Drawing Sound Inferences Concerning the Effects of Treatment on Dispersion in Outcomes: Bringing to Light Individual Differences in Response to Treatment

CSE Technical Report 710

Michael Seltzer
CRESST/University of California, Los Angeles

March 2007

National Center for Research on Evaluation,
Standards, and Student Testing (CRESST)
Center for the Study of Evaluation (CSE)
Graduate School of Education and Information Studies
University of California, Los Angeles
GSEIS Bldg., Box 951522
Los Angeles, CA 90095-1522
(310) 206-1532

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 (IES), U.S. Department of Education.

The findings and opinions expressed in this report are those of the author(s) and do not necessarily reflect the positions or policies of the National Center for Education Research, the Institute of Education Sciences (IES), or the U.S. Department of Education.

DRAWING SOUND INFERENCES CONCERNING THE EFFECTS OF TREATMENT ON DISPERSION IN OUTCOMES: BRINGING TO LIGHT INDIVIDUAL DIFFERENCES IN RESPONSE TO TREATMENT

Jinok Kim

Michael Seltzer

CRESST/University of California, Los Angeles

Abstract

Individual differences in response to a given treatment have been a longstanding interest in education. While many evaluation studies focus on average treatment effects (i.e., the effects of treatments on the levels of outcomes of interest), this paper additionally considers estimating the effects of treatments on the dispersion in outcomes. Differences in dispersion can, under certain circumstances, signal individual differences in response to a given treatment, thereby helping us identify factors that magnify or dampen the effects of treatments that might otherwise go unnoticed. Much of this paper focuses on quasi-experiments in nested settings, which are commonly encountered in multi-site evaluation studies. In such settings, studying differences in dispersion as well as in means (e.g., differences in levels of outcomes for treatment and control group students) entails jointly modeling mean and dispersion structures in a hierarchical modeling (HM) framework. This paper shows how a well-elaborated dispersion structure based on substantive theories mitigate the problem of confounding by cluster characteristics, while a well-elaborated mean structure helps avoid confounding by individual characteristics, with regard to inferences concerning dispersion. We illustrate these ideas with analyses of the data from a study of the effectiveness of two innovative instructional programs relative to traditional instruction in elementary mathematics classrooms. We employ a fully Bayesian approach and discuss its advantages in modeling dispersion. We further discuss possible extensions of the methodology to other evaluation settings, including longitudinal evaluation settings.

Introduction

Many questions and analyses in educational evaluation research focus on average treatment effects or expected differences in outcomes of interest given the assumption that possible confounding variables are controlled for: What is the expected difference in posttest scores between students assigned to an innovative program versus those assigned to a more traditional program? What is the expected difference in mathematics achievement between students attending public high schools versus those attending Catholic schools?

While estimating expected differences in outcomes is of primary interest in many studies, often overlooked in such analyses are potential differences in individual responses in outcomes to a given treatment. In settings where the true treatment effect is constant across individuals (i.e., the assumption of `constant effect' [Holland, 1986]), the average treatment effect is a reasonable estimate for every individual in the population. In contrast, when the true treatment effect varies substantially across individuals, the average treatment effect either represents only a partial picture of the true treatment effect, or can be misleading. In such situations, the entire distribution of the outcome around the means of treatment and control conditions will reflect the true variable effect across individuals.

Individual differences in response to a given treatment have been a long-standing interest in education. As noted in classic pieces in the educational research literature (Cronbach & Snow, 1977; Cronbach, 1975), when one considers `treatments,' such as a remedial reading interventions, opportunities to learn fractions, certain school policies, or particular types of parental involvement, it is very likely that the true effect of such educational experiences will vary substantially across students. What may be occurring is two-way or even three-way interactions among personal characteristics and treatment (e.g., Aptitudes and Treatment interaction [ATI]), with the personal characteristics being either identified or unidentified.

Attending to differences in dispersion in outcomes can be a useful way of bringing to light individual differences in responses to a given treatment. Differences between treatment and control conditions with respect to dispersion in outcomes can, under certain circumstances, signal that the treatment effect is not constant but varying across individuals. In particular, when the treatment effect varies systematically as a function of individual characteristics, unequal dispersion between conditions may indicate interactions between treatment and individual characteristics. Thus, addressing heterogeneity of dispersion and studying factors underlying the heterogeneity can help us identify important factors that magnify or dampen the effect of a treatment that might otherwise go unnoticed.

Toward this end, this paper examines the effects of treatments on dispersion in outcomes (i.e., expected differences in dispersion between individuals under treatment and control conditions) as well as the effects of treatments on mean in outcomes (i.e., expected differences in means between individuals under treatment and control conditions). Although the literature on causal inference has been proliferating in the last couple of decades, it has been concerned with assessing ``causal effects'' on means in outcomes but not on dispersion. The perspective of causal inference would be beneficial to drawing inferences concerning the effects of treatments on dispersion as well as inferences on means. Just as there has been much focus on obtaining unbiased estimates of average treatment effects on means in outcomes by assuring that possible confounding variables are controlled for via randomization, analytic adjustment (e.g., adjusting for covariates), or adjustment built in the design stage (e.g., matching or stratification), in inferences concerning dispersion, it is also important to control for possible confounding variables.

This paper first suggests a conceptual framework to think profitably about the effect of a treatment on dispersion. The framework helps us see how heterogeneity in dispersion can signal the presence of unspecified interactions between treatment and subject characteristics, and sensitize us to factors that make it difficult to draw sound inferences concerning the effects of treatments on dispersion in quasi-experimental settings. As will be seen, as we move from experiments to quasi-experiments, and particularly from quasi-experiments in non-nested settings to quasi-experiments in nested settings, it becomes more challenging to draw sound inferences concerning dispersion. Much of the focus in this paper is on quasi-experiments in nested or multilevel settings, which is a common framework in multisite evaluation studies.

The conceptual framework is based on multiple sources of literature. In particular, we draw on the work by Bryk and Raudenbush (1988) that examines heterogeneity in dispersion in experimental studies, primarily in non-nested settings (e.g., one-way ANOVA, or regression). They show that unequal dispersion between treatment and control conditions in randomized studies may be empirical evidence of interactions between treatment and subject characteristics. We also draw on the work by Raudenbush and Bryk (1987) that investigates how differences in cluster characteristics relate to differences in within-cluster dispersion. The main interest in Raudenbush et al. (1987) is in identifying cluster characteristics (e.g., within-school standard deviation in number of math courses taken) that predict within-cluster variability (e.g., within-school variability in math performance). They show that in randomized studies unequal dispersion between conditions arise entirely from the treatment, while in non-randomized studies unequal dispersion may stem from differences in dispersion prior to the treatment (e.g., pre-existing differences in dispersion in student SES).

Our conceptual framework combines and extends the work by Bryk et al. (1988) and Raudenbush et al. (1987) to incorporate nested settings, thereby addressing confounding by cluster characteristics in addition to confounding by subject characteristics. Furthermore, this paper constructively uses concepts from theories of potential outcomes (Holland, 1986; Rubin, 1974, 1978) and lays out sources of differences in dispersion respectively in experiments, quasi-experiments in non-nested settings, and quasi-experiments in nested settings. By enumerating the sources of differences in dispersion, we show that characteristics in different levels give rise to confounding in distinctive ways with regard to inferences concerning dispersion in the Hierarchical Modeling (HM) framework.

Secondly, this paper suggests modeling strategies and procedures to address the effects of treatment on dispersion as well as on mean, and to identify important interactions between treatment and individual characteristics, for commonly encountered designs in multi-site evaluation studies. This entails modeling dispersion structures as well as mean structures in hierarchical models (HMs). By mean structure, we mean hypothesized relationships between an outcome of interest and covariates (e.g., regression models), while by examining dispersion, we mean modeling the residual variance from the mean structure as a function of covariates (cf. McCullagh & Nelder, 1989). While HMs are widely used in educational research (for example, Goldstein, 1995; Raudenbush & Bryk, 2002), typically the residual variances at each level are assumed to be homogeneous, or independently and identically distributed (i.i.d.). Even in some applications of modeling residual variance (see, e.g., Browne, Draper, Goldstein, & Rasbash, 2000; Goldstein, Healy, & Rasbash, 1994; Kasim & Raudenbush, 1998; Littell, Milliken, Stroup, Wolfinger, 1996; Raudenbush & Bryk, 2002; Snijders & Bosker, 1999, chapter 8; Wolfinger, 1996), the applications are in general limited to specifying heterogeneous variances and pursuing better-fitting models.

Our approach makes connections between such HMs used in experimental or quasi-experimental studies and causal inferences concerning dispersion, and present modeling strategies and procedures that are useful especially in quasi-experiments in nested settings. A series of HMs will help us see how a well-elaborated dispersion structure based on substantive theories mitigate the problem of confounding by cluster characteristics, while a well-elaborated mean structure helps avoid confounding by individual characteristics, with regard to inferences concerning dispersion. In cases where significant effect of treatment on within-site dispersion is present after careful considerations of the confounding variables, both in the individual and cluster levels, we suggest to proceed to further elaboration of the mean structure by including possible interactions between treatment and individual characteristics.

We illustrate this framework and the modeling strategies and procedures with analyses of data from a study that assessed the effectiveness of two innovative instructional programs relative to traditional instruction in elementary mathematics classrooms in the domain of fractions (Gearhart, Saxe, Seltzer, Schlackman, & Ching, 1999). As will be seen, by modeling both the mean and dispersion structures in HMs, we found that classes of students in the two treatment conditions not only performed better in terms of problem-solving posttest scores in the domain of fractions, but also had larger within-classroom dispersion, relative to the classes of students in the control condition, after we controlled for possible confounding by a student characteristic and by a class characteristics, respectively in the mean and the dispersion structures. Larger dispersion in the treatment conditions helped us uncover interactions between treatment and student characteristics that might have otherwise gone unnoticed. We use the fully Bayesian (FB) approach implemented by WinBUGS (Spiegelhalter, Thomas, Best, & Lunn, 2003). The advantages using FB in inferences concerning dispersion will be discussed.

Conceptual Framework

Sources of Differences in Dispersion in Experiments

Let us start with a simple situation in which there is a treatment group and a control group and in which the variances around the mean levels are different for the treatment and control groups. Then one may ask why this difference arises. One quick answer might be that the outcome of interest is more similar or diverse across subjects in the presence of treatment than in the absence of treatment. Theories of potential outcomes (Holland, 1986; Rubin, 1978; Rubin, 1974) may help formalize this idea.

The key idea behind theories of potential outcomes is that potentially the outcome of interest y for an individual i, y_i, could have been observed either in the treatment T or in the control condition C. Let y_i^T denote the outcome for individual i that could potentially be observed in the treatment condition T; and y_i^C denote the outcome that could potentially be observed in the control condition C. In this framework, the effect of treatment T relative to control C on the outcome of interest y is defined as the difference between two potential outcomes, y_i^T - y_i^C, for subject i.

A fundamental difficulty arises in estimating this effect of treatment for individual i, because it is typically not feasible for a subject to receive both treatment and control conditions independently (i.e., it is typically impossible to observe both y_i^T and y_i^C). A `statistical solution' (Holland, 1986) to this problem is to obtain the expected treatment effect over subjects in the population, $E(y_i^T - y_i^C) = E(y_i^T) - E(y_i^C)$. While the average causal effect is the difference between two expected values $E(y_i^T)$ and $E(y_i^C)$, what one can estimate in practice is the difference between two expected values over such subjects as who are actually observed in either condition. Thus, the estimated quantity is

$$E(y_i^T \mid S_i=T) - E(y_i^C \mid S_i=C), \qquad (1)$$

where the assignment of individual i is denoted as S_i.

In experimental settings, the assignment S_i is independent of all variables including y_i^T and y_i^C (i.e., the assumption of `independence' [Holland, 1986]). Thus the two observed expectations in Equation 1 are respectively equivalent to the two expected values over subjects in the population. That is, $E(y_i^T) = E(y_i^T \mid S_i=T)$ and $E(y_i^C) = E(y_i^C \mid S_i=C)$. As such, the estimated quantity in Equation 1 becomes $E(\mathrm{y}_i^T\text{-}\mathrm{y}_i^C)$, which shows that experiments provide unbiased estimates of the expected effectiveness of treatments.

As for dispersion modeling, a relevant point in this account is that the `true' treatment effect (relative to the control condition) is originally defined for an individual on which the treatment exerts an effect (i.e., y_i^T- y_i^C). Based on this formalization, heterogeneity of variance between treatment and control groups implies that $Var(y_i^T | S_i=T)$ - $Var(y_i^C | S_i=C)$ is significantly different from zero. By the independence assumption, the above quantity is equivalent to $Var(y_i^T)$- $Var(y_i^C)$. This means that the variance in the outcome in the treatment condition is significantly different from that variance in the control condition. This means that the variance in the outcome in the presence of treatment is significantly larger or smaller than that variance in the absence of treatment. It may indicate that the true treatment effect relative to control, y_i^T- y_i^C, is variable across individuals.

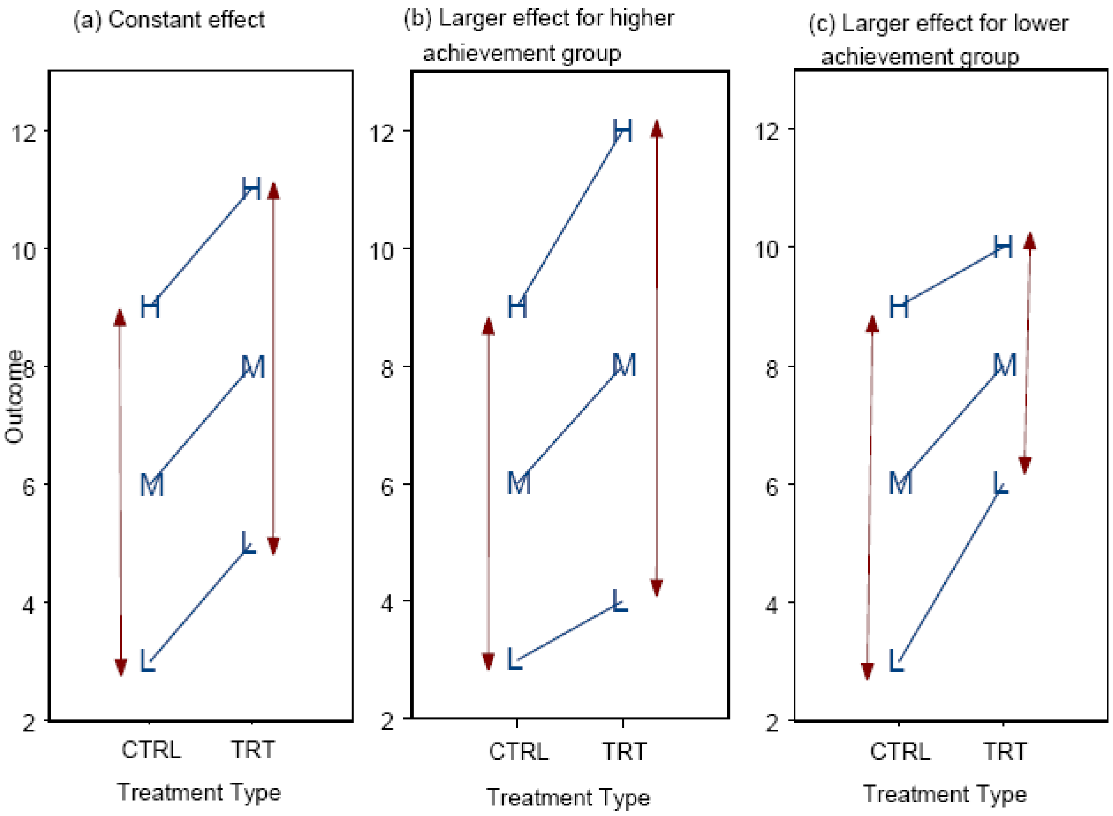

Figure 1. Summary plots of average performances of three subgroups. H, M, and L indicate high, medium, and low achievement groups, respectively. A vertical arrow represents dispersion in outcome for each treatment group.

Figure 1 displays three hypothetical settings in which the experiment is concerned with assessing the effectiveness of an innovative educational program (i.e., a treatment) on student achievement (i.e., an outcome of interest). For heuristic purposes, three subgroups are assumed in the population, high, medium, and low achievement groups, which are represented respectively by H, M, and L in the figure. Since the students are randomly assigned to either condition, we can assume that $E(y_i^T | S_i=T) = E(y_i^C | S_i=C)$ and $E(y_i^T | S_i=T) = E(y_i^C | S_i=C)$ for the sample, and also for the subgroups given a large enough sample size within the subgroups. Then, the lines in the figure represent the effectiveness of treatment, i.e., the increment in outcome from absence of treatment to presence of treatment.

In scenario (a) the treatment effect (relative to control) is constant across subjects and across subgroups, which is indicated by the parallel lines. The parallel lines convey that there is a constant increment from control to treatment. In this setting, the dispersions, represented by the vertical lines, of the two groups are equivalent.

The two other scenarios are examples of varying treatment effect across individuals. In these examples, one can see that heterogeneity in variance arises between the two groups. In scenario (b), the treatment effect is larger for the higher achievement subgroup, which is indicated by the steeper line for the higher achievement group. As a result, one can see that the dispersion of the treatment group is larger than that of the control group. In contrast, in scenario (c), the treatment effect is larger for the lower achievement subgroup, which is represented by a steeper slope for the lower achievement group. As can be seen, the dispersion of the treatment group is smaller.

Bryk and Raudenbush, in an independent line of work, provide further clarification of this issue. They conceive heterogeneity of variance in experimental studies as ``empirical evidence of an interaction of treatments with some unspecified subject characteristics'' (Bryk & Raudenbush, 1988, p.396). That is, the treatment effect varies, or the treatment exerts differential effects across subjects, because subjects with different characteristics respond to the fixed treatment differentially.

Sources of Differences in Dispersion in Quasi-Experiments in Non-Nested Settings

Experimental studies are not feasible in numerous situations in educational research or, more broadly, in social science research. When random assignment has not been achieved, groups might differ from each other in many ways prior to the exposure to treatments. Thus, in such settings as quasi-experimental studies, a number of confounding sources may exist in inferences concerning dispersion as well as in inferences concerning mean. Specifically, if the mean model is not fully specified, omitting some important variables that account for initial differences in dispersion between groups prior to the exposure to treatment, then the inference concerning dispersion may be confounded by the variables that are omitted from the mean model.

For illustrative purposes, let us assume that there are two covariates X_1 and X_2 that are related to the outcome *and* distributed with unequal variance across groups, prior to treatment (Raudenbush & Bryk, 1987). Then a proper mean model would be:

$$y_i = \beta_0 + \beta_1 Trt_i + \beta_2 X_{1i} + \beta_3 X_{2i} + e_i \,. \tag{2}$$

Assume now that the mean model is not properly specified omitting X_{2i}. The residual term in Equation 2 then becomes $\beta_3 X_{2i} + e_i$, which includes the covariate that has been omitted from the mean model. Therefore, the residual variance becomes a quadratic function of the omitted variable, X_{2i}. In this case, the dispersion inference is subject to a confounding problem, because the heterogeneity in the residual dispersion might be either due to differential treatment effects across subjects or due to different initial dispersions in X_{2i} between the groups, or due to a mix of both.

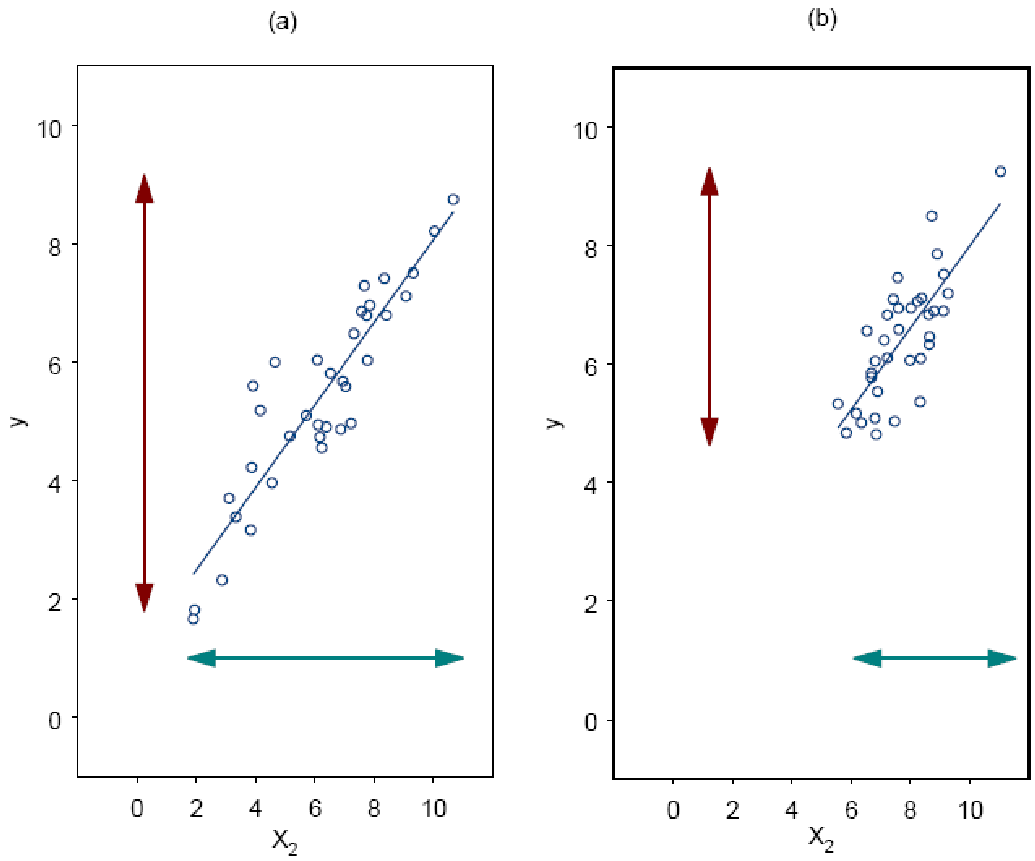

Figure 2. Scatter plots of outcome (y) and predictor (X2). Vertical Arrows represent outcome dispersion, while horizontal arrows represent predictor dispersion. OLS lines are superimposed.

Figure 2 displays how differences in dispersion of an outcome result from differences in dispersion of a predictor. In Group A shown in panel (a), the dispersion of a predictor X_2 is larger than that of Group B in panel (b). As can be seen by the vertical lines indicating the dispersions of the outcome, the larger dispersion in the predictor X_2 in Group A directly leads to the larger dispersion in the outcome in Group A. Therefore if this predictor is not included in the mean model, the residual dispersion will be different between Groups A and B due to the difference in the predictor dispersion.

Sources of Differences in Dispersion in Quasi-Experiments in Nested Settings

So far the framework has dealt with non-nested settings in which there is no nesting structure in the data such as multiple regressions. Studying differences in dispersion in quasi-experimentation can be more complicated in nested settings (e.g., settings where the individuals are nested within clusters). In addition to the sources of confounding in non-nested settings, differences in observation-level dispersion (i.e., within-cluster dispersion) may also arise from differences in cluster-level characteristics.

For example, let us think of a multilevel situation in which student i is nested within school j and in which different schools are assigned to different treatments. The following equation extends Equation 2 to this multilevel situation:

$$y_{ij} = \beta_0 + \beta_1 Trt_i + \beta_2 X_{1i} + \beta_3 X_{2i} + \beta_4 W_{1j} + u_j + e_{ij}\,, \qquad (3)$$

where the two subscripts i and j indicate the data nesting; u_j is a cluster-level residual; and W_{1j} denotes a cluster-level variable. A primary feature of multilevel models such as Equation 3 is to partition the outcome variability into different levels: student-level variability and cluster-level variability. As such, the variance of the student-level residuals e_{ij} represents the pooled within-school residual dispersion, while the variance of the cluster-level residuals u_j represents the between-school residual dispersion.

Note that, unlike observation-level characteristics (i.e., student characteristics), cluster-level characteristics (i.e., school characteristics) can be related to the within-school residual dispersion, regardless of whether they are omitted from the mean model or not. As one can see from Equation 3, omitting W_{1j} in the model would change the cluster-level residuals to $\beta_4 W_{1j} + u_j$, but would not directly change the within-school or student-level residuals. Thus, in nested settings, differences in within-school dispersion, which is captured by the variance of Level-1 residuals e_{ij}, might also arise from differences in any cluster-level characteristic W_j that are either included or not included in the mean structure.

Among cluster-level characteristics, intake characteristics of clusters can be related to within-cluster dispersion. For example, larger enrollment of schools may contribute to larger within-school dispersion in student achievement. Policies or practices of clusters may also be related. School policies such as tracking may lead to larger dispersion in student achievement, while policies that employ more unified curricula units for all students may lead to smaller dispersion.

Even some aggregates of observation-level characteristics may be related to within-cluster dispersion after the characteristics are specified in the mean model. For instance, the standard deviation of student academic backgrounds may be related to dispersion in student achievement, after including the variable in the mean structure.

In schools dealing with very diverse populations of students in terms of their academic backgrounds compared to schools with homogeneous populations in which virtually all students have very weak academic backgrounds, even students with similar academic backgrounds may be expected to perform in a different way in the outcome (e.g., posttest scores). A student in a very diverse population may interact with peers with diverse backgrounds and receive instruction that is geared toward students with a broad range of prior educational experiences. The expected outcome of this student would have more uncertainty since plausible outcomes could range from very low to very high. In contrast, a typical student in a homogeneous population composed of students with weak backgrounds is likely to do poorly on the test, as are his or her peers. The expected outcome thus has less uncertainty since plausible outcomes are apt to fall in a restricted range in such situations. Thus, diverse initial academic backgrounds may contribute to larger outcome dispersion, while homogeneous initial academic backgrounds may yield smaller outcome dispersion. This can be viewed as a contextual effect on dispersion.

Contextual or compositional effects are viewed as existing in settings where, even after controlling for a student characteristic at the student level, the aggregate of the student characteristics is related to the outcome. Contextual effects have normally been discussed in connection with means (e.g., relationships between school mean SES and school mean achievement), but not with dispersion. However, the same logic applies to variances. Even when student characteristics are controlled, the entire environment of schools in terms of the characteristic can have an impact on the dispersion in the outcome of interest as well as the central tendency.

Efforts to attend to relationships between within-cluster dispersion and various cluster characteristics can be found in Raudenbush and Bryk (1987). In analyzing the *High School and Beyond* data, Raudenbush et al. (1987) found that, within-school residual variability after accounting for various student characteristics was significantly related to various school characteristics, such as the within-school standard deviation of student SES, the within-school standard deviation of student academic backgrounds, whether the school is a Catholic or public school, school size, and the within-school standard deviation in the number of math courses taken.

Raudenbush et al. (1987) refer to these characteristics as “correlates of diversity,” in a sense that they help predict the outcome dispersion.

To summarize, in quasi-experimental studies that employ single-level regression models (e.g., Equation 2), a main possible source of confounding in estimating treatment effects on dispersion involves omission of variables in the mean structure, which is considered as a model misspecification problem. In quasi-experimental studies that use multilevel models (e.g., Equation 3), however, differences in cluster-level characteristics, which may be an additional source of confounding, is not associated with the problem of the misspecification of mean structure. Cluster-level characteristics can be related to differences in observation-level residual variance irrespective of the mean structure.

Summary on Confounding Variables in Inferences Concerning the Effects of Treatments on Dispersion

The conceptual framework lays out sources of differences in dispersion respectively in experiments, quasi-experiments in non-nested settings, and quasi-experiments in nested settings. As we move from experiments to quasi-experiments in non-nested settings, and from quasi-experiments in non-nested settings to quasi-experiments in nested settings, more sources are involved in differences in dispersion. Therefore, it is more challenging to make sound inferences concerning the effect of treatments.

In experimental settings where randomization assures that treatment assignment will be independent of all variables, the difference in dispersion in the outcome between treatment groups arises entirely from the effect of treatment on dispersion. In quasi-experimental settings, however, the effect of treatment on dispersion may be confounded by misspecification of the mean model, such as omitting important variables from the model (e.g., student characteristics that are related to outcomes and that have unequal dispersions across treatment groups). Thus, differences in dispersion between treatment groups may arise either from a treatment effect on dispersion or omitted variables from the model, or even a mix of both.

In addition to this, multilevel quasi-experimental studies confront another source of confounding. Differences in within-cluster or level-1 dispersion may also be due to differences in cluster characteristics. For example, in multisite evaluation studies, certain school policies (e.g., tracking) may be related to within-school dispersion. Thus, multilevel quasi-experimental studies, which are common in educational evaluation research, face the challenge of disentangling three sources of differences in variance: treatment effect on dispersion; model misspecification at the student level (e.g., omission of student characteristics in the mean structure); and differences in cluster characteristics that are related to within-cluster dispersion (e.g., school characteristics that are related to dispersion).

Hierarchical Models (HMs) for Modeling Mean and Dispersion Structures

In multisite evaluation studies, it is often the case that intact schools or classrooms are the unit to which different treatments are assigned. As a result, individuals are nested within different clusters (e.g., sites, schools, or classrooms). This characteristic of the data calls for Hierarchical Models (HMs) to account for the dependency among the observations within clusters. This paper focuses on HMs for continuous outcomes, with possible applications to multisite evaluation studies. HMs for continuous outcomes commonly assume normal assumptions in the observation level (Level 1) and normal assumptions for random effects that are associated with each nesting unit (Level 2). We first present a 2-level HM with homogeneous variance assumption. Next, we present how to extend these HMs to model dispersion as well as means simultaneously.

In design settings where intact clusters are nested within treatment types, treatment status becomes a Level-2 variable. Specifically, at level one, the outcome y_{ij} for student i in cluster j is a function of student characteristics, X_{qij}, $q = 1, \ldots, Q$.

$$y_{ij} = \beta_{0j} + \Sigma\, \beta_{qj} X_{qij} + r_{ij}, \qquad r_{ij} \sim N(0, \sigma^2), \qquad (4a)$$

where a key coefficient is the intercept, β_{0j}, which is the adjusted mean of the outcome for cluster j, given that the student characteristics are centered around their grand-means.

For the purpose of simplicity, let us assume that the coefficients of student characteristics do not vary across clusters. Then at level two the intercept can be modeled as a function of treatment status and other cluster-level characteristics, W_{sj}, where $s=2, \ldots, S$, such as

$$\begin{aligned} &\beta_{0j} = \gamma_{00} + \gamma_{01} Trt_j + \Sigma\, \gamma_{0s} W_{sj} + u_{0j}, \quad u_{0j} \sim N(0, \tau_{00}), \\ &\beta_{qj} = \gamma_{q0,} \qquad q = 1, \ldots, Q, \qquad (4b) \end{aligned}$$

where γ_{01} captures the expected relative effectiveness of the treatment; and τ_{00} is the residual variability of intercepts across sites after accounting for treatment status and other site characteristics.

Although the above HM can provide much useful information concerning the average effect of a treatment under certain conditions, relaxing the homogeneity assumption of the level-1 residual variance and modeling it as a function of key covariates opens up the possibility of examining another dimension of the outcome, which is dispersion in the outcome. We will refer to this part of the analysis as dispersion modeling, in contrast to mean structure modeling.

What follows in this section deals with dispersion modeling, in which we frame the procedure to consist of two stages: detecting and probing stages. Examining dispersion starts with checking to see if there is detectable heterogeneity of variance in the data: specifically, whether the residual variances are unequal across treatment type or across clusters. We refer to this as a "detecting" stage. Once heterogeneity of variance is detected, the next step would be to probe the current model and the data further in pursuit of investigating possible factors that underlie the heterogeneity. We refer to this as a "probing" stage.

Detecting Heterogeneity

Based on the discussion in the conceptual framework, in multisite evaluation studies, unless it is a perfectly randomized study with a large enough cluster sample size, one is confronted with at least three possible sources of the heterogeneity in level-1 residual variance. Since the focus in experimental or quasi-experimental studies is on estimating the effect of treatment on dispersion, which is one source of the heterogeneity, the other two sources become confounding variables. One confounding source is from differences in site characteristics, which were referred to as correlates of diversity following Raudenbush et al. (1987). The other confounding source is from model specification errors at Level 1, e.g., omitting student characteristics that are related to the outcome and have unequal dispersions across treatment type (see, e.g., Figure 2).

Relationships of key interest, i.e., the effect of a treatment on dispersion, can be examined by Level-1 dispersion modeling, i.e., modeling the Level-1 residual variance as a function of site characteristics as well as treatment indicator variables. With the mean model being specified as above in Equations 4a and 4b, the dispersion model can be specified in a log-linear model as follows:

$$\ln(\sigma^2_{ij}) = \alpha_0 + \alpha_1\, Trt_j + \Sigma\, \alpha_p W_{pj},\ p=2, \ldots, P\,. \qquad (5)$$

The coefficient α_1 captures the key parameter, i.e., whether the Level-1 residual variance depends on treatment membership. The variable W_{pj} are site characteristics such as site sample size, school policies or practices of sites, site dispersions of student characteristics (e.g., within-site standard deviations of student academic background). The coefficients of the site characteristics, α_p, represent the relationships of site characteristics to within-site dispersion in the outcome.

To obtain unbiased estimate of the key parameter indicating whether the Level-1 residual variance depends on treatment membership, it is important to adjust for the site-level confounding variables (W_j) in the dispersion structure. The adjustment is required especially in quasi-experimental settings, because the treatment indicator (Trt_j) is likely to be correlated with some site characteristics (W_j). Even in experimental studies, it is warranted to specify the site characteristics in the dispersion model to decrease within-treatment variability in the dispersion structure and thereby increasing the statistical power to detect differences associated with treatment membership, and to adjust for possible existing imbalance between treatment types.

Unlike site-level confounding variables (W_j), possible biases due to the student-level confounding variables (X_{ij}) should rather be removed by modifying the mean structure than modeling the dispersion structure. As detailed in the conceptual framework, inferences concerning the effect of treatment on dispersion can be misleading due to student-level confounding variables, if they are omitted from the mean structure. Thus, student characteristics that have unequal dispersions between treatment groups should be included in the mean structure to avoid associated biases.

Differences in level-1 dispersion may come from other kinds of model misspecification errors than omitting student-level confounding variables. For example, Raudenbush and Bryk (2002, chapters 7 & 9) suggest a statistical test, to see whether the Level-1 residual variances of each Level-2 unit are variable across Level-2 units. When heterogeneity in residual variance exists across clusters, it may indicate model specification errors at level one. According to Raudenbush et al. (2002), some possibilities are; 1) a slope (e.g., β_{1j}, the effect of treatment at site j) that varies appreciably across sites is erroneously specified as fixed; 2) there may be outliers in some sites; or 3) the distribution of outcome has heavier tails than normal.

Statistical tests of this kind and subsequent modifications or careful selections of the mean structure should be preceded to ensure the validity of the mean structure and to avoid confounding that arises from the mean model misspecification in inferences concerning the effect of a treatment on dispersion.

Probing Heterogeneity of Dispersion

Once heterogeneity of dispersion is detected between treatment and control groups despite efforts to take into account possible confounding sources, one may proceed to search for possible interactions empirically and substantively. The search should be guided by relevant theory and empirical work in the literature. One may also consult with developers of the treatment to get a sense of whether they think the treatment effect will be larger for students with certain characteristics. Empirical evidence is also important. It would be helpful to search carefully for measured student characteristics, to see whether any of them interacts with the treatment.

In design settings where intact clusters are nested within treatment types, the interaction of treatment with student characteristics is a cross-level interaction, that is, an interaction between a Level-2 variable (i.e., Trt_j) and a Level-1 variable (i.e., X_{ij}). For illustration, suppose that we have one covariate (X_{1ij}) in our Level-1 model, and that the slope for X_{1ij} does not significantly vary across sites. Suppose further that we fit the following model to the data in a detecting stage:

$$
\begin{aligned}
&y_{ij} = \beta_{0j} + \beta_{1j}X_{1ij} + r_{ij}, && r_{ij} \sim N(0, \sigma^2{}_j),\\
&\beta_{0j} = \gamma_{00} + \gamma_{01}Trt_j + \Sigma\, \gamma_{0s}W_{sj} + u_{0j}, && u_{0j} \sim N(0, \tau_{00}),\\
&\beta_{1j} = \gamma_{10},\\
&\ln(\sigma^2{}_j) = \alpha_0 + \alpha_1 Trt_j + \alpha_2 W_j, && Trt=0,\ 1, \qquad (6)
\end{aligned}
$$

where one finds that the estimate for α_1 is statistically significant, indicating a possible interaction between a Level-1 characteristic and treatment. To investigate whether there is an interaction between X_{1ij} and Trt_j we can expand the mean model as follows as part of a probing stage:

$$
\begin{aligned}
&y_{ij} = \beta_{0j} + \beta_{1j}X_{1ij} + r_{ij}, && r_{ij} \sim N(0, \sigma^2{}_j),\\
&\beta_{0j} = \gamma_{00} + \gamma_{01}Trt_j + \Sigma\, \gamma_{0s}W_{sj} + u_{0j}, && u_{0j} \sim N(0, \tau_{00}),
\end{aligned}
$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11}Trt_j \,. \qquad (7)$$

Note the non-randomly varying slope of the student characteristic. Rewriting this model as a mixed model helps us see the interaction:

$$y_{ij} = \gamma_{00} + \gamma_{01}Trt_j + \gamma_{10}\, X_{1ij} + \gamma_{11}X_{1ij}Trt_j + \Sigma\, \gamma_{0s}W_{sj} + u_{0j} + r_{ij} \,.$$

Iterations between the detecting and probing stages based on our conceptual framework may result in more elaboration both in the mean and dispersion structures. For example, heterogeneity in Level-1 residual variance may turn out to be due to omitting an important student characteristic from the mean model. This will lead to specifying the variable in the model, which is an elaboration in the mean structure. Likewise, heterogeneity in variance across treatment type may signal interactions between treatment and subject characteristics; and then the mean structure will be more elaborated by adding the interaction term to the model.

One may think of a situation where heterogeneity in Level-1 residual variance is related to a site characteristic (e.g., tracking). Then one includes the site characteristic in the dispersion model, which implies an elaboration of the dispersion structure. As mentioned above, in settings where the site characteristics is correlated with treatment type, specification of these site variables is required to get an unbiased estimate of a treatment effect on dispersion.

In principle, full elaboration of both the mean and dispersion structures should explain away the heterogeneity of level-1 residual variance between treatment types. By full elaboration, we mean identifying all main sources of heterogeneity in dispersion and specifying them in the mean and/or dispersion models. In multisite evaluation studies, we fully elaborate a hierarchical model, when we include in the mean structure all student characteristics that are related to the outcome dispersion and to treatment type; in the mean structure all significant interactions between treatment and student characteristics; *and* in the dispersion structure all site characteristics that are related to within-site dispersion and to treatment type.

Although, after the iterative procedures, it would be ideal to explain away the heterogeneity in residual dispersion between treatment types, one may not have enough information to do so in many studies. For instance, information about all important student and site characteristics may not have been collected or observed.

Illustrative Example

Study Background

The National Council of Teachers of Mathematics (NCTM) *Standards* calls for a conceptual, problem-solving approach to mathematics instruction. The implementation of the *Standards* is contingent on teachers' deep understandings of ``both mathematics and the ways that students interpret mathematical problems and build knowledge'' (Gearhart, Saxe, Seltzer, Schlackman, & Ching, 1999, p.287). In relation to this set of *Standards*, the state of California adopted and promoted curriculum units designed to supplement or replace chapters from traditional texts. The study focused on two such units, Seeing Fractions and My Travels with Gulliver. These units are designed to be aligned with the *Standards*: to support students' involvement with mathematical problem-solving and enhance their conceptual understanding. Gearhart et al. (1999) term this reform-minded instruction.

A pool of teachers was selected from volunteers from upper elementary schools. Teachers who had used and would be continuing with a traditional curriculum were assigned to the Traditional (TRAD) group. Teachers who had experience with the two state-adopted units were randomly assigned to one of two forms of professional development termed Integrating Mathematics Assessment (IMA) and Collegial Support (CS).

In the IMA program, teachers went through an intensive professional development program, a 5-day summer institute followed by 13 meetings during the year. The integrated series of workshops the IMA teachers attended dealt with the following activities: Teachers' Mathematics, Children's Mathematics, Children's Motivation, and Implementation of Integrated Mathematics Assessment. For more details on each activity, please refer to Gearhart et al. (1999). In contrast, in the CS program, the teachers did not receive any intensive workshops but met with other teachers in the program to discuss various issues concerning the implementation of the two curriculum units.

Researchers in the study believe that while reform-minded curricula have the potential to enhance student problem-solving skills and conceptual understanding of mathematics, teachers need extensive professional development in order to implement such curricula successfully. Thus the IMA and CS conditions provide an opportunity to study the effect of reform-minded curricula when teachers receive intensive training (i.e., IMA) and when they receive training that is much more typical (i.e., CS).

The assignment to either IMA or CS was done by a stratified random assignment procedure; a simple random assignment procedure was inappropriate due to the small sample size. Teachers were matched on the following characteristics: years of experience, experience with the problem-solving units, additional professional development, and student characteristics. As the assignment of teachers was conducted before the school year started, the students on which matching was based were not the ones in the sample, but the ones the teachers taught in the year prior to the study. This procedure resulted in 9, 7, and 5 teachers in the three groups, IMA, CS, and TRAD, respectively. For more details, refer to Gearhart et al. (1999).

A key feature of this study is that extensive observations were conducted in each classroom. One main aim of the study was to develop measures of opportunities in student learning that are aligned with reform-minded principles of instruction. In a slightly distinctive perspective from the original study, we view this study as a quasi-experimental intervention study in which two treatment groups (i.e., IMA and CS) and one control group (i.e., TRAD) are compared. We attempt to assess the effect of treatments on the dispersion as well as the mean level of student achievement in the domain of fractions. Specific research questions follow.

First, what is the expected difference in the outcome of interest (i.e., problem-solving posttest scores in the domain of fractions) between students assigned to the IMA program versus students assigned to traditional instruction? What is the expected difference in the outcome between students assigned to the CS program versus students assigned to traditional instruction?

The second question is: What is the expected difference in within-classroom dispersion in the outcome between IMA classrooms and traditional classrooms; and between CS classrooms and traditional classrooms? Is there significant heterogeneity in within-classroom dispersion across classrooms?

Third, in cases where we detect significant heterogeneity in within-classroom dispersion, what are the underlying sources of the heterogeneity? What would be the evidence of the presence of interactions between the IMA program and student characteristics, and between the CS program and student characteristics? In addition, what are the classroom characteristics that help predict within-classroom dispersion?

Variables and Exploratory Analyses

In terms of prior student characteristics that are expected to relate to the outcome (i.e., problem solving posttest in the domain of fractions), four student characteristics were the focal interest of the researchers studying reform-minded practices in the domain of fractions. The four variables include three pretest measures and a language status measure:

a) a computation pretest score (*Prep*), which is the sum of the computation items that a student got right; b) a problem-solving pretest score (*Prec*), which is the sum of the problem-solving items that a student got right; c) a binary indicator of incipient understanding of fractions (*Incip*), which indicates whether a student shows a rudimentary understanding of fractions; and d) a binary indicator of a student's language fluency (*Lang*), which measures whether a student is fluent in English. The information is provided from school data of the prior year. Table 1 presents the means and the standard deviations of all four student characteristics by treatment type. Although, as mentioned above, the stratified randomization in the design stage used student characteristics from the students who the teachers in the sample taught in the previous year, the results from the distributions of various student characteristics show substantial overlaps across groups prior to the implementation.

Table 1

Descriptives of Student Characteristics for all Sample and by Treatment Type.

Student Characteristics	Sample	IMA	CS	TRAD
Problem-solving pretest				
mean	2.67	3.46	1.97	2.12
sd	2.23	2.56	1.53	1.88
Computation pretest				
mean	2.95	4.00	1.81	2.55
sd	3.67	3.92	2.89	3.64
Language status				
mean	0.82	0.94	0.80	0.64
sd	0.38	0.24	0.40	0.48
Incipient understanding status				
mean	0.65	0.76	0.54	0.59
sd	0.48	0.43	0.50	0.49

As for the average level, in all four characteristics, the CS group and the control group are rather similar, while the IMA group appears to be more advantaged. The IMA group students performed better on the Problem-solving pretest and Computation pretest. Also a higher proportion of them was English proficient (94%) and had an incipient understanding of fraction (76%). According to ANOVA tests and Duncan post hoc comparisons, the average problem-solving pretest (*Prec*) in the IMA group is significantly higher than that of the other groups (Anova p-value=0.03). The other variables did not show differences in the average level across the treatment groups.

As for the dispersion, the Levene's test for homogeneity of the variance in English language proficiency status (*Lang*) is also rejected with borderline significance (p-value=0.05). The other variables did not show differences in dispersion across the treatment groups.

Since the problem-solving pretest (*Prec*) and English language proficiency status (*Lang*) show differences across groups, respectively in terms of the average level and the average dispersion, the adjustment of the differences in *Prec* is required to avoid confounding in inferences concerning the effects of treatments on the mean level; and the adjustment of the differences in *Lang* is required to avoid confounding in inferences concerning the effects on the dispersion.

Analyses and Results

The analysis consists of a series of three HMs. Model 1 is a base HM, in which we focus on modeling the mean structure, while assuming homogeneous variance at level 1 and level 2. This model provides estimates of differences in the mean level of the outcome between treatment and control groups. Model 2 examines differences in the dispersion of the outcome between treatment and control groups, and in relation to site characteristics. This is done by embedding a log-linear model for dispersion in the base HM (i.e., Model 1). Model 3 extends the mean structure of Model 2 by including interactions between treatment and student characteristics. Model 2 generally corresponds to the "detecting" stage in the dispersion modeling, and Model 3 to the "probing" stage, although the two stages are iterative and cannot be entirely separated.

All model estimation is done through fully Bayesian computation using the software WinBUGS (Spiegelhalter, Thomas, Best, & Lunn, 2003). A practical advantage of using fully Bayesian approach in this illustration is to allow us to compute of the posterior distributions of unknown quantities of more direct interest as well as the parameters in specified models, and thereby facilitating inferences based on Bayesian probabilistic statement. For example, from the log-linear models for the dispersion structure, we construct posterior distributions of residual variances of each treatment group. Also, for the interaction between student prior computational knowledge on fractions (*Prec*) and the CS treatment, we directly construct the posterior distribution of the slope of prior computational knowledge (*Prec*) of students in CS classes, and compute the Bayesian probability that the slope is greater than 0. Another advantage of FB approach is to provide a systematic way of checking and retrospectively improving the specification of the models, which is referred to as posterior model checks (e.g., Belin & Rubin, 1995).

Model 1: Base HM . The base model attempts to estimate the effects of treatments on the level in the outcome. Similarly to an exemplary HM in the previous section in Equations 4a and 4b, the base HM is specified as follows:

$$\begin{aligned}
y_{ij} &= \beta_{0j} + \beta_{1j}(\mathrm{Prec}_{ij} - \overline{\mathrm{Prec}_{..}}) + \beta_{2j}(\mathrm{Prep}_{ij} - \overline{\mathrm{Prep}_{..}}) + \beta_{3j}(\mathrm{Lang}_{ij} - \overline{\mathrm{Lang}_{..}}) + \beta_{4j}(\mathrm{Incip}_{ij} - \overline{\mathrm{Incip}_{..}}) + r_{ij}, \\
& \qquad\qquad\qquad\qquad\qquad\qquad r_{ij} \sim N(0, \sigma^2), \\
\beta_{0j} &= \gamma_{00} + \gamma_{01}\mathrm{IMA}_j + \gamma_{02}\mathrm{CS}_j + u_{0j}, \qquad u_{0j} \sim N(0, \tau_{00}), \\
\beta_{qj} &= \gamma_{q0}, \qquad\qquad\qquad\qquad\qquad q=1, \ldots , 4.
\end{aligned} \tag{8}$$

The outcome, y_{ij}, is the Problem-solving posttest score for student *i* in classroom *j*. The within-classroom relationships between student characteristics and the outcome, β_{1j}, β_{2j}, β_{3j}, and β_{4j}, are specified as non-varying across classrooms -- that is, there are no random components attached to the slopes, since there was no empirical evidence that the within-class slopes are varying across classrooms.

Given that all four predictors are grand-mean centered at Level 1, the intercept β_{0j} is the classroom mean for classroom *j* adjusted for the between-classroom differences in the Level-1 predictors included in the model. The same logic of ANCOVA applies to this. At Level 2, the adjusted classroom means are specified as a function of treatment indicators, of which the coefficients γ_{01} and γ_{02} are the key parameters. They are, respectively, the expected difference between the IMA and the control groups and the expected difference between the CS and the control groups.

Table 2 presents the results of the model. The expected difference between the IMA and the control classes is 2.2, while the difference between the CS and the control classes is 1.2. The CS-control contrast is barely significant - the lower end of 95% interval is only a little above zero.

The pooled within-class slopes are all positive, and also significant except for the ELP indicator, controlling for all other variables in the model. The slope of the Problem-solving pretest is 0.52, indicating that a 5-point difference in the pretest implies an expected difference of 2.6 points in the outcome, when all other variables in the model are held constant. The slope of the Computation pretest is 0.15, and so a 5-point difference in the pretest results in an expected difference of 0.75 points in the outcome, given that all other conditions are equal. Again controlling for all other predictors, the students with incipient understanding tend to perform better by 1 point.

Table 2

Results from Model 1.

	Estimate	95% Interval	Median
Fixed Effects:			
Model for adjusted class means			
CTL Grand Mean (γ_{00})	3.925	(2.985 4.872)	3.927
IMA/CTL contrast (γ_{01})	2.217	(1.026 3.402)	2.215
CS/CTL contrast (γ_{02})	1.218	(0.001 2.442)	1.216
Average Within-class slopes:			
Prec/Posttest slope (γ_{10})	0.516	(0.383 0.649)	0.516
Prep/Posttest slope (γ_{20})	0.152	(0.073 0.233)	0.152
ELP/Posttest slope (γ_{30})	0.077	(-0.645 0.792)	0.077
Incip/posttest slope (γ_{40})	0.987	(0.433 1.542)	0.987
Variance Components:			
Between class:			
Var. in adjusted mean (τ_{00})	0.812	(0.351 1.665)	0.744
Within class:			
Residual Var. (σ^2)	6.245	(5.485 7.118)	6.145

Model 2: HM with heterogeneity of residual dispersion. With the mean structure being identical to Model 1, Model 2 extends Model 1 by modeling the Level-1 dispersion structure. The natural logarithm of the Level-1 residual precision (i.e., the inverse of variance) is specified as a function of treatment indicators and the proportions of ELP students in classrooms, shown as follows:

$$
\begin{aligned}
y_{ij} &= \beta_{0j} + \beta_{1j}(\text{Prec}_{ij} - \overline{\text{Prec}_{..}}) + \beta_{2j}(\text{Prep}_{ij} - \overline{\text{Prep}_{..}}) + \beta_{3j}(\text{Lang}_{ij} - \overline{\text{Lang}_{..}}) + \beta_{4j}(\text{Incip}_{ij} - \overline{\text{Incip}_{..}}) + r_{ij}, \\
& \qquad\qquad\qquad\qquad r_{ij} \sim N(0, \sigma_j^2), \\
\beta_{0j} &= \gamma_{00} + \gamma_{01}\text{IMA}_j + \gamma_{02}\text{CS}_j + u_{0j}, \qquad u_{0j} \sim N(0, \tau_{00}), \\
\beta_{qj} &= \gamma_{q0}, \qquad\qquad q=1, \ldots, 4, \\
\ln(1/\sigma_j^2) &= \alpha_0 + \alpha_1\text{IMA}_j + \alpha_2\text{CS}_j + \alpha_3(\overline{\text{Lang}_{.j}} - \overline{\text{Lang}_{..}}) . \qquad (9)
\end{aligned}
$$

The parameters α_1 and α_2 in the dispersion model are the expected differences in dispersion, respectively between the IMA and control groups and between the CS and control groups, controlling for the classroom proportion of ELP students. The parameter α_3 is the expected change in dispersion when the ELP proportion changes from 0% to 100%, holding constant the treatment type. Table 3 presents the results.

The key parameters, the IMA/control and the CS/control contrasts in dispersion, are significant, which indicates that both the IMA and CS dispersions are significantly greater than the control group dispersion. The slope of the classroom proportion of ELP students is also significant holding constant treatment type.

Transformed back to the original scale as the last three lines of Table 3 shows, the posterior means of the residual variances (standard deviations) for the IMA, CS, and control groups are respectively 6.86, 6.55, and 4.13. The upper end of the 95% interval of the control group variance overlaps the lower ends of those of treatment groups. The classroom proportion of ELP students is also significant, with a coefficient of -0.69 holding constant treatment type. On the variance scale, it is 2.00.

Since student language status is already included in the mean model, the relationship between dispersion and the classroom proportions of ELP students can be viewed as a contextual effect on dispersion. Even students with the same level of language proficiency and other important characteristics may be expected to vary more in outcome, by being in a classroom with a larger proportion of ELP students.

Table 3

Results from Model 2

	Estimate	95% Interval	Median
Fixed Effects:			
Model for adjusted class means			
CTL Grand Mean (γ_{00})	3.941	(3.066 4.820)	3.940
IMA/CTL contrast (γ_{01})	2.175	(1.016 3.309)	2.181
CS/CTL contrast (γ_{02})	1.205	(0.027 2.394)	1.204
Average Within-class slopes:			
Prec/Posttest slope (γ_{10})	0.545	(0.411 0.679)	0.545
Prep/Posttest slope (γ_{20})	0.151	(0.071 0.233)	0.151
ELP/Posttest slope (γ_{30})	0.054	(-0.607 0.717)	0.051
Incip/posttest slope (γ_{40})	1.030	(0.517 1.543)	1.030
Variance Components:			
Between class:			
Var. in adjusted mean (τ_{00})	0.799	(0.341 1.655)	0.731
Within class:			
Ln. of CTL precision (α_0)	-1.410	(-1.722 -1.114)	-1.403
IMA/CTL contrast (α_1)	-0.513	(-0.900 -0.110)	-0.515
CS/CTL contrast (α_2)	-0.465	(-0.841 -0.083)	-0.466
ELP proportion slope (α_3)	-0.693	(-1.313 -0.054)	-0.696
IMA variance	6.856	(5.571 8.444)	6.805
CS variance	6.546	(5.193 8.204)	6.487
CTRL variance	4.133	(3.048 5.598)	4.069
Difference between:			
IMA and CTRL variance	2.722	(0.620 4.781)	2.730
CS and CTRL variance	2.413	(0.448 4.438)	2.402

Model 3: HM with both heterogeneity of residual dispersion and interactions. Model 3 extends Model 2 to include two interactions between treatment type and student characteristics: $IMA_j \times Lang_{ij}$ and $CS_j \times Prep_{ij}$. Since the treatment variables are Level-2 or cluster-level variables, their interactions with student characteristics constitute cross-level interactions. The model is specified as follows:

$$
\begin{aligned}
y_{ij} &= \beta_{0j} + \beta_{1j}(\mathrm{Prec}_{ij} - \overline{\mathrm{Prec}_{..}}) + \beta_{2j}(\mathrm{Prep}_{ij} - \overline{\mathrm{Prep}_{..}}) + \beta_{3j}(\mathrm{Lang}_{ij} - \overline{\mathrm{Lang}_{..}}) + \beta_{4j}(\mathrm{Incip}_{ij} - \overline{\mathrm{Incip}_{..}}) + r_{ij}, \\
& \qquad\qquad\qquad\qquad\qquad\qquad r_{ij} \sim N(0, \sigma_j^2), \\
\beta_{0j} &= \gamma_{00} + \gamma_{01}\mathrm{IMA}_j + \gamma_{02}\mathrm{CS}_j + u_{0j}, \qquad u_{0j} \sim N(0, \tau_{00}), \\
\beta_{1j} &= \gamma_{10}, \\
\beta_{2j} &= \gamma_{20} + \gamma_{21}\mathrm{CS}_j, \\
\beta_{3j} &= \gamma_{30} + \gamma_{31}\mathrm{IMA}_j, \\
\beta_{4j} &= \gamma_{40}, \\
\ln(1/\sigma_j^2) &= \alpha_0 + \alpha_1\mathrm{IMA}_j + \alpha_2\mathrm{CS}_j + \alpha_3(\overline{\mathrm{Lang}_{.j}} - \overline{\mathrm{Lang}_{..}}) . \qquad (10)
\end{aligned}
$$

The parameters, γ_{21} and γ_{31}, are the key coefficients in this model, capturing the magnitudes of the interactions. Note that the cross-level interactions make the slopes of the corresponding Level-1 variables (i.e., β_{2j} and β_{3j}) non-randomly varying, that is, varying depending on Level-2 covariates, but no random component is attached to the slopes. Table 4 presents the results.

Table 4

Results from Model 3

	Estimate	95% Interval	Median
Fixed Effects:			
Model for adjusted class means			
CTL Grand Mean (γ_{00})	3.891	(3.009 4.772)	3.893
IMA/CTL contrast (γ_{01})	2.155	(1.012 3.295)	2.157
CS/CTL contrast (γ_{02})	1.364	(0.176 2.547)	1.363
Average Within-class slopes:			
Prec/Posttest slope (γ_{0})	0.555	(0.419 0.691)	0.555
Prep/Posttest slope others (γ_{20})	0.119	(0.028 0.210)	0.119
difference from CS (γ_{21})	0.131	(-0.045 0.306)	0.131
ELP/Posttest slope others (γ_{30})	-0.144	(-0.878 0.576)	-0.142
difference from IMA (γ_{31})	1.064	(-0.654 2.781)	1.066
Incip/posttest slope (γ_{40})	1.015	(0.500 1.533)	1.015
Prep/Posttest slope CS ($\gamma_{20}+\gamma_{21}$)	0.250	(0.094 0.405)	0.250
ELP/Posttest slope IMA ($\gamma_{30}+\gamma_{31}$)	0.920	(-0.650 2.481)	0.922
Variance Components:			
Between class:			
Var. in adjusted mean (τ_{00})	0.795	(0.336 1.638)	0.729
Within class:			
IMA/CTL contrast (α_1)	-0.521	(-0.905 -0.127)	-0.523
CS/CTL contrast (α_2)	-0.456	(-0.828 -0.081)	-0.457
ELP proportion slope (α_3)	-0.628	(-1.248 0.013)	-0.632
IMA variance	6.889	(5.580 8.490)	6.839
CS variance	6.465	(5.129 8.117)	6.409
CTRL variance	4.118	(3.044 5.538)	4.055

Although both key parameters capturing the interactions are only border-line significant or insignificant, both are worth investigating. In multisite studies in which intact clusters are assigned to different treatment programs, the studies may suffer from lack of power for estimating the main effect of treatment unless they have many clusters in each treatment program. This problem may be aggravated when estimating interactions with the treatment. Given that the data have only 21 sites, and at most 9 sites within a treatment type, it is warranted to study the interactions further.

The coefficient capturing the interaction, i.e., the difference in slopes, between CS and the Computation pretest is 0.13. Although the 95% interval still does not exclude the value 0, note that the 96% of the mass of the posterior distribution is above 0 (Bayesian p-value=.96). The magnitude of slope seems to matter as well. While, for other groups, the posterior mean of the computation pretest slope is 0.12, the CS group slope is 0.25 which is more than two times the magnitude of other groups.

The coefficient capturing the interaction between IMA and language status is 1.2. As is the other interaction above, although the interaction is not significant, most of the mass of the posterior distribution lies above 0 (Bayesian p-value=.92). For other groups, the expected difference in outcome between ELP and non-ELP students is -0.14 which is insignificant. The expected difference in IMA group is 1.0, which is clearly significant (Bayesian p-value=1.0).

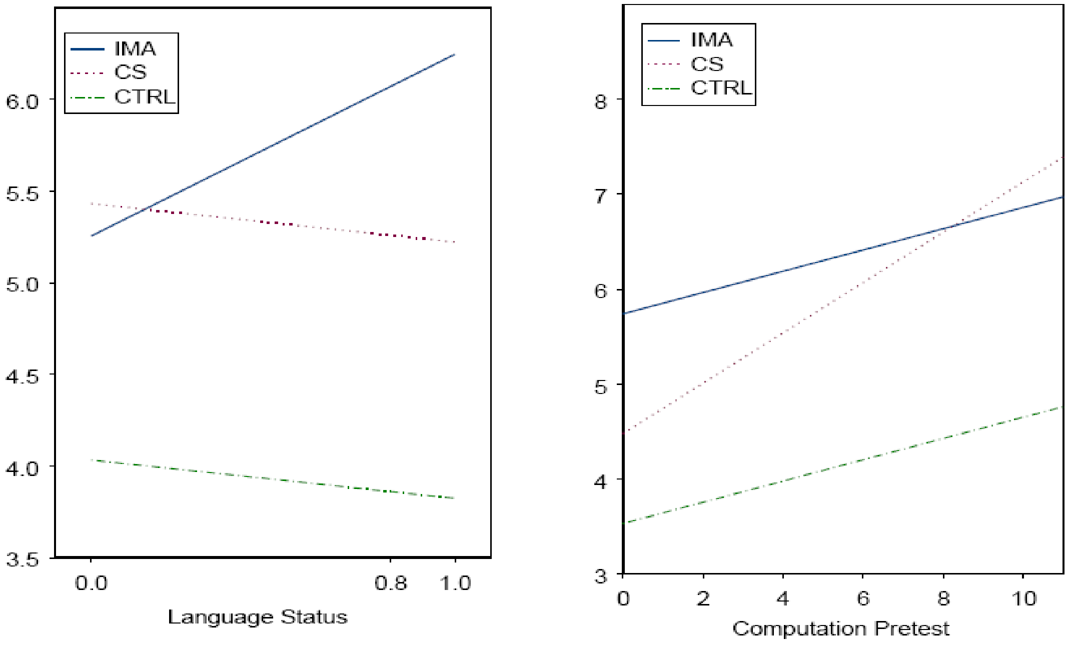

Figure 3. Summary plots of estimated relationships between outcome and student characteristics for each treatment type.

Figure 3 displays the estimated relationships between the two student characteristics and the outcome. Different lines display the relationships for different treatment types. Since treatment groups tend to perform better in the outcome given the value of the covariates, the lines for treatment groups (i.e., IMA and CS) are located higher for both plots.

In the left panel which plots the relationship between the ELP status and the posttest, the slope of the IMA group is positive and steeper compared to the other two groups. In the CS and the control groups, ELP and non-ELP students tend to perform at a similar level controlling for other student characteristics in the model, while in the IMA group ELP students tend to perform better than non-ELP students.

In the right panel which shows the relationship between the Computation Pretest and the posttest, the slope of the CS group is steeper than those of the other groups. One can see that, for students with very low computation skills on fractions prior to treatment, with other characteristics being equal, CS group students tend to perform much worse in the posttest than IMA group students. For students who have more computation knowledge prior to treatment, CS group students tend to perform as well as IMA group students.

The two treatment programs, IMA and CS, employed the same innovative curriculum units that are designed to encourage reform-minded instruction. As noted earlier, the IMA teachers received intensive training in reform-minded instruction in general and in the use of the innovative curricular units in particular, while the CS teachers had a chance to meet and discuss the challenges implementing curricular units with each other. One might wonder why these interactions would arise; different student characteristics interacting with different treatment programs.

Two measures from classroom observations, Conceptual/Assessment OTL measure and Numerics OTL measure, are extremely beneficial for understanding the interactions, by providing information about the dynamics of classroom instructions. The IMA classrooms have a high average score in the Conceptual/Assessment OTL measure, which means that in these classes there tended to be appreciably more whole-class discussion in the context of problem solving. Given this method of instruction, it is obvious that the ability to benefit more from the instruction will depend on student's English proficiency to some extent.

The CS classrooms have a very low average score in the Numerics OTL, while the IMA and the CTRL classrooms have high average scores alike. This means that in the CS classes there was insufficient instruction dealing with numerical representations and operations of fractions. Researchers believed that this would probably be due to the graphical focus of the innovative curricula used in the instruction (Gearhart et al., 1999). When students have low opportunity of dealing with computations of fractions, it appears that students who came with relatively low level of skill in the area were unable to learn much from the CS classes.

The coefficients, α_1 and α_2, estimate the expected differences in dispersion between IMA and the control groups, and between CS and the control groups, respectively, holding constant the proportion of ELP students. After including the interactions that, we hypothesize, caused the heterogeneity in the residual variance across treatment type when unspecified, we expect the degree of heterogeneity to decrease. However, the coefficients in the model indicating the degree of heterogeneity, α_1 and α_2, remain almost the same as those in the HM that does not contain the interactions (Model 2).

In order to understand the results, we can think of the variance reduction in terms of simplified statistical models in which there are only two conditions, treatment and control, in a non-nested setting (Bryk & Raudenbush, 1998). First, as for the comparison between the IMA and the control groups, when the interaction is unspecified, the residual term of the IMA group includes $\gamma_{31}Lang_{ij}$ while the residual term of the control group does not. This term yields the variance difference of γ_{31}^2 $Var(Lang_{ij})$, which is 0.21. Likewise, as for the comparison between the CS and the control groups, when the interaction is unspecified, the residual term of the CS group includes $\gamma_{21}Prep_{ij}$ whereas the residual term of the control group does not. This term gives rise to the variance difference of γ_{21}^2 $Var(Prep_{ij})$, which is 0.32. The actual estimated reductions do not exactly agree with these values in both comparisons, because the HMs are more complex and involve more terms in the residual variance.

However, even the simplified reductions in the residual variance, 0.21 and 0.32, are too small in comparison to the magnitudes of differences in the residual variance between the treatment groups and the control group. The differences reach more than 2.8 for the IMA vs. control, and about 2.3 for the CS vs. control. This implies that a considerable portion of the heterogeneity across treatment types may arise from sources other than unspecified interactions between treatment and student characteristics. There may be unmeasured site characteristics that are related to both treatment type and dispersion, in addition to the proportion of ELP students.

Summary and Discussion

This paper views dispersion in outcomes as an important dimension of a study, and provides a conceptual framework and statistical models for studying dispersion in experimental and quasi-experimental settings. While a primary interest in many studies and questions has been obtaining unbiased estimates of the effects of treatments on the mean levels in the outcome (Holland, 1986; Rubin, 1974, 1978; Shadish, Cook, & Campbell, 2002), this paper focuses on drawing sound inferences on the effects of treatments on dispersion and thereby suggesting a way to enhance understandings about the effects of treatments, e.g., uncovering interactions between treatment and subject characteristics (e.g., Cronbach & Snow, 1977; Cronbach, 1975).

We started with a conceptual framework laying out sources of differences in dispersion in various research settings: experiments, quasi-experiments in non-nested settings, and quasi-experiments in nested settings. The framework draws on the work by Bryk and Raudenbush (1988), Raudenbush and Bryk (1987), and theories of potential outcomes (Holland, 1986; Rubin, 1974, 1978). We then focused on the most complex setting, i.e., quasi-experiments in nested settings, and presented a series of HMs and procedures that are helpful in efforts to draw sound inferences concerning the effects of treatments on dispersion. By enumerating sources of differences in dispersion using the logic of potential outcomes, key statistical equations, and graphical representations, the conceptual framework helps us understand how sources of confounding at different levels (e.g., student, class) can be adjusted for with regard to inferences on dispersion, and how a significant effect of a treatment on dispersion may signal unspecified interactions between treatment and subject characteristics.

Our example illustrates procedures for drawing sound inferences concerning the effects of treatments on dispersion as well as on mean levels in the outcome of interest. The base HM (Model 1) addresses the basic questions of evaluation studies, i.e., it provides an estimate of the expected differences in the posttest scores between the IMA and control groups, and between the CS and control groups. By carefully controlling for possible differences in student characteristics prior to the treatments, this model addresses the effect of treatments on expected levels of outcomes.

Although many evaluation studies may end with estimates of expected differences in outcomes, the framework presented in this paper suggests attending to the heterogeneity in residual variance. Model 2 is set up for this purpose. In our example, we elaborated the dispersion structure and modeled the residual variance as a function of treatment indicators and a site characteristic, i.e., the classroom proportion of English Language Proficient (ELP) students. Based on our conceptual framework, Model 2 carefully makes covariance adjustments both for the student-level and for class-level confounding variables. The initial difference in dispersion in student status in proficiency in English (i.e., student-level or Level-1 variable) is adjusted for by elaborating the mean structure, while the initial difference in dispersion in the classroom proportion of ELP students (i.e., class-level or Level-2 variable) is adjusted for in the dispersion structure.

Since both the IMA and CS dispersions are significantly greater than the control group dispersion after adjusting for pre-existing differences, it suggests that the true treatment effect may be different across students. The rough interpretation would be that the treatments, IMA and CS, may be more effective for students with higher ``aptitude'' than lower ``aptitude'' (e.g., Scenario [b] in Figure 1; cf. Cronbach & Snow, 1977). We then ask: What would be the dimension of aptitude that is interacting with the treatment? Model 3 further elaborates the mean structure, by including two interactions between treatment type and student characteristics. In IMA classes, students who are proficient in English tend to benefit more from the instruction. In CS classes, students who came with better capability of solving more computational items on fractions tend to profit more than students who do not.

The modeling strategies and procedures we present result in greater elaboration of the mean and dispersion structures (e.g., Equation 10), which can be beneficial to evaluation studies in important ways. First, the elaboration with regard to interactions can provide critical insights into individual differences in responses to a given treatment and enhance understandings about the treatment of interest. For example, we gain understanding about the characteristics of students who might benefit more from the treatment, or the conditions (e.g., site characteristics, or level of implementation) under which the treatment might be more effective.

In our illustrative example, the interaction results yield far more detailed knowledge about the innovative curricula and professional development than simple results concerning which treatment produces a higher average level of achievement. Teaching with the innovative curricula involves much whole class discussion and thus it is possible that non-ELP students may be left out. Also, without professional development about how to properly use the innovative curricular units, it seems likely that teachers may implement them improperly, and neglect important but more traditional instructional foci, i.e., dealing with operations of fractions.

It is notable that these important interactions are more likely to be overlooked unless we pay attention to the signal arising from the differences in dispersion. Since the slopes of individual characteristics do not turn out to be significantly varying across Level-2 units (i.e., classrooms), it is easy to forget the possibility that the slopes may vary depending on characteristics of Level-2 units, which is the treatment type in this case. (See Raudenbush & Bryk's [2002] discussion of non-randomly varying slopes.) It appears that, by virtue of the link between heterogeneous variance and interactions, the interactions are possible to find.

Next, the elaboration with regard to the prior characteristics of students or clusters can increase the validity and the statistical power of the inferences as well as contributing to substantive knowledge about the study. Inclusion of the prior characteristics results in ANCOVA-like adjustments in the HMs both in the mean and dispersion structures (cf. Raudenbush & Bryk, 2002, Chapter 2). This helps us draw valid inferences concerning the effect of treatment, by controlling for possible pre-existing differences between treatment and control groups, especially in quasi-experimental studies. In addition, by decreasing the within-group variability in the outcome, it helps increase the power of the inferences concerning the effect of treatment.

Also the relationship between a prior characteristic and an outcome of interest can be of substantive interest in and of itself. In our illustrative example, students with a rudimentary understanding of the concept of fractions prior to treatment tended to perform significantly better on the outcome controlling for other important characteristics compared to students without such understanding, while ELP students tended to show no significant difference in their performance compared to non-ELP students. Although ELP status did not make a difference at the student level on average, the classroom proportion of ELP students was significantly related to the dispersion in outcomes. This indicated the contextual effect of ELP on dispersion in outcomes. By being in a class with more ELP students, the potential outcomes of students may have tended to be more variable around the expected value. The investigation of heterogeneity in the residual dispersion resulted in identifying a key classroom-level source of diversity: the classroom proportion of ELP students.

The conceptual framework, and the statistical modeling and procedures in this paper can be applicable to other types of evaluation studies, such as multisite evaluation studies that use a different design, longitudinal evaluation studies, or, under certain circumstances, observational studies (e.g., surveys such as High School and Beyond, or NELS) that may focus on the effect of certain practices or policies (Rosenbaum, 2002).

The modeling strategies are most readily applicable to another multi-site evaluation design. Commonly encountered designs in multi-site studies tend to fall into two broad categories (Seltzer, 2004; see also Kirk 1982 and Raudenbush, 1993). This paper has focused on one type of design, in which different treatment conditions are assigned to different intact clusters (e.g., classrooms), but no blocking is employed. This would give rise to a sample of clusters nested within each treatment condition. The other type of design involves blocking, e.g., forming matched pairs of schools or classrooms. Different treatment conditions are assigned to intact clusters within each matched pair. We refer to the first type as nested design and the second as crossed design.

One can use dispersion models such as Equation 9 to detect the effect of treatment on dispersion. For the crossed design, however, in settings where the treatment effect may be substantially variable across sites, it is still possible for variance heterogeneity to go unnoticed by applying the dispersion model presented in Equation 9, unlike the nested design. For example, if the variance of the treatment group is greater for some sites while the variance of the control group is greater for other sites, then they may cancel each other out, resulting in a null difference on average. In such settings, dispersion modeling may entail examining heterogeneity in dispersion between groups in each site and identifying site characteristics that relate to the extent of heterogeneity in dispersion between groups.

Specifically, the estimate of $\sigma^2_{j\ Trt=1} - \sigma^2_{j\ Trt=0}$, captures the difference between treatment and control groups in the within-site residual variance at site *j*, after accounting for treatment type and student characteristics. When the difference is significant, it implies that the treatment effect is variable across students at site *j*, and moreover signals the interaction between the treatment and particular student characteristics at site *j*. When the differences in the residual variance in different sites are related to certain site characteristics, one may also test if the relationship is significant by modeling the difference, $\sigma^2_{j\,Trt=1} - \sigma^2_{j\,Trt=0}$, as a function of relevant site characteristics. A significant relationship may reveal how differences in site-level factors relate to differences in the extent of interactions between treatment and student characteristics.

Furthermore, the logic and procedures presented in this paper apply to longitudinal evaluation studies, although the HMs require another level of nesting. In longitudinal evaluation studies, the focus is on the change trajectories before, during, or even after treatment, rather than on the outcome score at the end of the treatment in cross-sectional studies. Let us assume that intact classrooms are assigned to either treatment or control conditions, and that students within classrooms are measured on several occasions on the outcome of interest. For this longitudinal study, three-level HMs (i.e., HMs with repeated observations, nested within students, who in turn are nested within classrooms) may be used in contrast to the two-level HMs in pretest-posttest evaluation studies (Raudenbush & Bryk, 2002, Chapter 6).

Then the dispersion structure of interest would be the Level-2 residual dispersion instead of the Level-1 residual dispersion in pretest-posttest studies, especially the dispersion of student rates of change. The differences in the dispersion of student rates of change between treatment and control conditions can, under certain circumstances, indicate individual differences in responses to a given treatment. In addition, in inferences concerning dispersion, student-level possible confounding variables should be adjusted for by including in the mean structure, or specifically in the Level-2 (i.e., student-level) mean structure, while cluster-level possible confounding variables should be adjusted for by including them in the dispersion structure, or specifically modeling the dispersion in student rates of change.

Throughout, this paper assumes either experimental or quasi-experimental studies, with more focus on the quasi-experimental studies. Although quasi-experimental studies do not directly employ random assignment, they potentially offer more control compared with observational studies in terms of the subjects chosen for the study, methods of assignment, and measures of prior characteristics, etc. (Shadish, Cook, & Campbell, 2002). As a result, in well-conducted quasi-experimental studies, the treatment and control groups tend to be rather similar in important prior characteristics. In contrast, in observational studies (Rosenbaum, 2002), the group of individuals who were exposed to the 'treatment' of interest are likely to be very different in important prior characteristics than the group of individuals who were not exposed.

Direct application of the modeling strategy of this paper to such observational studies can be rather limited, mainly because the modeling strategy is based on covariance adjustment of confounding variables. We identify sources of confounding by individual or site characteristics, and make covariance adjustment of individual characteristics in the mean structure, and of cluster characteristics in the dispersion structure. This model-based adjustment may not be as effective to remove biases in observational studies. Much literature (e.g., Cochran, 1957; Cochran 1965; Rubin, 2001) show that, in settings where there is considerable disparity in the means or variances between treatment and control groups in covariates, the covariance adjustment may be either ineffective to remove biases or even increases the biases.

However, once comparable samples are constructed to draw sound inferences in observational studies, e.g., using matching, stratification, weighting based on propensity scores (Rosenbaum & Rubin, 1983; Rosenbaum, 2002), the conceptual framework and strategies are readily applicable just as in quasi-experimental studies. Propensity scores have been increasingly applied to many observational studies in recent years. By yielding comparable distributions in all observed variables, adjustment for propensity scores removes biases and thus can help us in efforts to draw valid causal inferences. Once the constructed comparison group sample becomes similar to the individuals in the treated or exposed group in terms of the distributions of all observed variables as a result of the use of propensity score methods, the modeling strategies of this paper such as covariance adjustment both in the mean and dispersion structures, or the elaboration with regard to interactions, remain a viable option for uncovering important individual differences to a given 'treatment' of interest.

References

Belin, T. R., & Rubin, D. B. (1995). The Analysis of Repeated-Measures Data on Schizophrenic Reaction Times Using Mixture Models. *Statistics in Medicine, 14,* 747-768.

Browne, W. J., Draper, D., Goldstein, H., & Rasbash, J. (2002). Bayesian and likelihood methods for fitting multilevel models with complex level-1 variation. *Computational Statistics and Data Analysis, 39,* 203-225.

Bryk, A. S., & Raudenbush, S. W. (1988). Heterogeneity of variance in experimental studies: a challenge to conventional interpretations. *Psychological Bulletin, 104(3),* 396-404.

Cochran W. G., Chambers S. P. (1965). The Planning of Observational Studies of Human Populations**.** *Journal of the Royal Statistical Society. Series A (General)*, 128(2), 234-266.

Cochran W. G. (1957). Analysis of Covariance: Its Nature and Uses. *Biometrics, 13(3),* 261-281.

Cronbach, L. J., & Snow, R. E. (1977). *Aptitudes and instructional methods: a handbook for research on interactions.* New York: Irvington Publishers.

Cronbach, L. J. (1975). Beyond the Two Disciplines of Scientific Psychology. *The American psychologist, 30(2),* 116-127.

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.

Goldstein, H. (1995). *Multilevel Statistical Models, 2nd ed.* London: Edward Arnold.

Goldstein, H., Healy, M.J.R., & Rasbash, J. (1994). Multilevel time series models with applications to repeated measures data. *Statistics in Medicine, 13,* 1643-1655.

Holland, P. W. (1986). Statistics and Causal Inference. *Journal of the American Statistical Association, 81,* 945-70.

Kasim, R. & Raudenbush, S. W. (1998). Application of Gibbs Sampling to nested variance components models with heterogeneous within-group variance. *Journal of Educational and Behavioral Statistics, 20(4),* 93-116.

Kirk, R. E. (1982). *Experimental design: procedures for the behavioral sciences 2nd ed.* California: Brooks/Cole.

Littell, R. C., Milliken, G. A., Stroup, W. W., & Wolfinger, R. D. (1996). *SAS system for Mixed Models.* Cary, NC: SAS Institute, Inc.

McCullagh, P., & Nelder, J. A. (1989). *Generalized Linear Models, 2nd ed.* London: Chapman and Hall.

Raudenbush, S. W. (1993). Hierarchical linear models and experimental design. In L. Edwards (Ed.), *Applied analysis of variance in behavioral science* (pp. 459-496). New York: Marcel Dekker.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods, 2nd ed.* Newbury Park, CA: Sage.

Raudenbush, S. W., & Bryk, A. S. (1987). Examining correlates of diversity. *Journal of Educational Statistics, 12,* 241-269.

Rosenbaum, P. R. (2002). *Observational Studies, 2nd ed.* New York, NY: Springer.

Rosenbaum, P. R. & Rubin, D. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70,* 41-55.

Rubin, D. B. (1974). Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. *Journal of Educational Psychology, 66,* 688-701.

Rubin, D. B. (1978). Bayesian Inference for Causal Effects: The Role of Randomization. *Annals of Statistics, 6,* 34-58.

Rubin, D. B. (2001). Using Propensity Scores to Help Design Observational Studies: Application to the Tobacco Litigation. *Health Services and Outcomes Research Methodology, 2(3-4),* 169-188.

Seltzer, M. H. (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.

Seltzer, M. H. (1994). Studying variation in program success: A multilevel modeling approach. *Evaluation Review, 18(3),* 342-361.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Boston: Houghton-Mifflin.

Snijders, T. A. B. & Bosker, R. J. (1999). *Multilevel analysis: an introduction to basic and advanced multilevel modeling*. Thousand Oaks, CA: Sage.

Wolfinger, R.D. (1996). Heterogeneous variance-covariance structures for repeated measures. *Journal of Agricultural, Biological, and Environmental Statistics, 1(2),* 205-230.