Using Hierarchical Growth Models
to Monitor School Performance Over Time:
Comparing NCE to Scale Score Results

CSE Report 618

Pete Goldschmidt, Kilchan Choi, and Felipe Martinez
CRESST/University of California, Los Angeles

January 2004

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

Project 3.1 Methodologies for Assessing Student Progress, Strand 2
Pete Goldschmidt, Project Director, CRESST/UCLA

Copyright © 2004 The Regents of the University of California

The work reported herein was supported under the Educational Research and Development Centers Program, PR/Award Number R305B60002, as administered by the Institute of Education Sciences, U.S. Department of Education.

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

USING HIERARCHICAL GROWTH MODELS TO MONITOR SCHOOL PERFORMANCE OVER TIME: COMPARING NCE TO SCALE SCORE RESULTS

Pete Goldschmidt, Kilchan Choi, and Felipe Martinez

CRESST/University of California, Los Angeles

Abstract

Monitoring school performance increasingly uses sophisticated analytical techniques and we investigate whether one such method, hierarchical growth modeling, yields consistent school performance results when different metrics are used as the outcome variable. Specifically we examine whether statistical and substantive inferences are altered when using normal curve equivalents (NCEs) vs. scale scores. The results indicate that the effect of the metric depends upon the evaluation objective. NCEs significantly underestimate absolute growth, but NCEs and scale scores yield highly correlated (.9) results based on mean initial status and growth estimates. Correlations between NCE and scale score rankings, based on fitted school initial status and growth values are generally over .94. Further, statistical and substantive results, using NCEs and scale scores, pertaining to school-wide program effects are highly correlated (.95) as well. NCEs and scale scores matched 99% of the time on whether or not the program indicator variable was statistically significant.

The use of longitudinal analysis of educational data encompasses a wide array of applications from individual growth curve modeling to program evaluation and school performance modeling. We focus on the latter of these uses, especially as they pertain to the general issue of school accountability. These models are often constrained by data availability, in terms of the number of time points available, the completeness of data, and the metric available for analysis. As school districts become more sophisticated in record keeping ability, it is now more common and less problematic to acquire multiple years of data on students, who are linked with a unique identifier across years. Recent advances in multilevel modeling techniques allow for the use of unbalanced data, make unequal timing of outcomes, and missing data less of a barrier to analysis than repeated measures ANOVA once presented (Hox, 2002; Raudenbush & Bryk, 2002). Given the recent trend towards accountability that is based on outcomes (Hanushek & Raymond, 2001), it is relevant to consider whether the metric used for the longitudinal analysis affects inferences. As Seltzer, Frank, and Bryk (1994) demonstrated, the metric matters when attempting to model individual growth trajectories. Seltzer et al. concluded that scale scores are the most appropriate metric for growth curve modeling. Theoretically, Item Response Theory (IRT)-based scale scores represent content mastery on a continuum and may be used to measure absolute academic progress over time; whereas, NCEs represent a relative standing compared to a norming population. Because NCEs represent a relative position from year to year, and not a change from the previous year, it cannot present a complete picture of student academic progress. Two issues give rise to this analysis: one, schools or school districts are often unable to provide scale scores and are, at best, able to provide normal curve equivalents (NCEs).[1] Two, school districts periodically change tests (publishers), but desire to conduct longitudinal evaluations across these different tests. Test publishers use different methods to scale their tests, and rarely is information available equating scores from one test to another; NCEs may be more comparable because they are relative scores. Within the longitudinal analysis framework, two questions arise: one, to what extent the metric matters when the focus is school performance over time; and two, to what extent the metric matters when the focus is program evaluation.

While the properties of both scale scores and NCEs are understood, the empirical evidence against the use of NCEs for longitudinal models utilizing individual student data to examine school performance over time is lacking. Hence, this analysis incorporates Monte Carlo methods and utilizes a three-level hierarchical model to compare the results of growth models using scale scores against growth models using NCEs on the same set of students. In general, the goal is to apply a three-level hierarchical linear model (HLM) with test scores nested within students at level one, students nested within schools at level two, and schools at level three, to determine the effect of the metric on school-level estimates of initial status, growth, and a school program participation indicator variable.

Background

Methodologically advanced methods of examining school performance or evaluating program effectiveness take both the nature of the data into consideration,

[1] Often districts report (understand) national percentile ranks (NPR), which can easily be converted to NCEs.

and attempt to mitigate the effects of potential confounding factors (PCF). The short-comings of simply examining school means (Aitkin & Longford, 1986), ignoring the nested nature of educational data (Raudenbush & Bryk, 2002), or the differing meaning of variables at different levels of aggregation (Burstein, 1980) are generally taken into account in recent sound efforts to discern school effectiveness. The effects of PCFs, in non-randomized, cross-sectional designs (Campbell & Stanley, 1963) and limitations of pre-post designs (Bryk & Wesiburg, 1977; Raudenbush, 2001; Raudenbush & Bryk, 1987) in making inferences about school or program effects (i.e., change in student outcomes due to a hypothesized cause) leads us to consider the advantages of examining growth trajectories to make inferences about change (Raudenbush & Bryk, 2002; Rogosa, Brandt, & Zimowski, 1982; Willet, Singer, & Martin, 1998). The usefulness of hierarchical longitudinal growth models that examine individual growth trajectories and make subsequent inferences about program effectiveness have been posited and applied for some time (Heck, 2000; Willms & Raudenbush, 1989); their use increasing with both computing power and data availability (see for example Goldschmidt, 2002b; Heck, 2000; Ramirez, Yuen, Ramey, & Pasta, 1991). Specifying an adequate model that correctly captures the structure of growth depends upon the nature of the data being modeled (Raudenbush, 2001).

Theoretical understanding of the optimum model is often confronted with the empirical realism of the data, however. The advantages of longitudinal growth models may be a mute point if proper data cannot be compiled. Assuming that student records have been accurately matched, and that there are at least three measurement occasions, one can consider a longitudinal[2] evaluation model. These models have examined diverse outcomes such as growth in mathematics grade equivalents (GE), (Wilms & Jacobson, 1990) early child vocabulary growth Huttenlocher, Haight, Bryk, and Seltzer (1991), early reading achievement growth (Goldschmidt, 2002a), and feelings of isolation (Osgood & Smith, 1995). While these models are often referred to as longitudinal growth models, they often, as in the case of Osgood and Smith, for example, evaluate an outcome that is intended to change, though not necessarily grow.

The scale is important in drawing conclusions from individual growth curves (Yen, 1986). The key element is that the outcome must have constant meaning over

[2] Which we differentiate from a pre-post, or simple gain score model.

time (Raudenbush, 2001). As Seltzer et al. (1994) demonstrated, the metric matters when looking at student progress over time. Seltzer et al. focused on the inadequacy of grade equivalents (GEs). Theoretically, the optimal metric to use when examining change is a vertically equated IRT based scale score because it is on an interval scale and is comparable across grades (Hambleton & Swaminathan, 1987). Thus a change in a scale score from year to year is an absolute measure of academic progress, irrespective of grade. However, equating is generally designed to compare contiguous grade pairs (Yen, 1986) and scales may be less meaningful as the grade span increases. The NCE are based on national percentile ranks, but are standardized to have a mean of 50 and a standard deviation of 21.06, and also have an equal interval scale (Worthen, White, Fan, & Sudweeks, 1999). As noted, NCEs cannot adequately describe absolute achievement growth because it places students at a relative position at each test occasion. A change in this relative position does not necessarily functionally correspond to absolute achievement gains. One argument against the use of NCEs is that it places students on a relative position, which guarantees "winners" and "losers." But, it should be noted that both NCEs and IRT based scale scores are based on a representative norming sample that is generally conducted approximately every five to eight years. This implies that all scores are potentially based on a fixed standard for some period of time.

The focus of this discussion has centered on the correct metric and ensuing interpretation of individual student growth trajectories, but not on whether results pertaining to school performance would be either statistically or substantively altered by using NCE scores rather than scale scores. The questions that arise are whether the results using NCEs or scale scores would substantively change the inferences made about school performance or program effectiveness, within the context of a hierarchical longitudinal evaluation. This is relevant in that school accountability systems compare school performance, but accurate models do not simply evaluate school aggregate performance because that suffers biases brought about due to aggregation (Aitkin & Longford, 1986), and incorrectly mixes inferences concerning student and school level variables (Burstein, 1980). Further, program evaluation may not simply consist of comparing two groups of students—treatment vs. control—rather it may consist of evaluating the context in which students are placed (i.e., individual students are not assigned to a group, rather the entire school may be part of a reform effort). As in general school performance modeling, the question of interest is to what extent the program school performs better than non-program school, with the same notion of value added as in accountability models. Given that NCE scores are available for analysis substantially more often than scale scores, it is important to compare school level results using both NCEs and scale scores.

Data

The data are from a large, racially integrated, urban school district that enrolls approximately 65,000 students. We have four years of panel data beginning with the 1997-98 (1998) school year and ending with the 2000-2001 (2001) school year. The outcome measures are reading and mathematics Stanford Achievement Test, v9, (SAT-9) scores, for which we have both NCEs and scale scores. In order to focus this analysis on the effect of the metric and not confound school-effect results with issues of cross-classification (Rashbash & Goldstein, 1994; Raudenbush, 1993) we limit our sample to students who attended the same school between 1998 and 2001. We further reduce our sample by excluding students with missing demographic or outcome information (although these are certainly not requirements for longitudinal analyses). Table 1 presents the SAT-9 reading and mathematics means for each of the years that we have data. Although not inherently comparable, both metrics demonstrate an increase between 1998 and 2001 of between 12% and 15%. Simple zero-order correlations between students' scores measured by NCEs and scale scores indicate that they are only moderately correlated in each year (range r = .59 to .68).

Table 1

Means of NCEs and Scale Scores by Year

	Reading		Mathematics	
Year	NCE	SS	NCE	SS
1998	38.5	561.2	43.3	560.8
	(20.1)	(40.7)	(22.3)	(41.0)
1999	40.8	595.1	49.2	596.7
	(20.5)	(44.4)	(22.1)	(43.8)
2000	41.2	619.7	47.0	617.6
	(20.6)	(42.1)	(21.3)	(41.3)
2001	43.2	640.6	48.7	644.3
	(19.7)	(37.8)	(21.9)	(40.6)

Note. N = 7,856 students; standard deviation in parentheses.

The final sample that we utilize consists of 7,856 students, representing 31,424 test scores for each content area. Student demographic characteristics are presented in Table 2. The sample that forms the basis for the simulation study (forming the population from which we draw samples) matches fairly closely the district as a whole, but more importantly is a substantially diverse sample, intimating that results are not biased by an unrealistically homogeneous sample.

We utilize school context measures that are constructed from the aggregate student characteristics for each school. We also use an indicator variable that identifies whether schools participated in a school-wide school reform effort (program schools). Fifteen percent of schools are classified as program schools.

Methodology

We employ a Monte Carlo method to examine the consequences of using NCE scores instead of scaled scores when using a three level hierarchical growth model in monitoring school performance. Comparing the key parameters of interest allows us to examine whether the results are statistically or substantively similar. Monte Carlo techniques allow us to simulate key factors of interest so that we can obtain more general information about the differences between the results pertaining to each test metric, when those measures are used for monitoring school performance over time.

Table 2

Student Characteristics

	Proportion	
Student characteristic	District	Sample
Female	0.50	0.49
African-American	0.21	0.21
Asian	0.17	0.13
Hispanic	0.41	0.45
Other	0.03	0.03
White	0.19	0.18
ELL	0.37	0.49
Free/reduced lunch	0.67	0.88
Special ed.	0.07	0.08

Note. N = 7,856 students

In general, the Monte Carlo method involves randomly generating data under specific conditions. However, in this study, we consider the data described above as the population from which we will sample repeatedly for a Monte Carlo study. Each school has an average sample size of 745 students. Given that the number of students per school is extremely large, we can comfortably say that repeated samples hold the i.i.d assumption. The main factors of the Monte Carlo study are the following: the number of students within schools, the number of schools, grade levels, and content areas. Table 3 summarizes the conditions. In each of the sampling conditions, we sample the population data 2,000 times. We use SAS™ to sample and generate the datasets and HLM™ to run the hierarchical growth models.

Depending upon content area, the choice of NCE or scale scores might lead to different consequences in terms of inferences concerning school performance. For example, in a content area in which students show remarkable progress over time, scale scores may be a better metric in detecting progress than NCE scores, because NCE scores might wash out the magnitude of absolute growth over time.

In this study, we focus on the four key parameters estimated in a three-level hierarchical model (described below) for measuring school performance over time:

1. school mean initial status for non-program schools;
2. the difference between non-program schools and program schools in initial status;

Table 3

Sampling Conditions for Monte Carlo Study

Total number	Students sampled	
of schools	Percent	Mean *n*
60	25%	31.3
60	50%	65.6
60	75%	98.5
60	100%	130.9

3. the school mean rate of change for non-program schools; and,
4. the difference between non-program and program schools in the mean rate of change.

For each simulation condition, we estimate Pearson correlations between each of the parameters based on NCE scores and the corresponding parameters based on scale scores (see Table 4). In addition, school rankings based on the magnitudes of school mean initial status and school mean rate of change[3] are calculated for the cases of NCE and scale scores. We compare Spearman and Kendall's Tau rank-order correlations based on the estimated school rankings between the two metrics (see Table 5).

Briefly, the three-level model is as follows:

$$Y_{tij} = \pi_{0ij} + \pi_{1ij}\_{tij} + e_{tij}, \qquad (1)$$

where Y_{tij} is the outcome at time t for student i in school j with α as a time parameter measured in school years. Since growth trajectories are assumed to vary across students, at level 2 for the initial status at time = 0:

$$\pi_{0ij} = \_{00j} + \_{01j}X_{1ij} + \ldots + \_{0Pj}X_{Pij} + r_{0ij}, \qquad (2)$$

where there are p = 1 to P student-level predictors (e.g. student background characteristics). Similarly, for the growth trajectories

$$\pi_{1ij} = \_{10j} + \_{11j}X_{1ij} + \ldots + \_{1Pj}X_{Pij} + r_{1ij}, \qquad (3)$$

The effects of student characteristics (X_{ij}'s) are assumed to vary across schools at level 3. For example, for mean initial status for school j:

$$\_{00j} = \_{000} + \_{001}Z_{1j} + u_{00j}, \qquad (5)$$

where Z is a 1/0 indicator variable denoting school program membership. For student-level effects on initial status for school j:

$$\_{0pj} = \_{0p0} + \_{0p1}Z_{1j} + u_{0pj}. \qquad (6)$$

For the mean rate of change for school j:

$$\_{10j} = \_{100} + \_{101}Z_{1j} + u_{10j}, \qquad (7)$$

and for the student level effects on the student rate of change for school j:

[3] That is based on the fitted values for school mean initial status and rate of change.

$$\text{—}_{1pj} = \text{—}_{1p0} + \text{—}_{1pq} Z_{1j} + u_{1pj}, \qquad (8)$$

Within each condition we run four models. Model one is the unconditional model used to partition the unconditional variation in the outcome among time, students, and schools, and is used as a basis for comparing the fit of subsequent more complex models (Raudenbush & Bryk, 2002). Model two includes only student level covariates, as these are potentially associated with changes in achievement and account for between school differences in enrollment; thereby potentially reducing the between school variation in the outcome. Model three includes the student level covariates at level two and the program indicator variable at level three. This model attempts to capture differences between program and non-program schools, accounting for differences in student enrollment, but excluding other school contextual factors. Model four includes all of the variables included in model three,

Table 4

Summary Parameter Estimates Compared

Question	Scale scores		NCEs
1)	—_{000}	vs.	—_{000}
2)	—_{001}	vs.	—_{001}
3)	—_{100}	vs.	—_{100}
4)	—_{101}	vs.	—_{101}

Note. NCE = normal curve equivalent.

Table 5

Fitted Values Generating School Ranks

School	Scale scores	NCEs
Initial status	b_{ooj}	b_{ooj}
Rate of change	b_{10j}	b_{10j}

Note. NCE = normal curve equivalent.

and additional school context variables to examine whether the program effect is moderated by school context. The full model (model 4) is displayed in the Appendix.

Results

The results of the simulations demonstrate that the legitimacy of using NCEs instead scale scores depends on the intended objective. The results for each of the conditions, models, and content areas are presented below. Comparisons that involve the actual estimated coefficients are recast into effect sizes, given the obvious difference in scale between NCEs and scale scores. Although not of primary interest, we briefly present the substantive results of the tested models. Tables 6 and 7 summarize the results of the full model for each of the sampling conditions. The estimated coefficients are presented in effect sizes that we define as the estimated coefficient divided by the outcome standard deviation (Cooper & Hedges, 1994).[4] Generally the magnitude of the estimated effects, by content area are consistent between NCEs and scale scores—with one major exception, growth, which we discuss in more detail below. Although we do not present detailed results concerning the parameters estimates here, it is important to note that the estimates were normally distributed with standard deviations decreasing as sample size increased. The results in Table 8 give one indication of how well the models perform, in terms of reducing the unconditional between-school variation in growth. Overall, the full model accounts for approximately 43% to 52% of the variation in the Reading growth rate and approximately 16% to 17% in the mathematics growth rate. The marginal impact of adding school context variables (after accounting for student covariates)—that is the variance reduction from model 2 to model 4—is much more tightly aligned between the NCE and scale score models; although they do differ by content area. The addition of school context has a consistent effect across sampling conditions and metric, within content area.

Except for the growth parameter estimates, within content area, the outcome metric and the sampling condition would lead to consistent interpretations of the associations of the covariates and their ability to reduce unconditional between-school variation. However, our focus is whether the results would lead to

[4]For indicator variables this is strictly correct given these are group differences. Subsequently we will utilize both the above definition for the effect size as well as another identified in Raudenbush and Feng (2001).

substantively consistent inferences about school and/or program performance, when comparing models using NCEs and scale scores.

Table 6

Summary of Results Describing SAT-9 Reading Achievement – in Effect Size

	25%		50%		75%	
SAT-9 Reading Achievement	NCE	SS	NCE	SS	NCE	SS
Mean Initial status (g_{000})						
Student Predictors						
Special Education (g_{010})	-0.47	-0.44	-0.47	-0.44	-0.47	-0.44
Low SES (g_{020})	-0.36	-0.40	-0.35	-0.40	-0.35	-0.39
LEP (g_{030})	-0.34	-0.35	-0.33	-0.34	-0.32	-0.33
Minority (g_{040})	-0.48	-0.54	-0.48	-0.54	-0.48	-0.53
Girl (g_{050})	0.10	0.10	0.10	0.10	0.10	0.10
School Predictors						
LAAMP Effect (g_{001})	0.03	0.04	0.02	0.03	0.02	0.02
Minority (g_{002})	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01
Low (g_{003})	0.13	0.10	0.17	0.15	0.20	0.17
Mean Growth (g_{100})	0.07	0.64	0.07	0.63	0.07	0.63
Student Predictors						
Special Education (g_{110})	0.00	-0.03	0.00	-0.03	0.00	-0.03
Low SES (g_{120})	0.05	0.06	0.05	0.06	0.05	0.06
LEP (g_{130})	0.07	0.07	0.07	0.07	0.07	0.07
Minority (g_{140})	-0.03	-0.02	-0.03	-0.02	-0.03	-0.02
Girl (g_{150})	0.01	0.01	0.01	0.01	0.01	0.01
School Predictors						
LAAMP Effect (g_{101})	0.01	0.01	0.01	0.01	0.01	0.01
Minority (g_{102})	0.11	0.14	0.12	0.15	0.12	0.16
Low (g_{103})	-0.08	-0.08	-0.08	-0.08	-0.08	-0.08

Note. NCE = normal curve equivalent, SS = scale scores; SES = socio-economic status, LEP = limited English proficient, LAAMP = Los Angeles Annenberg Metropolitan Project.

Table 7

Summary of Results Describing SAT-9 Mathematics Achievement – in Effect Size

	r25		r50		r75	
SAT-9 Reading Achievement	NCE	SS	NCE	SS	NCE	SS
Mean Initial status (g_{000})	59.659	593.58	60.61	595.47	61.01	596.2
Student Predictors						
Special Education (g_{010})	-0.604	-0.546	-0.516	-0.554	-0.613	-0.555
Low SES (g_{020})	-0.244	-0.243	-0.241	-0.239	-0.236	-0.235
LEP (g_{030})	0.029	0.038	0.040	0.051	0.046	0.057
Minority (g_{040})	-0.582	-0.575	-0.578	-0.569	-0.576	-0.568
Girl (g_{050})	-0.040	-0.044	-0.040	-0.045	-0.039	-0.044
School Predictors						
LAAMP Effect (g_{001})	-0.022	-0.021	-0.030	-0.033	-0.032	-0.035
Minority (g_{002})	-0.010	-0.010	-0.011	-0.012	-0.012	-0.012
Low (g_{003})	-0.006	-0.023	0.036	0.021	0.046	0.031
Mean Growth (g_{100})	0.032	0.638	0.026	0.023	0.026	0.632
Student Predictors						
Special Education (g_{110})	0.038	0.010	0.040	0.011	0.040	0.013
Low SES (g_{120})	0.021	0.014	0.021	0.023	0.021	0.014
LEP (g_{130})	0.040	0.033	0.041	0.007	0.040	0.033
Minority (g_{140})	0.000	0.001	-0.001	0.009	-0.019	-0.005
Girl (g_{150})	0.036	0.038	0.035	0.006	0.035	0.038
School Predictors						
LAAMP Effect (g_{101})	0.0252	0.0275	0.028	0.031	0.029	0.032
Minority (g_{102})	0.0009	0.0009	0.001	0.001	0.001	0.001
Low (g_{103})	-0.0367	-0.0323	-0.04	-0.037	-0.04	-0.034

Note. NCE = normal curve equivalent, SS = scale scores; SES = socio-economic status, LEP = limited English proficient, LAAMP = Los Angeles Annenberg Metropolitan Project.

Table 8

Percent Reduction in Between School Variation in Growth

Sampling	Reading		Math	
Condition	NCE	SS	NCE	SS
Model 2 to 4				
25%	24.5	23.7	9.3	9.2
50%	24.4	25.5	9.6	9.7
75%	24.5	26.4	9.2	9.3
Model 1 to 4				
25%	43.8	52.2	16.8	16.8
50%	42.7	51.9	16.4	16.5
75%	42.9	52.3	16.1	16.1

Note. NCE = normal curve equivalent, SS = scale scores.

Hence, we turn to the correlations between mean initial status and growth rates for models using scale scores vs. models using NCEs, the rank order correlations among the fitted values for mean school initial status and mean school rates of change for models using scale scores vs. models using NCEs, and the correlations for the estimated difference between program and non-program schools for models using scale scores vs. models using NCEs. Tables 9a through 9d present the correlations for each of the models for each of the sampling conditions. We present both Spearman rank-order correlations and Kendall's Tau, which are both appropriate for use with ordinal ranks (Allen & Yen, 1979). The correlations range from a low of about .75 to a high of about .98 and average about .9. The Spearman Correlations overall average is about .97 for initial status and .94 for growth. The

Table 9a

Correlations Between Estimated Coefficients – Model 1

		Correlation		Kendall (Tau) correlation	
Sample	Test type	Initial status	Growth	Initial status	Growth
R25	Read	0.988	0.936	0.925	0.806
	Math	0.987	0.963	0.925	0.863
R50	Read	0.990	0.932	0.931	0.798
	Math	0.988	0.964	0.929	0.870
R75	Read	0.991	0.932	0.935	0.798
	Math	0.989	0.964	0.932	0.871

Table 9b

Correlations Between Estimated Coefficients – Model 2

		Correlation		Kendall (Tau) correlation	
Sample	Test type	Initial status	Growth	Initial status	Growth
R25	Read	0.964	0.914	0.857	0.779
	Math	0.975	0.955	0.887	0.849
R50	Read	0.970	0.910	0.872	0.775
	Math	0.978	0.956	0.898	0.857
R75	Read	0.974	0.908	0.881	0.776
	Math	0.981	0.955	0.905	0.857

Table 9c

Correlations Between Estimated Coefficients – Model 3

		Correlation		Kendall (Tau) correlation	
Sample	Test type	Initial status	Growth	Initial status	Growth
R25	Read	0.963	0.916	0.856	0.781
	Math	0.975	0.956	0.887	0.849
R50	Read	0.971	0.912	0.873	0.777
	Math	0.978	0.958	0.897	0.858
R75	Read	0.974	0.910	0.882	0.776
	Math	0.980	0.956	0.904	0.857

Table 9d

Correlations Between Estimated Coefficients – Model 4

		Correlation		Kendall (Tau) correlation	
Sample	Test type	Initial status	Growth	Initial status	Growth
R25	Read	0.939	0.897	0.817	0.754
	Math	0.969	0.954	0.873	0.847
R50	Read	0.942	0.895	0.821	0.750
	Math	0.972	0.956	0.880	0.856
R75	Read	0.943	0.896	0.826	0.748
	Math	0.973	0.955	0.878	0.858

Kendall Tau correlations are slightly lower, with an overall average of .89 for initial status and .82 for growth. The pattern across sampling conditions and models is consistent for both correlation measures. That is, the correlations increase with sample size and decrease with model complexity. However, it should be noted that these changes are relatively small—generally about 4%, at the maximum, in either direction. Figure 1 summarizes this relationship.[5]

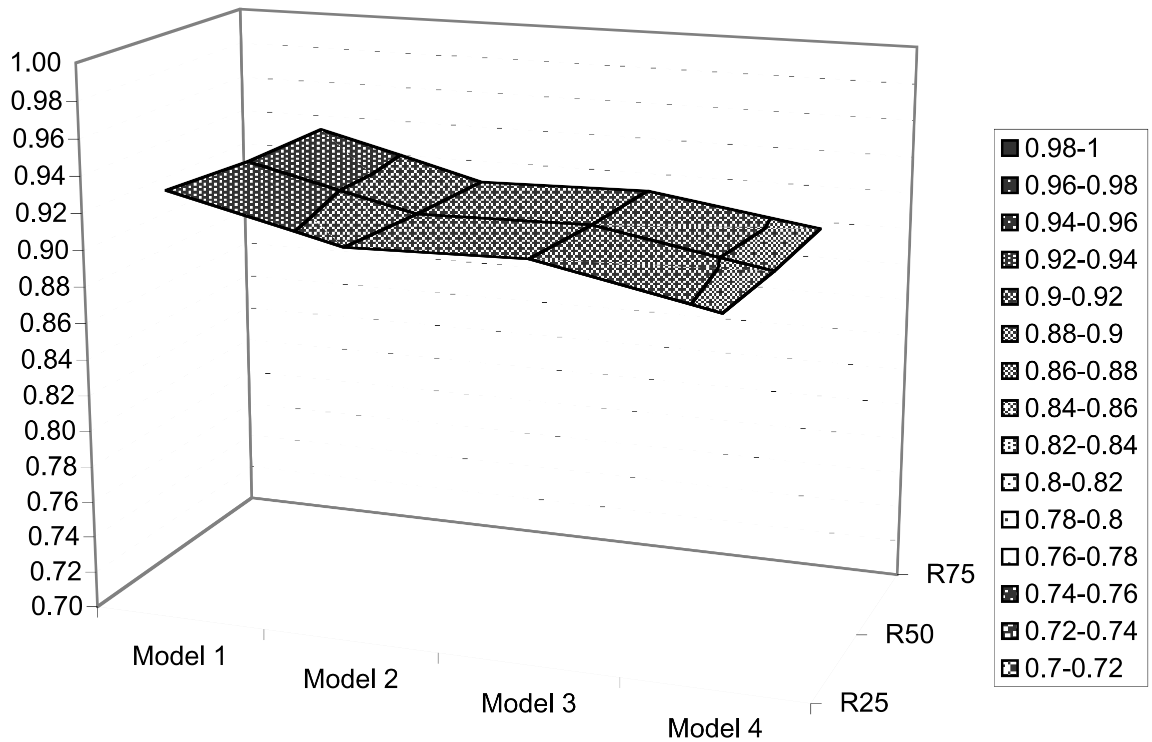

Figure 1. Correlation pattern between sampling condition and model – Reading SAT-9 growth.

We are, of course, most interested in the effects of the metric on growth as this is the parameter estimate upon which the evaluation of school performance or program effectiveness will be based. Hence, we next turn to the correlations of the fitted values and effect sizes for growth. Tables 10a through 10d summarize the results for each of the models and sampling conditions. The first column of Table 10 presents the correlation of the fitted school mean growth rates estimated using NCEs and scale scores. These range between about .89 and .96, with a mean of about .93. These correlations do not exhibit any pattern across sampling conditions or model. We generate two effect sizes to standardize annual growth.

---

[5]We present only one of the possible figures as an example. The remaining conditions are available from the authors.

Table 10a

Summary of Correlations Among Fitted Values and Effect Sizes for Growth – Model 1

Sample	Test type	Correlation		ES1		ES2	
		FV	*Raudenbush* eff. size	Ratio	Difference	Ratio	Difference
R25	Read	0.936	0.812	0.113	-0.567	0.148	-8.792
	Math	0.963	0.749	0.088	-0.596	0.095	-8.113
R50	Read	0.931	0.822	0.114	-0.568	0.147	-8.738
	Math	0.964	0.736	0.089	-0.596	0.095	-7.970
R75	Read	0.932	0.823	0.114	-0.568	0.147	-8.725
	Math	0.963	0.734	0.090	-0.596	0.096	-7.941

Note. FV = Fitted Value, ES = Effect Sizes.

Table 10b

Summary of Correlations Among Fitted Values and Effect Sizes for Growth – Model 2

Sample	Test type	Correlation		ES1		ES2	
		FV	*Raudenbush* eff. size	Ratio	Difference	Ratio	Difference
R25	Read	0.914	0.788	0.116	-0.570	0.139	-11.388
	Math	0.954	0.757	0.091	-0.595	0.098	-8.474
R50	Read	0.909	0.822	0.117	-0.570	0.139	-11.057
	Math	0.956	0.743	0.092	-0.596	0.098	-8.277
R75	Read	0.908	0.859	2.023	0.672	2.413	18.419
	Math	0.954	0.931	1.834	0.545	1.949	8.640

Note. FV = Fitted Value, ES = Effect Sizes.

Table 10c

Summary of Correlations Among Fitted Values and Effect Sizes for Growth – Model 3

Sample	Test type	Correlation		ES1		ES2	
		FV	*Raudenbush* eff. size	Ratio	Difference	Ratio	Difference
R25	Read	0.916	0.780	0.114	-0.571	0.138	-11.430
	Math	0.955	0.749	0.088	-0.595	0.094	-8.518
R50	Read	0.911	0.811	0.115	-0.571	0.137	-11.085
	Math	0.958	0.733	0.088	-0.595	0.094	-8.318
R75	Read	0.909	0.860	2.023	0.670	2.424	18.509
	Math	0.956	0.931	1.834	0.543	1.950	8.653

Note. FV = Fitted Value, ES = Effect Sizes.

Table 10d

Summary of Correlations Among Fitted Values and Effect Sizes for Growth – Model 4

Sample	Test type	Correlation		ES1		ES2	
		FV	*Raudenbush* eff. size	Ratio	Difference	Ratio	Difference
R25	Read	0.896	0.703	0.160	-0.562	0.194	-12.761
	Math	0.953	0.621	0.096	-0.593	0.103	-8.863
R50	Read	0.895	0.664	0.163	-0.563	0.193	-12.542
	Math	0.955	0.635	0.099	-0.593	0.106	-8.678
R75	Read	0.895	0.685	0.164	-0.562	0.192	-12.574
	Math	0.954	0.934	0.099	-0.593	0.105	-8.621

Note. FV = Fitted Value, ES = Effect Sizes.

The first (ES1) is based on the general effect size presented in Cooper and Hedges (1994), while the second (ES2) is based on Raudenbush and Feng (2001). We define ES1 as the estimated growth parameter (γ_{100}), divided by the sample standard deviation of the outcome. ES2 is defined as the estimated growth parameter (γ_{100}) divided by the standard deviation of true change ($\tau_{11}^{1/2}$) and has the advantage of excluding the sampling variance (Raudenbush & Feng). The correlations of these effect sizes are presented in column two. These range from approximately .62 to .93, and average about .78. These correlations, while still high, are expectantly lower as they are based on parameter estimates for growth and for the standard deviation of growth, which vary with each of the 2,000 simulations for each condition. The remaining four columns present the ratios and absolute differences for ES1 and ES2, respectively. We present both ratios and absolute differences as estimates near zero tend to disproportional effects on the ratio. The results in Tables 10a through 10d further corroborate the results presented above in terms of actual estimates of achievement growth. That is, NCE scores significantly and consistently under-estimate annual growth. This result is consistent across sampling conditions and model. The ratios of effect sizes for both ES1 and ES2 clearly demonstrate that actual growth is under-estimated using NCEs. In fact we calculate an index of relative bias (RB), as in Krull and MacKinnon (2001). In this case we define

$$RB = (ES_{100}^{NCE} - ES_{100}^{ss}) / \ ES_{100}^{ss},$$

where each ES (1 and 2) is calculated as noted above. This can be used to estimate the proportional under/over estimate using NCEs vs. scale scores. Figure 2[6] presents association between RB and the actual magnitude of annual growth (as measured by scale scores) in effect size units. We use this chart to demonstrate that the NCE metric under-estimates growth by about 85% and that this proportion is inversely related to the magnitude of growth. However, it is important to note that the range of growth and the corresponding range of RB are relatively small.

Hence, the metric does not change the substantive inferences made from ranking schools based on their fitted growth estimates, but matters when the inferences concern inferences about actual absolute growth.

Finally, we present the simulation results for evaluating school-wide program effects. Tables 11a and 11b display the relevant results. The first two columns present the correlations among the IS and growth estimates using NCEs and scale scores. In every case these correlations were at least .94. The next four columns present the ratios and absolute differences in effect sizes. In general, the program effect sizes were relatively small, which tends to make our comparison criteria take on extreme values in some

[6]The results in Figures 2, 3, and 4 are based on model 4.

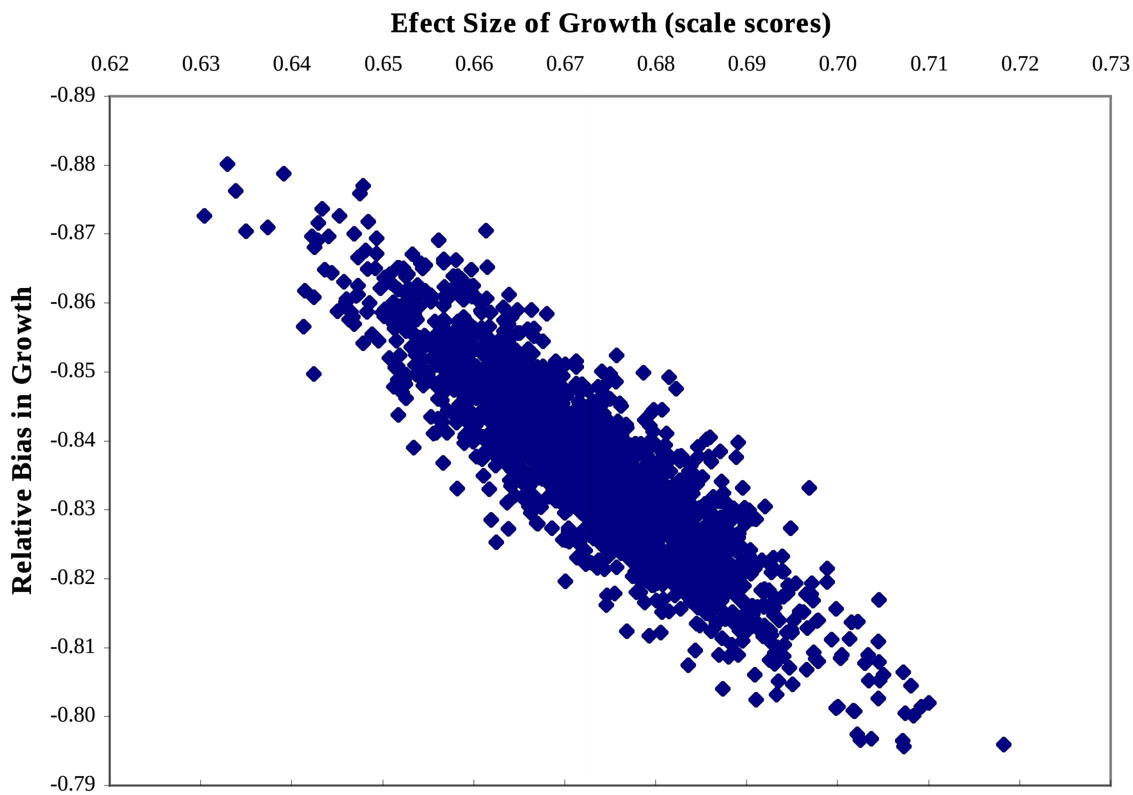

Figure 2. Comparison of relative bias to the effect size of growth – SAT-9 Reading (in scale scores).

Table 11a
Summary of Program Effect Sizes for SAT-9 Achievement – Model 3

Sample	Test type	Correlation		Effect size ratio		Effect size difference		Agreement (growth)
		Initial status	Growth	Initial status	Growth	Initial status	Growth	
R25	Read	0.982	0.943	0.686	0.934	-0.016	0.006	0.993
	Math	0.988	0.950	1.155	-0.267	-0.007	-0.001	0.998
R50	Read	0.982	0.941	0.843	1.417	-0.015	0.006	1.000
	Math	0.989	0.943	0.884	0.891	-0.006	-0.002	1.000
R75	Read	0.984	0.942	0.873	2.220	-0.014	0.006	1.000
	Math	0.990	0.939	0.910	0.904	-0.005	-0.002	1.000

Table 11b

Summary of Program Effect Sizes for SAT-9 Achievement – Model 4

Sample	Test type	Correlation		Effect size ratio		Effect size difference		Agreement (growth)
		Initial status	Growth	Initial status	Growth	Initial status	Growth	
R25	Read	0.988	0.954	1.149	0.348	-0.008	0.004	0.985
	Math	0.982	0.940	0.793	0.812	0.000	-0.003	0.992
R50	Read	0.990	0.950	0.965	0.693	-0.004	0.003	0.994
	Math	0.984	0.940	0.966	0.880	0.003	-0.004	0.999
R75	Read	0.985	0.940	0.265	1.146	-0.002	0.002	1.000
	Math	0.990	0.948	0.710	0.878	0.004	-0.004	1.000

instances. That is estimated ES1's that are very small may have small absolute differences, but have relatively large ratios. Substantively estimated ES1 that are close to zero for NCEs are close to zero for scale scores as well—leading to the same inferences regarding program effectiveness. Tables 11a and 11b, therefore, present both the ratio of $ES1_{NCE}$ to $ES1_{SS}$, but also the mean differences. The tables also display the proportion of the time that there is agreement in statistical significance program indicator variable between models using NCE and models using scale scores. The results indicate that the two metrics agree almost 100% of the time.

As noted, due to the relatively small ES1 for growth, we again use the RB measure. Figures 3 and 4 plot the relative bias as a function of the actual program

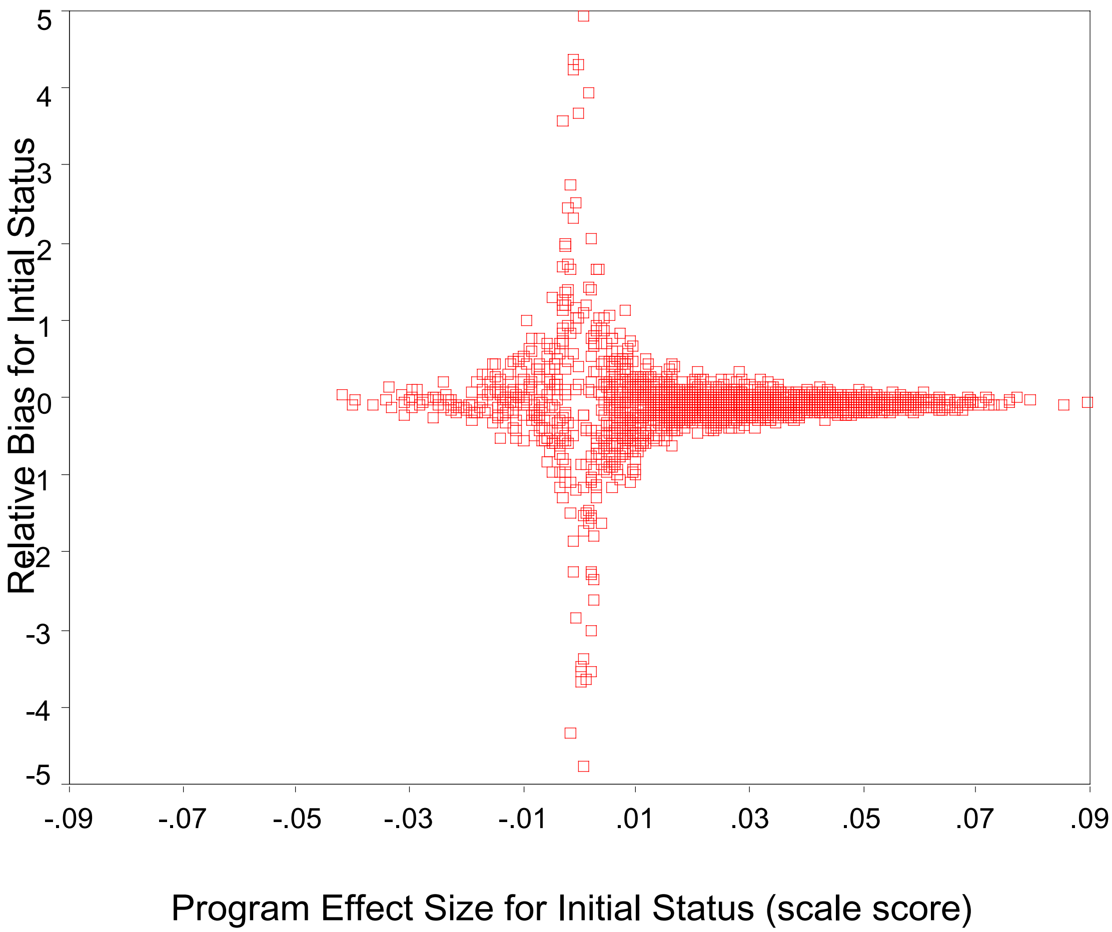

Figure 3. Relationship between relative bias in NCEs for initial status.

effect estimates, based on scale scores. These figures demonstrate two important points. One, the RB is largest when the actual scale score effect is close to zero (which, as noted, often generates very small actual differences in effect sizes, but very large ratios). And two, The RB decreases with the increase of the absolute value of the program effect size. This result holds true for both initial status and growth. The correlation, for SAT-9 reading, between $ES1_{NCE}$ and $ES1_{SS}$ is .986 and .940 for initial status and growth respectively

Hence, the simulations indicate that metric does not change inferences of whether or not a program is statistically or substantively significant. It is interesting to note that while NCEs significantly under estimate actual absolute growth, they accurately represent the difference between program and non-program schools in that growth.

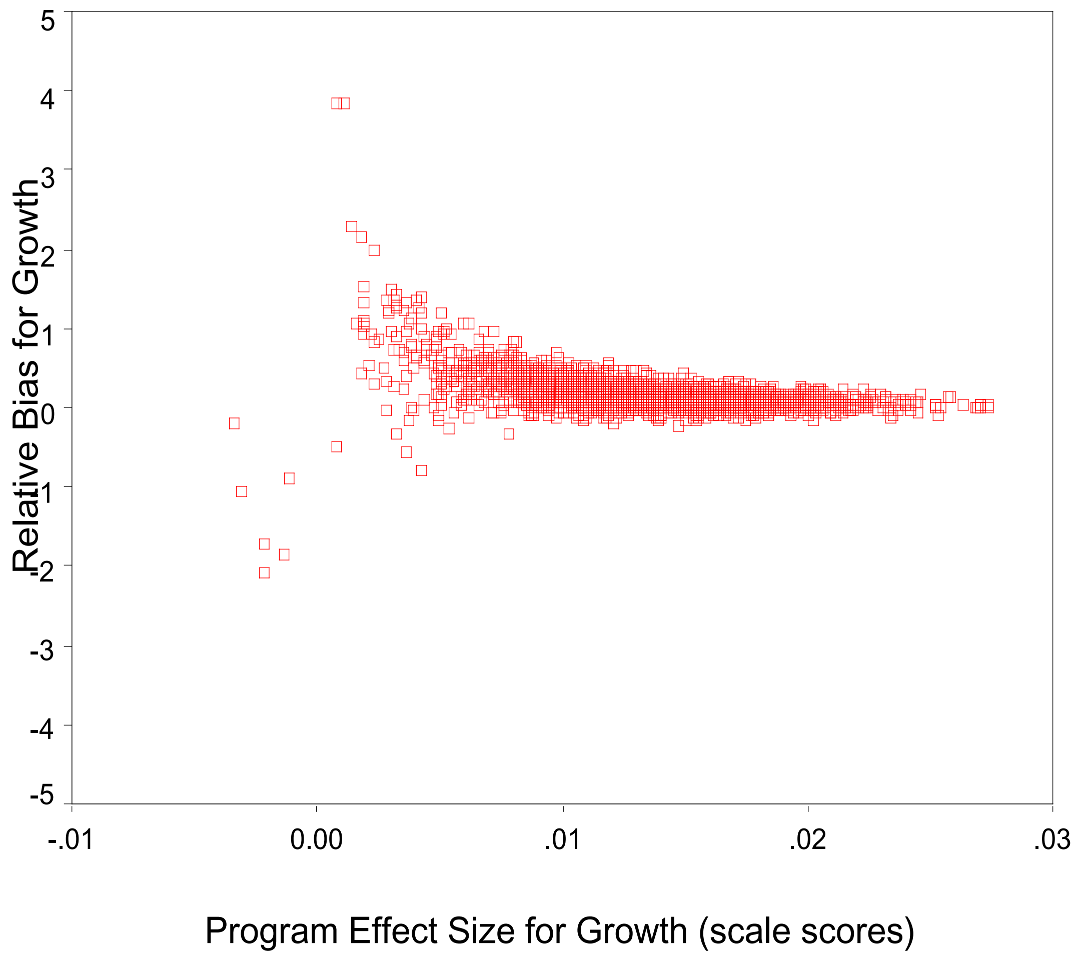

Figure 4. Relationship between relative bias in NCEs for growth.

Discussion

Recent legislation has increased the need to accurately evaluate school performance. School accountability has moved beyond simply comparing school means in league tables to more advanced techniques, such as hierarchical linear growth models. Coinciding educationists' increased focus on accountability is increased public interest in accountability, as demonstrated by the growing demand for school quality information. Given both the ubiquitousness of NCE scores and the ease of interpretation, it is valuable to ascertain the sensitivity of school level results to the choice of the metric. Further, given the volatility in the use of tests, it is important to ascertain whether NCE scores produce estimates that result in consistent policy implications compared to those derived using scale scores. The simulation results indicate that the effect of the metric is tied to the evaluation objective. We find that the correlation between mean initial status and growth rates for models using scale scores vs. models using NCEs is strong. The simulation results also suggest that this relationship is affected, to a small, extent, by the sample size and the complexity of the model used. The effect of model complexity is stronger than the effect of sample size. The results for the rank order correlations between fitted values for mean school initial status and mean school rates of change, for models using scale scores vs. models using NCEs, are also consistently strong. These results are invariant across sample size and model complexity. This demonstrates that if the objective of the evaluation is to rank schools then the choice of NCE or scale score will not change the ensuing school ranking. NCEs can accurately rank schools.

Further we find that the correlation of program effect sizes for models using scale scores vs. NCEs are strong and consistent. This means that evaluations based on NCEs and using hierarchical longitudinal models will be able to accurately estimate both initial differences among program and non-program schools, and whether the program has an effect on growth—both in terms of statistical and substantive significance.

Still, it is important to reiterate that when estimating mean school growth, NCEs will yield misleading absolute achievement growth estimates. In fact the simulation results indicate that using NCEs under-estimate growth by about 85%. Interpreting growth using NCEs, without additional information, is likely to be difficult, despite the fact that students maintaining the same NCE score from one year to the next must have demonstrated some absolute growth in order to maintain their relative standing. Even with additional information, the estimated range of growth in the scale score metric, derived from NCEs, will be imprecise. However, ranking schools by the amount of growth they exhibit does yield consistent results.

The results of this analysis provide some guidance in the planning of evaluations or monitoring of school performance; this is particularly relevant for program evaluations or school performance systems that attempt to take advantage of longitudinal data sources, but are limited to conducting analyses with NCE scores. These results may be particularly relevant as school districts change tests, but desire to conduct longitudinal evaluations across the different tests—as NCEs may be more comparable across tests than IRT based scale scores. As Linn (2000) demonstrated there will clearly be effects from switching tests, but these can be handled within the model (Goldschmidt & Swigert, 2001). The simulations also demonstrate that the results based on a 25% sample are very consistent with results based on the full sample. This finding may be particularly relevant for longitudinal methods that do not rely on panel data, such as HLM models for estimating school effects proposed by Willms and Raudenbush (1989). The effects of NCEs vs. scale scores in these types of models are unknown and warrant further research. These results allow program evaluators and school accountability analysts more flexibility in designing evaluations—especially when cost is an issue. While the results indicate that schools can be ranked and programs evaluated consistently using longitudinal methods and NCE scores, ranks should be interpreted carefully. Confidence intervals of estimated effects generally overlap substantially, which means that caution should be exercised when comparing schools in this manner (Goldstein et al., 1993). Further, we are by no means advocating that using a single standardized measure be the sole criteria upon which school or program quality ought to be judged.

References

Aitkin, M., & Longford, N. (1986). Statistical modeling issues in school effectiveness studies. *Journal of the Royal Statistical Society, 149*(1), 1-43.

Allen, M., & Yen, W. (1979). *Introduction to measurement theory*. Monterey, CA: Brooks / Cole Publishing.

Bryk, A. S., & Weisberg, H. I. (1977). Use of the nonequivalent control group design when subjects are growing. *Psychological Bulletin, 84*(5), 950-962.

Burstein, L. (1980). The analysis of multi-level data in educational research and evaluation. *Review of Research in Education, 4,* 158-233.

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research.* Boston: Houghton Mifflin.

Cooper, H., & Hedges, L. (1994). *The handbook of research synthesis.* New York: Sage.

Goldschmidt, P. (2002a). *Evaluation of Alaska Beginning Literacy Institutes*. Los Angeles: University of California, Center for the Study of Evaluation (CSE).

Goldschmidt, P. (2002b). *A comparison of student achievement in LAAMP and Non-LAAMP schools in Los Angeles county: Longitudinal analysis results 1997-'98–1999-2000.* Los Angeles: University of California, Center for the Study of Evaluation (CSE).

Goldschmidt P., & Swigert, S. (2001). *Oxymoronic program evaluation: The short-term longitudinal analysis dilemma.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Goldstein, H., Rasbash, J., Yang, M., Woodhouse, G., Pan, H., Nuttall, D., et al. (1993). A multilevel analysis of school examination results. *Oxford Review of Education*, *19*(4).

Hambleton, R. K., & Swaminathan, H. (1987). *Item Response Theory: Principles and applications*, Boston, MA: Kluwer.

Hanushek, E. A., & Raymond, M. E. (2001, June). The confusing world of educational accountability. *National Tax Journal, 54*(2), 365-384.

Heck, R. H. (2000, October). Examining the impact of school quality on school outcomes and improvement: A value-added approach. *Educational Administration Quarterly, 36*(4), 513-552.

Hox, J. (2002). *Multilevel analysis: Techniques and applications.* Mahwah, New Jersey: Erlbaum.

Huttenlocher, J., Haight, W., Bryk, A., & Seltzer, M. (1991). Early vocabulary growth: Relation to language input and gender. *Developmental Psychology, 27,* pp. 236-248.

Krull, J., & MacKinnon, D. P. (2001). Multivariate modeling of individual and group level mediated effects. *Multivariate Behavioral Research, 36*(2), 249-277.

Linn, R.L. (2000). Assessments and accountability. *Educational Researcher, 20*(2), 4-16.

Osgood, W. D., & Smith, G. (1995). Applying hierarchical linear modeling to extended longitudinal evaluations. *Evaluation Review, 19*(1), 3-39.

Ramirez, D., Yuen, R., Ramey, R., & Pasta, D. (1991). *The immersion study* (final report). Washington, DC, Office of Educational Research and Improvement.

Rasbash, J., & Goldstein, H. (1994). Efficient analysis of mixed hierarchical and cross-classified random structures using a multilevel model. *Journal of Educational and Behavioral Statistics, 19*(4), 337-350.

Raudenbush, S. (2001). Comparing personal trajectories and drawing causal inferences from longitudinal data. *Annual Review of Psychology, 52*, 501-525.

Raudenbush, S. (1993). A crossed random effects model for unbalanced data with applications in cross-sectional and longitudinal research. *Journal of Educational Statistics, 18*(4), 321-349.

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

Raudenbush, S. W., & Bryk, A. S. (1987). Application of hierarchical linear models to assessing change. *Psychological Bulletin, 104*(3), 396-404.

Raudenbush, S., & Xiao-Feng, L. (2001). Effects of study duration, frequency of observation, and sample size on power in studies of group differences in polynomial change. *Psychological Methods, 6*(4), 387-401.

Rogosa, D. R., Brandt, D., & Zimowski, M. (1982). A growth curve approach to the measurement of change. *Psychological Bulletin*, *92,* 726-74.

Seltzer, M., Frank, K., & Bryk, A. (1994). The metric matters: The sensitivity of conclusions about growth in student achievement to the choice of metric. *Educational Evaluation and Policy Analysis, 16*, 41-49.

Willett, J. B., Singer, J. D., & Martin, N. C. (1998). The design and analysis of longitudinal studies of development and psychopathology in context: Statistical models and methodological recommendations. *Development and Psychopathology, 10,* 395-426.

Willms, J. D., & Jacobsen, S. (1990). Growth in mathematics skills during the intermediate years: Sex differences and school effects. *International Journal of Educational Research, 14,* 157-174.

Willms, J. D., & Raudenbush, S. (1989). A longitudinal hierarchical linear model for estimating school effects and their stability. *Journal of Educational Measurement, 26*(3), 209-232.

Worthen, B. R., White, K. R., Fan, X., & Sudweeks, R. R. (1999). *Measurement and assessment in the schools* (2nd ed.). New York: Longman.

Yen, W. M. (1986). The choice of scale for educational measurement: An IRT perspective. *Journal of Educational Measurement, 23*(4), 299-325.

Appendix

Level-1 Model

Y = P0 + P1*(TIME) + E

Level-2 Model

P0 = B00 + B01*(SPED1) + B02*(LOW1) + B03*(LEP1) + B04*(MINOR)

+ B05*(GIRL_1) + R0

P1 = B10 + B11*(SPED1) + B12*(LOW1) + B13*(LEP1) + B14*(MINOR)

+ B15*(GIRL_1) + R1

Level-3 Model

B00 = G000 + G001(MEDLAMP) + G002(MMINOR2) + G003(MLOW1) + U00

B01 = G010

B02 = G020

B03 = G030

B04 = G040

B05 = G050

B10 = G100 + G101(MEDLAMP) + G102(MMINOR2) + G103(MLOW1) + U10

B11 = G110

B12 = G120

B13 = G130

B14 = G140

B15 = G150

Where:

Time = years Coded as 0 = 1998, 1 = 1999….)

SPED1 = Special Education (0 = no 1 = yes)

LOW1 = (proxy is free / reduced lunch status) (0= no 1 = yes)

LEP1 = English Language Learner (0= no 1 = yes)

MINOR = Minority (i.e., non-White) (0= no 1 = yes)

GIRL = gender = female (0= no 1 = yes)

MEDLAAMP = school participated in school-wide reform (0= no 1 = yes)

MINOR2 = school mean percentage of MINOR

MLOW1 = school mean percentage of LOW1.