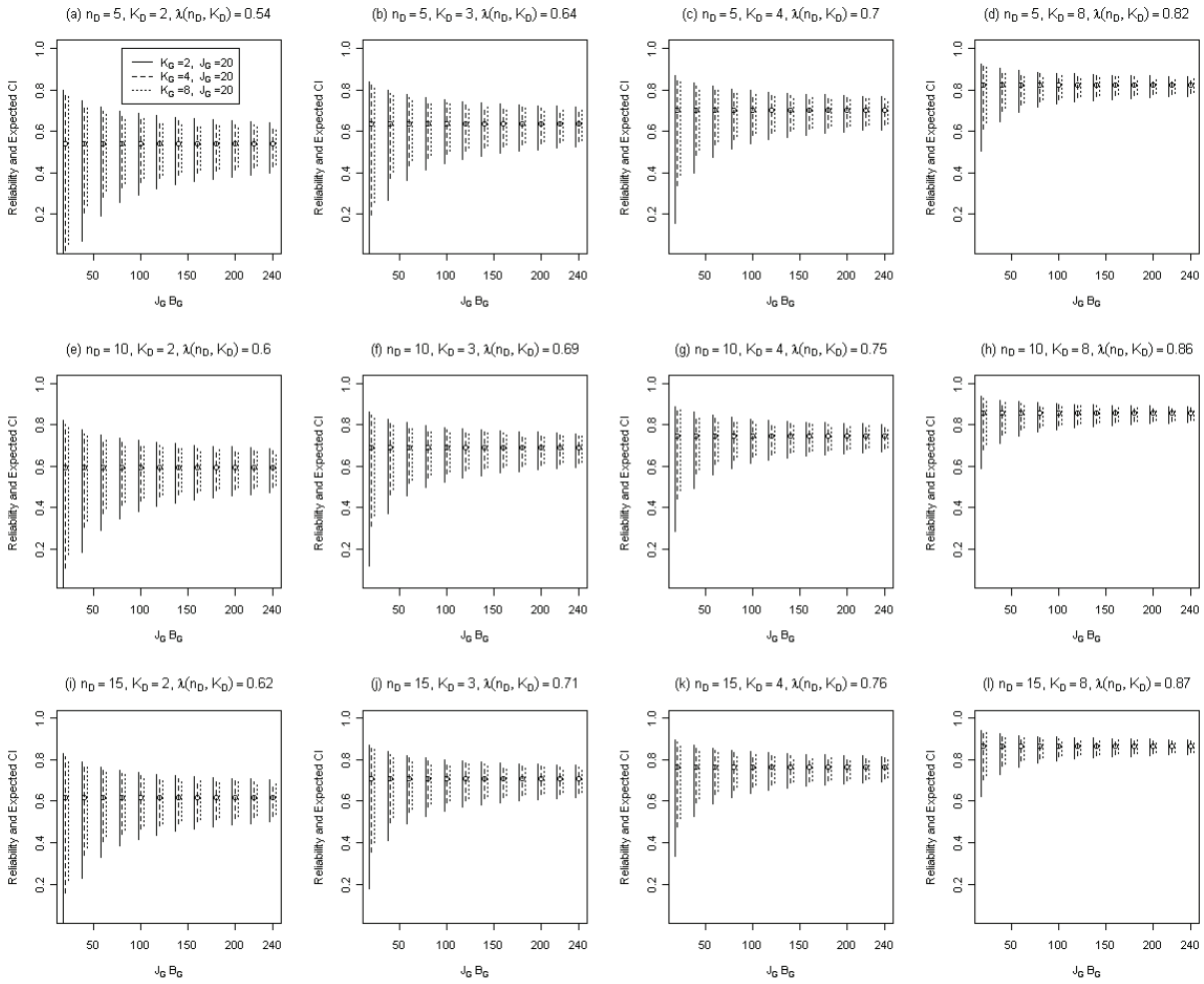

Figure 5: Each graph draws expected confidence intervals (24) and reliability sizes as dots on the vertical axis. Otherwise, G-study and D-study sample sizes, the legend, horizontal axis labels and the heading on top of each graph are identical to those of Figure 3.

D-study design (n_D, K_D) and G-study design $(n_G, K_G, J_G B_G)$ are positively associated with the expected uncertainly (24).

5 Discussion

In this paper, we expressed the reliability for the measure of teaching quality of a classroom as the correlation between the observed effects of the classroom over a pair of randomly parallel realizations of the measurement procedure. We derived a confidence interval for the

reliability, a test for the hypothesis that the reliability meets a minimum standard, and the power of the test against alternative hypotheses for the design of a study. Then, we showed how to inform the design of a D study based on the inferences drawn with uncertainty from a G study. Illustrative examples showed how to compare alternative D-study designs in terms of their reliability sizes, uncertainty and powers that will be achieved in the D study given the variance components estimated with error in the G study. Because the variance component estimates and their uncertainty estimates in the G study depend on the sample sizes of the G study, the estimates as well as the G study design are informative about the design of a D study.

Illustrative examples revealed diminishing return of increasing sample sizes in improving reliability size and power and reducing the uncertainty of a reliability estimate. Therefore, it seems wise to capitalize on the strong impact of low-range sample sizes in designing a study.

The lower bound of the confidence interval (23) for λ_α may be negative in particular when the frequency of rating each classroom in the G or D study is small. In Figure 4 where the frequency $K_G = 4$ is modest, the simulation generates 2.3% of the lower bounds negative for $(n_D, K_D) = (5, 2)$, and 0.1 to 0.4% of lower bounds negative for others except for $(n_D, K_D) = (15, 4)$ and $K_D = 8$ where all lower bounds are positive. The likelihood of λ_α is defined between 0 and 1. The negative lower bound was set to zero to produce Figure 4. Designs with the smaller frequency of rating each classroom produce more negative lower bounds. To reduce the lower bounds going out of the proper range, the confidence interval may be based on a 95% highest density interval. In our simulation involving multiple study designs with 2 raters where more than 20 % of the generated lower bounds may be negative in some severe cases, the highest-density 95% confidence interval for reliability reduced the negative lower bounds by approximately 30%.

The developed methods and their illustrations in this paper are based on a model that may be too simple for some cases. First, they do not involve logistics and cost in sample size determination. For example, high cost may limit the number of raters available for a study.

Next, a more elaborate model may have interaction effects involving items; rating scores depending on measurement time of the day; and raters rating a classroom inconsistently across measurement times of the day. Consequently, a study may be longitudinal involving multiple occasions nested within each classroom which adds an additional facet to the model (14). Moreover, the outcome variable may be nonnormal. Furthermore, selection of an optimal design for a field study may involve an unbalanced design due to, for example, attrition and non-responses in the model (14) where n_{jkb} items are rated in J_b classrooms by K_b raters within block b for $b = 1, \cdots, B$. Therefore, useful future extensions of the methods presented in this paper may involve cost, logistics and a series of more elaborate models.

We analyzed data from Classroom Assessment Scoring System (CLASS) that uses a 7 point scale for each item where the median responses tend to be around 3 or 4 with quite symmetric distributions (La Paro et al. 2004). With nonnormal sample data, the robustness of Theorems 3.1 and 3.2 against the departures from the assumed normality of the model (14) should be examined. Note that reliability (2) is not associated with variances in the rated scores due to rater or block differences and that the derivation of the distribution (18) in the Appendix does not depend on rater and block effects. Consequently, the Theorems are quite robust against the departures from the assumed normality of the rater or block effects. However, the violation of the model assumptions may come from other effects which could be checked by, for example, plotting the estimates such as the classroom effect estimates $\bar{Y}_{\cdot j\cdot b} - \bar{Y}_{\cdots b}$. Then, such a departure from the model assumptions may be simulated, and the robustness of the Theorems against such a violation may be assessed. Therefore, developing sensitivity analysis for violations of our model assumptions, including normality assumptions, is a valuable topic for future research.

Appendix: Derivation of Equation (18)

Let $\bar{\gamma}_{\cdot} = \sum_b \gamma_b/B$, $\bar{\alpha}_{\cdot b} = \sum_j \alpha_{jb}/J$, $\bar{\alpha}_{\cdot\cdot} = \sum_b \alpha_{\cdot b}/B$, $\bar{\beta}_{\cdot b} = \sum_k \beta_{kb}/K$, $\bar{\beta}_{\cdot\cdot} = \sum_b \beta_{\cdot b}/B$, $\overline{(\alpha\beta)}_{j\cdot b} = \sum_k (\alpha\beta)_{jkb}/K$, $\overline{(\alpha\beta)}_{\cdot kb} = \sum_j (\alpha\beta)_{jkb}/J$, $\overline{(\alpha\beta)}_{\cdot\cdot b} = \sum_k (\alpha\beta)_{\cdot kb}/K$, $\overline{(\alpha\beta)}_{\cdots} = \sum_b (\alpha\beta)_{\cdot\cdot b}/B$,

$\bar{\epsilon}_{.jkb} = \sum_i \epsilon_{ijkb}/n$, $\bar{\epsilon}_{..kb} = \sum_j \bar{\epsilon}_{.jkb}/J$, $\bar{\epsilon}_{.j.b} = \sum_k \bar{\epsilon}_{.jkb}/K$, $\bar{\epsilon}_{...b} = \sum_j \bar{\epsilon}_{.j.b}/J$ and $\bar{\epsilon}_{....} = \sum_j \bar{\epsilon}_{...b}/B$. Reasonable estimators for γ_b, α_{jb}, β_{kb}, $(\alpha\beta)_{jkb}$ and ϵ_{ijkb} may be expressed as

$$\begin{aligned}
\hat{\gamma}_b &= (\gamma_b + \bar{\alpha}_{.b} + \bar{\beta}_{.b} + \overline{(\alpha\beta)}_{..b} + \bar{\epsilon}_{...b}) - (\bar{\gamma}_{.} + \bar{\alpha}_{..} + \bar{\beta}_{..} + \overline{(\alpha\beta)}_{...} + \bar{\epsilon}_{....}) \quad (25)\\
\hat{\alpha}_{jb} &= (\alpha_{jb} + \overline{(\alpha\beta)}_{j.b} + \bar{\epsilon}_{.j.b}) - (\bar{\alpha}_{.b} + \overline{(\alpha\beta)}_{..b} + \bar{\epsilon}_{...b})\\
\hat{\beta}_{kb} &= (\beta_{kb} + \overline{(\alpha\beta)}_{.kb} + \bar{\epsilon}_{..kb}) - (\bar{\beta}_{.b} + \overline{(\alpha\beta)}_{..b} + \bar{\epsilon}_{...b})\\
\hat{\epsilon}_{ijkb} &= \epsilon_{ijkb} - \bar{\epsilon}_{.jkb}\\
(\alpha\beta)_{jkb} + \bar{\epsilon}_{.jkb} &= \widehat{(\alpha\beta)}_{jkb} + \{[\overline{(\alpha\beta)}_{j.b} + \bar{\epsilon}_{.j.b}] - [\overline{(\alpha\beta)}_{..b} + \bar{\epsilon}_{...b}]\} + [\overline{(\alpha\beta)}_{.kb} + \bar{\epsilon}_{..kb}]
\end{aligned}$$

where $\sqrt{nJK}(\gamma_b + \bar{\alpha}_{.b} + \bar{\beta}_{.b} + \overline{(\alpha\beta)}_{..b} + \bar{\epsilon}_{...b}) \sim N(0, nJK\sigma_\gamma^2 + nK\sigma_\alpha^2 + nJ\sigma_\beta^2 + n\sigma_{\alpha\beta}^2 + \sigma^2)$ independent across blocks b, $\sqrt{nK}(\alpha_{jb} + \overline{\alpha\beta}_{j.b} + \bar{\epsilon}_{.j.b}) \sim N(0, nK\sigma_\alpha^2 + n\sigma_{\alpha\beta}^2 + \sigma^2)$ independent across classrooms j and blocks b, $\sqrt{nJ}(\beta_{kb} + \overline{(\alpha\beta)}_{.kb} + \bar{\epsilon}_{..kb}) \sim N(0, nJ\sigma_\beta^2 + n\sigma_{\alpha\beta}^2 + \sigma^2)$ independent across raters k and blocks b, $\sqrt{n}[(\alpha\beta)_{jkb} + \bar{\epsilon}_{.jkb}] \sim N(0, n\sigma_{\alpha\beta}^2 + \sigma^2)$ independent across classrooms j, raters k and blocks b, $\sqrt{nK}[\overline{(\alpha\beta)}_{j.b} + \bar{\epsilon}_{.j.b}] \sim N(0, n\sigma_{\alpha\beta}^2 + \sigma^2)$ across j and b, and $\sqrt{nJ}[\overline{(\alpha\beta)}_{.kb} + \bar{\epsilon}_{..kb}] \sim N(0, n\sigma_{\alpha\beta}^2 + \sigma^2)$ across k and b. When squared and summed over $nJKB$ units, the last equation has $\sum_b \sum_k \sum_j n\{(\alpha\beta)_{jkb} + \bar{\epsilon}_{.jkb} - [\overline{(\alpha\beta)}_{..b} + \bar{\epsilon}_{...b}]\}^2 \sim (n\sigma_{\alpha\beta}^2 + \sigma^2)\chi^2_{(JK-1)B}$ on the left hand side, and $\sum_b \sum_k \sum_j n\widehat{(\alpha\beta)}_{jkb}^2 + \sum_b \sum_j nK\{[\overline{(\alpha\beta)}_{j.b} + \bar{\epsilon}_{.j.b}] - [\overline{(\alpha\beta)}_{..b} + \bar{\epsilon}_{...b}]\}^2 + \sum_b \sum_k nJ\{[\overline{(\alpha\beta)}_{.kb} + \bar{\epsilon}_{..kb}] - [\overline{(\alpha\beta)}_{..b} + \bar{\epsilon}_{...b}]\}^2$ on the right hand side where $\sum_b \sum_j nK\{[\overline{(\alpha\beta)}_{j.b} + \bar{\epsilon}_{.j.b}] - [\overline{(\alpha\beta)}_{..b} + \bar{\epsilon}_{...b}]\}^2 \sim (n\sigma_{\alpha\beta}^2 + \sigma^2)\chi^2_{(J-1)B}$ and $\sum_b \sum_k nJ\{[\overline{(\alpha\beta)}_{.kb} + \bar{\epsilon}_{..kb}] - [\overline{(\alpha\beta)}_{..b} + \bar{\epsilon}_{...b}]\}^2 \sim (n\sigma_{\alpha\beta}^2 + \sigma^2)\chi^2_{(K-1)B}$. The three terms on the right hand side are independent and must add to have $(n\sigma_{\alpha\beta}^2 + \sigma^2)\chi^2_{(JK-1)B}$, which implies that $\sum_b \sum_k \sum_j n\widehat{(\alpha\beta)}_{jkb}^2 \sim (n\sigma_{\alpha\beta}^2 + \sigma^2)\chi^2_{(J-1)(K-1)B}$. As $\hat{\gamma}_b$, $\hat{\alpha}_{jb}$, $\hat{\beta}_{kb}$, $\widehat{(\alpha\beta)}_{jkb}$ and $\hat{\epsilon}_{ijkb}$ are independent, so are

$$\begin{aligned}
\frac{SSG}{nJK\sigma_\gamma^2 + nK\sigma_\alpha^2 + nJ\sigma_\beta^2 + n\sigma_{\alpha\beta}^2 + \sigma^2} &\sim \chi^2_{B-1}, \quad (26)\\
\frac{SSA}{nK\sigma_\alpha^2 + n\sigma_{\alpha\beta}^2 + \sigma^2} &\sim \chi^2_{(J-1)B},\\
\frac{SSB}{nJ\sigma_\beta^2 + n\sigma_{\alpha\beta}^2 + \sigma^2} &\sim \chi^2_{(K-1)B},
\end{aligned}$$

$$\begin{aligned} \frac{SSAB}{n\sigma^2_{\alpha\beta}+\sigma^2} &\sim \chi^2_{(J-1)(K-1)B}, \\ \frac{SSE}{\sigma^2} &\sim \chi^2_{(n-1)JKB}. \end{aligned}$$

Consequently, we have

$$\frac{MSA/(nK\sigma^2_\alpha + n\sigma^2_{\alpha\beta}+\sigma^2)}{MSAB/(n\sigma^2_{\alpha\beta}+\sigma^2)} = \frac{1-\lambda_\alpha}{1-\hat{\lambda}_\alpha} \sim F_{(J-1)B,(J-1)(K-1)B}. \tag{27}$$

References

[1] Bartlett, M.S. and Kendall, D.G. (1946). The Statistical Analysis of Variance-Heterogeneity and the Logarithmic Transformation, *Journal of the Royal Statistical Society*, Suppl. **8**, 128-138.

[2] Borman,G ., Slavin, R. E., Cheung,A ., ChamberlainA, ., Madden,N . A., and Chambers, B. (2005). The national randomized field trial of Success for All: Second-year outcomes. *American Educational Research Journal*, **42**, 673-696.

[3] Burdick, R.K. ana Graybill, F.A. (1992). *Confidence Intervals on Variance Components*. New York: Dekker.

[4] Brennan, R.L. (2001). *Generalizability theory*. New York: Springer-Verlag.

[5] Hirsch, B.J., and Wong, V. (2005). After-school programs. In D.L. DuBois and M.J. Karcher (Eds.), Handbook of youth mentoring (pp. 364-375). Thousand Oaks, CA: Sage.

[6] Kinzie, M., Whitaker, S., Neesen, K., Kelley, M., Matera, M. and Pianta, R. (2005). State-wide Web-based Professional Development & Curricula for Early Childhood Educators: Design & Infrastructure. In G. Richards (Ed.), *Proceedings of World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education 2005* (pp. 814-821). Chesapeake, VA: AACE.

[7] La Paro, K., Pianta, R. and Stuhlman, M. (2004). The Classroom Assessment Scoring System: Findings from the Prekindergarten Year, *The Elementary School Journal*, **104**, 409-426.

[8] Pianta, R., Howes, C., Burchinal, M., Bryant, D., Clifford, R., Early, D. and Barbarin, O. (2005). Features of Pre-Kindergarten Programs, Classrooms, and Teachers: Do They Predict Observed Classroom Quality and ChildTeacher Interactions? *Applied Developmental Science*, **9**, 144-159.

[9] Raudenbush, S.W. and Bryk, A.S. (1987). Examining Correlates of Diversity, *Journal of Educational Statistics*, **12**, 241-269.

[10] Raudenbush, S.W., Martinez, A., Bloom, H., Zhu, P. and Lin, F. (2010). Studying the Reliability of Group-Level Measures with Implications for Statistical Power: A Six-Step Paradigm, University of Chicago Working Paper.

[11] Raudenbush, S.W. and Sadoff, S. (2008). Statistical Inference When Classroom Quality is Measured With Error, *Journal of Research on Educational Effectiveness*, **1**, 138-154.

[12] Shin, Y. and Raudenbush, S. W. (2010). A Latent Cluster Mean Approach to The Contextual Effects Model with Missing Data. *JEBS*, **35**, 26-53.

[13] Smith, R. E., Smoll, F. L., and Cumming, S. P. (2007). Effects of a motivational climate intervention for coaches on young athletes' sport performance anxiety. *Journal of Sport and Exercise Psychology*, **29**, 39-59.