Studying the Sensitivity of Inferences to Possible Unmeasured Confounding Variables in Multisite Evaluations

CSE Technical Report 701

Michael Seltzer and Jinok Kim,
CRESST/University of California, Los Angeles

Ken Frank, Michigan State University

October 2006

National Center for Research on Evaluation,
Standards, and Student Testing (CRESST)
Center for the Study of Evaluation (CSE)
Graduate School of Education & Information Studies
University of California, Los Angeles
GSE&IS Building, Box 951522
Los Angeles, CA 90095-1522

(310) 206-1532

Project 3.1: Methodologies for Assessing Student Progress
Michael Seltzer and Kilchan Choi, CRESST/UCLA, Project Directors

The work reported herein was supported under the Educational Research and Development Centers Program, PR/Award Number R305B960002, as administered by the Institute of Education Sciences, U.S. Department of Education.

The findings and opinions expressed in this report do not reflect the positions or policies of the National Institute on Student Achievement, Curriculum, and Assessment, the Institute of Education Sciences, or the U.S. Department of Education.

Studying the Sensitivity of Inferences to Possible Unmeasured Confounding Variables in Multisite Evaluations[1]

Michael Seltzer and Jinok Kim CRESST/University of California, Los Angeles

Ken Frank, Michigan State University,

Abstract

In multisite evaluation studies, questions of primary interest often focus on whether particular facets of implementation or other aspects of classroom or school environments are critical to a program's success. However, the differences with which teachers implement programs can depend on an array of factors, including differences in their training and experience, in the prior preparation of their students, and in the degree of support they receive from school administrators. As such, a crucially important implication is that in studying connections between various aspects of implementation and the effectiveness of programs, we need to be alert to factors that may be confounded with differences in implementation. Despite our best efforts to anticipate and measure possible confounding variables, teachers who differ in terms of the quality and frequency with which they implement various program practices use particular program materials and the like, may differ in important ways that have not been measured, giving rise to possible hidden bias. In this paper, we extend Frank's (2000) work on assessing the impact of omitted confounding variables on coefficients of interest in regression settings to applications of HMs in multiste settings in which interest centers on testing whether certain aspects of implementation are critical to a program's success. We provide a detailed illustrative example using the data from a study focusing on the effects of reform-minded instructional practices in mathematics (Gearhart et al., 1999; Saxe et al., 1999).

[1] Acknowledgements: We wish to thank Maryl Gearhart and Geoff Saxe for permission to use the data from their study Integrating Mathematics Assessment with Instruction in Elementary Mathematics, which was supported by NSF grand MDR 9154512. Many thanks to Hyekyung Jung for valuable discussions about this work, and for her help with various calculations, constructing tables and figures, and formatting equations.

Introduction

Much emphasis in multisite studies of educational programs over the years has centered on single-number summaries of overall differences in outcomes between those assigned to treatment and those assigned to comparison conditions (i.e., intent-to-treat (ITT) estimates). Such estimates, however, can mask potentially substantial variability in program effects across sites, the possibility that implementation may have been good in some sites but poor in others, and that the students with certain prior educational experiences may have benefited from the program far more than others (see, e.g., Burstein, 1980; Cronbach, 1976, 1982; see, also, Seltzer, 2004).

To make studies of the effects of educational programs as useful as possible to an array of stakeholders, including policy makers, school administrators, teachers, program developers, and parents, it is necessary to extend our investigations beyond estimating ITT effects. It is important to study how differences in level of implementation magnify or dampen the effects of programs, to test whether certain program components are critical to a program's success, and to investigate whether a program is more beneficial for students with certain background characteristics and prior skills. For discussion of the use of Hierarchical Models in addressing such questions, see, for example, Hong (2004; 2006), Raudenbush and Willms (1991), Seltzer (1994; 2004), and.

However, efforts to draw sound conclusions concerning the kinds of questions outlined above can be extremely challenging. In particular, the lack of random assignment of classes of schools, for example, to treatment or control conditions in multisite studies can give rise to numerous potential confounding variables that we must attend to.

Even when assignment of teachers, for example, to treatment and control conditions is random, it is important to note that teachers are not randomly assigned to different levels of implementation. The differences with which teachers implement programs (e.g., differences in the skill with which, and extent to which, they are able to implement particular instructional practices), can depend on a whole host of factors, including differences in their training and experience, in the prior preparation of their students, the degree of support teachers receive from school administrators and the like. As such, a crucially important implication is that

in studying connections between various aspects of implementation and the effectiveness of programs, we need to be alert to factors that may be confounded with differences in implementation.

Despite our best efforts to anticipate and measure possible confounding variables, teachers who differ in terms of the quality and frequency with which they implement various program practices, use particular program materials and the like, may differ in important ways that have not been measured, giving rise to possible hidden bias. In this paper, we extend Frank's (2000) work on assessing the impact of omitted confounding variables on coefficients of interest in regression settings to applications of HMs in multisite settings in which interest centers on testing whether certain aspects of implementation are critical to a program's success.

We are especially interested in two commonly employed designs in multisite field studies. One type of design entails forming matched pairs of organizational units (e.g., classrooms or schools), and assigning one organizational unit within a pair to an innovative program of interest and the other to a comparison condition. This essentially gives rise to a series of mini-experiments or mini-quasi-experiments. The within-pair (level-1) parameter of primary interest in such designs would be a treatment/control contrast for each pair (e.g., β_{1j}). In a between-pair model, the β_{1j} are modeled as function measures of implementation and other site-level characteristics.

The second type of design, in contrast, does not involve forming blocks (e.g., pairs) before assigning organizational units to treatment or control conditions. Rather than yielding a series of mini-experiments, such designs provide us with a sample of treatment classrooms, for example, that are compared with a sample of control classrooms. The level-1 (e.g., within-class) parameter of primary interest in such settings is typically a mean outcome score for each class adjusted for differences among classes in their student intake characteristics (e.g., β_{0j}). In a between-class model, the β_{0j} are modeled as a function of various class-level characteristics, which might include treatment group indicator variables, measures of implementation, and various compositional characteristics. Given the prevalence of this type of design in multisite evaluations, we provide a detailed illustrative example of our strategy for sensitivity analysis using the data from a study focusing on the effects of reform-minded instructional practices in mathematics (Gearhart et al., 1999; Saxe et al., 1999). A brief example using the data from a study that

employed a paired design—a study of an innovative pre-algebra program called Transition Mathematics (TM; see Seltzer, 2004)—is sketched in Appendix B.

Frank's approach to assessing the impact of a confounding variable was initially developed for use in settings in which one is working with linear models in analyses of non-nested data, i.e., settings in which OLS provides a sensible strategy for obtaining estimates of regression coefficients and their standard errors. Suppose, for example, that the OLS estimate of the regression coefficient for a predictor of interest (X) is statistically significant, i.e., $t = \hat{\beta}/se(\hat{\beta})$ exceeds the relevant t critical value. Suppose further that there is an unmeasured confounding variable (CV). How large would the correlation need to be between the outcome (Y) and CV, and between X and CV to impact the estimate of β and its standard error to the point where it is no longer significant at the chosen α level? Frank's approach addresses questions of this kind. As will be seen, a key facet of his approach capitalizes on the fact that OLS regression estimates and their standard errors can be re-expressed in terms of standard deviations of Y and X and various relevant correlations (e.g., $r_{Y\bullet X}$).

To help set the stage for extending Frank's work on sensitivity analysis to multilevel settings, we begin by discussing the estimation of fixed effects in means-as-outcomes HMs in cases where the number of observations per cluster is equal. We first focus on a model with a single level-2 predictor (W_j), and then expand this model to include an observed level-2 covariate (CV_j) that we wish to control for. We then consider the situation in which CV_j is unobserved and outline a strategy for assessing its impact based on Frank's approach (2000). Next we discuss the extension of our strategy to HMs in which adjusted means (β_{0j}) are treated as level-2 outcomes, and models in which level-1 treatment/control contrasts (β_{1j})—more generally level-1 slopes—are treated as level-2 outcomes. We then turn to our illustrative examples.

Note that pioneering work in the area of sensitivity analysis has been conducted by Paul Rosenbaum and his colleagues (e.g., Gastwirth, Krieger & Rosenbaum, 1998; Rosenbaum, 2002; Rosenbaum & Rubin, 1983). Some advantages of employing Frank's approach to sensitivity analysis in our work in HM settings is discussed in the concluding section of our paper.

Estimation of Fixed Effects in Means-as-Outcomes Models

Means-As-Outcomes Models that Contain a Single Level-2 Predictor

To motivate the extension of Frank's approach to multilevel settings, we first focus on the estimation of fixed effects in means-as-outcomes models in cases in which the number of observations per cluster is equal. At level 1, the outcome score for each student ($i = 1,\mathrm{L}\ ,n$) in cluster j ($j = 1,\mathrm{L}\ ,J$) is modeled as a function of the mean outcome score for cluster j (β_{0j}):

$$Y_{ij} = \beta_{0j} + r_{ij},\ r_{ij} \sim N(0, \sigma^2) \qquad (1)$$

where r_{ij} is a residual assumed normally distributed with mean 0 and variance σ^2. At level-2, cluster outcome means are modeled as a function of various cluster characteristics. We begin by focusing on a simple level-2 model containing a single predictor (W_j):

$$\beta_{0j} = \gamma_{00} + \gamma_{01}W_j + u_{0j},\ u_{0j} \sim N(0, \tau) \qquad (2)$$

where γ_{01}, our parameter of primary interest, is a fixed effect capturing the relationship between W_j and cluster means, and u_{0j} is a random effect assumed normally distributed with mean 0 and variance τ.

The error variance connected with the sample mean outcome score ($\overline{Y}_j$) for each cluster is $V_j = V = \sigma^2/n$. For heuristic purposes we begin by employing a REML/Empirical Bayes (EB) estimation approach (see chapters 3 and 13 in Raudenbush & Bryk, 2002). Setting the variance components equal to their REML estimates (i.e., $\sigma^2 = \hat{\sigma}^2$; $\tau = \hat{\tau}$), we construct the weights $\hat{\Delta}_j^{-1} = \hat{\Delta}^{-1} = 1/(\hat{\tau} + \hat{V})$, where $\hat{V}$ = $\hat{\sigma}^2/n$; since the error variances of the $\overline{Y}_j$ are equal, each cluster receives equal weight. We then obtain an estimate of γ_{01} and its standard error using the following formulas:

$$\hat{\gamma}_{01} = \sum_{j=1}^{J} \frac{\hat{\Delta}^{-1}(\overline{Y}_{.j} - \overline{Y})(W_j - \overline{W})}{\hat{\Delta}^{-1}(W_j - \overline{W})^2} \qquad (3)$$

and

$$SE(\hat{\gamma}_{01}) = \left[\sum_{j=1}^{J} \hat{\Delta}^{-1} (W_j - \bar{W})^2 \right]^{-1} \qquad (4)$$

$$= \frac{1}{\hat{\Delta}^{-1} \sum_{j=1}^{J} (W_j - \bar{W})^2}$$

$$= \frac{\hat{\tau} + \hat{V}}{\sum_{j=1}^{J} (W_j - \bar{W})^2}$$

Note that the weights in the numerator and denominator in Equation 3 cancel out. This helps us see that in the balanced case, estimating fixed effects in means-as-outcomes models can be viewed as an OLS regression of sample mean outcome scores on level-2 predictors. Thus in the case of our example:

$$\bar{Y}_{.j} = \gamma_{00} + \gamma_{01} W_j + e_j \ , \ e_j \sim N(0, D) \qquad (5)$$

where $D = \tau + V$. The formulas for the OLS estimate of γ_{01} and its standard error are as follows:

$$\hat{\gamma}_{01} = \sum_{j=1}^{J} \frac{(\bar{Y}_{.j} - \bar{Y})(W_j - \bar{W})}{(W_j - \bar{W})^2} \qquad (6)$$

and

$$SE(\hat{\gamma}_{01}) = \frac{\sqrt{\hat{D}}}{\sqrt{\sum_{j=1}^{J} (W_j - \bar{W})^2}} \qquad (7)$$

where

$$\hat{D} = \frac{\sum_{j=1}^{J} (\bar{Y}_{.j} - [\hat{\gamma}_{00} + \hat{\gamma}_{01} W_j])^2}{J-2} \qquad (8)$$

Note that the resulting estimate of D will be equal to $\hat{\tau} + \hat{\sigma}^2/n$. Thus in means-as-outcomes settings in which there are equal numbers of observations per cluster we see that the OLS estimate of γ_{01} and its standard error based on a cluster-level regression of $\bar{Y}_j$ on W_j will be identical to those obtained via a REML/EB estimation strategy in this setting.

To help lay the groundwork for implementing Frank's sensitivity analysis approach, note that the OLS estimate of γ_{01} and its standard error can be expressed in terms of the standard deviations of the outcome and predictor variables in Equation 5, and the correlation between the outcome and predictor variables:

$$\hat{\gamma}_{01} = \frac{S_{\bar{Y}_j}}{S_{W_j}} \times r_{\bar{Y}_j \bullet W_j} \qquad (9)$$

where

$$S_{\bar{Y}_j} = \sqrt{\frac{\sum_{j=1}^{J} (\bar{Y}_{.j} - \bar{Y})^2}{J-1}}, \; S_{W_j} = \sqrt{\frac{\sum_{j=1}^{J} (W_j - \bar{W})^2}{J-1}} \text{ and } r_{\bar{Y}_j \bullet W_j} = \frac{\sum_{j=1}^{J} (\bar{Y}_{.j} - \bar{Y})(W_j - \bar{W})}{\sqrt{\sum_{j=1}^{J} (\bar{Y}_{.j} - \bar{Y})^2} \sqrt{\sum_{j=1}^{J} (W_j - \bar{W})^2}} \qquad (10)$$

and

$$SE(\hat{\gamma}_{01}) = \frac{\frac{\sqrt{\sum_{j=1}^{J} (\bar{Y}_{.j} - \bar{Y})^2} \sqrt{1 - r^2_{\bar{Y}_j \bullet W_j}}}{\sqrt{J-2}}}{\sqrt{\sum_{j=1}^{J} (W_j - \bar{W})^2}} \qquad (11a)$$

$$= \frac{S_{\bar{Y}_j}}{S_{W_j}} \times \frac{\sqrt{1 - r^2_{\bar{Y}_j \cdot W_j}}}{\sqrt{J-2}} \qquad (11b)$$

Furthermore, utilizing the above equations, we see that dividing the OLS estimate of γ_{01} by its standard error yields the following t ratio:

$$t = \frac{\frac{S_{\bar{Y}_j}}{S_{W_j}} \times r_{\bar{Y}_j \cdot W_j}}{\frac{S_{\bar{Y}_j}}{S_{W_j}} \times \frac{\sqrt{1 - r^2_{\bar{Y}_j \cdot W_j}}}{\sqrt{J-2}}} \qquad (12)$$

$$= \frac{r_{\bar{Y}_j \cdot W_j}}{\frac{\sqrt{1 - r^2_{\bar{Y}_j \cdot W_j}}}{\sqrt{J-2}}}$$

Equation 12 shows that the resulting t ratio reduces to the correlation between $\bar{Y}_j$ and W_j divided by its standard error. Hence drawing inferences concerning the OLS estimate of γ_{01} is tantamount to drawing inferences regarding $r_{\bar{Y}_j \cdot W_j}$. This is central to the application of Frank's sensitivity analysis approach in multilevel settings.

For later reference, note also that given the value of the resulting t ratio, it is possible to compute $r_{\bar{Y}_j \cdot W_j}$:

$$r_{\bar{Y}_j \cdot W_j} = \frac{t}{\sqrt{(J-2) + t^2}} \qquad (13)$$

Controlling for an Observed Level-2 Covariate

Suppose that we now add a predictor that we wish to control for to our level-2 model:

$$\beta_{0j} = \gamma_{00} + \gamma_{01} W_j + \gamma_{02} CV_j + u_{0j} \ , \ u_{0j} \sim N(0, \tau) \qquad (14)$$

Again assuming equal numbers of observations in each cluster, an OLS regression of the $\overline{Y}_j$'s on W_j and CV_j will, analogous to the single level-2 predictor setting, result in estimates of the fixed effects and their standard errors identical to those obtained via a REML/EB estimation strategy. Thus we write:

$$\overline{Y}_{.j} = \gamma_{00} + \gamma_{01} W_j + \gamma_{02} CV_j + e_j,\ e_j \sim N(0, D) \qquad (15)$$

Paralleling Equations 9 and 11b, we can re-express the OLS estimate of γ_{01} and its standard error in terms of standard deviations and correlation coefficients. The formula for the OLS estimate of γ_{01} is as follows:

$$\hat{\gamma}_{01} = \frac{S_{\overline{Y}_j}\sqrt{1 - r^2_{\overline{Y}_j \cdot CV_j}}}{S_{W_j}\sqrt{1 - r^2_{W_j \cdot CV_j}}} \times r_{\overline{Y}_j \cdot W_j | CV_j} \qquad (16)$$

where

$$r_{\overline{Y}_j \cdot W_j | CV_j} = \frac{r_{\overline{Y}_j \cdot W_j} - r_{\overline{Y}_j \cdot CV_j} r_{W_j \cdot CV_j}}{\sqrt{1 - r^2_{\overline{Y}_j \cdot CV_j}}\sqrt{1 - r^2_{W_j \cdot CV_j}}} \qquad (17)$$

Thus a key statistic in Equation 16 is $r_{\overline{Y}_j \cdot W_j | CV_j}$, i.e., the partial correlation between $\overline{Y}_j$ and W_j holding constant CV_j. We now inspect the formula for the partial correlation more closely. Note that the numerator in Equation 17 consists of the simple correlation between $\overline{Y}_j$ and W_j minus the product of the simple correlation between $\overline{Y}_j$ and CV_j ($r_{\overline{Y}_j \cdot CV_j}$), and the simple correlation between W_j and CV_j ($r_{W_j \cdot CV_j}$). In addition, the simple correlations involving CV_j appear in the denominator as well. As can be seen, if $r_{\overline{Y}_j \cdot CV_j}$ = 0, and similarly $r_{W_j \cdot CV_j}$ = 0, then $r_{\overline{Y}_j \cdot W_j | CV_j}$ = $r_{\overline{Y}_j \cdot W_j}$. Conversely, if CV_j is strongly correlated with $\overline{Y}_j$ and W_j then inclusion of CV_j in the model will appreciably impact the resulting partial correlation.

The formula for the standard error is as follows:

$$SE(\hat{\gamma}_{01}) = \frac{S_{\bar{Y}_j}\sqrt{1-r^2_{\bar{Y}_j \cdot CV_j}}}{S_{W_j}\sqrt{1-r^2_{W_j \cdot CV_j}}} \times \frac{\sqrt{1-r^2_{\bar{Y}_j \cdot W_j | CV_j}}}{\sqrt{J-3}} \qquad (18)$$

To help see that Equation 18 translates to a form that is more familiar, note that:

$$\sqrt{1-r^2_{\bar{Y}_j \cdot CV_j}} \times \sqrt{1-r^2_{\bar{Y}_j \cdot W_j | CV_j}} = \sqrt{1-R^2_{\bar{Y}_j | W_j, CV_j}} \qquad (19)$$

Thus Equation 18 can be re-written as follows:

$$\frac{S_{\bar{Y}_j}\sqrt{1-R^2_{\bar{Y}_j | W_j, CV_j}}}{S_{W_j}\sqrt{1-r^2_{W_j \cdot CV_j}}\sqrt{J-3}} \qquad (20a)$$

$$= \frac{\sqrt{\sum_{j=1}^{J}(\bar{Y}_{.j}-\bar{Y})^2}\sqrt{1-R^2_{\bar{Y}_j | W_j, CV_j}}}{\sqrt{J-3}} \Big/ \sqrt{\sum_{j=1}^{J}(W_j-\bar{W})^2}\sqrt{1-r^2_{W_j \cdot CV_j}} \qquad (20b)$$

Note that the numerator of Equation 20b provides us with $\sqrt{\hat{D}}$, where $\hat{D}$ is the estimate of the residual variance term in Equation 15. And analogous to the previous example, $\hat{D}$ will be equal to the REML estimate of the random effects variance parameter in Equation 14 plus the error variance term $\hat{\sigma}^2/n$.

Dividing the OLS estimate of γ_{01} (Equation 16) by its standard error (Equation 18), yields the following t ratio:

$$t = \frac{\frac{S_{\bar{Y}_j}\sqrt{1-r^2_{\bar{Y}_j \cdot CV_j}}}{S_{W_j}\sqrt{1-r^2_{W_j \cdot CV_j}}} \times r_{\bar{Y}_j \cdot W_j | CV_j}}{\frac{S_{\bar{Y}_j}\sqrt{1-r^2_{\bar{Y}_j \cdot CV_j}}}{S_{W_j}\sqrt{1-r^2_{W_j \cdot CV_j}}} \times \frac{\sqrt{1-r^2_{\bar{Y}_j \cdot W_j | CV_j}}}{\sqrt{J-3}}} \qquad (21a)$$

$$= \frac{r_{\bar{Y}_j \cdot W_j | CV_j}}{\frac{\sqrt{1-r^2_{\bar{Y}_j \cdot W_j | CV_j}}}{\sqrt{J-3}}} \qquad (21b)$$

Paralleling the single-predictor setting, we see that the t ratio reduces to $r_{\bar{Y}_j \cdot W_j | CV_j}$ divided by its standard error. Thus drawing inferences concerning γ_{01} holding constant CV_j is tantamount to drawing inferences concerning $r_{\bar{Y}_j \cdot W_j | CV_j}$.

It can also be shown that:

$$r_{\bar{Y}_j \cdot W_j | CV_j} = \frac{t}{\sqrt{(J-3)+t^2}} \qquad (22)$$

Assessing the Impact of an Unobserved Level-2 Covariate

Suppose that in fitting the means-as-outcomes model containing only W_j as a level-2 predictor, we find that γ_{01} is statistically significant, which, from above, implies that $r_{\bar{Y}_j \cdot W_j}$ is statistically significant as well. Now suppose that rather than being an observed covariate, CV_j is an unmeasured potential confounding variable, and we wish to assess its impact on inferences concerning γ_{01}. Frank's approach capitalizes on the fact that drawing inferences concerning γ_{01} holding constant CV_j would be equivalent to drawing inferences concerning the partial correlation $r_{\bar{Y}_j \cdot W_j | CV_j}$. Focusing on the formula for $r_{\bar{Y}_j \cdot W_j | CV_j}$ in Equation 17 we can ask: How large must $r_{\bar{Y}_j \cdot CV_j}$ and $r_{W_j \cdot CV_j}$ be to result in a partial correlation such that the t ratio formed by dividing $r_{\bar{Y}_j \cdot W_j | CV_j}$ by its standard error is less than the relevant t critical value? To implement Frank's approach, we employ the following steps:

1. For a chosen α value, find the relevant critical value based on a t distribution with $J-3$ degrees of freedom, which we will term $t^{\#}$, i.e., the threshold t value.

2. Employing Equation 22, find the value of $r_{\bar{Y}_j \bullet W_j | CV_j}$ that corresponds to $t^{\#}$:

$$r^{\#}_{\bar{Y}_j \bullet W_j | CV_j} = \frac{t^{\#}}{\sqrt{(J-3)+(t^{\#})^2}} \qquad (23)$$

We will refer to $r^{\#}_{\bar{Y}_j \bullet W_j | CV_j}$ as the threshold value of the partial correlation.

3. Substituting $r^{\#}_{\bar{Y}_j \bullet W_j | CV_j}$ for $r_{\bar{Y}_j \bullet W_j | CV_j}$ in Equation 17 we have:

$$r^{\#}_{\bar{Y}_j \bullet W_j | CV_j} = \frac{r_{\bar{Y}_j \bullet W_j} - r_{\bar{Y}_j \bullet CV_j} r_{W_j \bullet CV_j}}{\sqrt{1-r^2_{\bar{Y}_j \bullet CV_j}}\sqrt{1-r^2_{W_j \bullet CV_j}}} \qquad (24)$$

4. Drawing from Frank (2000), the impact of CV_j will be maximized when $r^2_{\bar{Y}_j \bullet CV_j} = r^2_{W_j \bullet CV_j} = r_{\bar{Y}_j \bullet CV_j} \times r_{W_j \bullet CV_j} = k$. Thus we rewrite Equation 24 as follows:

$$r^{\#}_{\bar{Y}_j \bullet W_j | CV_j} = \frac{r_{\bar{Y}_j \bullet W_j} - k}{\sqrt{1-k}\sqrt{1-k}} \qquad (25a)$$

$$= \frac{r_{\bar{Y}_j \bullet W_j} - k}{1-k} \qquad (25b)$$

5. Since $r_{\bar{Y}_j \bullet W_j}$ is estimated from the data, the only unknown in Equation 25 is *k*, and so we now solve for k:

$$k = \frac{r_{\bar{Y}_j \bullet W_j} - r_{\bar{Y}_j \bullet W_j | CV_j}}{1 - r_{\bar{Y}_j \bullet W_j | CV_j}} \qquad (26)$$

6. The resulting value for *k* provides us with the desired threshold correlations:

$r_{\bar{Y}_j \bullet CV_j} = r_{W_j \bullet CV_j} = \sqrt{k}$

Substituting the value of these correlations into Equation 24 will result in a partial correlation between $\overline{Y}_j$ and W_j controlling for CV_j equal to $r^{\#}_{\overline{Y}_j \cdot W_j|CV_j}$, which results in a t ratio equal to the threshold t value:

$$t^{\#} = \frac{r^{\#}_{\overline{Y}_j \cdot W_j|CV_j}}{\frac{\sqrt{1-(r^{\#}_{\overline{Y}_j \cdot W_j|CV_j})^2}}{\sqrt{J-3}}} \qquad (27)$$

Note that if $r_{\overline{Y}_j \cdot CV_j}$ and $r_{W_j \cdot CV_j}$ exceed $\sqrt{k}$, then the resulting partial correlation will be less than $r^{\#}_{\overline{Y}_j \cdot W_j|CV_j}$, and as such the corresponding t ratio will be less that $t^{\#}$.

7. Note further that $r^{\#}_{\overline{Y}_j \cdot W_j|CV_j}$ and the resulting values for $r_{\overline{Y}_j \cdot CV_j}$ and $r_{W_j \cdot CV_j}$ can be substituted into Equations 16 and 18 to obtain the corresponding values for $\hat{\gamma}_{01}$ and its standard error holding constant CV_j. Note also that substituting these values into Equations 19 and the numerator of 20b, provides us with a corresponding R-squared value, and an estimate of the MSE that would reflect the amount of random effects variance that remains after taking into account W_j and CV_j plus the error variance in the $\overline{Y}_j$'s. As will be seen, the resulting MSE can be used to obtain rough estimates of τ based on the inclusion of CV_j in the analysis, and we will also discuss a way of assessing the reasonableness of such estimates.

Note that in addition to finding the value of k that will result in a partial correlation equal to the threshold t value, we can also find the value of k that will reduce the estimate of γ_{01} by an amount deemed substantively significant. Furthermore, Frank (2000) shows, the above approach can be extended readily to situations in which we have a model that includes a predictor of interest along with other covariates (e.g., Z_j), and we want to assess the impact of an unobserved confounding variable (CV_j).

Extending the Above Strategy for Sensitivity Analysis to More Complex Settings

How widely applicable is the approach to sensitivity analysis that we outlined above? Answering this question entails answering the following question: When is it sensible to re-cast HM analyses as cluster-level regression analyses employing

OLS estimates of level-1 intercepts or slopes as outcomes? That is, in what situations or settings will cluster-level regression analyses yield point estimates and standard errors for fixed effects of interest that are highly similar to those produced via a REML/EB estimation strategy?

Means-as-outcomes settings

In means-as-outcomes settings in which the data are balanced, we saw that an OLS regression of $\overline{Y}_j$ on cluster-level predictors will produce estimates of fixed effects and their standard errors identical to those produced via a REML/EB approach. It is also likely that an OLS cluster-level regression approach will produce fairly similar results when cluster sample sizes vary but not to a large degree, or when the variability in n_j is appreciable but $\hat{\tau}$ is large relative to the error variances of the $\overline{Y}_j$'s. In such situations, the weight accorded to each cluster should be somewhat similar.

Note that a Weighted Least Squares (WLS) cluster-level regression approach provides a valuable option when cluster sample sizes vary. Consider, for example, a means-as-outcomes setting in which there is a single level-2 predictor and n_j varies across clusters. In this case we could use weights based on REML estimates of the variance components (i.e., $1/(\hat{\tau}+\hat{V}_j)$, where $\hat{V}_j = \hat{\sigma}^2/n_j$) to compute weighted versions of $S_{\overline{Y}_j}$, S_{W_j} and $r_{\overline{Y}_j \bullet W_j}$, and then substitute these quantities into Equations 9 and 11b to obtain a WLS estimate of γ_{01} and its standard error (see Appendix A). Note that the resulting t ratio would reduce to the WLS estimate of $r_{\overline{Y}_j \bullet W_j}$ divided by its standard error. Such an approach will yield point estimates for fixed effects identical to those produced via a REML/EB approach in means-as-outcomes settings, and standard errors that will be extremely close, though not necessarily equal to, those produced via REML/EB. (This is due to a subtle difference in how standard errors are computed under these two approaches, which will be discussed below. As will be seen, a WLS approach analogous to the one depicted in Appendix A provides certain advantages when we wish to incorporate weights in our sensitivity analysis.)

Adjusted means as outcomes

In evaluation studies in which clusters (e.g., classes) are nested within treatment type, the level-1 parameter of primary interest is typically an adjusted

mean outcome score for cluster j. Consider, for example, the following two-level model:

$$Y_{ij} = \beta_{0j} + \beta_{1j}(X_{ij} - \bar{X}_{..}) + r_{ij} \qquad (28a)$$

where X_{ij} is a level-1 covariate that is centered around its grand mean, and at level-2:

$$\beta_{0j} = \gamma_{00} + \gamma_{01}W_j + u_{0j} \qquad (28b)$$

$$\beta_{1j} = \gamma_{10}$$

Analogous to ANCOVA models, β_{0j} represents an adjusted mean outcome score for cluster j. Suppose that X_{ij} is a pretest score that is positively related to outcome scores. If, for example, pretest scores in class j are, on average, lower than the grand mean, then the expected outcome score for class j will be adjusted upwards.

In this example we see that the adjusted means are modeled as a function of a single level-2 predictor, but of course the model may contain multiple predictors. Treatment group indicator variables, measures of implementation, and various compositional characteristics are possible predictors that might be included in the level-2 equation for β_{0j}.

Now suppose primary interest centers on inferences concerning γ_{01}. In programs for multilevel modeling, a mixed-modeling strategy such as the one depicted in Raudenbush and Bryk (2002, pp. 44-45) is used to obtain estimates of the fixed effects in "adjusted-means-as-outcomes" models such as the one depicted above. But conceptually, we can view the level-2 model for β_{0j} as a regression model relating differences in W_j to β_{0j}. Thus to estimate γ_{01}, an alternative approach would entail regressing OLS estimates of the β_{0j} (i.e., $\bar{Y}_{(ADJ)_j}$) on W_j:

$$\bar{Y}_{(ADJ)_j} = \gamma_{00} + \gamma_{01}W_j + e_j$$

where

$$\bar{Y}_{(ADJ)_j} = \bar{Y}_{.j} - \hat{\gamma}_{10}(\bar{X}_{.j} - \bar{X}_{..})$$

Note that $\hat{\gamma}_{10}$ can be obtained in an analysis in which we fit the HM defined by Equations 28a and b using a REML/EB estimation strategy. Estimates of the error variances of the $\overline{Y}_{(ADJ)_j}$, i.e., $\hat{V}_j$, can be obtained from REML/EB analyses as well. This will be discussed further in the context of our illustrative examples.

Thus, for example, in Equations 9 and 11b, we could replace $S_{\overline{Y}_j}$ with the standard deviation of the $\overline{Y}_{(ADJ)_j}$ (i.e., $S_{\overline{Y}_{(ADJ)_j}}$), and similarly replace $r_{\overline{Y}_j \bullet W_j}$ with the correlation of $\overline{Y}_{(ADJ)_j}$ and W_j (i.e., $r_{\overline{Y}_{(ADJ)_j} \bullet W_j}$). Note that we could also employ a WLS strategy analogous to the means-as-outcomes case described above. That is, we can use weights based on REML estimates of the variance components (i.e., $1/(\hat{\tau} + \hat{V}_j)$), compute weighted versions of $S_{\overline{Y}_{(ADJ)_j}}$, S_{W_j} and $r_{\overline{Y}_{(ADJ)_j} \bullet W_j}$, and substitute these quantities into Equations 9 and 11b. Both the un-weighted and weighted approaches extend readily to settings in which adjusted means are modeled as a function of two or more predictors.

We have found these strategies to produce estimates and standard errors for fixed effects in level-2 equations for adjusted means that are extremely similar to those produced via programs such as HLM. As will be seen in our first illustrative example below, there are simple checks that one can do to gauge if one is in a situation where cluster-level regressions employing the $\overline{Y}_{(ADJ)_j}$ as outcomes produce estimates of fixed effects of interest and their standard errors that are similar to those obtained via a REML/EB strategy. (Note that there is one situation we have encountered where such an approach may result in underestimation of standard errors. See Endnote 1.)

Slopes as outcomes

Such models come into play in settings in which within site treatment/control group contrasts (β_{1j}) are treated as outcomes at level 2 and modeled as a function of measures of implementation. This involves re-casting the estimation of fixed effects in the level-2 equation for β_{1j} as a cluster-level regression of $\hat{\beta}_{1j}$ on various cluster-level predictors. Similar to cluster-level regressions involving $\overline{Y}_j$ or $\overline{Y}_{(ADJ)_j}$ as outcomes, the estimate of a fixed effect of interest and its standard error can be re-expressed as a function of standard deviations and correlations (e.g., $S_{\hat{\beta}_{1j}}$ and $r_{\hat{\beta}_{1j} \bullet W_j}$). Such quantities would be substituted into the various equations presented above (e.g., Equations 9 and 11b).

As in cluster-level regressions involving $\overline{Y}_j$ or $\overline{Y}_{(ADJ)_j}$ discussed above, one can carry out weighted (WLS) or un-weighted (OLS) regressions. In the analyses we have conducted, we have found that cluster-level regressions employing the $\hat{\beta}_{1j}$ as outcomes yield point estimates and standard errors for fixed effects of interest that are very similar to those obtained via the HLM program.

A cluster-level regression approach employing $\hat{\beta}_{1j}$ as an outcome can be expected to produce results fairly similar to those obtained via REML/EB in settings in which the same set of predictors is used to model cluster intercepts and slopes (i.e., β_{0j} and β_{1j}) at level 2. As Raudenbush and Bryk note (2002, pp. 272-3), this can help mitigate possible dependencies between the estimates of the fixed effects in the level-2 equation for β_{0j} and in the level-2 equation for β_{1j}. Group-mean centering the level-1 treatment/comparison group indicator variable will help mitigate such dependencies as well, since the resulting sampling covariance between the OLS estimates of β_{0j} and β_{1j} will be equal to 0. Also, as will be shown in the analyses of the Transition Mathematics data – our second example – there are simple checks we can do to see if we are in a situation where it is sensible to employ a cluster-level regression approach

Potential advantages of WLS

WLS has a couple of features that make it particularly useful in situations in which there is some uncertainty about the magnitude of τ. (For a valuable overview of WLS, see Neter, Kutner, Nachtscheim & Wasserman, 1996.) Let's begin by extending the cluster-level regression model shown in Equation 5 to unbalanced settings:

$$\overline{Y}_{.j} = \gamma_{00} + \gamma_{01} W_j + e_j \ , \ e_j \sim N(0, D_j)$$

The residual term e_j essentially consists of two components: A random effect (u_{0j}) connected with cluster j and an error component stemming from the sampling variance of $\overline{Y}_j$ (r_j), where u_{0j} ~ N(0, τ) and r_j ~ N(0, V_j). Thus $D_j = \tau + V_j$. Note that when τ and V_j are known, and we multiply the left- and right-hand sides of the above equation by $1/\sqrt{(\tau + V_j)}$, then $Var(1/\sqrt{(\tau + V_j)} \times e_j) = 1/(\tau + V_j) \times Var(e_j) = 1/(\tau + V_j) \times D_j = \sigma^2_{WLS} = 1$. Thus in a WLS setting in which τ and the sampling variances are known, we could simply set the residual variance parameter σ^2_{WLS} equal to a value of 1.

But generally we will be working with estimates of τ and V_j as shown in the formulas in Appendix A. Rather than fixing σ^2_{WLS} equal to a value of 1, a reasonable thing to do is to obtain an estimate of σ^2_{WLS} based on the data. Note that an estimate of the mean square error appears in the numerator of the standard error for $\hat{\gamma}_{01(WLS)}$ (see Equation A.4). We can see that the squared residual for each cluster (i.e., $(\overline{Y}_{.j} - [\hat{\gamma}_{00(WLS)} + \hat{\gamma}_{01(WLS)} W_j])^2$) is divided by the quantity $\hat{\tau} + \hat{V}_j$. Note that if we are working with an estimate of τ that is a good reflection of the amount of random effects variance that remains after taking into account W_j —for example, the REML estimate of τ —then the resulting estimate of the MSE (i.e., $\hat{\sigma}^2_{WLS}$) will be approximately equal to 1. But if we are working with an estimate of τ that is too big, then we will obtain an estimate of the MSE that is appreciably less than 1; with weights of the form $1/(\hat{\tau} + \hat{V}_j)$, we would essentially be dividing the squared residuals that appear in the formula for the MSE by amounts that are too large, which would diminish the magnitude of the MSE. Similarly, if we are employing an estimate of τ that is too small, then we will obtain an estimate of the MSE that is appreciably larger than 1, since we would be dividing the squared residuals by quantities that are too small, thus inflating our estimate of the MSE. For a discussion of this role of the estimated MSE in connection with using WLS in analyses of non-nested data, see Neter et al. (1996).

When we consider the impact of an unmeasured confounding variable (CV_j) on inferences concerning γ_{01}, theoretically the inclusion of CV_j in the analysis should result in a reduction of τ. As we noted on p. 13, the resulting MSE from a sensitivity analysis based on an OLS approach can be used to obtain rough estimates of τ in such situations. As will be seen in our illustrative example, weights based on such estimates of τ along with estimates of sampling variances can be used to conduct sensitivity analyses via WLS. The resulting MSE from the WLS analysis provides a kind of gauge regarding the adequacy of our estimate of τ. Thus, for example, if the resulting MSE is substantially smaller than a value of 1, this signals that our estimate of τ is overestimating the amount of random effects variance that remains after taking into account CV_j; if the resulting MSE is approximately equal to 1, this would signal that our estimate of τ seems to adequately reflect that amount of random effects variance which remains.

Another nice feature of WLS is that even if we might be dividing the squared residuals by quantities that are, for example, too big, we will also be dividing the squared deviations of the predictor values in the denominator of the standard error

by quantities that are too big as well (see, e.g., Eq. A.4), and so these two things will tend to counterbalance each other to some extent. Thus the resulting standard error may not be too adversely affected when employing estimates of τ that are too large or too small; that is, this kind of counterbalancing may provide us with some robustness. Again, this can be useful when we are conducting analyses to assess the impact of unmeasured confounding variables. Our example helps illustrate these ideas. Situations in which estimates of fixed effects and their standard errors may be somewhat sensitive to the estimate of τ that we employ are discussed below.

Illustrative Example One: The Integrated Mathematics Assessment Study

In this example, we focus on the data from a study of mathematics curricula and instructional practices called for in such reform-minded documents as the NCTM standards (see Gearhart et al., 1999; Saxe, Gearhart & Seltzer, 1999). In particular, this study focused on a mathematics instructional unit for upper elementary students called *Seeing Fractions*, and on the development of teaching practices and skills thought to be integral to the successful implementation of such units (see Gearhart et al., 1999; Saxe et al., 1999). *Seeing Fractions* was designed to enhance engagement with mathematical concepts and the development of problem-solving skills. However, implementing these materials successfully is very challenging, and requires considerable skill and training. For example, teachers must be adept at eliciting and building on student thinking in the context of whole-class discussions of problem-solving.

The sample for this study consisted of 21 upper elementary teachers and their students in the greater Los Angeles area. Sixteen of the teachers in the sample had prior experience using *Seeing Fractions*, and used *Seeing Fractions* during the course of study. Based on a random assignment procedure discussed in Gearhart et al. (1999) and Seltzer (2004), 9 of these teachers were assigned to an intensive professional development program called Integrated Mathematics Assessment (IMA), which was intended to help teachers develop the kinds of instructional skills viewed as being central to the successful implementation of *Seeing Fractions*, and 7 were assigned to a program called Collegial Support (SUPP), which provided teachers with opportunities to discuss and reflect on their instructional practices. In addition, 5 teachers committed to using traditional texts were recruited for the study, and used traditional instructional materials with their students during the course of the study; we term this condition TRAD. Note that the three groups of

teachers were similar in terms of years of teaching experience, and the IMA and SUPP teachers were similar, on average, in terms of participation in relevant prior professional development programs.

As noted in Gearhart et al. (1999) and Seltzer (2004), the assignment procedure for this study resulted in a situation in which the IMA students tended to be more advantaged in terms of their pretest skills and knowledge and their English language proficiency. Therefore adjustment for important ways in which the students in these 21 classes differ at the outset is essential.

We now pose the following level-1 model:

$$Y_{ij} = \beta_{0j} + \beta_{1j}(PREC_{ij} - \overline{PREC}_{..}) + \beta_{2j}(PREP_{ij} - \overline{PREP}_{..}) + \beta_{3j}(INCIP_{ij} - \overline{INCIP}_{..}) + r_{ij} \ ,$$

$$r_{ij} \sim N(0, \sigma^2) \qquad (30)$$

where Y_{ij} is the score for student i in class j on a 13 item problem-solving posttest in the domain of fractions; the items on this test could not be solved employing routine, algorithmic approaches. $PREC_{ij}$ and $PREP_{ij}$ are student i's scores on a problem-solving pretest and a procedural-item pretest, respectively, and $INCIP_{ij}$ is an indicator variable that takes on a value of 1 if student i began the school year with an incipient understanding of fractions (0 otherwise). Note that we initially included an indicator variable in our level-1 model capturing a student's English language proficiency (ELP). However, the resulting point estimate of the ELP coefficient was extremely close to a value of 0 and not statistically significant; hence, we excluded this covariate from our analyses in this example. When ELP is included, the results are extremely similar to those presented below. Though not quite statistically significant, there is some suggestion of a cross-level interaction between IMA and ELP (see Kim and Seltzer, 2006).

As a result of grand-mean centering the level-1 predictors, β_{0j}, analogous to ANCOVA models, represents the adjusted problem-solving posttest mean for class j. In a between-class (level-2) model, we model adjusted class posttest means as a function of IMA and $SUPP$ indicator variables:

$$\beta_{0j} = \gamma_{00} + \gamma_{01} IMA_j + \gamma_{02} SUPP_j + u_{0j},\ u_{0j} \sim N(0, \tau) \qquad (31)$$

$$\beta_{1j} = \gamma_{10}$$

$$\beta_{2j} = \gamma_{20}$$

$$\beta_{3j} = \gamma_{30}$$

where *IMA* is coded 1 if class j was taught by a teacher who participated in the *IMA* program (0 otherwise), and $SUPP_j$ is coded 1 if class j was taught by a teacher who participated in *SUPP* (0 otherwise). Thus γ_{01} represents the expected difference in problem-solving posttest scores between students taught by *IMA* and *TRAD* teachers holding constant the covariates in our level-1 model, and similarly, γ_{02} captures the expected difference between students taught by *SUPP* and *TRAD* teachers.

In Table 1 we see that the expected difference between *IMA* and *TRAD* classes in problem-solving posttest scores is substantial (i.e., 2.24 points), and highly significant. (Note that the expected score for *TRAD* classes is approximately 3.9 points.) In addition, the *SUPP* vs. *TRAD* contrast is statistically significant as well, though the magnitude of the point estimate is approximately 1 point less than the *IMA* vs. *TRAD* contrast.

Table 1.
Estimates of the Effects of the IMA and SUPPORT Programs Before and after Adding Conceptual/Assessment OTL to the Level-2 Model

	Coefficients (SE)		
Predictors	1. Model without $CNCPT_j$	2. Model with $CNCPT_j$	3. Reduced model with $CNCPT_j$
IMA vs. TRAD	2.24 (.56) t = 4.03	.59 (.81) t = .72	
SUPP vs. TRAD	1.23 (.58) t = 2.13	-.13 (.73) t = -.18	
CNCPT		.84 (.33) t = 2.55	.98 (.20) t = 4.95
	$\hat{\tau}$ = .69	$\hat{\tau}$ = .47	$\hat{\tau}$ = .49

A goal of reform-minded curricula in combination with professional development programs such as IMA or SUPP, is to help teachers provide students with certain kinds of potentially powerful learning opportunities. Thus an important facet of this study focused on collecting data on teachers' instructional practices via videotape and field notes. A key measure developed from these data captures the extent to which a teacher provides opportunities for engagement with mathematical (i.e., fractions) concepts in discussions of problem-solving in ways that build on students' thinking. Gearhart et al. (1999) refer to this measure as Conceptual/Assessment Opportunity to Learn (C/A OTL).

We employ the same level-1 model in Equation 30 and expand the level-2 model as follows:

$$\beta_{0j} = \gamma_{00} + \gamma_{01} IMA_j + \gamma_{02} SUPP_j + \gamma_{03} CNCPT_j + u_{0j},\ u_{0j} \sim N(0, \tau) \qquad (32)$$

where $CNCPT_j$ is the C/A OTL score for the teacher of class j. Note that this measure has been standardized. As can be seen in Table 1, adding this measure of teacher practice to the model reduces the $IMA/TRAD$ contrast to approximately .60 points ($t = .72$), and the $SUPP/TRAD$ contrast to a value that is quite close to 0 ($t = -.18$). The estimate of the fixed effect connected with $CNCPT$, however, is more that twice its standard error. The magnitude of the point estimate suggests an increase in expected class posttest performance of a little under a point when C/A OTL increases one standard deviation.

The question arises, however, whether C/A OTL is in fact a key mechanism –-a key constellation of practices—through which *Seeing Fractions* and similar reform-minded materials impact student problem-solving skills. It may be the case, for example, that other types of OTL or instructional practices are confounded with C/A OTL. Or it may be the case that certain class compositional characteristics are confounded with C/A OTL.

Another measure of teacher practice constructed by Gearhart et al. is termed Numerics OTL, which measures the extent to which a teacher provides opportunities to use and interpret numeric representations of fractions that capture important underlying concepts. Despite very low degrees of power (i.e., the data set

contains only J=21 classes, and there is appreciable collinearity among the predictors in Equation 32), the results change very little when we add Numerics OTL to Equation 32. The pattern of results is similar when we add less proximal measures one variable at a time (e.g., removing Numerics OTL and adding, for example, years of teaching), or add various class mean intake characteristics one at a time, including class mean pretest scores and the proportion of students in a class who are English language proficient. Note that in all of these analyses except one, the p-value connected with the fixed effect estimate for C/A OTL was less than .05. In an analysis in which class-mean procedural pretest performance was added to Equation 32, the p-value for the resulting fixed effect estimate of C/A OTL was .12. The fixed effect estimates for IMA, SUPP and the various covariates that were added never reached statistical significance.

We now re-fit our model with the treatment group indicators set aside:

$$\beta_{0j} = \gamma_{00} + \gamma_{01} CNCPT_j + u_{0j} \ , \ u_{0j} \sim N(0, \tau) \qquad (33)$$

As can be seen in Table 1, the estimate of the level-2 random effects variance component based on the reduced model (Model 3) is nearly as small as the estimate based on Model 2 (.49 vs. .47), and both of these estimates are substantially smaller than the estimate based on Model 1 (.69), which does not contain *CNCPT*. The reduced model results in a point estimate for *CNCPT* that is a little larger than the estimate based on Model 2 (i.e., .98 vs. .84). One difference in results that stands out a bit more is that the standard error for γ_{01} based on the reduced model is substantially smaller (i.e., .197 vs. .330). As Kenny points out, possible mediating variables will be highly correlated with treatment indicators, which will result in large standard errors for the predictors in the model (see, for example, the notes on mediation on David Kenny's website, and Judd & Kenny [1981]). In this connection, note that regressing *CNCPT* on the *IMA* and *SUPP* indicators results in an R-squared value of .65.

But questions concerning unmeasured confounding variables arise. Are there various unmeasured aspects of practice that are confounded with C/A OTL? For example, Gearhart et al. attempted to construct a measure of Graphics OTL but levels of rater agreement were found to be unacceptable. If such a measure were available and added to Equation 33, one might wonder whether the impact would be such that the estimate of the fixed effect for C/A OTL is no longer statistically

significant. In addition, the students in the *IMA*, *SUPP* and *TRAD* classes perhaps may also have differed in important ways at the outset of the study not fully captured by the available student intake/pretest measures. Suppose that such a measure were available, aggregated to the classroom level, and added to Equation 33. The coefficient for this covariate would capture the between-class relationship between the covariate and adjusted class posttest means, which can be viewed as a combination of the within-class relationship between the covariate and the outcome (e.g., β_w) and the contextual effect of the covariate (e.g., β_c) (i.e., $\beta_w + \beta_c$). Again the question is what impact might this covariate have on results concerning C/A OTL?

Extending Frank's work (2000) to multilevel settings, we now illustrate our strategy for assessing the impact of an unobserved confounding variable on results concerning a fixed effect of interest in a level-2 equation for adjusted cluster mean outcome scores. We first consider whether we are in a situation where it is sensible to re-cast multilevel analyses as cluster-level regression analyses employing OLS estimates of adjusted means (i.e., $\overline{Y}_{(ADJ)_j}$) as outcomes. Figure 1 presents a plot of the adjusted means versus C/A OTL, and we see a roughly linear upward trend.

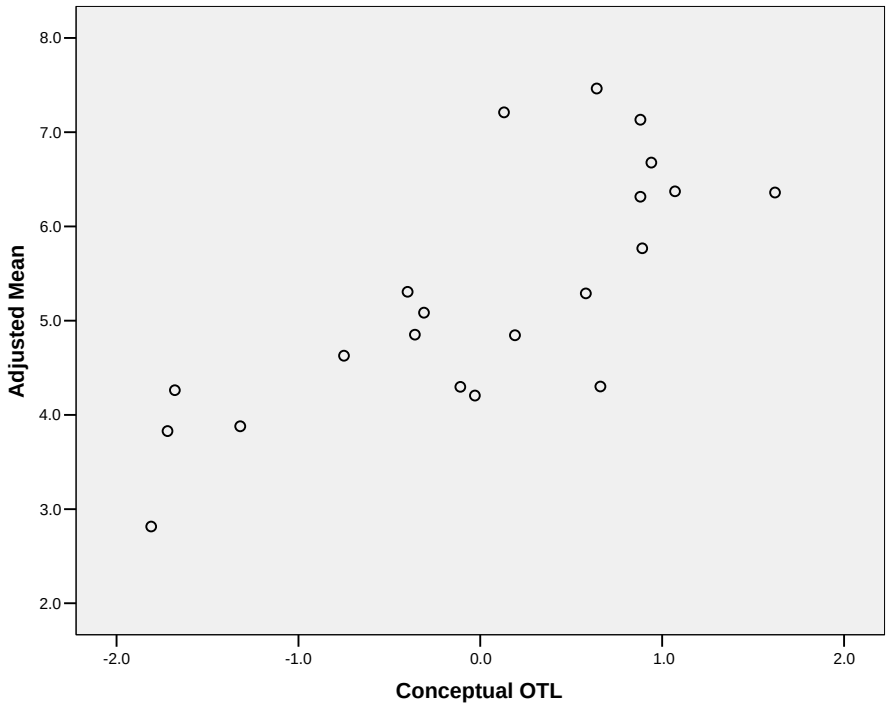

Figure 1. OLS estimates of adjusted class mean outcome scores versus conceptual OTL values.

Note that an OLS regression of $\overline{Y}_{(ADJ)_j}$ on $CNCPT_j$ results in a point estimate, standard error and t ratio for γ_{01} (i.e., the fixed effect for C/A OTL) that are extremely similar to those produced by a REML/EB strategy using the HLM program (see Table 2).

Table 2.
Fixed Effect Estimates ($\hat{\gamma}_{01}$) for Conceptual OTL Based on the HM Defined in Equations 30 and 3, and Corresponding Correlations

	$\hat{\gamma}_{01}$ (SE)	t	$r_{\overline{Y}_{(ADJ)}\bullet CNCPT}$
HLM	.976 (.197)	4.949	.750
OLS	.975(.190)	5.140	.763
WLS	.976(.192)	5.068	.758

Employing weights based on the HLM analysis (i.e., $1/[\,\hat{\tau}+\hat{V}_j\,]$), a WLS regression of $\overline{Y}_{(ADJ)_j}$ on $CNCPT_j$ yields results that are even closer to those produced by HLM. Thus it appears that we are in a situation where employing a cluster-level regression approach via OLS or WLS is quite sensible.

Note further that the un-weighted (OLS) correlation between $\overline{Y}_{(ADJ)_j}$ and $CNCPT_j$ is .763, while the weighted (WLS) correlation is .758. Note also that transforming the t ratio produced by HLM (4.949) to a correlation (i.e., $4.949\Big/\sqrt{(J-2)+(4.949)^2}$) yields a value of .750. As can be seen, the values of these three correlations are extremely close.

We now follow steps 1 – 7 on pages 11-13 above, substituting $\overline{Y}_{(ADJ)_j}$ for $\overline{Y}_j$ throughout. We will be conducting several sensitivity analyses, and for each one the threshold t value ($t^{\#}$) is the upper .025 critical value based on a t distribution with $J-3=18$ degrees of freedom, i.e., $t^{\#}=2.101$. The corresponding threshold partial correlation value is

$$r^{\#}_{\overline{Y}_{(ADJ)_j}\bullet CNCPT_j|CV_j} = \frac{t^{\#}}{\sqrt{(21-3)+(t^{\#})^2}} \qquad (34)$$
$$= .444$$

We first employ an OLS approach. Using the un-weighted correlation coefficient ($r_{\bar{Y}_{(ADJ)_j} \bullet CNCPT_j} = .763$) and the threshold partial correlation (i.e., $r^{\#}_{\bar{Y}_{(ADJ)_j} \bullet CNCPT_j | CV_j} = .444$), we solve for k as in Step 5 above:

$$k = \frac{r_{\bar{Y}_{(ADJ)_j} \bullet CNCPT_j} - r^{\#}_{\bar{Y}_{(ADJ)_j} \bullet CNCPT_j | CV_j}}{1 - r^{\#}_{\bar{Y}_{(ADJ)_j} \bullet CNCPT_j | CV_j}} \qquad (35)$$

$$= .574$$

Taking the square root of k, we obtain $r_{\bar{Y}_{(ADJ)_j} \bullet CV_j} = .758$, and $r_{CNCPT_j \bullet CV_j} = .758$.

Substituting $S_{\bar{Y}_{(ADJ)_j}}$, S_{CNCPT_j}, $r^{\#}_{\bar{Y}_{(ADJ)_j} \bullet CNCPT_j | CV_j}$, $r_{\bar{Y}_{(ADJ)_j} \bullet CV_j}$ and $r_{CNCPT_j \bullet CV_j}$ into Equations 16 and 18, gives us the corresponding estimate of γ_{01} and its standard error holding constant CV_j (see Table 3). Using these quantities, we can also obtain an estimate of the resulting R-squared value and Mean Square Error based on the inclusion of CV_j in the model. (See Table 3; see, also, Equations 19 – 20b).

Table 3.
Sensitivity of Inferences Concerning the Effects of Conceptual OTL to the Impact of an Unmeasured Confounding Variable

	OLS	WLS I	WLS II
$r_{\bar{Y}_{(ADJ)_j} \bullet W_j}$	.763	.758	.757
$S_{\bar{Y}_{(ADJ)_j}}$	1.278	1.465	1.627
S_{W_j}	1.000	1.139	1.262
k	.574	.565	.563
$\sqrt{k} = r_{\bar{Y}_{(ADJ)_j} \bullet CV_j}$ and $r_{CNCPT_j \bullet CV_j}$	.758	.752	.750
$\hat{\gamma}_{01(CV)}$	.567	.571	.572
$SE(\hat{\gamma}_{01(CV)})$	.270	.272	.272
R^2	.658	.650	.649
MSE	.621	.833	1.032

This analysis suggests that the correlation between an unmeasured confounding variable CV_j and $\overline{Y}_{(ADJ)_j}$, and between CV_j and $CNCPT_j$, would need to exceed values of approximately .76 to result in a partial correlation between $\overline{Y}_{(ADJ)_j}$ and $CNCPT_j$ (holding constant CV_j) that yields a t ratio below $t^{\#} = 2.101$. Similarly, correlations of this magnitude would be needed to result in a fixed effect estimate for $CNCPT_j$ (holding constant CV_j) and its standard error that yields a t ratio below the threshold t value.

Correlations exceeding .76 are large by social science standards. Note also that among the various observed covariates in this study (see Table 4), only one has a correlation with $\overline{Y}_{(ADJ)_j}$ or $CNCPT_j$ that exceeds .76, i.e., Relevant Prior Professional Development (see Table 4). But while this covariate has a correlation with $CNCPT_j$ of .77, its correlation with $\overline{Y}_{(ADJ)_j}$ is only .41. As can be seen, many of the correlations in Table 4 are small to medium in size. We also see that the correlations between IMA and $\overline{Y}_{(ADJ)_j}$, and between IMA and $CNCPT_j$ are fairly substantial. Note further that when we compute the partial correlations between $\overline{Y}_{(ADJ)_j}$ and $CNCPT_j$ holding constant the observed covariates one at a time, we can see in Table 4 that all of the resulting partial correlations far exceed the threshold value of .444. The smallest partial correlation that we obtain (i.e., .667) arises when we hold constant *IMA*, and clearly the magnitude of that correlation far exceeds .444.

Table 4.
Correlations of observed covariates with adjusted problem-solving posttest means and with Conceptual OTL. Each entry in the last column is the product of the correlations for a given predictor with adjusted problem-solving posttest means and with Conceptual OTL.

	Adjusted Prob.-Solving Posttest Mean	Conceptual OTL	$r_{\bar{Y}_{(ADJ)_j} \cdot CNCPT_j \mid obs.cov.}$
Conceptual OTL	0.763**		
Numerics OTL	0.109	-0.001	0.768
IMA Indicator	0.578**	0.514*	0.667
Support Indicator	-0.058	0.173	0.786
Relevant Prof. Dev.	0.408	0.768**	0.769
Math Specialist	0.247	0.189	0.753
Years Teaching	0.052	-0.230	0.797
Prop. English Lang. Proficient	0.393	0.381	0.721
Prop. W. Incip. Understanding	0.194	-0.016	0.781
Procedural Pretest Mean	-0.104	-0.109	0.760
Problem-solving Pretest Mean	0.243	0.136	0.760

N=21; **: p<.01; *: p<.05

Table 4 provides an opportunity to make a more general point. Namely, examining correlations of observed covariates (e.g., Numerics OTL) with a level-2 outcome and predictor of interest (e.g., $\bar{Y}_{(ADJ)_j}$ and $CNCPT_j$) provides a valuable frame of reference for thinking about the magnitudes of correlations with unobserved covariates (e.g., $r_{\bar{Y}_{(ADJ)_j} \cdot CV_j}$ and $r_{CNCPT_j \cdot CV_j}$) needed to impact the partial correlation or fixed effect estimate of interest to the point where it is below a threshold t value. That is, we can begin to get a sense of how likely it might be to obtain correlations of such magnitude. In particular, one would hope to have a number of covariates that provide measures of various aspects of instructional practice (proximal factors) and of various compositional characteristics that might, based on the relevant research literature or on pilot work, be confounded with one's predictor of interest (e.g., $CNCPT_j$), and pay particular attention to the correlations of such observed covariates with one's level-2 predictor and outcome of interest. In

short, we need to do more than simply examine the magnitude of quantities such as the square root of *k*.

In Table 3, we also present results from two analyses employing weighted correlations between $\overline{Y}_{(ADJ)_j}$ and $CNCPT_j$, and weighted versions of $S_{\overline{Y}_{(ADJ)_j}}$ and S_{CNCPT_j}. As can be seen, the resulting values for $r_{\overline{Y}_{(ADJ)_j} \cdot CV_j}$ and $r_{CNCPT_j \cdot CV_j}$ are extremely similar to those obtained via an OLS (un-weighted) approach. The first weighted analysis (WLS I) employs weights based on the REML estimate of τ obtained in the HLM analysis in which β_{0j} was modeled as a function of $CNCPT_j$ (see Equation 33), i.e., $1/(\hat{\tau} + \hat{V}_j)$, where $\hat{\tau} = .49$. But note that the inclusion of a confounding variable should, theoretically, result in a reduction in τ. Thus to obtain a rough estimate of τ that reflects the inclusion of a confounding variable in the level-2 model, we subtracted the average value of the $\hat{V}_j$'s from the estimate of the MSE obtained in the sensitivity analysis in which no weights were employed (i.e., .621). This resulted in a value of .35, which we used as our estimate of τ in the second weighted analysis. As can be seen in Table 3, both weighted analyses yield results that are quite similar to the first analysis.

As noted above, WLS has a couple of features that make it useful in situations in which there is some uncertainty about the magnitude of τ. In general, if we are working with an estimate of τ that is a good reflection of the amount of random effects variance that remains after taking into account the level-2 predictors in a given model, then the estimate of the MSE obtained via WLS should be close to a value of 1. If our estimate of τ is too big, then WLS will produce an MSE that is appreciably less than 1; with weights of the form $1/(\hat{\tau} + \hat{V}_j)$, we would essentially be dividing the squared residuals that appear in the formula for the MSE by amounts that are too large, which would diminish the magnitude of the MSE. Note that in the WLS I analysis, in which we employ an estimate of $\hat{\tau} = .49$—an estimate that does not take into account the impact of CV_j — the resulting MSE is .833. In contrast, in the WLS II analysis, in which we employ an estimate of $\hat{\tau} = .35$, the resulting MSE is 1.032. Another nice thing about WLS is that even if we might be dividing the squared residuals by quantities that are too big, we will also be dividing the squared deviations of the predictor values in the denominator of the standard error by quantities that are too big (see, e.g., Equation A.4), and so these two things may tend to counterbalance each other in numerous settings. In this particular example, the SEs obtained in the WLS I and WLS II analyses are the same. Note, however, that the V_j's are not too heterogeneous in this application.

Differences in values of τ that we employ can be consequential in settings in which the number of clusters is small and the sample contains an influential outlying cluster whose sampling variance is very large or small in relation to the sampling variances of the other clusters. In such situations, differences in values for τ can increase or diminish the influence of the outlying cluster on estimates of the fixed effects, which in turn can impact the magnitudes of the squared residuals, which are key elements of the MSE.

The above results in combination with various studies of reform-minded mathematics instruction cited Saxe et al. (1999) would seem to point to Conceptual/Assessment OTL (or perhaps a particular facet of it) as being a key mechanism through which *Seeing Fractions* impacts student problem-solving skills. But this may not be the only mediating factor. For example, in future studies, one might want to investigate whether graphics OTL or the extent to which particular kinds of examples are employed might also, along with Conceptual/Assessment OTL, be mediating factors.

Discussion

In many multisite evaluation studies, questions of primary interest often focus on whether particular facets of implementation or other aspects of classroom or school environments are critical to a program's success. From a statistical analysis standpoint, addressing such questions entails drawing inferences concerning fixed effects that capture how differences in implementation, for example, relate to differences in adjusted class mean outcome scores (β_{0j}) or in site treatment/control contrasts (β_{1j}). In carrying out such analyses, it is crucial that we attend to both measured and unmeasured variables that may be confounded with the effects of implementation. While measured potential confounding variables can be included as covariates in our analyses, attending to the possible effects of unmeasured confounding variables on our inferences is clearly very challenging.

In this paper we have outlined a strategy for studying the sensitivity of inferences concerning fixed effects of interest in analyses of multisite data to unmeasured confounding variables. Our strategy builds on Frank's (2000) approach to sensitivity analysis, which was developed for use in settings in which one is working with linear models in analyses of non-nested data. The crux of our strategy entails re-casting focal cluster-level (level-2) equations in Hierarchical Models as weighted or un-weighted cluster-level regressions in which OLS estimates of β_{0j} or

β_{1j} (i.e., $\overline{Y}_{(ADJ)_j}$; $\hat{\beta}_{1j}$) are employed as outcomes and modeled as a function of key cluster-level predictors. The weighted approach that we outlined entails constructing weights using estimates of variance components obtained from HLM analyses.

The example that we focused on above provides an illustration of our approach in situations in which intact organizational units are nested in different levels or types of treatment. Multilevel analyses in such cases typically involve modeling adjusted class means, for example, as a function of various class-level characteristics, including treatment type, key aspects of implementation, and various observed covariates. In matched-pairs designs, treatment/comparison contrasts for the pairs (i.e., blocks) in our sample are modeled as a function of key pair characteristics (e.g., particular facets of implementation, observed site-level covariates). For a sketch of our approach in matched-pairs settings, please see Appendix B.

In our illustrative example, we focused on inferences concerning a fixed effect (γ_{01}) relating differences in Conceptual/Assessment OTL ($CNCPT_j$) to adjusted class-mean problem-solving posttest scores (see Equation 33). Specifically, we used our approach to investigate the sensitivity of inferences concerning γ_{01} to an unmeasured potential confounding variable (CV_j), and found that the correlations between adjusted mean outcome scores and CV_j , and between $CNCPT_j$ and CV_j, would both need to exceed a value of .76 to result in an estimate of $CNCPT_j$ *holding constant* CV_j that is no longer statistically significant.

Note that the level-2 (between-class) model in this example contained only one observed predictor (i.e., $CNCPT_j$). As pointed out above, our approach can easily be extended to settings in which level-2 models contain more that one observed predictor. Suppose, for example, that we were to include *IMA* as a covariate in Equation 33. The resulting fixed effect relating $CNCPT_j$ to $\overline{Y}_{(ADJ)_j}$, holding constant whether a class was taught by a teacher who participated in the IMA program or not, is .80 (t = 3.732). To assess the sensitivity of inferences concerning $CNCPT_j$ in this setting to an unobserved confound (CV_j), we focus on the partial correlation between adjusted mean outcome scores and CV_j holding constant *IMA*, and the partial correlation between $CNCPT_j$ and CV_j holding constant *IMA*, and ask how large must these partial correlations be to render our results for $CNCPT_j$ no longer statistically significant? Extending our approach to this setting, we find that these partial correlations must exceed values of .61.

Employing Frank's work on sensitivity analysis in developing our strategy for conducting sensitivity analysis in HM settings offers several advantages. First, it provides a fairly direct way of addressing the question: How strongly correlated must a potential confounding variable (CV_j) be with an outcome and predictor of interest such that if we were able to control for CV_j in our analysis, the fixed effect estimate connected with the predictor of interest would no longer be statistically significant at a chosen alpha value? As noted above, we can also find the magnitude of the correlations that would reduce the estimate of the fixed effect by an amount deemed substantively meaningful. Secondly, as noted in the previous paragraph, our approach can be easily extended to situations in which the focal level-2 equation in our model includes a predictor of interest (W_j) along with one or more observed covariates (Z_j). In this connection, our approach enables us to alter assumptions concerning the degree to which CV_j and Z_j are associated. For example, we might conduct a sensitivity analysis in which we assume that CV_j and Z_j are independent; in a subsequent analysis, we might assume that a moderately large proportion of the variability in CV_j (e.g., 50%) is accounted for by Z_j.

While we have focused on the use of our approach to sensitivity analysis in multisite studies of educational programs, note that our approach can be employed in an array of hierarchical modeling settings and applications. For example, the estimation of key fixed effects in many growth modeling applications could be recast as regressions of OLS estimates of individual growth parameters (e.g., growth rates) on person-level predictors. In addition, many meta-analyses could be recast as regressions of effects size estimates on study-level characteristics. Furthermore, our approach could be extended to settings in which time-series observations (level-1) are nested within students (level-2) who, in turn, are nested within different schools (level-3), and where interest centers on inferences concerning how certain school-level policies or programs relate to differences in school-mean rates of change. In such cases, the estimation of key fixed effects at the school level could be recast as regressions of estimates of school-mean rates of change on sets of school-level predictors.

Endnote 1: We have found that cluster-level regressions employing OLS estimates of adjusted means (i.e., $\overline{Y}_{(ADJ)_j}$) as outcomes can produce underestimates of standard errors for contextual effects—that is, situations in which the level-2 predictor of interest is, for example, the mean of X_{ij} aggregated to the cluster level (i.e., $\overline{X}_{.j}$). In such cases, γ_{01} in Equation 28b is the contextual effect of the pretest, which is in effect a contrast between the estimates of the between-cluster and within-cluster pretest fixed effects (i.e., $\hat{\beta}_b$ - $\hat{\beta}_w$ using the notation in Raudenbush and Bryk [2002, chpt. 5]). In the case of the HSB example on p. 140 in Raudenbush and Bryk (2002), one can see that the standard error of the contextual effect of SES is equal to the square root of the quantity $SE(\hat{\beta}_b)^2 + SE(\hat{\beta}_w)^2$. (As Raudenbush and Bryk [2002] point out, the standard error of the contextual effect can be determined from the sampling variance/covariance matrix for the fixed effect estimates based on a model in which SES is group-mean centered at level 1, and school-mean SES is employed as a predictor of β_{0j}.) Using a WLS strategy in which the $\overline{Y}_{(ADJ)_j}$ are regressed on school mean SES, the resulting point estimate of the contextual effect is identical to the value produced by HLM. However, the standard error that we obtain is approximately equal to $SE(\hat{\beta}_b)$, i.e., it does not reflect the fact that the point estimate we have is in essence a contrast between the between-cluster and within-cluster fixed effects of SES. Note that if we include SECTOR as well as MEAN SES at level 2, the point estimate and standard error that we obtain for SECTOR using a WLS strategy are extremely similar to the values produced by HLM. Thus the underestimation of standard errors seems to be confined to the estimate of the contextual effect.

The extent to which the standard errors for contextual effects via regressions involving $\overline{Y}_{(ADJ)_j}$ are too small will depend on the magnitudes of $SE(\hat{\beta}_b)^2$ and $SE(\hat{\beta}_w)^2$. Typically $SE(\hat{\beta}_w)^2$ will be substantially smaller than $SE(\hat{\beta}_b)^2$.

References

Burstein, L. (1980). The analysis of multi-level data in education research and evaluation. *Review of Research in Education, 8*, 158-233.

Cronbach, L. (1976). *Research on classrooms and schools: Formulations of questions, design and analysis*. Occasional paper. Stanford, CA: Stanford Evaluation Consortium.

Cronbach, L. (1982). *Designing evaluations of educational and social programs*. San Francisco: Jossey-Bass.

Frank, K. (2000). Impact of a confounding variable on a regression coefficient. *Sociological Methods and Research*, 29(2), 147-194.

Gastwirth, J., Krieger, A. & Rosenbaum, P. (1998). Dual and simultaneous sensitivity analysis for matched pairs. Biometrika, 85, 907-920.

Gearhart, M. Saxe, G. B., Seltzer, M. H., Schlackman, J. ,Ching C. C., Nasir, N., Fall, R., Bennett, T., Rhine, S. & Slon,T. F. (1999). Opportunities to learn fractions in elementary mathematics classrooms. *Journal for research in Mathematics Education,* 30, 287-315.

Hong, G. (2004). *Causal inference for multi-level observational data with application to kindergarten retention*. Unpublished Doctoral Dissertation, University of Michigan, Ann Arbor.

Hong, G. (2006). *Multi-level experimental designs and quasi-experimental approximations for studying intervention implementation as a mediator*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Kim, J. & Seltzer, M. (2006). *Examining heterogeneity in residual variance in experimental and quasi-experimental settings*. Paper presented at the annual meeting of American Educational Research Association, San Francisco.

Neter, J., Kutner, M., Nachtscheim, C. & Wasserman, W. (1996). *Applied linear statistical models,* 4th ed. Chicago: Irwin.

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

Raudenbush, S. & Willms, D. (1991). *Pupils, classrooms and schools: International studies of schooling from a multilevel perspective.* New York: Academic Press.

Rosenbaum, P. (2002). *Observational studies*, 2nd ed. New York: Springer-Verlag.

Rosenbaum, P. & Rubin, D. (1983). Assessing sensitivity to an unobserved binary covariate in an observation study with binary outcome. *Journal of the Royal Statistical Society, Series B*, 45, 212-218.

Saxe, G., Gearhart, M. & Seltzer, M. (1999). Relations between classroom practices and students' learning in the domain of fractions. *Cognition and Instruction*, 17, 1-24.

Seltzer, M. (1994). Studying variation in program success: A multilevel modeling approach. *Evaluation Review*, 18, 342-361.

Seltzer, M. (2004). The use of hierarchical models in analyzing data from experiments and quasi-experiments conducted in field settings. In D. Kaplan (Ed.), *The Handbook of Quantitative Methods for the Social Sciences* (pp. 259-280). Thousand Oaks, CA: Sage Publications.

Appendix A: WLS Formulae for Means-as-Outcomes Models with a Single Level-2 Predictor

$$\hat{\gamma}_{01(WLS)} = \frac{S_{\bar{Y}_j}}{S_{W_j}} \times r_{\bar{Y}_j \cdot W_j} \tag{A.1}$$

where

$$S_{\bar{Y}_j} = \sqrt{\frac{\sum_{j=1}^{J} \hat{\Delta}_j^{-1} (\bar{Y}_{.j} - \bar{Y}_{..})^2}{J-1}}, \; S_{W_j} = \sqrt{\frac{\sum_{j=1}^{J} \hat{\Delta}_j^{-1} (W_j - \bar{W}_{.})^2}{J-1}},$$

$$r_{\bar{Y}_j \cdot W_j} = \frac{\sum_{j=1}^{J} \hat{\Delta}_j^{-1} (\bar{Y}_{.j} - \bar{Y}_{..})(W_j - \bar{W}_{.})}{\sqrt{\sum_{j=1}^{J} \hat{\Delta}_j^{-1} (\bar{Y}_{.j} - \bar{Y}_{..})^2} \sqrt{\sum_{j=1}^{J} \hat{\Delta}_j^{-1} (W_j - \bar{W}_{.})^2}} \tag{A.2}$$

and

$$\hat{\Delta}_j^{-1} = (\hat{\tau} + \hat{V}_j)^{-1}, \; \hat{V}_j = \frac{\hat{\sigma}^2}{n_j}, \; \bar{Y}_{..} = \frac{\sum_{j=1}^{J} \hat{\Delta}_j^{-1} \bar{Y}_{.j}}{\sum_{j=1}^{J} \hat{\Delta}_j^{-1}}, \text{ and } \bar{W}_{.} = \frac{\sum_{j=1}^{J} \hat{\Delta}_j^{-1} W_j}{\sum_{j=1}^{J} \hat{\Delta}_j^{-1}} \tag{A.3}$$

$$SE(\hat{\gamma}_{01(WLS)}) = \frac{\sqrt{\frac{\sum_{j=1}^{J} \hat{\Delta}_j^{-1} (\bar{Y}_{.j} - [\hat{\gamma}_{00(WLS)} + \hat{\gamma}_{01(WLS)} W_j])^2}{J-2}}}{\sqrt{\sum_{j=1}^{J} \hat{\Delta}_j^{-1} (W_j - \bar{W}_{.})^2}} \tag{A.4}$$

$$= \frac{\frac{\sqrt{\sum_{j=1}^{J} \hat{\Delta}_j^{-1} (\bar{Y}_{.j} - \bar{Y}_{..})^2} \sqrt{1 - r^2_{\bar{Y}_j \cdot W_j}}}{\sqrt{J-2}}}{\sqrt{\sum_{j=1}^{J} \hat{\Delta}_j^{-1} (W_j - \bar{W}_{.})^2}} = \frac{S_{\bar{Y}_j}}{S_{W_j}} \times \frac{\sqrt{1 - r^2_{\bar{Y}_j \cdot W_j}}}{\sqrt{J-2}}$$

$$t = \frac{r_{\bar{Y}_j \cdot W_j}}{\frac{\sqrt{1 - r^2_{\bar{Y}_j \cdot W_j}}}{\sqrt{J-2}}} \tag{A.5}$$

Appendix B: Sketch of the TM Example

To help sketch the application of our approach to studies in which blocks are crossed with treatment type we consider an example from Seltzer (2004) that focuses on an analysis of the data from an evaluation of Transition Mathematics (TM), which is an innovative pre-algebra curriculum developed by the University of Chicago School Mathematics Project. Details concerning the design of the study can be found in Seltzer (2004). Briefly our sample consists of 20 carefully matched pairs of classrooms located within various school districts throughout the U.S. Within each pair, the students in one class were taught by a teacher who utilized the TM text, while the students in the other class were taught by a teacher who used the materials already in place at that particular school. The decision as to which teacher at a site would use TM and which would use the materials already in place was based on random assignment in the case of 10 sites; logistical reasons precluded this in the case of the 10 other pairs. (An analysis in Seltzer [2004] shows that the effects of TM appear to be similar on average at sites in which teachers were randomly assigned and at sites where the assignment was not random)

We now consider an analysis of one of the outcomes of interest, i.e., a measure of geometry readiness based on a 19-item test. We pose the following level-1 (within-site) model:

$$Y_{ij} = \beta_{0j} + \beta_{1j}(TRT_{ij} - \overline{TRT}_{\cdot j}) + \beta_{2j}(PRE_{ij} - \overline{PRE}_{\cdot j}) + r_{ij},\ r_{ij} \sim N(0, \sigma^2)$$

where Y_{ij} is the geometry readiness score for student i in site j, TRT_{ij} is a treatment indicator variable that takes on a value of 1 if student i in site j is a member of the TM class (0 otherwise), and PRE_{ij} is the score for student i in site j on a general math pretest. The level-1 parameter of primary interest is β_{1j}, which represents the expected TM/Comparison class contrast for site j holding constant pretest performance. By virtue of group-mean centering, β_{0j} is the mean geometry readiness score for site j. (Note also that group-mean centering removes the sampling covariance between OLS estimates of β_{0j} and β_{1j}.)

Employing a between-site model containing no predictors, we find that the average TM effect is a little over 1 point, which is basically worth 1 item on the 19-item geometry readiness test. Furthermore, we find that there is substantial heterogeneity in the effects of TM. To begin to get a sense of this, note that the OLS estimates of the site TM effects ($\hat{\beta}_{1j}$) range from -2.15 to 4.67.

The developers of TM view daily discussion of the reading passages in the TM text as a key element of the program. As such, information regarding the usage of reading in the text was obtained through a teacher questionnaire administered at the end of the school year. The TM teachers fell into two categories, i.e., those who indicated they discussed the reading in the text on a daily basis, which we term high implementation (i.e., $IMPLRDG_j$ = 1), and those who indicated that reading was discussed frequently but was not part of the daily routine, which we term low implementation ($IMPLRDG_j$ = 0).

At level-2 we model β_{0j} as a function of site-mean pretest scores, and we model site TM effects as a function of $IMPLRDG_j$:

$$\beta_{0j} = \gamma_{00} + \gamma_{01}(\overline{PRE}_{\cdot j} - \overline{PRE}) + u_{0j},\ u_{0j} \sim N(0, \tau_{00})$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11} IMPLRDG_j + u_{1j},\ u_{1j} \sim N(0, \tau_{11})$$

$$\beta_{2j} = \gamma_{20}$$

and $Cov(u_{0j}, u_{1j}) = \tau_{01} = \tau_{10}$. Note that the equation for β_{1j} is our focal level-2 equation. The base or intercept in this model represents the expected effect of TM at low implementation sites, and γ_{11} represents the expected increment in the effectiveness of TM when the level of implementation is high.

Note that the HLM estimate for γ_{11} is $\hat{\gamma}_{11}$ =2.03 and its standard error is .76 (t=2.68). An OLS regression of the $\hat{\beta}_{1j}$ on $IMPLRDG_j$ yields extremely similar results: $\hat{\gamma}_{11}$ =2.05, SE=.78 (t=2.61). Point estimates and SEs for the base are very similar as well: .12 (SE=.53) for HLM, and .16 (SE=.55) based on a cluster-level regression analysis. This points to the reasonableness of re-casting the estimation of the fixed effects in the level-2 equation for β_{1j} as a cluster level regression.

To consider the impact of an unmeasured confounding variable on inferences concerning γ_{11}, we can essentially follow the same steps as in the analysis for the

IMA data but substituting $\hat{\beta}_{1j}$ for $\overline{Y}_{(ADJ)_j}$ and $IMPLRDG_j$ for $CNCPT_j$. To implement our approach we need the correlation between $\hat{\beta}_{1j}$ and $IMPLRDG_j$, the standard deviation of the $\hat{\beta}_{1j}$'s, and the standard deviation of $IMPLRDG_j$. Note that the correlation that we obtain without employing weights is .524. Employing a threshold t value of 2.11, which corresponds to the upper .025 critical value based on a t with 17 degrees of freedom, we obtain a value of k =.126, and so $r_{\hat{\beta}_{1j} \cdot CV_j} = r_{W_j \cdot CV_j}$ = .35, and the corresponding estimate of γ_{11} holding constant CV_j is 1.78. (Note that here W_j represents $IMPLRDG_j$.) Thus the correlation between CV_j and $\hat{\beta}_{1j}$, and between CV_j and $IMPLRDG_j$ would need to exceed a value of .35 to result in a partial correlation between $\hat{\beta}_{1j}$ and $IMPLRDG_j$ (holding constant CV_j), or to result in an estimate of γ_{11} holding constant CV_j, that yields a t ratio below the threshold t value. Note that if we set the threshold t value equal to the upper .05 critical value, we obtain $r_{\hat{\beta}_{1j} \cdot CV_j} = r_{W_j \cdot CV_j}$ = .47, and the corresponding estimate of γ_{11} holding constant CV_j is 1.52.

Further work needs to be done in terms of interpreting the results of these analyses, and comparing values for $r_{\hat{\beta}_{1j} \cdot CV_j}$ and $r_{W_j \cdot CV_j}$ with correlations based on observed covariates as in Table 4 for the IMA example. But this sketch is intended to provide a sense of the applicability of our approach to a commonly encountered design in evaluation studies.