The Impact of Including Predictors and Using Various Hierarchical Linear Models on Evaluating School Effectiveness in Mathematics

Nicole M. Traxel, PhD
Associate Scientist, Center for Addiction and Behavioral Health Research
University of Wisconsin - Milwaukee

Cindy Walker, PhD
Associate Professor, Educational Psychology
University of Wisconsin - Milwaukee

ABSTRACT

2-level, 2-level gain score, and 3-level hierarchical linear models were fit to data containing 2 years of standardized test results from a large Midwestern urban school district. First the models were fit including only initial score as a student level predictor of final score. Then, each model was fit adding race, socioeconomic status, and race and SES as student level predictors. Third, each model was fit adding race, SES, and race and SES as student and school level predictors. School effectiveness rankings varied by the type of model but not the predictors that were included. The 2-level gain and 3-level models were most valid; however, because of its power and flexibility, the 3-level model is best for examining school effectiveness.
A fair and accurate way to assess the effectiveness of individual schools is to determine how much value a particular school adds to student learning by examining student progress in their individual achievement levels over time (Andrejko, 2004; Crane, 2002; Kupermintz, 2003; Olson, 1998; Popham, 2005; Tekwe, et al., 2004). Value added accountability systems can be used to more accurately identify which schools are failing or succeeding, with success or failure being determined by student growth over time, as opposed to the change in proficiency classification rates for different cohorts of students over time (Crane, 2002). These value added accountability systems permit schools to be evaluated based on how much progress their students have made, regardless of their achievement levels when they entered the school (Ballou, et al., 1999; Raudenbush, 2004a). They provide a way of recognizing that schools serve students who start at different places and progress at different rates (Heck, 2006).

Hierarchical Linear Models

One particular way to implement a value added accountability system is to make use of hierarchical linear modeling. By its very nature, research pertaining to the effectiveness of a school requires exploration of hierarchical data because students are nested within classrooms and schools (Raudenbush & Bryk, 1986). Hierarchical models simultaneously model student level relationships and take into account the way students are grouped into individual schools and/or classrooms (Goldstein, 1997). By modeling both the student level and the school level simultaneously, the annual academic growth of students is separated into that which can be attributed to the student and that which can be attributed to the school (Hershberg, 2005; Meyer, 2003; Raudenbush & Bryk, 2002; Yeagley, 2007). Therefore, hierarchical linear modeling is a
way to isolate the impact of instruction on student learning. One advantage of using hierarchical linear modeling is that these models allow for improved estimation of effects within individual units, such as schools, which is accomplished through Empirical Bayes estimation. Empirical Bayes estimates are also called shrinkage estimates because in these estimates the group mean is pulled toward the grand mean proportionately to the group sample size and the reliability of the group mean (Raudenbush & Bryk, 2002). Several different hierarchical linear models can be used for these purposes and there have been no empirical studies examining which model is best. Thus, the current study compared and contrasted the results obtained from 2-level, 2-level gain score, and 3-level individual growth hierarchical linear models to determine which model is best when trying to ascertain the effectiveness of a school.

According to Meyer (1995), achievement growth from one grade to the next can be adequately characterized by a 2-level hierarchical linear model, with student characteristics represented by level-1 of the model and school characteristics represented by level-2 of the model. In a model such as this, a student’s final achievement test score is entered as the dependent variable of the level-1 model and the initial score on the test is entered as one of the independent variables in the level-1 model. Then, each of the regression coefficients in the level-1 model are conceived as outcome variables of the level-2 model that are hypothesized to depend on specific organizational characteristics. For example, if the initial achievement test score is included as a predictor in level-1 it becomes an outcome variable in one of the level-2 equations. Level-2 (school level) predictors such as race and SES could then be included as predictors of initial score. A unique set of predictors can be specified for each regression coefficient. These level-2 predictors show the net effects of the predictor after controlling for all of the level-1 predictors (Raudenbush & Bryk, 2002).
In general, the equations for the 2-level model are as follows:

Level-1: \[ Y_{ij} = \beta_{0j} + \sum_{q=1}^{Q} \beta_{qj} X_{qij} + r_{ij}, \] (1)

Level-2: \[ \beta_{qj} = \gamma_{q0} + \sum_{s=1}^{S} \gamma_{qs} W_{sj} + u_{qj}, \quad q \geq 0, \] (2)

where \( \beta_{0j} \) is the level-1 intercept coefficient for school \( j \), \( \beta_{qj} \) are the \( q \) level-1 slope coefficients for school \( j \), \( X_{qij} \) are the \( q \) level-1 predictors for student \( i \) in school \( j \), \( r_{ij} \) is an error term that is a level-1 random effect for student \( i \) in school \( j \), \( \gamma_{q0} \) is the level-2 intercept coefficient for predictor \( q \), \( \gamma_{qs} \) are the \( s \) level-2 slope coefficients for the \( q \) predictors, \( W_{sj} \) are the \( s \) level-2 predictors for school \( j \), and \( u_{qj} \) is the level-2 random error term for school \( j \) on predictor \( q \). For example, \( Y_{ij} \) could be the final achievement test score for student \( i \) in school \( j \). This final score could be predicted by an intercept which reflects the average scale score for school \( j \) and by other predictors for this student including his/her initial score and SES. There would be some random error for this student in this school as well. These would all be represented by the terms in level-1. Then each of the coefficients from the level-1 equation become outcomes in the level-2 equations. For example, the coefficient associated with the student’s initial score would be predicted by an intercept and by other predictors at the school level like the percentage of students receiving free or reduced price lunch in the student’s school or by the percentage of minority students in that school. There would be some random error for the school as well.

Measuring student progress, as is done in value added models, requires controlling for initial achievement, which can either be done by including the pretest score as a predictor or by basing the analysis on the simple difference or gain score. To do the latter, a difference or gain score from adjacent grades can be calculated and then used as the dependent variable in a 2-level hierarchical linear model, as long as the test is vertically scaled (Bracey, 2004; Kupermintz,
The Impact of Including 

2003; McCaffrey, et al., 2004; Tekwe, et al., 2004). One potential disadvantage of using gain scores is that these models require that observations are available from both years. Thus, students with incomplete data cannot be used (McCaffrey, et al., 2004; Tekwe, et al., 2004). Nevertheless, the equations for gain score models are the same as for the 2-level model shown above, except the gain score replaces the final score as the dependent variable in the level-1 equation.

In a 3-level individual growth model, the level-1 model is a repeated-observations model. At this level, each person’s development is represented by an individual growth trajectory that depends on a unique set of parameters. The parameters in this level of the model are considered to vary over time. That is, variables like age or grade vary with the timepoints at which the tests were administered. For example, initial test scores could be coded with 0 for the grade level while final scores could be coded with 1 for the grade level. Then the grade level variable would be entered as a predictor in level-1 of the model. The coefficient associated with this predictor would then be interpreted as the change in test scores from time 0 to time 1. Other variables that vary over time may also be included at this level. Level-2 of this model is the person level model. In this level, the outcome variables are the individual growth parameters in the level-1 model. Returning to the example, the coefficient for the time variable, which was interpreted as the change in test scores from time 0 to time 1, would become the outcome variable in level-2. Thus, variables included in level-2 of this model are predictors of the change in test scores from time 0 to time 1. This model can also include person level characteristics, such as race or socioeconomic status. This level of the model shows the variation among growth parameters between level-2 units within level-3 units. Level-3 of this model is the organization level which includes organization level variables. For example, the organization level could be schools, and this level would then model the variation among schools (Raudenbush & Bryk, 2002).
In general, the equations for the 3-level model are as follows:

**Level-1:**

\[ Y_{tij} = \pi_{0ij} + \sum_{p=1}^{P} \pi_{p0j} a_{ptij} + e_{tij}, \]  

(3)

**Level-2:**

\[ \pi_{p0j} = \beta_{p0j} + \sum_{q=1}^{Q_p} \beta_{pqj} X_{qij} + r_{p0j}, p \geq 0, \]  

(4)

**Level-3:**

\[ \beta_{pqj} = \gamma_{pq0} + \sum_{s=1}^{S_p} \gamma_{pq0} W_{sij} + u_{pqj}, q \geq 0, \]  

(5)

where \( Y_{tij} \) is the outcome at time \( t \) for child \( i \) in school \( j \), \( a_{ptij} \) is an indicator of which predictor \( p \) is reflected by child \( i \) in school \( j \) at time \( t \), \( \pi_{0ij} \) reflects the initial score of child \( i \) in school \( j \), \( \pi_{p0j} \) is a coefficient for predictor \( p \) for student \( i \) in school \( j \), and \( e_{tij} \) is random error. The remainder of the equations are the same as those from the 2 level model, with the exception that now there is a third subscript which reflects time.

The main advantage to fitting a 3-level model is the flexibility of the approach in handling missing data. This model readily incorporates all participants who have been observed at least once (Doran, 2004; Raudenbush & Bryk, 2002). The results can be interpreted as if no missing data were present under the assumption that the data are missing at random (Doran, 2003; Raudenbush & Bryk, 2002). Even if the nonignorable missingness assumption is not met, the results are generally robust to missingness to the extent that all of the data are efficiently used. Three-level hierarchical linear models make efficient use of all data, even using data from individuals with responses at only one time point. Two-level models with students at level-1 and schools at level-2 only include students with complete data (Raudenbush & Bryk, 2002).

**Using HLM to Study School Effectiveness**

Policy makers want to know how well students in one school fare compared with those in other schools (Willms, 1999). The impact that a school has on the progress or lack of progress in
educational advancement of learning, or growth, of a student is referred to as the effect that the
school has on the educational progress of its students (Kupermintz, 2003). Thus, results from
school effectiveness studies are used to hold schools accountable for their role in increasing
student learning (Pituch, 1999). To identify effective schools, researchers often use a regression
based approach that captures the effect of an individual school in the model residual term
(Pituch, 1999). These residuals are estimates for each school of the deviation of the estimated
slope or intercept from its predicted value based on the model (Betebenner, 2004; Raudenbush &
Bryk, 2002). That is, the effect of attending a school is to adjust the growth rate from the average
slope to the average slope plus the effect of the school (Doran, 2004; Olson, 2004). Schools with
higher residuals, that is, schools that score higher than predicted, are viewed as effective
(Goldstein, 1997; Raudenbush & Bryk, 2002). Thus, schools can be ranked based on their
residuals (Thompson, 2004). In hierarchical linear modeling, Empirical Bayes residuals are
obtained for each slope and intercept in each level of the model. Recall that the Empirical Bayes
estimator is viewed as a shrinkage estimator because it pulls the group mean toward the grand
mean (Raudenbush & Bryk, 2002). Shrinkage estimates are conservative in the sense that where
there is little information on any one school, the estimate is close to the average over all schools
(Goldstein, 1997). That is, all schools are assumed to be at the grand mean except when gains of
their students exceed or are less than the grand mean by a significant amount and are based on
significant numbers of students (Kain, 1998).

Many researchers include predictors such as race, gender, and socioeconomic status in
the models from which they obtain the residuals which are then used as indicators of the
effectiveness of a school (Goldstein, 1997; Olson, 2004; Olson, 1998; Pituch, 1999; Raudenbush,
2004; Raudenbush & Bryk, 2002). They claim that comparisons of schools must be based on
suitable adjustments for these and other factors that influence achievement levels or else biased estimates of school effects will result (Goldstein, 1997; Pituch, 1999). However, others argue that the effectiveness of a school in helping students to make gains should not be predicted based on its racial or economic makeup (TVAAS, web document). Effective teachers or schools raise the test scores of their students regardless of the racial or economic status of the students.

Some evidence does show that including covariates may have a notable impact on the value-added assessment of school performance (Kupermintz, 2003; Tekwe, et al., 2004). In one study, researchers fit a 2-level HLM both with and without covariates and the correlations between the school effects from these two models ranged from .61 to .95, which indicates that the two models were somewhat different, thus including the covariates did have an impact (Tekwe, et al., 2004). Yet another researcher concluded that inclusion of SES and demographic controls makes a considerable difference in effectiveness estimates of schools and teachers (Ballou, Sanders, & Wright, 2004).

Another decision that researchers must make concerns whether the effect of covariates should be included for the slope or for the intercept. That is, does each covariate affect the initial score of a student or their growth rate from one year to the next? Gong (2002) concluded that SES and race/ethnicity may be more strongly correlated with initial achievement test score than with growth. In addition, Braun (2005) concluded that while student characteristics are strongly correlated with initial status, it appears that the correlation is much weaker with changes in attainment. Furthermore, Ballou (2002) concluded that while covariates are partially controlled for by including pretest scores, they may still have an effect on the rate of progress, and thus they still need to be accounted for in the model.
Researchers must also decide whether to include covariates at the student level, at the school level, or both. In addition to the previously described individual level predictors, school level predictors can also be entered into a hierarchical linear model. For example, student level SES can be entered into level-1 of a 2-level model, and the percentage of students who qualify for free/reduced price lunch within a school can be entered into level-2 of the model. These predictors are called compositional or contextual effects when the aggregate of a person-level characteristic is related to the outcome even after controlling for the effect of the individual characteristic. This compositional effect is interpreted as the increment to learning that accrues to a student by virtue of being educated in one school versus another (Raudenbush & Bryk, 2002). These compositional effects may occur because of normative effects associated with an organization. For example, the racial makeup of the school could affect the culture of the school and cause some peer effects above and beyond that caused by the individual race of the child.

In addition, researchers must decide whether to use the residual for the intercept or the residual for the slope as an estimator of school performance. Several researchers use the predicted mean outcome, the residual for the intercept, as an estimator of school performance (Raudenbush & Bryk, 2002). This results in the ranking of schools based on the residuals for their final scores. Thus, the ranking actually reflects the average scale score of a school, which is strongly influenced by variables like socioeconomic status and race, resulting in the use of a status-based accountability system. However, as has already been argued, the effectiveness of a school is better evaluated using a measure of growth, not overall average (Kupermintz, 2003). Thus, the residual for the slope should be used as an indicator of school effectiveness, rather than the residual for the intercept, if one wants to use a value added accountability system rather than a status-based system.
The credibility of the effectiveness rankings can be enhanced by providing other types of evidence to accompany the results provided by this quantitative modeling approach (Pituch, 1999; Raudenbush, 2004a). For example, student, teacher, and parent satisfaction surveys or independent evaluations of the schools may provide evidence that reinforces the effectiveness estimates. Combining the effectiveness estimates with other information such as this can stimulate useful discussions about how to improve schools by helping to determine what makes effective schools different from ineffective schools (Raudenbush, 2004; Raudenbush, 2004a).

Research Questions

Various 2-level, 2-level gain score, and 3-level individual growth models were fit to determine whether effectiveness rankings differed among the models. Specifically, each model was fit in seven different ways. First, the 2-level and 3-level models were fit including only initial score as a predictor, while the 2-level gain score model was fit including no predictors. Then, each model was fit adding race, socioeconomic status, and race and socioeconomic status as student level predictors to the original models. Finally, each model was fit adding race and/or socioeconomic status as both student level and school level predictors to the original models. At the school level, predictors were entered for both the slope and the intercept. The significance of these predictors was compared across the three models to determine if predictor significance remained constant regardless of the model being fit. However, the larger focus of this study was to determine whether the inclusion of predictors changes the effectiveness rankings of schools. Specifically, the effectiveness rankings from the six models which included predictors were compared to the original model within each model type to determine whether including predictors changed effectiveness rankings. Then the effectiveness rankings from the original models of each model type were compared to determine whether the three types of models
resulted in different effectiveness rankings. Finally, the effectiveness rankings were validated using a gain in math focus measure which was completed by each school. Conclusions were then made on which model provided the most valid effectiveness rankings for schools.

Method

Participants

The sample consisted of 7,232 students from 128 schools from a large urban school district in the Midwest. Each student was in 3rd grade in the 2005-06 school year and 4th grade in the 2006-07 school year. Table 1 shows the demographics of the sample.

<table>
<thead>
<tr>
<th>Table 1. Demographic information.</th>
<th>Cohort 2015</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Schools</td>
<td>128</td>
</tr>
<tr>
<td>Number of Students</td>
<td>7232</td>
</tr>
<tr>
<td>Number of Males</td>
<td>3691</td>
</tr>
<tr>
<td>Number of Students Receiving Free/Reduced Lunch</td>
<td>5745</td>
</tr>
<tr>
<td>Number of Minority Students</td>
<td>6321</td>
</tr>
</tbody>
</table>

Measures

The primary measure in this study was the Mathematics scale score from a state mandated standardized test given to all public school students in grades 3 through 8 and 10. The Mathematics test in each grade level consisted of multiple choice and short answer questions with 80-90% of the total score on the test coming from the multiple choice questions, while 10-20% came from the short answer questions. The Mathematics tests in grades 3 and 4 covered the same sub-skill areas in roughly the same proportions, making scores from each test comparable. The Mathematics scale scores used in this study are vertically aligned so that student level growth can also be measured from one year to the next.

The standardized test results were accompanied by information on the race and socioeconomic status of each student. The race variable was dichotomized with minority groups
in one category and Caucasians in the other category as is commonly done in social science research. Similarly, the socioeconomic status variable was also dichotomous. Students receiving free or reduced price lunch were classified as having low socioeconomic status and students not receiving free or reduced price lunch were classified as not having low socioeconomic status.

The school level race and socioeconomic status predictors were determined by calculating the proportion of students at each school that were minorities and low socioeconomic status, respectively. Table 2 gives demographic information on the sample.

Table 2. Demographic information on the sample.

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>N (complete data for both years)</th>
<th>Mean Gain</th>
<th>Mean Initial Score</th>
<th>% Free/Reduced Lunch</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minority</td>
<td>6321</td>
<td>4372</td>
<td>38.05</td>
<td>395.22</td>
<td>84.5%</td>
</tr>
<tr>
<td>Caucasian</td>
<td>911</td>
<td>720</td>
<td>36.24</td>
<td>429.13</td>
<td>44.4%</td>
</tr>
</tbody>
</table>

The final measure that was used was a measure of math focus. This measure was obtained from a questionnaire which was administered to all of the mathematics teachers and administrators at each school. Each participant responded to an item reading, “There is a strong focus on increasing student achievement in mathematics at my school.” Participants responded on a 4 point Likert type scale with a response of 1 indicating Strongly Disagree, 2 indicating Disagree, 3 indicating Agree, and 4 indicating Strongly Agree. Math focus scores were averaged across all individuals in each school to compute a math focus score for each school in each year. Then a gain in math focus score was created by subtracting the math focus score from the first year from the math focus score from the second year for each school.

Procedure

Predictors were entered into the three types of hierarchical models considered in this research (i.e. 2-level, 2-level gain, 3-level growth). For this study, race and socioeconomic status were used as predictors. The purpose of the current study was to compare the effectiveness
rankings from models including predictors to those obtained from models including only initial
status (2-level and 3-level) or no predictors (2-level gain) to determine if including predictors in
a model changed the effectiveness rankings. This was determined for each of the 3 model types
(2-level, 2-level gain, and 3-level growth). Thus, the same predictors were included in each
model to allow for equitable comparisons of the models. Table 3 shows the 21 models that were
fit in this study. It should be noted that each of the 2-level and 3-level models included initial
score as a predictor, while none of the 2-level gain models included initial score as a predictor.
This was done because the initial score is included as one component of the gain score (gain
score being the final score minus the initial score), so it would be redundant, and perhaps
introduce some collinearity to the model, if the initial score was also included in the model.

Table 3. The models being fit.

<table>
<thead>
<tr>
<th></th>
<th>2-Level</th>
<th>2-Level Gain</th>
<th>3-Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial Score</td>
<td>None</td>
<td>Initial Score</td>
<td>Student Race</td>
</tr>
<tr>
<td>Student Race</td>
<td>Student Race</td>
<td>Student Race</td>
<td>Student Race</td>
</tr>
<tr>
<td>Student SES</td>
<td>Student SES</td>
<td>Student SES</td>
<td>Student SES</td>
</tr>
<tr>
<td>Student Race &amp; SES</td>
<td>Student Race &amp; SES</td>
<td>Student Race &amp; SES</td>
<td>Student Race &amp; SES</td>
</tr>
<tr>
<td>Student &amp; School Race</td>
<td>Student &amp; School Race</td>
<td>Student &amp; School Race</td>
<td>Student &amp; School Race</td>
</tr>
<tr>
<td>Student &amp; School SES</td>
<td>Student &amp; School SES</td>
<td>Student &amp; School SES</td>
<td>Student &amp; School SES</td>
</tr>
<tr>
<td>Student &amp; School Race &amp; SES</td>
<td>Student &amp; School Race &amp; SES</td>
<td>Student &amp; School Race &amp; SES</td>
<td>Student &amp; School Race &amp; SES</td>
</tr>
</tbody>
</table>

Whenever school level predictors were included they were included as predictors of both
the slope and the intercept as Raudenbush and Bryk (2002) suggest that a common set of school
level predictors should be introduced in all of the school level equations. The school level
variables of race and socioeconomic status were percentages of minority students within each
school and percentages of students receiving free/reduced price lunch within each school. These
percentages were determined for the individual cohort being examined in this study.

According to Raudenbush and Bryk (2002), the most common estimator of school
effectiveness is the predicted mean growth for each school. Schools that grow more than
predicted are viewed as effective while schools that grow less than predicted are viewed as ineffective, where growth is defined as the increase in test scores from one year to the next on a vertically scaled test. Typically, the effectiveness score is just the actual mean growth minus the predicted mean growth for each school. An advantage of HLM in determining school effectiveness is that the residuals (actual minus expected growth) in HLM are empirical Bayes residuals which provide a stable indicator for judging individual school effectiveness. These school level growth residuals were obtained from each of the models being fit (3 models x 7 predictor combinations), yielding 21 different sets of effectiveness ratings. To obtain growth residuals as opposed to achievement residuals, the residuals from the slope were used as opposed to the residuals from the intercept in each model.

To determine whether including predictors changed the effectiveness rankings of schools, the effectiveness rankings from the 2-level model containing only the initial score were then compared using Spearman correlations to the effectiveness rankings from the other six 2-level models listed in column 1 of Table 3. Again, it was expected that the magnitude of the correlations between the rankings from the model including only initial status and the other 6 models would be low, which would indicate that the rankings from the model containing only initial score as a predictor were different from those including other predictors. The same comparisons were made within the second and third columns of Table 3 with the same expected results. That is, in the 2-level gain model, the rankings from the model including no predictors were compared to the rankings from the remaining six models listed in column 2 of Table 3 using Spearman correlations. Similarly, in the 3-level growth model, the rankings from the model including only initial status was compared to the rankings from the remaining six 3-level growth models listed in column 3 of Table 3 using Spearman correlations.
While it was interesting to determine whether adding predictors changed the effectiveness rankings, the main purpose of this study was to determine whether the effectiveness rankings differed among the three models being examined and also to determine which model produced the most valid effectiveness rankings. Theoretically, effectiveness rankings should not be determined from models including predictors (Willms, 1999); therefore, the effectiveness rankings from each of the three different models including either no predictors or only initial score were compared and analyzed to determine which model produced the most valid rankings. These models are listed in the first row of Table 3. To determine whether the effectiveness rankings from these three models differed, Spearman correlations were obtained. It was anticipated that the magnitude of the correlation between the 2-level gain model and the 3-level model would be higher than the correlations between the 2-level model and the 2-level gain model and between the 2-level model and the 3-level model. This is because the slopes, and thus the effectiveness rankings, from the 2-level model are contingent on the initial score, which is not true for the other 2 types of models.

After determining whether or not the effectiveness rankings from the three models differed, it was still to be shown which model produced the most valid set of effectiveness rankings. The rankings were validated using the change in math focus score previously described. While math focus clearly is not the sole determining factor in whether a school is effective or not in mathematics, it would not be possible for a school to be effective in mathematics without focusing on increasing student achievement in math. In fact, the math focus variable was found to be statistically significantly related to achievement in mathematics after controlling for initial scores (MMP Year 4 Evaluation Report, 2007). Here, effectiveness was operationally defined not as achievement in mathematics but as growth in scale scores from one
year to the next on the standardized test. Therefore, theoretically, if a school increased its focus on increasing student achievement in math, scale scores in math should have also increased. Thus, schools with large increases in Math Focus should have also displayed large increases in scale scores and thus have been deemed effective. Therefore, it was expected that schools deemed to be effective would have larger increases in Math Focus than schools deemed to be less effective. Thus, the effectiveness estimates from each of the three models were validated by obtaining the Pearson correlation between the effectiveness estimates and the gain in Math Focus from each of the schools. Here the most valid effectiveness estimates were the ones that were most highly correlated with the gain in Math Focus.

Results

The first results which were obtained in the current study were correlations among the predictors which were included in the various models. It was found that the phi coefficient between the student level race and student level socioeconomic status variables was $.28 (df = 7232, p < .01). While this result is clearly statistically significant, the interpretation of a phi coefficient of this magnitude would be that there is little or no association between these variables. Therefore, collinearity should not be a problem in all further analyses using these variables. The Pearson correlation between the school level race and school level socioeconomic variables was found to be $.61 (df = 126, p < .01). This is a relatively strong correlation, so caution must be taken in interpreting the results from the model that includes both the school level race and school level socioeconomic variables as collinearity may be a problem. That is, both variables may appear to be non-significant predictors when both are in the model because they are taking up the same portion of the variance, when individually each variable may significantly predict the outcome variable being examined.
Next, the significance of the predictors in each model was compared across the three model types. Table 4 displays the t-values, degrees of freedom, and p-values for each predictor in each model. It should be noted that the 3-level individual growth model was the only model which provided estimates for all possible variable combinations. Predictors in the first level of the 2-level model predicted the outcome in the model, which is the final score. Thus, these variables are thought of as predictors of the achievement status of a school. On the other hand, predictors in the first level of the 2-level gain model predicted the gain score. Thus, these variables are thought of as predictors of the growth of a school. Because the first level of the 3-level model models the repeated observations and thus contains an intercept and a slope, the second level of the 3-level model is the student level (comparable to the first level in the 2-level models). However, as the first level of the 3-level model contains both an intercept and a slope, the 3-level model contains 2 equations in the second level, one for the intercept and one for the slope, thus allowing for simultaneous prediction of both the achievement and the growth which is unavailable from either of the 2-level models.

For example, model 5 contained both student level and school level race as predictors. For the 2-level model, the level 1 equation was \(Y = \beta_0 + \beta_1(\text{Initial Score}) + \beta_2(\text{Race}) + \text{error}\), where \(Y\) was the Final Score. Thus there were three equations in level 2 of this model: one to predict \(\beta_0\) (achievement), one for \(\beta_1\) (growth), and one for \(\beta_2\) (race), each of which predicted the final score. School level race was then included in each of these three level-2 equations. Results in Table 4 show that in the 2-level model, student level race was a statistically significant predictor of achievement within schools \((t = -2.33, p = .02)\). In addition, school level race was a statistically significant predictor of achievement between schools \((t = -9.33, p < .01)\), but not of growth between schools \((t = -0.58, p = .56)\). Finally, school level race was not a statistically significant
Table 4. Predictor significance across model types.

<table>
<thead>
<tr>
<th>Model</th>
<th>Predictor</th>
<th>Predicting What</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>2-Level Model</td>
<td>t</td>
<td>df</td>
<td>p</td>
<td>t</td>
<td>df</td>
<td>p</td>
<td>t</td>
<td>df</td>
<td>p</td>
</tr>
<tr>
<td>1</td>
<td>Slope</td>
<td>Achievement</td>
<td>41.64</td>
<td>127</td>
<td>.00</td>
<td>29.60</td>
<td>127</td>
<td>.00</td>
<td>29.40</td>
<td>127</td>
<td>.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Growth</td>
<td>-3.60</td>
<td>127</td>
<td>.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Race-Student</td>
<td>Achievement</td>
<td>-3.04</td>
<td>127</td>
<td>.00</td>
<td>1.30</td>
<td>127</td>
<td>.20</td>
<td>0.14</td>
<td>127</td>
<td>.89</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Growth</td>
<td>-3.12</td>
<td>127</td>
<td>.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>SES-Student</td>
<td>Achievement</td>
<td>-3.04</td>
<td>127</td>
<td>.00</td>
<td>1.30</td>
<td>127</td>
<td>.20</td>
<td>0.14</td>
<td>127</td>
<td>.89</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Growth</td>
<td>-3.12</td>
<td>127</td>
<td>.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Race-Student</td>
<td>Achievement</td>
<td>-2.64</td>
<td>127</td>
<td>.01</td>
<td>1.22</td>
<td>127</td>
<td>.23</td>
<td>0.12</td>
<td>127</td>
<td>.90</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Growth</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Race-Student</td>
<td>Achievement</td>
<td>-2.30</td>
<td>126</td>
<td>.02</td>
<td>-0.19</td>
<td>126</td>
<td>.85</td>
<td>-0.35</td>
<td>126</td>
<td>.72</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Growth</td>
<td>-9.33</td>
<td>126</td>
<td>.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Race-School</td>
<td>Achievement</td>
<td>-0.58</td>
<td>126</td>
<td>.56</td>
<td>0.52</td>
<td>126</td>
<td>.61</td>
<td>0.52</td>
<td>126</td>
<td>.61</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Growth</td>
<td>-0.58</td>
<td>126</td>
<td>.56</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>SES-Student</td>
<td>Achievement</td>
<td>-2.43</td>
<td>126</td>
<td>.02</td>
<td>1.24</td>
<td>126</td>
<td>.22</td>
<td>0.49</td>
<td>126</td>
<td>.63</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Growth</td>
<td>-4.03</td>
<td>126</td>
<td>.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>SES-School</td>
<td>Achievement</td>
<td>-0.41</td>
<td>126</td>
<td>.69</td>
<td>0.60</td>
<td>126</td>
<td>.55</td>
<td>0.40</td>
<td>126</td>
<td>.69</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Growth</td>
<td>-0.41</td>
<td>126</td>
<td>.69</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Student Race</td>
<td>Achievement</td>
<td>0.88</td>
<td>126</td>
<td>.38</td>
<td>0.51</td>
<td>125</td>
<td>.61</td>
<td>0.97</td>
<td>126</td>
<td>.34</td>
</tr>
</tbody>
</table>
## The Impact of Including Model Type  

<table>
<thead>
<tr>
<th>Model</th>
<th>Type</th>
<th>Achievement</th>
<th>Growth</th>
<th>Student Race (Achievement)</th>
<th>Student SES (Achievement)</th>
<th>Student Race (Growth)</th>
<th>Student SES (Growth)</th>
<th>t</th>
<th>df</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>Race-Student</td>
<td>-2.14</td>
<td>125</td>
<td>.03</td>
<td>-0.36</td>
<td>125</td>
<td>.72</td>
<td>-6.42</td>
<td>125</td>
<td>.00</td>
</tr>
<tr>
<td></td>
<td>SES-Student</td>
<td>-2.12</td>
<td>125</td>
<td>.04</td>
<td>1.20</td>
<td>125</td>
<td>.23</td>
<td>-6.31</td>
<td>125</td>
<td>.00</td>
</tr>
<tr>
<td></td>
<td>Race-School</td>
<td>-6.96</td>
<td>125</td>
<td>.00</td>
<td>-1.37</td>
<td>125</td>
<td>.17</td>
<td>-5.42</td>
<td>125</td>
<td>.00</td>
</tr>
<tr>
<td></td>
<td>Student Race (Achievement)</td>
<td>-0.53</td>
<td>125</td>
<td>.60</td>
<td>0.19</td>
<td>125</td>
<td>.85</td>
<td>1.65</td>
<td>125</td>
<td>.10</td>
</tr>
<tr>
<td></td>
<td>Student SES (Achievement)</td>
<td>-0.39</td>
<td>125</td>
<td>.70</td>
<td>0.19</td>
<td>125</td>
<td>.85</td>
<td>0.37</td>
<td>125</td>
<td>.71</td>
</tr>
<tr>
<td></td>
<td>Growth</td>
<td>-0.07</td>
<td>125</td>
<td>.95</td>
<td>-0.07</td>
<td>125</td>
<td>.95</td>
<td>-0.44</td>
<td>125</td>
<td>.65</td>
</tr>
<tr>
<td></td>
<td>Student Race (Growth)</td>
<td>-0.47</td>
<td>125</td>
<td>.64</td>
<td>-0.47</td>
<td>125</td>
<td>.64</td>
<td>-1.03</td>
<td>125</td>
<td>.31</td>
</tr>
<tr>
<td></td>
<td>Student SES (Growth)</td>
<td>-0.77</td>
<td>125</td>
<td>.44</td>
<td>1.46</td>
<td>125</td>
<td>.15</td>
<td>-0.72</td>
<td>125</td>
<td>.47</td>
</tr>
<tr>
<td></td>
<td>Achievement</td>
<td>0.77</td>
<td>125</td>
<td>.44</td>
<td>1.46</td>
<td>125</td>
<td>.15</td>
<td>2.39</td>
<td>125</td>
<td>.02</td>
</tr>
<tr>
<td></td>
<td>Student Race (Achievement)</td>
<td>-0.01</td>
<td>125</td>
<td>.99</td>
<td>0.37</td>
<td>125</td>
<td>.72</td>
<td>-1.23</td>
<td>125</td>
<td>.22</td>
</tr>
<tr>
<td></td>
<td>Student SES (Achievement)</td>
<td>-0.88</td>
<td>125</td>
<td>.38</td>
<td>-0.88</td>
<td>125</td>
<td>.38</td>
<td>0.08</td>
<td>125</td>
<td>.94</td>
</tr>
<tr>
<td></td>
<td>Growth</td>
<td>1.07</td>
<td>125</td>
<td>.29</td>
<td>1.07</td>
<td>125</td>
<td>.29</td>
<td>1.58</td>
<td>125</td>
<td>.12</td>
</tr>
<tr>
<td></td>
<td>Student Race (Growth)</td>
<td>0.77</td>
<td>125</td>
<td>.44</td>
<td>1.46</td>
<td>125</td>
<td>.15</td>
<td>-0.37</td>
<td>125</td>
<td>.71</td>
</tr>
<tr>
<td></td>
<td>Student SES (Growth)</td>
<td>0.77</td>
<td>125</td>
<td>.44</td>
<td>1.46</td>
<td>125</td>
<td>.15</td>
<td>1.58</td>
<td>125</td>
<td>.12</td>
</tr>
</tbody>
</table>
The Impact of Including 21 predictor of the student level race as it predicted achievement ($t = -0.30, p = .76$). The level-1 equation for the 2-level gain model was similar to that of the 2-level model except that initial status was not included as a predictor. In addition, the Y variable was the gain score instead of the final score. Since this was the case, nothing in the 2-level gain model predicted the achievement level, only the growth. Results for this model showed that student level race was not a statistically significant predictor of growth within schools ($t = 0.52, p = .61$). In addition, school level race was not a statistically significant predictor of either growth ($t = -0.19, p = .85$) or student level race as it predicted growth between schools ($t = -0.72, p = .48$). Level 2 (the student level) of the 3-level model contained two equations, one for the average achievement and one for the slope (growth). These equations were $\Pi_0 = \beta_{00} + \beta_{01}\text{(minority)} + \text{error}$ and $\Pi_1 = \beta_{10} + \beta_{11}\text{(minority)} + \text{error}$, where $\Pi_0$ was the intercept in the level 1 equation and $\Pi_1$ was the slope in the level 1 equation. Thus, student level minority status predicted both the achievement level and the growth within schools. Furthermore, in the four level 3 equations, school level minority percentage predicted achievement, growth, student level minority status as it predicted achievement, and student level minority status as it predicted growth. Student level minority status was found to be a statistically significant predictor of achievement within schools ($t = -6.68, p = .00$) but not of growth within schools ($t = -0.35, p = .75$). School level minority percentage was also found to be a statistically significant predictor of achievement between schools ($t = -8.62, p = .00$), but not of growth ($t = 0.52, p = .61$), student level minority status as it predicted achievement ($t = -0.53, p = .60$), or student level minority status as it predicted growth ($t = -0.83, p = .41$).

With few exceptions, conclusions about the statistical significance of predictors were consistent across the three models. In general, student level race and student level socioeconomic
status were statistically significant predictors of the achievement level within schools, but they were not statistically significant predictors of the growth within schools. However, the only type of hierarchical linear model that gave researchers all of this information within one modeling schema was the 3-level model. That is, the 2-level model provided information on achievement, and the 2-level gain model provided information on growth, but the only model that provided information on both achievement and growth was the 3-level model. School level race and school level socioeconomic status were also found to be statistically significant predictors of the achievement level of schools, but they were not statistically significant predictors of the growth of schools. As expected, when both school level race and school level socioeconomic status were included in the model, one of them (socioeconomic status) was no longer statistically significant, although it was shown to be a statistically significant predictor of the intercept when school level race was not included. This finding is probably due to the collinearity between school level race and school level socioeconomic status.

The next goal of the current study was to determine whether including predictors in each type of model changed the effectiveness rankings of the schools produced by each model. To test this research question, Spearman correlations were obtained between the models listed within each column of Table 3. It was expected that including predictors would change the effectiveness rankings so that the magnitude of the Spearman correlations would be low. However, Table 5, which shows the correlations between the first model listed under each model type in Table 3 and the remaining models within each column, indicates that this is not the case. As indicated by the Spearman correlations, including predictors in the models did not change the effectiveness rankings for any of the models. This was contrary to the expected results; however, a possible explanation is explored in the discussion section.
It was also stated that correlations among effectiveness rankings from each type of null model (that is, the models across the first row in Table 3) would be examined. It was expected that the rankings from the 2-level gain and 3-level models would be similar since both of these models predicted growth, however, that these rankings would be different from those obtained from the 2-level model which predicted achievement rather than growth. Spearman correlations were obtained between the models listed in the first row of Table 3. It was found that the Spearman correlation between the 2-level and 2-level gain models was .09, between the 2-level and 3-level models was .10, and between the 2-level gain and 3-level models was .99. Only the third of these correlations was statistically significant ($p < .01$). These results confirmed the expectation that the 2-level gain and 3-level models would display similar effectiveness rankings, and that these rankings would differ from the rankings obtained in the 2-level model.

Recall that the effectiveness rankings are based on residuals of the slope in the model, thus reflecting growth. Intercept residuals from the 2-level and 3-level models can also be correlated to determine whether rankings of achievement level are the same for these two models. Since the 2-level gain model does not predict the achievement level, there is no equivalent achievement level residual from that model. It was found that the Spearman
correlation between intercept residuals for the 2-level and 3-level models shown in the first row of Table 3 was .77, which is statistically significant with a p-value less than .01. These results make it clear that examining the slope residuals from the 2-level gain and 3-level models is an examination of the effectiveness of schools, while examining the intercept residuals from the 2-level and 3-level models is an examination of the achievement of schools.

To validate the effectiveness ratings from the 3 models, Pearson correlations were obtained between the effectiveness estimates (that is, the school level slope residuals) for each of the 3 models and the gain in math focus for each school. It was found that the correlation between the gain in math focus and the effectiveness estimates from the 2-level model was .03 (p = .75). It was also found that the correlations between the gain in math focus and the effectiveness estimates from the 2-level gain and 3-level models were .20 (p = .06) and .18 (p = .09), respectively. While these correlations were low, and none of them reached statistical significance, this did serve to validate the effectiveness ratings from each of the models.

Specifically, the effectiveness ratings from the 2-level gain and 3-level models were found to be statistically significantly correlated with one another, and the ratings from the 2-level gain and 3-level models displayed correlations of a higher magnitude with the gain in math focus than the rankings from the 2-level model. Therefore, it was concluded that the ratings from the 2-level gain and 3-level models were more valid than the ratings from the 2-level model.

Discussion

Two of the results of the analyses were unexpected and deserve further explanation. First, it was expected that the effectiveness rankings would change when predictors were added in each of the 3 types of models. However, the rankings were quite consistent, even when race and socioeconomic status were added as variables at both the student and school levels. It was found
that student level race and socioeconomic status were statistically significant predictors, but they only statistically significantly predicted the intercept, not the slope. The effectiveness rankings were based on the slope residuals, for which the predictors were not statistically significant. This may be why the effectiveness rankings did not change when predictors were added, because the predictors that were added were not statistically significant predictors of the slope. It is possible that if rankings were made on the intercept residuals instead, their rankings would change when student level race and socioeconomic status predictors were added because these variables were statistically significant predictors of the intercept. This was tested by finding the correlations between the intercept residuals from the 3-level model with only the intercept as a predictor and from the other models in column 3 of Table 3. These correlations are shown in Table 6.

Table 6. Spearman correlations of the intercept residuals from the null model and models including predictors.

<table>
<thead>
<tr>
<th>Student Level Race</th>
<th>1.00</th>
</tr>
</thead>
<tbody>
<tr>
<td>Student Level SES</td>
<td>1.00</td>
</tr>
<tr>
<td>Student Level Race &amp; SES</td>
<td>1.00</td>
</tr>
<tr>
<td>Student &amp; School Level Race</td>
<td>0.80</td>
</tr>
<tr>
<td>Student &amp; School Level SES</td>
<td>0.91</td>
</tr>
<tr>
<td>Student &amp; School Level Race &amp; SES</td>
<td>0.80</td>
</tr>
</tbody>
</table>

Since the magnitude of the correlations, particularly from the models which contained student level and school level predictors simultaneously, were lower than those shown in Table 5 it does appear that including statistically significant predictors changed the intercept residuals. Therefore, if more or different predictors, particularly statistically significant predictors, were included at the student and/or school levels than were included here, the intercept and/or slope residuals of those models may differ significantly from models which include no such predictors.
The Pearson correlations indicated that the math focus gain and effectiveness estimates were more highly correlated for the 2-level gain and 3-level models than for the 2-level model, demonstrating that these models gave more valid effectiveness estimates than the 2-level model. It should be noted that the correlations were quite low (only .20 and .18, respectively); however, this was expected as gain in math focus is only one component of school effectiveness.

As stated, the effectiveness estimates from the 2-level gain and 3-level models were the most valid, as compared to the estimates from the 2-level model, and the estimates from these two models were quite similar. However, there are several reasons to prefer the 3-level model over the 2-level gain model. First, the 3-level model includes all available data, even students who are missing data, which leads to more powerful analyses. The 2-level gain model can only include students who were tested in both years; otherwise not enough data is available to create the gain score. In this study, 7,206 students were included in the analyses for the 3-level model, while only 5,092 students were included in the analyses for the 2-level and 2-level gain models. Thus, roughly 30% of the students in the complete data set for this study were missing test scores from either 2005 or 2006. It was found that the rankings from the 2-level gain and 3-level models were statistically significantly correlated; however, this may not be the case if the overall sample was smaller and/or a larger proportion of students were missing data. With smaller sample sizes and more missing data, the 3-level model may be more powerful than the 2-level gain model.

Second, the 3-level model can include data from multiple time points in the same analysis. That is, data from two to an almost infinite number of time points can be included in the analysis. Students do not necessarily need to be tested at the same time either, because the “time” variable that is included in the first level of the model is continuous. For example, one student may be tested at time 1 which could be October of the first school year while another student
The Impact of Including 27

may not be tested until March of the first school year, which would be time 1.5. That is, the time variable is a continuous variable and data from students who tested at any time may be included. On the other hand, gain scores should be taken from tests which are evenly spaced for all students. In addition, gain scores can only be created from two tests. To include three years of test data it is unclear how gain scores would be created. Taking the simple difference between the third year and the first year would cause a significant loss in data, while including all three years would necessitate creating multiple gain scores to be used as the dependent variable.

Third, as demonstrated, results from the 3-level model provide more information than results from the 2-level and 2-level gain models. That is, variables can be included as predictors of both the slope and the intercept at the student and school levels. Therefore, information on whether the variables predict the initial score or the growth can be obtained from the same model. On the other hand, the 2-level model can only assess the variables as they predict the intercept, while the 2-level gain score model can only assess the variables as they predict the slope. Therefore, the 3-level model is preferred over the 2-level gain model, except in studies in which researchers are only including two time points and they are not interested in examining the significance of predictors of the intercept. If only two time points are being used and predictors of the intercept are unnecessary, the 2-level gain model may be preferred because of its simplicity as compared to the 3-level model.

The results from this study have contributed to knowledge about the differences between these three types of models in determining the significance of predictors, and more importantly the effectiveness of schools. The current study found that student and school level race and socioeconomic status variables were statistically significant predictors of the achievement level, but not the growth, of students and schools in each of the models. Furthermore, the inclusion of
predictors did not change the effectiveness estimates in any of the three model types. Finally, the 2-level gain and 3-level models gave the most valid effectiveness ratings as shown by the correlation between the estimates and the gain in math focus. However, the 3-level model is generally more powerful and flexible than the 2-level gain model and is therefore preferred over the other two models except in the very limited circumstances that were previously mentioned when looking at school effectiveness and the statistical significance of predictors in hierarchical linear models.
The Impact of Including 29

References


Doran, H. C. (2004). *Value-added models and Adequate Yearly Progress: Combining growth...*
The Impact of Including 
and adequacy in a standards-based environment. Paper presented at the Annual CCSSO 
Large-Scale Assessment Conference, Boston, MA, June 14, 2004.

School Improvement, 8 (4), 369-395.

Gong, B. (2002). Designing school accountability systems: Towards a framework and process. 
Accountability Systems and Reporting: Council of Chief State School Officers.

students. HEPC Policy Brief, October, 2006.


achievement. Paper prepared for the Association of Public Policy Analysis and 

Tennessee Value Added Assessment System. Educational Evaluation and Policy 

Value-Added Modeling of Teacher Effects. Journal of Educational and Behavioral 


