EDUCATION POLICY ANALYSIS ARCHIVES

English Editor: Sherman Dorn
College of Education
University of South Florida

Spanish Editor: Gustavo Fischman
Mary Lou Fulton College of Education
Arizona State University

Volume 17 Number 22 **November 20, 2009** **ISSN 1068–2341**

Background Variables, Levels of Aggregation, and Standardized Test Scores

Sharon E. Paulson
Gregory J. Marchant
Ball State University

Citation: Paulson, S. E., & Marchant, G. J. (2009). Background variables, levels of aggregation, and standardized test scores. *Education Policy Analysis Archives, 17*(22). Retrieved [date] from http://epaa.asu.edu/epaa/v17n22/.

Abstract
This article examines the role of student demographic characteristics in standardized achievement test scores at both the individual level and aggregated at the state, district, school levels. For several data sets, the majority of the variance among states, districts, and schools was related to demographic characteristics. Where these background variables outside of the control of schools significantly affected averaged scores, and test scores result in high stakes consequences, benefits and sanctions may be inappropriately applied. Furthermore, disaggregating the data by race, SES, limited English, or other groupings ignores the significant confounding and cumulative effects of belonging to more than one disadvantaged group. With these approaches to evaluation being fundamental to the No Child Left Behind mandates, the danger of misinterpretation and inappropriate application of sanctions is substantial.

Keywords: statistical methodology; accountability; student achievement; demographic factors.

Variables de antecedentes, niveles de agregación, y resultados en pruebas estandarizadas

Resumen
Este artículo examina el papel de las características demográficas de los estudiantes en los resultados de los pruebas estandarizadas tanto a nivel individual, como agregado a nivel de estado, distrito, y niveles escolares. Para varios conjuntos de datos, la mayoría de las variación entre los estados, distritos y las escuelas se relaciona con características demográficas. Cuando las variables de antecedentes, que están fuera del control de las escuelas, afectan de manera significativa el promedio de calificaciones y los resultados de las pruebas, premios y sanciones pueden ser utilizados de manera inadecuada. Además, desglosando los datos por raza, situación socioeconómica, limitaciones en el manejo del idioma inglés, u otras características importantes presenta el problema de ignorar los efectos acumulativos de pertenecer a más de un grupo desfavorecido. En estos métodos de evaluación que son fundamentales para la ley "No Child Left Behind", los problemas y riesgos de malas interpretaciones y aplicación indebida de sanciones son muy importantes.
Palabras clave: metodología estadística; responsabilidad; rendimiento de los estudiantes; factores demográficos.

Introduction

The celebration of success is important in every field, as is the inevitable identification of failure. In the field of education, the valued outcomes used to identify winners and losers have not always been well defined. Unlike business where the production of the most, best, and cheapest widgets may contribute to the bottom-line measure of success or profit, education has had no one-size-fits-all bottom line number to identify winners and losers. For some students, success was receiving a diploma, for others it was entrance into college, and for others success might have been developing skills to benefit them in life and in the workplace. Failures were those who did not complete high school, who left the process without identifiable skills or knowledge that would serve to benefit them.

Recently, however, the percentage of students passing their state's standardized achievement test has become the universal indicator of success in education. It is the defining measure for winners and losers for individual students, for schools and districts, and even for states. To determine success and failure beyond the student level, scores are aggregated or averaged at different levels, and these aggregated scores often have very tangible consequences for teachers, schools, and districts. The practice of comparing averaged scores as an approach to evaluation is not new or unusual. The comparison of groups receiving an intervention or special program to other groups can provide insight into the effectiveness of the intervention. However, when the defining of success or more importantly failure carries serious consequences, special care must be taken to assure that the comparisons are valid or meaningful. The purpose of this study is to examine the validity of aggregating test scores at various levels of comparison: classroom, school, district, and state. In particular, does aggregation change the meaning of the scores, thereby rendering many of the associated high stakes decisions invalid?

Validity of Accountability Systems

Although accountability systems have been used to evaluate the effectiveness of schools for decades, these systems vary widely from state to state and from district to district (Hamilton & Koretz, 2002; Linn, 2006a). The No Child Left Behind Act of 2001 (NCLB, 2002) created a dramatic movement towards standardizing accountability systems across the nation. Under NCLB, every state is required to develop a set of learning standards and a statewide test to assess whether or not students meet the standards. By 2014, every school must demonstrate that 100% of its students have reached proficiency on the standardized state test (where each state determines its own cutoff for proficiency). This system requires also that each school show adequate yearly progress (AYP) in meeting its goals, by demonstrating that there are improvements in the percentage of students meeting proficiency each academic year. Schools that fail to meet AYP two years in a row are deemed failures, and parents can move their children to a school that has met AYP. Inherent within the system are similar high-stakes consequences for students and for teachers. Students who do not reach proficiency may be held back or not allowed to graduate; teachers who do not show improvements in their students' scores from year to year may receive no pay raise or be let go. These are indeed high-stakes tests.

In the accountability system established by No Child Left Behind, annual standardized testing in every state has been heralded as *the* universal tool for accountability. However, to be effective, the system must be based on a number of assumptions: It assumes that the tests reflect important standards of learning that are being taught in the schools. It assumes that students who do not reach proficiency are inadequate in their knowledge and skills, regardless of their performance on other forms of assessment; and it assumes that these tests are better indicators of students' ability than the judgments of the teachers. It assumes that the collective scores of teachers' students reflect the quality of their instruction and it assumes that the collective scores of schools and districts reflect the quality of their educational programs. It even assumes that the collective scores of test-takers from a state represent the quality of education and educational policies of the state. With little if any evidence that these assumptions are valid, policy makers have chosen to use these scores as the sole means of accountability for student, teacher, school, district, and even state-level performance.

Even before the implementation of NCLB but especially in recent years, a number of researchers in education called into question such assumptions (e.g., Baker & Linn, 2002; Linn, 2005, 2006a,b; Koretz, 2002; Meyer, 2000). They have questioned the validity of this system, based mainly on whether or not the test scores themselves are valid and on what scores aggregated at the classroom- (teacher) or school-level really mean. Validity is a concept used to indicate that a test or measure indeed assesses what it purports to test or measure. Although there has been much discussion about whether states' standardized tests are valid assessments of their standards (Linn, 1998; Miller & Linn, 2000), more debatable is whether the status of the percentage of students reaching proficiency on the test is a valid assessment of either a teacher's or school's educational performance (Kane, Staiger, & Geppert, 2002; Kirby, McCaffrey, Lockwood, McCombs, Naftel, & Barney, 2002; Le & Klein, 2002; Linn, 2006a, 2006b; Raudenbush, 2004)

The inherent assumptions that the aggregate scores of students reflect their achievement in that classroom and in that school are wrong on a number of counts. First, no matter how valid the test may be in assessing a state's learning standards, an individual score is also a reflection of only one type of learning (mostly concrete knowledge), using only one format (paper and pencil; usually multiple choice questions), in only one testing condition (timed, high pressure, anxiety producing). Further, scores may become inflated by practice effects, teaching to the test, and coaching (Kirby et

al., 2002; Koretz, 2002, 2005). Moreover, it has been argued that students' individual scores include their initial cognitive skill, their history of achievement from previous years (and teachers), their family status, and the peer-orientation of the school (Hanushek, Kain, Markman & Rivkin, 2003; Linn, 2005; Meyer, 2000; Rouse, 2005). Consequently, aggregates of these individual scores are a reflection of achievement from many different sources, most of which are outside of the control of either the teacher or the school (Koretz, 2002; Linn, 2006a). When scores are aggregated, validity problems from the individual level become magnified (Kane & Staiger, 2002; Kane et al., 2002; Kirby et al., 2002). Therefore, in current status accountability systems, where classroom- and school-level performance are based on aggregates of students' scores at one point in time, the assessment of teachers' and schools' instructional and educational quality are not valid: the aggregate scores do not measure what the system purports they do (Hamilton, 2003; Linn, 2005, 2006a,b).

More recently, accountability systems based on growth, or changes in scores from one year to the next, have been developed to overcome these validity issues. However, when AYP is calculated as the change in students' aggregate scores from one year to the next, a new layer of validity problems are added. Similar to current status accountability systems, growth models also are based on a number of assumptions. Most egregious is that growth models are based on the inference that changes in students' scores within a classroom or school are caused by the quality of instruction or education provided by teachers or schools; an assumption that is not valid on a number of counts (Hamilton, 2003; Koretz, 2005; Linn, 2006a,b). Comparing the percentage of students reaching proficiency on a state's standardized achievement test from one year to the next to determine AYP is equivalent to a quasi-experimental research design that tests the effects of an intervention by comparing the scores of groups that have not been randomly assigned. Consequently, there are numerous threats to the validity of accountability systems based on growth (see Campbell & Stanley, 1966; Cook & Campbell, 1979).

The scores from one year to the next (one group to the next) may differ not because of any increased learning but for any number of the following reasons: testing – increased familiarity with a test (e.g., practice, practice, practice) can lead to improved performance not related to learning; instrumentation – changes in the measuring instrument (e.g., eighth grade test vs. ninth grade test, or changes in the criterion for proficiency) across testing times can yield changes in results not related to performance; statistical regression – on unreliable measures, high scorers tend to score lower and low scorers tend to score higher on subsequent testing (Kirby et al., 2002). In addition, the comparison of aggregate scores of successive cohorts from one year to the next is not a comparison of the same sample or even a matched sample [i.e., selection bias] (Linn, 2006a; Linn, Baker, & Betebenner, 2002). The students in a teacher's classroom change from year to year and students within a school change from year to year (a function of both mobility [i.e., attrition] and new groups of younger students taking the place of the older students who have moved on).

Finally, just as static scores may be inflated or confounded by selection factors (e.g., cognitive skill, race, SES) out of control of teachers and schools, change scores become inflated or confounded by selection even more. Although measuring change scores may control for some initial differences between students, these factors are still responsible for a large part of the growth (Kane & Staiger, 2002; Zvoch & Stevens, 2008). For example, more cognitively capable students may be primed for greater learning, and students from more advantaged homes may gain support from outside the school that is responsible for the learning. Therefore, to say that growth (or test-score change) is caused by educational practices in the school or classroom is invalid (Hamilton, 2003; Kane, 2006; Koretz, 2005; Linn, 2006a, 2006b; Raudenbush, 2004) from both a measurement and an experimental perspective. Although it is unclear whether those establishing or advocating current educational policy are aware of the inherent flaws in their approach to accountability, the threats to the validity of the evaluations cannot be ignored.

Yet another type of accountability system has been advocated as a means to overcome most or all of these threats to validity. Value-added models track individual student scores over time, and they have been touted as the most valid means for controlling for background characteristics of students (Ballou, Sanders, & Wright, 2004; Sanders & Horn, 1998). Raudenbush (2004) agreed that value-added models are better than current status models, but he also pointed out that the confounding effects of student background factors could never be completely eliminated. Furthermore, selection biases can become magnified as these change scores begin to become confounded with classroom- or school-level changes (Kane & Staiger, 2002; Linn et al., 2002). These problems become even greater at the middle and secondary levels where students have multiple teachers and teachers' class assignments change from year to year. Even the use of the most sophisticated statistical models cannot totally remove alternative explanations for learning, and it has been argued that value-added models should not be seen as causal models but as descriptive ones (e.g., Rubin, Stuart, & Zanutto, 2004). Additionally, many of the longitudinal changes in students, classrooms, and schools create violations of the assumptions of traditional hierarchical linear models, such that the "causes" of student learning can never be fully understood (McCaffrey, Lockwood, Koretz, Louis, & Hamilton, 2004).

One further attempt to overcome the influence of demographic traits on aggregated scores is to disaggregate scores, particularly at the school-level. Disaggregation is when (aggregated) scores are reported separately for different demographic groups, an approach mandated by the *No Child Left Behind* legislation. NCLB requires schools to report scores for seven subgroups: students with disabilities, limited English proficient, free/reduced lunch, African American, Asian/Pacific islander, Hispanic, and White (or alternative groups as defined by a state). The premise is that schools with students whose demographic or background characteristics may lower their aggregate scores can establish whether or not growth has occurred in each group separately. However, because schools must show AYP for each subgroup, schools who may have met AYP overall may fail if just one subgroup fails to reach the required level of proficiency. The more subgroups a school has, the more likely the school will fail to meet AYP (Baker & Linn, 2002; Linn, 2005). Furthermore, with many disadvantages for children being correlated, it is likely that many students will have multiple disadvantages and fall into more than one category. For example, a child of minority status likely will score lower on a standardized achievement test. If that same child is eligible for free lunches, the child might be expected to score even lower. If that poor minority student also has limited English proficiency or a disability, the child should be expected to score even lower. Disaggregation based on only one subgroup ignores the compounding effects of multiple disadvantages. So, if one school has ten percent minority students, and another school has fifty percent minorities, one would not necessarily expect the two schools' disaggregated scores based on minority status to be different from each other, provided all of the students from both schools otherwise had the same characteristics. However, if that larger subgroup of minority students also had a larger percentage of students in multiple categories with multiple disadvantages, then that school will show lower scores for its minority students (perhaps not reaching AYP for that group) despite the same quality of education at the two schools. Therefore, the confounding effects of multiple demographic disadvantages make simple disaggregation both incomplete and potentially deceptive.

Validity of Different Levels of Aggregation

Despite validity problems when calculating AYP (aggregation at the school level) or when assessing teachers (aggregation at the classroom level), the common practice of aggregating measures to create averages or other indices for comparison and evaluation remains the cornerstone of the

education accountability movement. Many schools, districts, or even states do not have the resources or personnel needed to use the sophisticated statistical models of value-added accountability systems that may correct some but not all of the problems associated with assessing the quality of instruction or education. For these units, aggregate scores may be all they have to assess their own practices. Perhaps an even larger problem is the misuse of aggregated data to inform the public of the quality of their state's education system. Local newspapers are quick to publish the average scores or proficiency percentages for area schools when achievement test data are released. Local television news highlights the meager gains and minor losses, and principals, superintendents, and state officials talk about the need to improve education and to raise standards. The general public uses these comparisons to make judgments about the quality of schools in their district and state. The public assumption that these scores are valid assessments of educational quality exacerbates the already weighty consequences of NCLB. And it is at the public level that the validity of different levels of aggregation becomes important. Is aggregation at the classroom level any more or less valid than aggregation at the school, district, or state levels? It has been suggested that higher levels of aggregation magnify the validity problems already inherent at lower levels (Kane et al., 2002; Kane & Staiger, 2002; Raudenbush, 2004). Although there is much theoretical evidence to question the validity of aggregated test scores, the empirical evidence that aggregated test scores are invalid assessments of educational quality at any level is limited.

The studies that have been conducted have shown that the most prominent threat to the validity of group comparisons across all levels of aggregation (classroom, school, district and state) is selection. Because students are not randomly assigned to their groups (states, school districts, school buildings, or even classrooms), the demographic characteristics of groups of students have a major effect on aggregated scores at any level. Two studies highlight the effects of selection on aggregated scores; one using state aggregated SAT scores and one using state aggregated NAEP scores. For years, the College Board released the rank-order of state aggregated SAT scores to the public, who in turn interpreted the rankings to imply variation in states' education quality. When its practice came under fire (Grissmer, 2000; Marchant & Paulson, 2001), the College Board acknowledged that the percentage of students taking the test from each state influences the averaged scores. The College Board now publishes the state data in rank-order based on the proportion of students who are test-takers from each state; however, they fail to acknowledge that simply reporting the proportion taking tests masks how the demographic characteristics of the test takers might impact aggregated scores. Indeed, these test-taker characteristics (e.g., race, parent education, and family income) were found to account for as much as 94% of the variance in scores among states (Marchant & Paulson, 2001). In a similar study using the National Assessment of Educational Progress (NAEP) data, 80% of the variation among states' scores in reading and math were attributable to demographics, and changes in test scores from year to year were related significantly to changes in demographics (Marchant, Paulson, & Shunk, 2006). Therefore, to say that differences between group means are caused by differences in educational quality is invalid and inappropriate. The purpose of this study is to examine the role of selection in aggregated test scores at multiple levels; are the effects of selection apparent at every level of aggregation, thereby influencing the validity of the aggregate scores and the accountability judgments that can be made at different levels of aggregation?

The Current Study

The purpose of the current study is to examine the effects of aggregation (and disaggregation) at various levels. Two sources of data were used, because data for multiple levels of aggregation were not available from any one source. First, we used data from the 2001 SAT

database, because these data allowed for analysis of individual-, school-, and state-level data from a single source. Although the SAT assesses a slightly different aspect of learning (reasoning rather than achievement), not unlike state standardized tests, it too has high stakes consequences. At the individual level, students may or may not be admitted to certain colleges based on their scores. In turn, students' aggregate scores at the high school level are seen as an indicator of schools' educational quality, and at the college or university level as an indicator of the quality of admissions standards. Similarly, SAT scores aggregated at the state level are often viewed as an indication of the quality of a states' education system. Furthermore, in previous work, it was shown that the effect of demographic characteristics on aggregated scores at the state-level were similar for both the SAT and the NAEP, despite differences in the tests' purposes (Marchant & Paulson, 2001; Marchant et al., 2006). Moreover, when comparing state- or school-level aggregate scores, the SAT allows for comparison across states using the same test. Although state-standardized test data are available from each state's department of education, each state uses a different test, making aggregate comparisons even less valid.

Second, we used the data from the ISTEP+, the Indiana State Test of Educational Progress, for the assessment of both school- and district-level aggregation within a single state. We used the language arts and math test scores only, because those are areas comparable to those assessed by the SAT. Also similar to the SAT, the test is administered in a multiple-choice format. The ISTEP+ assesses both knowledge and critical thinking skills and has been found to be well aligned with the Indiana standards (Edwards, 2001). By examining school-level aggregation using two different data sources (the SAT and the Indiana state test), we could further support our use of the SAT for other levels of aggregation. Finally, we used state SAT data from two randomly chosen states, Louisiana and Delaware, in our assessment of the validity of disaggregation.

In this study, we addressed two research questions. First, to what degree do demographic characteristics of test-takers influence aggregate test scores at different levels of aggregation (individual, school, district, and state); and second, is the practice of disaggregating scores for different demographic subgroups at the school level a valid means for controlling for demographic differences in score comparisons?

Method

Participants

Participants for this study came from two databases. First, the College Board provided the database of all SAT test takers from 2001 (N = 1,219,550). From this database, SAT scores could be examined for individual students and aggregated by state (n = 51, including the District of Columbia) and by school (n = 14,432). The second database provided by the Indiana Department of Education contained state standardized test scores from 2001 aggregated for all 295 school districts and their schools in the state of Indiana.

Measures

SAT scores: State, school, and individual level data. The SAT is a reasoning test administered annually by the College Board to students who self-select to take the test. Nonetheless, the SAT is a high-stakes test in that it holds consequences for college admissions for those who take it. The voluntary nature of the test creates sampling issues unique to the test; however, SAT scores

aggregated by state behave similarly to other aggregated standardized achievement test scores in its relations to demographic characteristics of the test takers (i.e., the proportion of variance between states that can be attributed to demographics such as family income and parent education is similar; Marchant et al., 2006). Furthermore, state aggregated scores are more comparable across states with the SAT, than with individual state's standardized test scores, because everyone is taking the same test. Similarly, data could be aggregated at the school level and comparisons of schools in different states are based on the same test. Moreover, individual level data were available from this dataset.

For the purposes of this study, students' total scores on both the math and verbal portion of the test were used. Individual SAT scores range from 400–1600; in this sample, 464 students achieved a perfect score of 1600 and 330 students achieved the lowest score of 400.

ISTEP scores: District and school level data. The Indiana Standardized Test of Educational Progress (ISTEP+) is the standardized achievement test given to students in grades three through ten in all school districts in Indiana. The test was developed as an assessment of the Indiana State Standards of Learning in accordance with No Child Left Behind. For the purposes of this study, data on both the math and language arts portions of the test were used, reported as the percentage of students who reached proficiency on both portions, aggregated by district and by high school. We used percentage of students reaching proficiency, because that is the standard by which AYP under NCLB is determined. The average pass rate across the state is about 60%.

Demographic data. Demographic characteristics of the test takers that have been found to be confounded with aggregated test scores were included. Academic skills of students were measured from the SAT data using students' high school grade point average and their high school class rank, assessed as the percentage of test-takers who ranked academically in the top ten percent of their class. From the ISTEP+, we used the Cognitive Skill's Index (an assessment similar to an intelligence score thought to reflect non-subject specific ability) and the percentage of special education students. The SAT provided the family demographics of income, assessed as the percentage of families with income over $80,000, and parent education level, assessed as the percentage of parents with a bachelors degree or higher. The ISTEP+ provided the percent of students eligible for free lunch. Finally, race was assessed using the percent of African American students for both the SAT and the ISTEP+. In addition, in state-level analyses with the SAT, the percent of students taking the test from each state was included to provide a moderate control for the selection bias inherent in the test.

Results

State-Level Data

Multiple regression analyses were used to assess the proportion of variance in scores aggregated at the state-level that could be attributed to the demographic characteristics of the test-takers. Using the SAT data aggregated by state, regression showed that the percentage of test-takers from each state, high school grade point average, high school rank, family income, and parent education level predicted a large proportion of the variance in SAT scores *among* states, $R^2 = .91$, $F(5,45) = 88.99$, $p < .001$). In a second regression analysis, the addition of the race factor boosted the proportion of variance predicted, $R^2 = .94$, $F(6, 44) = 115.62$, $p < .001$. The contributions of each demographic factor are shown in Table 1.

In a follow-up analysis, all of the variables that did not account for a unique proportion of variance were removed from the regression equation, leaving parent education level and race. In this

regression, these two variables accounted for 90 percent of the variance in aggregated SAT scores among states, $R^2 = .90$, $F(2, 48) = 89.00$, $p < .001$. Both parent education level (standardized $\beta = 0.85$, $t = 18.45$, $p < .001$) and race ($\beta = -0.32$, $t = -6.92$, $p < .001$) accounted for significant proportions of unique variance.

Table 1
Regression analysis of demographics predicting state-aggregated SAT scores

Variable	*B*	*SE*	β	*t*	*p*
(Constant)	738.32	144.86		5.10	.001
Percent taking SAT	-0.17	0.32	-0.07	-0.54	.589
High school GPA	49.69	43.37	0.16	1.15	.258
High school rank	1.18	1.05	0.18	1.12	.267
Family income	1.19	0.60	0.16	1.98	.054
Parent education	2.80	0.84	0.41	3.32	.002
Race	-1.37	0.28	-0.21	-4.87	.001

$R^2 = .94$, $df = 6, 44$, $F = 115.62$, $p < .001$

District-Level Data

Multiple regression analyses were used to assess the proportion of variance in percentage of students reaching proficiency on the ISTEP+ aggregated at the district-level that could be attributed to the demographic characteristics of the test-takers. Four variables were entered into a regression equation: the average Cognitive Skills Index for 2001–2002, percentage of special education students, percentage of minority students (race), and percentage of students eligible for free lunch (family income). These factors predicted 70% of the differences among school districts ($R^2 = .70$, $F[4, 256] = 146.65$, $p < .001$) with the Cognitive Skills Index, percentage of special education student, and family income accounting for significant proportions of unique variance (see Table 2). Removal of the Cognitive Skills Index from the equation reduced the R^2 to .51 ($p < .001$).

Table 2
Regression analysis of demographics predicting district-aggregated ISTEP+ proficiency

Variable	*B*	*SE*	β	*t*	*p*
(Constant)	-74.14	12.48		-5.94	.001
Cognitive Skills Index	1.38	0.12	0.53	12.00	.001
Special Education	-0.43	0.11	-0.15	-4.01	.001
Family income	-0.27	0.05	-0.29	-5.58	.001
Race	-0.06	0.03	-0.08	-1.85	.065

$R^2 = .70$, $df = 4, 256$, $F = 146.65$, $p < .001$

In an effort to make the district-level analyses using the ISTEP+ more equivalent to those with the SAT data, only high school ISTEP+ data (tenth grade) were retained in the final analysis for districts. In this multiple regression, demographic factors predicted 77% of the variability in district-level pass rates for high school students ($R^2 = .77$, $p < .001$), with Cognitive Skills and income both predicting unique proportions of variance. The total predicted proportion of variance dropped to 68% when the Cognitive Skills Index was removed from the equation and both race ($\beta = -0.17$,

t = -3.43, p < .01) and income (β = -0.70, t = -14.21, p < .001) predicted significant portions of unique variance.

School-Level Data

Using the SAT database again, with data aggregated at the school level, multiple regression analyses were used to assess the proportion of variance in scores that could be attributed to the demographic characteristics of the test-takers. At the school level, 63% of the variance among 14,432 high schools in the United States was predicted with grade point average, high school rank, parent education and income, and race (R^2 = .63, F[5, 14426]= 4968.28, p < .001) with all of the variables predicting unique portions of the variance (see Table 3). In a follow-up analysis using only parent education and race (in keeping with the final analysis run with the state-level data), 51% of the variance in school-aggregated scores continued to be predicted, R^2 = .51, F(2, 14429) = 7651.88, p < .001. Both parent education level (β = 0.53, t = 88.95, p < .001) and race (β = - 0.37, t = -62.64, p < .001) accounted for significant proportions of unique variance.

Table 3
Regression analysis of demographics predicting school-aggregated SAT scores

Variable	*B*	*SE*	β	*t*	*p*
(Constant)	493.04	8.54		57.75	.001
High school GPA	130.85	2.71	0.33	48.28	.001
High school rank	19.59	4.03	0.03	4.86	.001
Family income	115.05	3.94	0.21	29.20	.001
Parent education	205.38	4.23	0.34	48.60	.001
Race	-144.88	3.39	-0.24	-42.75	.001

R^2 = .63, df = 5, 14426, F = 4968.28, p < .001

In addition, SAT scores aggregated at the high school level were examined for schools in Indiana only. A multiple regression found that the same variables accounted for 60% of the variation in school-aggregated SAT scores in Indiana (R^2 = .60, F[5, 400] = 121.82, p < .001), with all of the demographic factors predicting unique proportions of the variance among schools (see Table 4). Even after removing all of the factors except parent education and race from the equation, 53% of the variance continued to be predicted (p < .001).

Table 4
Regression analysis of demographics predicting school-aggregated SAT scores for Indiana

Variable	*B*	*SE*	β	*t*	*p*
(Constant)	666.01	39.20		16.99	.001
High school GPA	89.53	13.18	0.27	6.79	.001
High school rank	54.34	24.24	0.08	2.24	.025
Family income	104.31	23.72	0.22	4.40	.001
Parent education	172.47	25.10	0.34	6.87	.001
Race	-141.68	15.68	-0.32	-9.04	.001

R^2 = .60, df = 5, 400, F = 121.82, p < .001

Finally, school-level aggregation was examined using the ISTEP+ data from Indiana high schools. These data provided two parallels: we were able to examine high school-level data using two different tests (SAT and ISTEP+) and we were able to compare both school- and district-level data using the same test (ISTEP+). The multiple regression showed that the demographic variables of Cognitive Skills Index, race, and family income significantly predicted 65% of the variance in school aggregated state achievement, $R^2 = .65$, $F(3, 246) = 58.69$, $p < .001$ (see Table 5). With Cognitive Skills Index removed from the equation, 56% of the variance in the percent passing the ISTEP+ at the high school level was accounted for by family income and race.

Table 5
Regression analysis of demographics predicting school-aggregated ISTEP+ proficiency

Variable	*B*	*SE*	β	*t*	*p*
(Constant)	-54.66	21.65		-2.43	.017
Cognitive Skills Index	1.13	0.20	0.47	5.78	.001
Family income	-0.49	0.12	-0.40	-4.01	.001
Race	-0.06	0.03	-0.08	-0.47	.639

$R^2 = .65$, $df = 3, 246$, $F = 58.69$, $p < .001$

Individual-Level Data

At the individual student level, high school grade point average, high school rank, family income and parent education, and race were significant predictors of SAT scores ($R^2 = .41$, $p < .001$). In a second regression with only parent education and race, the amount of variance predicted dropped to 16%.

Comparisons of Top- and Bottom-Performing States, Districts, Schools, and Individuals

To further demonstrate the role of demographics in predicted variability at all levels, we compared the means for each of the demographic factors for top and bottom performing units (states, districts, schools) at each level. At the state level, we compared the contribution of each demographic factor for the top ten and the bottom ten SAT-scoring states (see Table 6). The top ten scoring states had students with significantly higher GPA, class rank, parent education and parent income and significantly fewer African American students than did the bottom ten states.

Demographic differences between high and low performing school districts were highlighted further by comparing the characteristics of the ten Indiana school districts with the highest ISTEP+ scores to the ten school districts with the lowest scores (see Table 7). To highlight the school-level differences in demographics, the characteristics of the 15 highest scoring schools and the 15 lowest scoring schools on the SAT were compared (see Table 6). As with states, students from the highest scoring schools had significantly higher GPA, class rank, parent income, and parent education, and significantly fewer African Americans. Similar comparisons for both the SAT (Table 6) and the ISTEP+ (Table 7) were made for the top 15 and bottom 15 scoring schools in Indiana.

Table 6
Means, standard deviations of demographic predictors of SAT scores: State, school, individual

Unit of analysis and set of interest	SAT total	GPA	Class rank	Parent educ	Family income	Race
States in United States (R^2 = .94, n = 51)	1,066 (67)	3.41 (0.22)	31 (10)	37 (10)	31 (09)	10 (10)
Top 10 states	1,165 (20)	3.63 (0.09)	42 (05)	50 (03)	43 (07)	5 (04)
Bottom 10 states	989 (14)	3.22 (0.13)	21 (02)	28 (04)	24 (02)	20 (17)
High schools, nation (R^2 = .63; N = 14,432)	1,018 (124)	3.38 (0.31)	33 (20)	30 (21)	26 (22)	11 (21)
Top 15 high schools	1,407 (20)	3.51 (0.25)	43 (19)	74 (14)	64 (18)	8 (07)
Bottom 15 high schools	628 (33)	2.92 (0.28)	24 (19)	8 (10)	2 (04)	54 (33)
Indiana high schools (R^2 = .60; n =406)	987 (67)	3.16 (0.20)	24 (10)	22 (13)	20 (14)	5 (15)
Top 15 high schools	1,148 (58)	3.36 (0.16)	29 (15)	49 (20)	39 (23)	5 (08)
Bottom 15 high schools	791 (48)	2.76 (0.24)	22 (12)	12 (11)	6 (09)	62 (39)
Individuals (R^2 = .41; N = 1,219,550)	1,018 (207)	3.28 (0.65)	23 (42)	31 (46)	28 (45)	12 (32)
Perfect score (n = 464)	1,600 (0)	4.09 (0.27)	93 (26)	83 (38)	66 (47)	0 (00)
Lowest score (n = 330)	400 (0)	2.66 (0.65)	5 (22)	9 (29)	2 (15)	51 (50)

Class rank = percentage of test-taking students ranking in top decile of class; parent education = percentage of parents with bachelors degree or higher; family income = percentage with income above $80,000;
race = percentage African American

Table 7
Means, standard deviations of demographic predictors of ISTEP+ scores: District, school

Unit of analysis and set of interest	ISTEP+ Pass %	Cognitive Skills	Family Income	Race
Districts in Indiana (R^2 = .70, n = 261)	61 (12)	106 (05)	12 (09)	5 (11)
Top 10 states	87 (05)	114 (04)	3 (01)	3 (03)
Bottom 10 states	31 (04)	95 (03)	39 (13)	34 (25)
High schools in Indiana (R^2 = .65; N = 406)	59 (16)	106 (05)	14 (12)	10 (20)
Top 15 high schools	89 (05)	114 (03)	2 (02)	3 (03)
Bottom 15 high schools	10 (08)	93 (05)	45 (25)	36 (29)

Family income = percentage with free lunch; race = percentage African American

Finally, at the individual level, demographic characteristics of those test-takers receiving the highest possible score on the SAT (1600) and those test-takers receiving the lowest possible score on the SAT (400) were compared (see Table 6). Differences between the proportions of variance explained at the state and individual level were remarkable. Figure 1 emphasizes the extreme role of demographics in predicting higher-level aggregated scores.

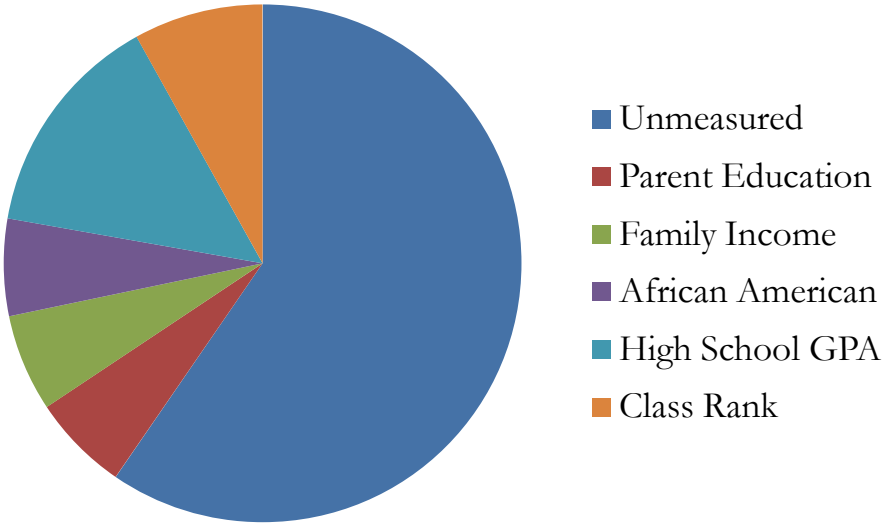

Figure 1a. Predicted variance distribution by factors among individual SAT scores.

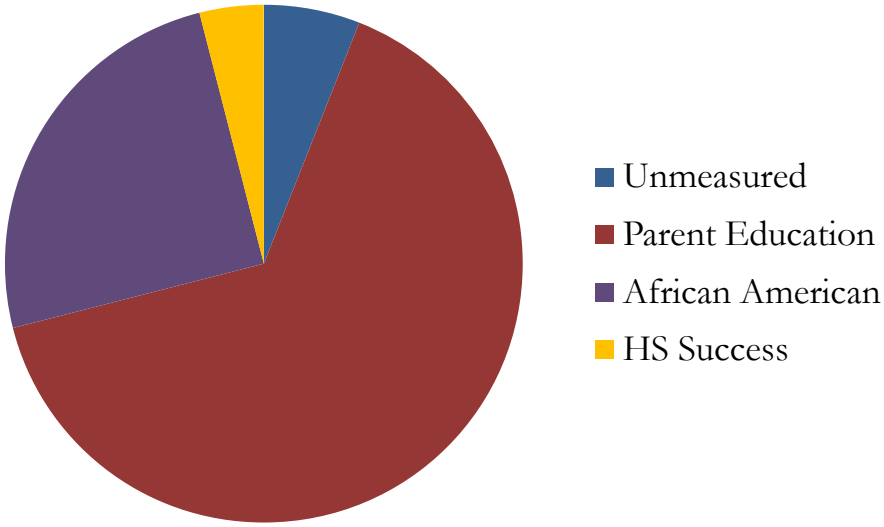

Figure 1b. Predicted variance distribution by factors among state aggregated (mean) SAT scores.

Disaggregation

One problem with disaggregation is that it ignores the confounding and cumulative effects that occur when students belong to more than one disaggregation category. This problem can be seen best by looking at the cumulative percentages of students across all demographic categories that are known to be related to lower SAT scores for each state (see Figure 2): percentage African American, percentage from families with incomes less than $80,000, percentage with parents without a college education, percent from the bottom 90 percent of their high school class, and lower GPAs. Note that cumulative percentages are greater than 100 percent in all states, given that most students fall into more than one disadvantaged category. As would be expected, the states with the lower SAT scores have a higher combination of disadvantages. Although it is possible that disadvantaged

test-takers in the lower scoring states are each falling into only one discrete category, it is more likely that they are falling into multiple categories. In fact, the variable created by the combination of test-disadvantaged characteristics was more highly related to SAT scores ($r = .89$, $p < .001$) than any single characteristic.

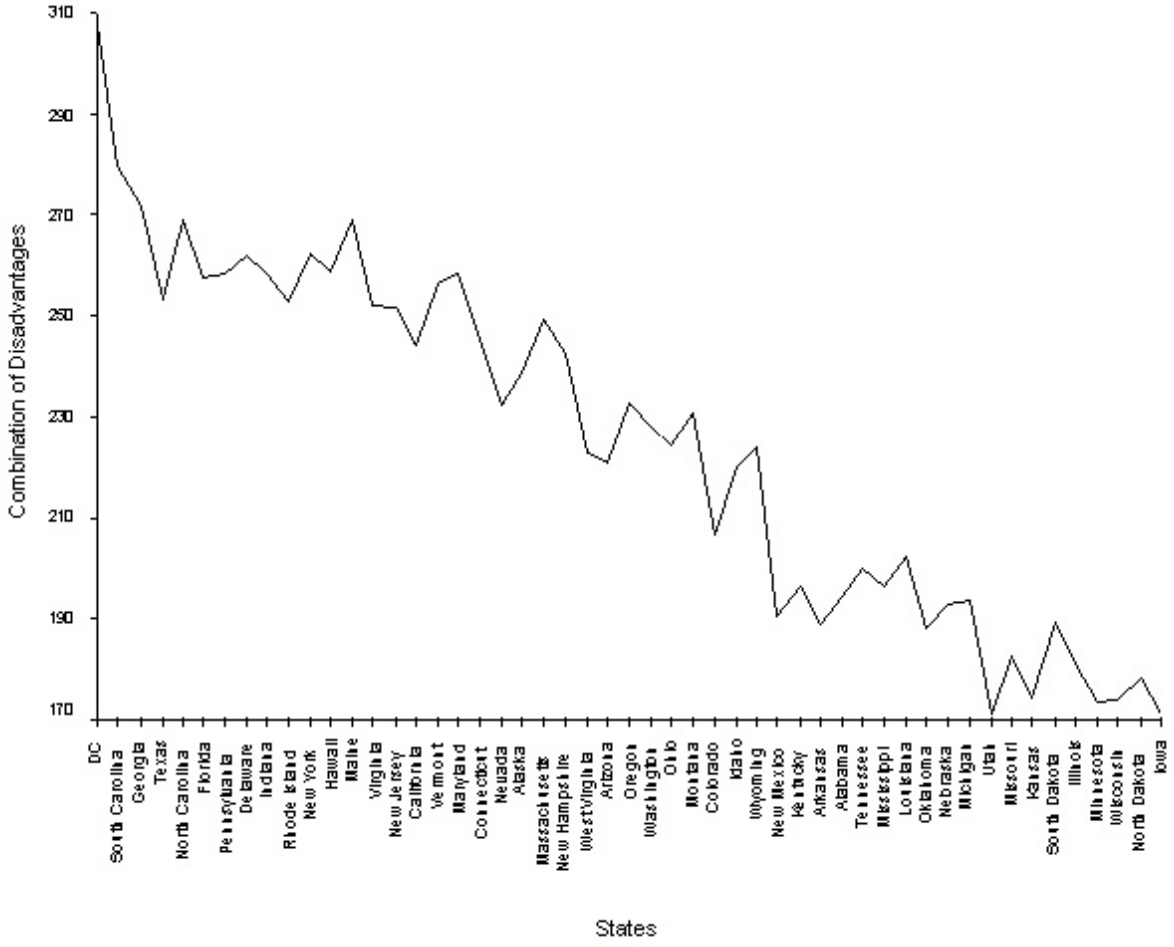

Figure 2. Total percentage of students in all demographic categories by state ranked from lowest to highest aggregated mean SAT scores.

Statistically, disaggregation assumes that each disadvantaged category of students has a main effect on achievement outcomes. Although this may indeed be true, examining only main effects ignores interactions between categories. Using the SAT data, an analysis of variance that examined the effects of family income and race on SAT scores yielded significant differences on both factors (see Table 8). However, the interaction also was significant, showing that students who belong to both disadvantaged categories have significantly lower scores than either individual category.

Table 8
Interaction of race and family income in predicting SAT scores: mean and standard deviations

	Lower income	High income	Total
African American	843.50 (175.27)	964.99 (190.69)	855.79 (180.64)
Not African American	1,003.47 (192.14)	1,111.57 (187.11)	1,037.04 (197.05)
Total	980.01 (198.02)	1,105.53 (189.51)	1,015.93 (203.68)

Note. ANOVA showed main effects and interaction of income and race on SAT scores to be significant, $p < .001$.

The problem of students belonging to multiple categories is demonstrated further by comparing SAT scores from individual states. Here, we focus on SAT scores from Delaware and Louisiana disaggregated for African American test-takers (about 18% of the test-takers from both Delaware and Louisiana were African American). Louisiana's mean SAT score was 126 points higher than for Delaware and 140 points higher for African American test-takers. These results were not surprising considering that Delaware had half as many of their African American test-takers from the top decile of their high school class as Louisiana, and Delaware's African American test-takers had lower GPAs. Louisiana had a greater percentage of their African American test-takers from families with incomes over $80,000 and a larger proportion of their test-takers' parents had college degrees. To say that Louisiana is preparing its African American test takers better than Delaware would be invalid, because their membership in advantaged income and education groups has been ignored in such an assumption.

Discussion

Numbers are seductive. Numbers put concrete values on abstract things. Numbers create meaning when none is apparent. It is not surprising that when the post-Sputnik United States experienced beliefs of educational inferiority, numbers showed up offering a deceptively simple answer to a vague and complex problem. Paired with new approaches to education in general and instruction in particular, ideas of measuring outcomes, rewarding the good, and apparently more importantly, punishing the bad found their way into our schools. Over time, this simple behavioral approach thought to improve achievement found its way into federal law. The current system of assessing outcomes as a means of holding schools accountable for their performance, established by NCLB, has become the educational intervention itself. The instruments for assessing change have become the instruments for producing change. Inherent within NCLB is the belief that testing will produce improved achievement, which in turn is measured by the tests. Inherent as well are the beliefs that teachers' instructional practices and schools' educational policies are the causes of the improved achievement; it is assumed that test scores are valid assessments of teaching and learning.

Contrary to these assumptions, this is what we know: The meaning of aggregated test scores changes at different levels of aggregation because the factors that are inherent within individual scores interact when they are summarized at an aggregate level. Individual students' scores have prior academic skills and family background imbedded in them. As much as 41% of the variation among students can be accounted for by innate and contextual factors that students bring with them when they come to school. When these data are summarized at increasingly higher levels of aggregation, these background differences account for greater proportions of variance. From a statistical perspective, this would be expected. What this means in practice, however, is that the meaning of the scores has changed and the conclusions that can be drawn from the scores has changed. Unfortunately, aggregated scores are being used at every level of aggregation to draw

similar conclusions about quality of instruction and education. Arguments concerning the meaning (i.e., the validity) of individual test scores were posed even before the establishment of NCLB, yet they are still used to make high stakes decisions for teachers, schools, and districts.

The results of this study showed that the variance among states' test scores could be predicted by knowing the demographic characteristics of the students within each state. As much as 94% of the state variation in SAT scores can be explained by students' academic standing, parent income and education, and race. In fact, 90% of the variance can be predicted with parent education and race. To assume that differences in states' SAT scores are an indication of differences in their educational quality is not valid, based almost solely on selection bias. Using states' standardized test scores developed under NCLB would be even more invalid on the basis of differences in instrumentation among states (each state's test is different). However, it has not historically been unusual to see the media make inferences about the quality of states' education systems based on their test scores. Similarly, state lawmakers use such comparisons to either praise or admonish their state's ability to educate its children.

Similarly, a major portion of the differences among schools and school districts is also a function of student demographic characteristics that are not under the control of schools or teachers. In Indiana, as much as 70% of the variance among school districts and 65% of the variance among individual schools on the state standardized achievement test (ISTEP+) can be predicted by students' inherent academic skills, family income, and race. Similarly, 63% of the variance in schools' SAT scores among high school nationwide (and 60% in Indiana alone) can be explained by students' academic skills, parent income and education, and race. To use these scores to pass judgment about a school or district's quality of education is simply not valid. And to use these scores to infer differences in the educational quality of schools or districts is unethical. To assume that these test scores are valid indicators of teachers' instructional skills or schools' educational policies is not justifiable on either theoretical or empirical/experimental grounds.

Perhaps knowing the proportion of variance in test scores contributed by demographic characteristics would provide for better judgments or comparisons within accountability systems. For example, in a hypothetical school, Teacher A has 20 students: 10 are African American, 5 are disabled, and 5 are from low SES families. Her aggregated test scores show that 60% of her students reached proficiency. In another school, teacher B has 20 students: all 20 are White, no students are disabled, and all students are from middle/upper income families. Her aggregated test scores show that 75% of her students reached proficiency. Under current accountability systems, Teacher B would be deemed a better teacher; and her school would meet AYP, because all of its students have similar demographic traits. However, by controlling for the known contribution of certain demographic characteristics to average test scores at varying levels of aggregation, expected aggregated test scores could be computed. In this example (using state-wide data from Wisconsin), we found that Teacher A had more students reach proficiency *than would be expected* based on their demographics; whereas Teacher B had fewer students reach proficiency than would be expected. In a recent study, a comparison of the "best" and "worst" schools in Indiana based on aggregated scores showed that when expected scores were computed by controlling for demographics, several of the "top" schools performed well below expectations and several of the "bottom" schools performed well above expectations (Marchant, Grizzle, Morales Ordonez, & Paulson, 2008). The current judgments being made under NCLB using static accountability systems are not valid.

Despite the development of accountability systems (e.g., growth models and value-added models) to control for factors that are beyond the control of teachers and schools, the influence of students' inherent cognitive skills and students' family background on their school performance cannot be eliminated completely. In addition, value-added systems tend to be costly and may be difficult for even the most technically sophisticated school districts to implement without

consultants. Moreover, they are most useful at the lowest levels of aggregation (student or classroom) and will do little to eliminate invalid comparisons that continue to be made at school, district, and even state levels.

Perhaps even more troubling is that information about the demographic factors that are known to explain the greatest proportion of variance in scores has become increasingly more difficult to get. Educational policies are moving away from making that information available. For example, Indiana no longer requires the Cognitive Skills component of the state achievement test, despite our knowing that one of the most important variables included in value-added assessments during the elementary years is the cognitive ability of the students (Raudenbush, 2004). Similar movements have been seen at the national level. In September of 2003, the board that was established as a result of *No Child Left Behind and* that now oversees the National Assessment of Educational Progress (NAEP) decided to sharply curtail the survey data that is collected with the tests (Schemo, 2003). The board stated that the surveys were too intrusive and burdensome and had little to do with the mission of providing a snapshot of achievement.

Another means of overcoming arguments regarding the influence of demographic factors on aggregated test scores is to disaggregate scores for individual groups of students based on family income, race, or special education designations. However, while these data provide information about the main effects of a single demographic trait, they ignore the additive and interactive effects that multiple factors working in concert have on scores. Although identifying categories of students in need of assistance may be helpful, variations in highly correlated factors that influence achievement (e.g., race and income) render simple comparisons inappropriate.

There is increasing evidence that, as part of a punitive accountability system, numbers are taking a toll on the qualities previously valued in our education system. The case against high-stakes testing on the basis of what it does to the nature of instruction, the impact on children, and the cost resulting in the depletion of other educational resources is fairly well established (Marchant, 2004). The unfortunate negative consequences associated with the use of aggregated scores that drive this accountability system are even less tolerable given the knowledge that these numbers are not valid representations of educational quality. The potential negative effect of using simple numbers in educational decision-making should come as no surprise, a warning is found in a well established principle of social science known as Campbell's law (Campbell, 1975 as cited in Nichols & Berliner, 2005, p. 4): "The more any quantitative social indicator is used for social decision making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor." Thirty years later, similar arguments concerning the validity of using single numbers as the basis for educational accountability continue to be made.

> High-stakes decisions based on school-mean proficiency are scientifically indefensible. We cannot regard differences in school mean proficiency as reflecting differences in school effectiveness…To reward schools for high mean achievement is tantamount to rewarding those schools for serving students who were doing well prior to school entry…The unjustifiable use of school-mean proficiency for high-stakes decisions will disparately affect schools serving poor children. (Raudenbush, 2004, p. 35)

In the face of inescapable validity concerns, continued use of aggregated and disaggregated scores to assess educational quality begs the question: Are those advocating the current educational accountability policy unaware of these issues, or purposefully ignoring them? Given the information available to them, educational policy makers cannot claim ignorance anymore and schools cannot ignore the potential evil inherent within the policies that govern them. The use of invalid accountability systems to make high-stakes decisions must stop; our children's education and our nation's future depend on it.

References

Baker, E. L, & Linn, R. L. (2002). Validity issues in accountability systems. *CSE Technical Report 585*. Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Ballou, D., Sanders, W., & Wright, P. (2004). Controlling for student background in value-added assessment of teachers. *Journal of Educational and Behavioral Statistics, 29*(1), 37–65.

Campbell, D. T. (1975). Assessing the impact of planned social change. In G. Lyons (Ed.), *Social research and public policies: The Dartmouth/OECD Conference*. (pp. 3–45). Hanover, NH: Dartmouth College Public Affairs Center.

Campbell, D. T., & Stanley, J. C. (1966). *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally.

College Board. (2001). SAT state comparisons: A response to Marchant & Paulson (2001). *NASSP Bulletin, 85*(627), 75–78.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues in field settings*. Chicago: Rand McNally.

Edwards, V. B. (Ed.) (2001, January). Quality counts 2001: A better balance. An Education Week/Pew Charitable Trusts report. *Education Week, 20*(17).

Grissmer, D. W. (2000). The continuing use and misuse of SAT scores. *Psychology, Public Policy, and the Law, 6*, 223–232.

Hamilton, L. S., & Koretz, D. M. (2002). About tests and their use in test-based account-ability systems. In L. S. Hamilton, B. M. Stecher, & S. P. Klein (Eds.), *Making sense of test-based accountability in education* (pp. 10–35). Santa Monica, CA: RAND.

Hamilton, L. (2003). Assessment as a policy tool. In R. L. Floden (Ed.), *Review of Research in Education, 27*, 25–68.

Hanushek, E. A., Kain, J. F., Markman, J. M., & Rivkin, S. G. (2005). Does peer ability affect student achievement? *Journal of Applied Economics, 18*(5), 527–544.

Kane, M. (2006). *Validation*. In R. L. Brennan (Ed.), *Educational measurement* (4th ed.) (pp. 17–64). Westport, CT: American Council on Education/Praeger.

Kane, T. J., & Staiger, D. O. (2002). Volatility in school test scores: Implications for test-based accountability systems. In D. Ravitch (Ed.), *Brookings papers on education policy, 2002* (pp. 235–283). Washington, DC: Brookings Institution.

Kane, T. J., Staiger, D. O., & Geppert, J. (2002). Randomly accountable. *Education Next, 2,* 56–61.

Kirby, S. N., McCaffrey, D. F., Lockwood, J. R., McCombs, J. S., Naftel, S., & Barney, H. (2002). Using state school accountability data to evaluate federal programs: A long uphill road, *Peabody Journal of Education*, *7*(4), 122–145.

Koretz, D. (2002). Limitations in the use of achievement tests as measures of educators' productivity. *Journal of Human Resources, 37,* 752–777.

Koretz, D. (2005). *Alignment, high stakes, and the inflation of test scores*. In J. L. Herman & E. H. Haertel (Eds.), Uses and misuses of data in accountability testing. *Yearbook of the National Society for the Study of Education* (Vol. 104, Part I, pp. 99–118). Boston, MA: Blackwell Publishing.

Le, V., & Klein, S. P. (2002). Technical criteria for evaluating tests. In L. S. Hamilton, B. M. Stecher, & S. P. Klein (Eds.), *Making sense of test-based accountability in education* (pp. 36–55). Santa Monica, CA: RAND.

Linn, R. L. (1998). Validation inferences from national assessment of educational progress achievement-level reporting, *Applied Measurement in Education, 11,* 23–47.

Linn, R. L., Baker, E. L., & Betebenner, D. W. (2002). Accountability systems: Implications of requirements of the No Child Left Behind Act of 2001. *Educational Researcher, 31*(6), 3–16.

Linn, R. L. (2005). Issues in the design of accountability systems. *CSE Technical Report 650*. Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing (CRESST). (ERIC Reproduction Document No. ED 488 720).

Linn, R. L. (2006a). Educational accountability systems. *CSE Technical Report 687*. Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing (CRESST). (ERIC Reproduction Document No. ED 488 720).

Linn, R. L. (2006b). Validity of inferences from test-based educational accountability systems *Journal of Personnel Evaluation in Education*, *19*, 5–15.

Madaus, G., & Clarke, M. (2001). The adverse impact of high-stakes testing on minority students: Evidence from one hundred years of test data. In G. Orfield & M. L. Kornhaber (Eds.). *Raising standards or raising barriers? Inequality and high-stakes testing in public education.* New York: The Century Foundation Press.

Marchant, G. J. (2004). What is at stake with high stakes testing? *The Ohio Journal of Science: A Multidisciplinary International Journal, 104*(2), 2–7.

Marchant, G. J., Grizzle, R., Morales Ordonez, O. & Paulson, S. E., (2008, April). *Elementary schools performing above expectations: Who are those guys?* Paper presented at the American Psychological Association (Boston, MA).

Marchant, G. J., & Paulson, S. E. (2001). State comparisons of SAT scores: Who's your test taker? *NASSP Bulletin, 85*(627), 62–74.

Marchant, G. J., Paulson, S. E., & Shunk, A. (2006). Relationships between high-stakes testing policies and student achievement after controlling for demographic factors in aggregated data. *Educational Policy Analysis Archives, 14*(30*)*. Retrieved November 11, 2009, from http://epaa.asu.edu/epaa/v14n30/.

Meyer, R. H. (2000). Value-added indicators: A powerful tool for evaluating science and mathematics programs and policies. *NISE Brief, 3*(3), 1–7.

Miller, D. M., & Linn, R. L. (2000). Validation of performance-based assessments, *Applied Psychological Measurement, 24,* 367–378.

McCaffrey, D. F., Lockwood, J. R., Koretz, D., Louis, T. A., & Hamilton, L. (2004). Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics, 29*, 67–101.

No Child Left Behind Act of 2001. (2002). Public Law No. 107–110.

Nichols, S. L., & Berliner, D. C. (2005, March). *The inevitable corruption of indicators and educators through high –stakes testing*. Arizona State University: Educational Policy Studies Laboratory. Retrieved April 1, 2005, from http://www.asu.edu/educ/epsl/EPRU/documents/EPSL-0503-101-EPRU.pdf.

Raudenbush, S. (2004). Schooling, statistics, and poverty: Can we measure school improvement? *The Ninth Annual William H. Angoff Memorial Lecture*. Princeton, NJ: Educational Testing Service. Retrieved March 30, 2005, from http://www.ets.org/research/pic/angoff9.pdf.

Rouse, C. E. (2005). Accounting for schools: Economic issues in measuring school quality. In C. A. Dwyer (Ed.), *Measurement and research in the accountability era* (pp. 275–298). Mahwah, NJ: Erlbaum.

Rubin, D. B., Stuart, E. A., & Zanutto, E. L. (2004). A potential outcomes view of value-added assessment. *Journal of Educational and Behavioral Statistics, 29*(1), 103–116.

Sanders, W., & Horn, S. (1998). Research findings from the Tennessee value added assessment system (TVAAS) database: Implications for educational evaluation and research. *Journal of Personnel Evaluation in Education, 12*(3), 247–256.

Schemo, D. J. (2003, Sept. 15). U.S. officials pull questions from surveys about children. *New York Times*, p. A22.

Zvoch, K., & Stevens, J. J. (2008). Measuring and evaluating school performance: An investigation of status and growth-based achievement indicators, *Evaluation Review*, *32*, 569–595.

About the Authors

Sharon E. Paulson
Ball State University

Greg Marchant
Ball State University

Email: spaulson@bsu.edu

Sharon E. Paulson is a Professor of Psychology—Educational Psychology at Ball State University. She specializes in adolescent development, advocating the importance of understanding developmental principles to teaching and learning. In addition to her work on standardized test scores, her research examines contextual factors related to adolescent achievement.

Greg Marchant is a Professor of Psychology—Educational Psychology at Ball State University. His current research focuses on high-stakes testing and the uses and misuses of aggregated test scores in accountability systems. In addition, his work examines the impact of current educational policies on schools, teachers, and students.