education policy analysis archives

A peer-reviewed, independent,
open access, multilingual journal

Arizona State University

Volume 22 Number 5 February 3rd, 2014 ISSN 1068-2341

Focusing on Short-term Achievement Gains Fails to Produce Long-term Gains

David W. Grissmer
University of Virginia
USA
David R. Ober & John A. Beekman
Ball State University
USA

Citation: Grissmer, D., Ober, D. & Beekman, J. (2014). Focusing on short-term achievement gains fails to produce long-term gains. *Education Policy Analysis Archives, 22* (5).
http://dx.doi.org/10.14507/epaa.v22n5.2014

Abstract: The short-term emphasis engendered by No Child Left Behind (NCLB) has focused research predominantly on unraveling the complexities and uncertainties in assessing short-term results, rather than developing methods and assessing results over the longer term. In this paper we focus on estimating long-term gains and address questions important to evaluating schools and identifying educational policies and practices that produce long-term sustained gains. Estimates are made of annual pass rates on state exams using fixed effect models for six years of pass rates at grades 3, 6, 8 and 10; the percentages of schools making statistically significant gains, gains, losses, and statistically significant losses in pass rates are determined. Estimates are contrasted using models that include and exclude demographic characteristics. The percentages of schools with statistically significant gains varied markedly from 38 to 6 at grades 6 and 10, respectively; the percentage of schools with statistically significant declines ranged from less than 8

Journal website: http://epaa.asu.edu/ojs/
Facebook: /EPAAA
Twitter: @epaa_aape

Manuscript received: 10/15/2012
Revisions received: 7/2/2013
Accepted: 8/5/2013

percent at grades 3, 6, and 8, to 23 percent at grade 10. Including demographics increased the percentages of schools with statistically significant gains and lowered the percentages with statistically significant declines. The results suggest that schools with higher proportions of free-reduced lunch and minority students are more likely to have statistically significant gains with demographic controls. Estimates of pass rate trends are made using Monte Carlo simulations; from these simulations the percentages of schools that may be mislabeled as having statistically significant gains and losses are determined. Even with six years of trend data, results suggest that chance can still play a significant role in mislabeling school performance, especially in grades having weak overall trends.
Keywords: accountability; longitudinal achievement; changing demographics.

Centrándose en las mejoras de rendimiento a corto plazo no produce mejoras a largo plazo .
Resumen: El énfasis de corto plazo generada por la ley NCLB ha centrado principalmente la investigación en desentrañar las complejidades e incertidumbres en la evaluación de resultados a corto plazo, en lugar de desarrollar métodos y evaluación de los resultados a más largo plazo. En este artículo nos centramos en la estimación de ganancias a largo plazo y tratamos cuestiones importantes para la evaluación de las escuelas y la identificación de las políticas y prácticas educativas que producen mejoras sostenidas a largo plazo. Se realizan estimaciones de los índices de aprobación en exámenes estatales anuales utilizando modelos de efectos fijos para los seis años en los grados 3 , 6, 8 y 10 con índices de aprobación. Se determinaron los índices de aprobación de los porcentajes de escuelas con mejoras estadísticamente significativas, mejoras, empeoramiento y empeoramiento estadísticamente significativo. Las estimaciones se contrastaron utilizando modelos que incluyen y no incluyen características demográficas. Los porcentajes de las escuelas con las mejoras estadísticamente significativas variaron notablemente 38-6 en los grados 6 y 10 respectivamente , y el porcentaje de escuelas con empeoramientos estadísticamente significativos varió de menos del 8 por ciento en los grados 3, 6 y 8 , hasta 23 por ciento en el grado 10. Incluyendo datos demográficos se aumentó los porcentajes de las escuelas con mejoras estadísticamente significativas y se disminuyó los porcentajes con empeoramiento estadísticamente significativos. Los resultados sugieren que las escuelas con mayores proporciones de estudiantes que reciben subsidios de almuerzo y con estudiantes de minorías son más propensos a tener ganancias estadísticamente significativas con los controles demográficos. Las estimaciones de la evolución del tipo de paso se realizan utilizando simulaciones de Monte Carlo, a partir de estas simulaciones se determinan los porcentajes de las escuelas que pueden ser mal identificadas como teniendo ganancias y pérdidas significativa . Incluso con seis años de datos sobre tendencias, los resultados sugieren que el azar todavía puede jugar un papel significativo en la medición del rendimiento escolar, especialmente en los grados que tienen tendencias generales de mayor debilidad
Palabras clave: responsabilidad; logro longitudinal; cambios demográficos.

Focando em melhorias de desempenho a curto prazo não produz melhorias a longo prazo
Resumo: O foco no curto prazo gerado pela NCLB fez que a pesquisa seja focada principalmente para desvendar as complexidades e incertezas na avaliação de resultados de curto prazo, em vez de desenvolver métodos e avaliação de resultados no longo prazo. Neste artigo vamos nos concentrar na estimativa de ganhos a longo prazo e tratar questões importantes para a avaliação das escolas e a identificação de políticas e práticas educacionais que produzem melhorias sustentaveis a longo prazo. Estimativas dos índices de aprovação foram realizadas em testes anuais estaduais, utilizando

modelos de efeitos fixos para os seis anos nas classes 3, 6, 8, e 10 com índices de aprovação. Foram determinadas os índices de aprovação do percentual de escolas com melhorias estatisticamente significativas, melhorias, retrocessos, e retrocessos estatisticamente significativos. As estimativas foram comparadas com modelos que incluem e excluem características demográficas. Os percentuais de escolas com melhorias estatisticamente significativas variaram acentuadamente de 38 a 6 nas classes 6 e 10, respectivamente, bem como a percentagem de escolas com retrocessos significativos variou de menos de 8 por cento em notas 3, 6 e 8-23 por cento no grau 10. Incluindo dados demográficos os percentuais de escolas com melhorias estatisticamente significativas foram aumentados e percentagens retrocessos estatisticamente significativos diminuíram. Os resultados sugerem que as escolas com maior proporção de estudantes que recebem subsídios de almoço e estudantes de minorias são mais propensos a ter ganhos estatisticamente significativos com controles demográficos. As estimativas da evolução da etapa são realizadas por meio de simulações de Monte Carlo, e com base nessas simulações o percentual de escolas que podem ser erroneamente identificado como tendo ganhos e perdas significativas foram determinados. Mesmo com seis anos de tendência de dados, os resultados sugerem que a sorte ainda pode desempenhar um papel significativo na medição de desempenho escolar, especialmente nas series que tem tendências gerais a serem mais fracas.
Palavras-chave: Responsabilidade; medidas longitudinais; mudanças demográficas.

Introduction

Whether schools have statistically significant long-term trends and whether those estimates are reliable should be of primary interest to policy makers when evaluating schools and teachers. Moreover, these long-term trends rather than short-term performance should occupy a central position when setting future education policies.

The No Child Left Behind (NCLB) Act of 2001 had as a central tenet that all children become proficient in math and reading literacy by 2014. In order to hold schools and states accountable, each state independently developed a strategy for trying to meet this long-term goal by setting a path of Adequate Yearly Progress (AYP) targets that would attain this goal. Some states set lower, more achievable goals in the short term leaving larger gains to later years. Other states projected a more linear path of similar gains over the years. From the beginning, AYP became a major focus of efforts to evaluate and compare schools, and a major preoccupation of teachers, principals, district administrators, and policymakers across states, as well as researchers who focused on assessing the reliability and interpretation of short-term results. Almost all schools in the nation received annual ratings based on AYP. Accountability was embedded in the legislation by mandating that each of several student groups identified by demographic, family income, and special education status would have to meet AYP goals in order for a school to be successful. Failure to repeatedly meet these goals triggers mandated policies for schools that included offering parents more school choices and tutoring of students.

NCLB measures have been criticized in four ways. First, the long-term performance goals have been characterized as implausible given the underlying normal distribution of scores unless the proficiency standards are set very low. Second, assessing whether AYP is met annually can often be problematical given annual score changes and statistical uncertainties in score changes can often be similar in magnitude to AYP, making AYP a poor measure on which to base rewards or sanctions. Third, the variation between states in their standards and strategies for setting AYP make the standards and strategies difficult to interpret and compare. Finally, the use of AYP may place high poverty and racially diverse schools at a disadvantage (see Mintrop & Trujillo, 2005; Kane & Staiger,

2002; Kim & Sunderman, 2005; Raudenbush, 2004; Rothstein, 2008; Linn & Haug, 2002; Linn, Baker, & Herman, 2002; Stecher, Hamilton, & Gonzalez, 2003).

Given that education is a cumulative process, short-term gains at each grade are only important if they are part of a pattern that leads to sustained long-term gains in later grades. True gains at each grade will accumulate across grades to increase high school graduation and college entrance rates. AYP in each grade has had little success in predicting and promoting longer-term gains in later grades; in fact the percentage of the nation's schools making AYP has declined from 71 percent in 2006 to 52 percent in 2011 (Usher, 2012). Focusing on AYP and using it to drive new policy have not generated practices that lead to sustained and cumulative long-term gains. Instead the short-term emphasis engendered by NCLB has focused research predominantly on unraveling the complexities and uncertainties in assessing short-term results, rather than on developing methods and assessing results over the longer term.

Short-term gains are used in Indiana's accountability system to measure improvement in performance. A grid of pass-rate performance and improvement based on year-to-year changes assigns schools and school corporations/districts to the following improvement categories: Exemplary, Commendable, Academic Progress, Academic Watch and Academic Probation; these categories have been changed recently by adding to the above designations the easy-to-understand letter grades of A, B, C, D, and F, respectively (Indiana DOE, 2011).

Problem Statement

In this paper we address a series of questions that are important to evaluating and identifying schools that produce statistically significant long-term gains and declines. First, are the same schools identified when controls for socioeconomic status and ethnicity are incorporated? Second, what percentages of schools register trends that are statistically significant (gains and declines) due to inherent randomness? Third and most important, are schools that have statistically significant long-term trends of improvement (and decline) being properly identified with short-term annual measures?

The current study analyzes Indiana's test performance across four grade levels (grades 3, 6, 8, and 10) and over a six-year time period (fall 2002 through fall 2007). We estimate the number and proportion of schools at each grade making statistically significant (95 percent confidence) gains and losses over this six-year period, and assess how these estimates change if changing family characteristics are included in the estimation. We take account of uncertainty in achievement scores that can mislabel schools by using Monte Carlo simulations (see Winston, 2004; Metropolis & Ulam, 1949) that estimate the number of such mislabeled schools, thereby providing an indicator of the reliability of the state's system used to label a school's performance. Such an indicator can better guide educational policies especially those that provide rewards or sanctions to schools. When evaluating schools and teachers, the reliability of the state's system for determining school performance and whether schools have statistically significant long-term trends should be of primary interest to policy makers.

Background and Evolution of State Accountability System

Since 1988 Indiana has been administering at multiple grade levels the Indiana Statewide Testing for Education Progress (ISTEP) in English/Language Arts and Mathematics to assess and improve student learning. In 1995, the ISTEP exams were redesigned to measure student achievement of the state content standards. In 1998 legislation was passed that required the grade 10 ISTEP exam to be used as an additional requirement for graduation beginning in 2000. With the implementations of Public Law 221 (PL 221) in Indiana in 1999 and NCLB at the federal level in 2001, the purpose of these exams was expanded to include use of 3rd to 10th grade scores both as a

measure of AYP and as a measure of school performance and improvement under PL 221. Thus the achievement scores used in this analysis at 3rd, 6th, 8th and 10th grade are currently being used in Indiana for PL 221 accountability purposes in grades 3 through 8.

Literature Review

Many researchers have presented evidence and argued that NCLB was flawed for at least four reasons. First, the long-term performance goals have been characterized as implausible given the underlying normal distribution of scores unless the proficiency standards are set very low. Second, meeting AYP can be a poor measure on which to base rewards or sanctions because of the inherent uncertainty in annual score changes. Third, the variation between states in their standards and strategies for setting AYP make them difficult to interpret and compare. Finally, the use of AYP may place high poverty and racially diverse schools at a disadvantage (as noted earlier see Mintrop & Trujillo, 2005; Kane & Staiger, 2002; Raudenbush, 2004; Rothstein, 2008; Linn & Haug, 2002; Linn et al., 2002; Kim & Sunderman, 2005; Stecher et al., 2003). Rogosa (2005) has provided some cogent response to some of this criticism, and argues that not all blame should reside with NCLB, but with flawed estimation, application and interpretation of statistical results by the research community resulting often in poor advice to policymakers and ineffective policies.

Policymakers inevitably return to two long-term goals–first, closing international score gaps, and, second, closing national achievement gaps between racial/ethnic groups and advantaged/disadvantaged students. Gaps of these two types typically can be in the range of 0.5 to 1.25 standard deviation depending on the test, subject, and grade; in the various international tests, the comparison group of countries influences the gaps. Empirical evidence across NAEP, PISA and TIMMS suggests that the largest annual sustained gains from any country or state in any subject over the last 20–25 years tend to be about 0.07 standard deviation. For instance, the largest annual gains in NAEP scores from 1990–2007 occur for 4th grade math with annual gains of 0.05 standard deviation. A few individual states with low beginning scores in 1990 made annual gains as large as 0.07 standard deviation or about two percentile points per year. Perhaps the largest sustained NAEP gains occurred for cohorts of Black students entering school from 1970–1980 where annual gains were as large as 0.07 standard deviation a year for about 10 years. These occurred in the reading scores for 17-year-old students and in the math scores for 9- and 13-year-old students (Grissmer, Kawata, & Williamson, 1998).

Experimental evidence from interventions suggests that annual gains of 0.07 are unusual. For instance, the Project Star experiment of lowering class size by approximately seven students over the first four years of school showed overall effects of about 0.20 standard deviation and effects of about 0.30 standard deviation for Black students (Finn & Achilles, 1999; Krueger, 1999). Combining such studies yielded average annual gains over four years of 0.05 to 0.07 standard deviation units as a result of a very substantial and costly reduction in class size (Brewer, Krop, Gill, & Reichardt, 1999**)**. These gains were approximately equal for reading and math. However, the gains coming from reduced class sizes were not fully sustained in the long term, but were reduced by about one-half by 8th grade (Krueger & Whitmore, 2001). Even with very highly sustained annual gains of 0.07 standard deviation, it would take 10–20 years to eliminate the gaps desired by policymakers. In summary, the authors believe it is more important to focus on measuring and explaining historical changes in longer-term trends as opposed to trying to assess, measure, and interpret short-term achievement gains.

Perhaps the major flaw of NCLB was that the focus on annual gains took the public attention off research measuring long-term gains and explaining the pattern of long-term gains. If

short-term gains could reliably predict long-term gains and if the policies and practices that produce short-term gains are the same ones that produce long-term gains, this approach would not be problematical. However, short-term gains can be the result of four misleading causes. First, random variations provide false signals to teachers, schools and policymakers. Second, emphasis on short-term interventions and policies encourage memorization rather than critical thinking. Third, teaching test-taking techniques disturbs the true test scores both positively and negatively. Fourth, narrowed curricula create gains at the expense of knowledge in other subjects (see Marion et al., 2002; Wiley, Mathis, & Garcia, 2005; Yeh, 2005: Hamilton & Stecher, 2006; Stecher & Hamilton, 2002).

The policies that drive short-term gains may be very different from those driving long- term gains. Moreover, the policies for mathematics may differ from those for reading. For instance, the long-term large gains in 4th and 8th grade NAEP math scores from 1990–2009 of 1.5 and 1.2 percentile points a year, respectively, were in contrast to much smaller 4th grade reading gains of 0.2 percentile points a year and no 8th grade reading gains. Thus, the policies that would be expected to affect both reading and math similarly could not explain these large differentials in trends. Policies that might be largely expected to affect both subjects might be class size reductions, standards' based accountability, improving teacher quality, and increasing pre-school attendance. However, these policies would be unable to explain the large differential between math and reading gains. These math gains would have to be explained by subject specific factors like changes in curriculum, better and more widely accepted math standards, or greater alignment between math standards and NAEP tests. A focus on analyzing long-term trends rather than short-term gains would likely identify different successful policies, and these policies would have the advantage of being linked to long-term sustained gains. Policies identified through short-term analysis must still be empirically tested over the long term in order to be viable, and many such policies may fail to be sustainable. Raudenbush (2004) suggests that three years (or longer) are needed to determine whether newly implemented strategies (supported with appropriate assessment data) have been successful.

One issue that arises in estimating long-term trends is whether the inclusion of socio-demographic variables provides better estimates for policymaking when comparing schools than does their exclusion. Research has long established since the Coleman report (1966) that socio-demographic characteristics account for most of the explainable variance in scores, and thus if the socio-demographic characteristics change across years, the scores will be affected. The argument favoring their inclusion is that schools cannot control their student population, and so comparison across schools should remove these effects before comparing trends. The argument against inclusion is that a component of quality schools is their capacity to adjust and accommodate changes such as student demographics. In any case, an important consideration is to estimate how much inclusion of demographic changes alters the number and characteristics of schools that have statistically significant gains and losses.

Brown (2008) compares results with and without socio-demographic characteristics from North Carolina's accountability system. That study suggests there are significant changes in growth rates and school ratings with the inclusion of demographic data. Thompson (2004) analyzed five years of achievement data from Milwaukee elementary schools to assess the importance of incorporating demographic characteristics and the stability of school rankings over time. The author substantiates the stability of school ratings by using an earlier rating to predict ratings four years later. While there are positive and significant relationships between the two ratings, only 18 percent of the variance is explained indicating that short-term gains are weak predictors of long-term gains. The author uses a poverty measure to adjust ratings and concludes that including the poverty measure can have significant effect on the ranking of schools.

A second issue is that use of long-term trends does not protect against randomness or luck affecting a school's ranking. The analytical question is how much different the school ratings would be if the scores were known with no error. Luck works in both directions placing some schools' trends in the statistically significant category when perfectly accurate scores would rate them as insignificant. However, these schools are at least partly, if not wholly, offset by schools with insignificant trends when accurate scores would show that their trends are significant. Teachers and the public want to know what proportions of schools are misclassified, i.e., what proportions of schools that are rated as having statistically significant trends might be there due to luck and what proportion actually had significant trends, but luck placed them in the insignificant category. This proportion is primarily dependent on at least two factors: the amounts of random errors in the scores and the length of the time series underlying the trends. Less score error and a longer time series will produce more reliable ratings. One factor underlying the amount of random error is the number of students at each grade in the school taking the test. Since elementary schools and rural schools have smaller grade specific populations compared to middle and high schools and urban and suburban schools, misclassification will more often occur in elementary grades and rural schools. Awareness of the reliability of the ratings will help policymakers determine how many years of data to use in applying school sanctions and rewards. We estimate the expected proportion of misclassified schools by Monte Carlo simulations and discuss their implications.

For the purpose of this study, short term will refer to using data that includes the most recent year and the data from the previous year (or an average of two or more previous years). It then follows that long term refers to using data that is from three or more years in the study by Raudenbush (2004), five years in the study by Thompson (2004), and six years for the current study.

Methodology and Data

Publicly available pass-rate data (http://mustang.doe.state.in.us/SAS/sas1.cfm) from the Indiana Statewide Testing for Educational Progress (ISTEP) were analyzed for grades 3, 6, 8, and 10 for exams that were administered from the fall of 2002 through the fall of 2007. Pass rates for English/Language Arts (ENLA), Mathematics (Math), and BOTH subject areas were investigated. Schools that had 30 or more students in classes at each grade level during the six-year period were included in the current study. Table 1 is a summary of the number of schools, range of school sizes, average school size, and number of students at each grade level in this study. Demographic and school level data are taken from Indiana statistics at the school level that partially relies on U.S. Census data.

Table 1
Summary of the Indiana Public School Populations Included in the Analysis

Grade	Number of Schools	Range of School Sizes[1]	Average Size	Students per Year in Study
3	862	30–190	77	67,380
6	455	30–701	135	62,250
8	378	30–663	192	74,750
10	334	30–1084	232	74,350

[1] Schools with less than 30 students were eliminated from the study.

Methodology

The regression analyses that were carried out in this investigation followed the methodology used by Grissmer, Flanagan, Kawata, and Williamson (2000) and Grissmer and Flanagan (2006) on

state NAEP data. In the current study the following estimations and/or predictions have been made using pass rates for BOTH (students passing both ENLA and Math) at grades 3, 6, 8, and 10 at the state and school levels for the six-year time period. We estimate with models using fixed effects and panel data sets by school for their pass rates from 2002–2007. We make separate estimates by grade. We estimate two versions for each model, no family controls and with family controls.

State-wide Gains from Base Year

The equation to estimate state-wide gains while controlling for family variables is as follows:

$$y_{ij} = a + \sum f_k F_{ijk} + g_m d_{2002+m} + e_{ij} \quad (1)$$

where y_{ij} is a percentage pass rate on a z-scale that has been normalized to the fall 2002 pass rates for the i-th school (i = 1,N schools) in the j-th year (j = 1,6); F_{ijk} is the k-th family variable for the i-th school in the j-th year; d_{2002+m} is the m-th dummy state gain variable (m = 2,6) measured from the fall 2002 baseline year to year m; e_{ij} is the error term for the i-th school in the j-th year; and a, f_k, and g_m are coefficients of the regression analysis.

School-level Trends

Annualized school trends that control for family variables are estimated by

$$y_{ij} = a + g_i T_j + \sum b_k F_{ijk} + u_i + e_{ij} \quad (2)$$

where g_i is the annualized estimated gain for school i, T_j is the trend variable (j = 1,6), u_i is the fixed effect for school i, and the remaining variables are defined above. It should be noted that u_i is an unobserved factor for each school that does not vary over time (six years).

It is seen that the above models do not make use of the performances of demographic subgroups. Therefore, unusual improvement (or decline) by a single subgroup at the school level can only be identified through the state AYP measures required by NCLB.

Data

Table 1 shows the numbers and sizes of the schools and student populations included in the analysis by grade. Average school size approximately triples from elementary schools (3rd grade) to high schools (10th grade) making school trend estimates more uncertain at the lower grades.

Table 2 presents the family and school variables used in the study and their source. The variables that were included in the models of Equations 1 and 2 were chosen on the basis of their significance in adding predictive strength to the models. Even though school districts, administrators, and teachers have no control over these variables, the state does not control for any of these demographics when measuring school performance or improvement under law PL 221.

Figure 1 shows pass rates for ENLA, Math, and students passing both ENLA and Math (called BOTH) for grades 3, 6, 8, and 10 between 2002 and 2007. Since the fall of 2002 there have been significant differences by grade in annualized rates of gain/loss in percentage pass rates. The BOTH annualized rates of change are 0.6, 1.6, 0.9 and -0.6 percent/yr, respectively, for grades 3, 6, 8, and 10 (2002–2007).

The corresponding pass rates for BOTH during the previous six years (1996–2001) are 0.7, -1.1, -0.1, and 0.9 percent/yr for grades 3, 6, 8, and 10, respectively. The large annualized change for grade 6 (-1.1 to +1.6 percent/yr) was due to a rescaling of grade 6 exams by the state; the BOTH pass rates between 2001 and 2002 changed from 46.0 to 59.0 percent, respectively. These two sets of six-year rates of gains and declines for grades 3, 6, 8, and 10 demonstrate the lack of sustained improvement over the 12-year period 1996–2007; the six-year period of 2002–2007 was after the

dates that accountability measures associated with NCLB and Indiana's PL 221 became effective in 2001 and 1999, respectively.

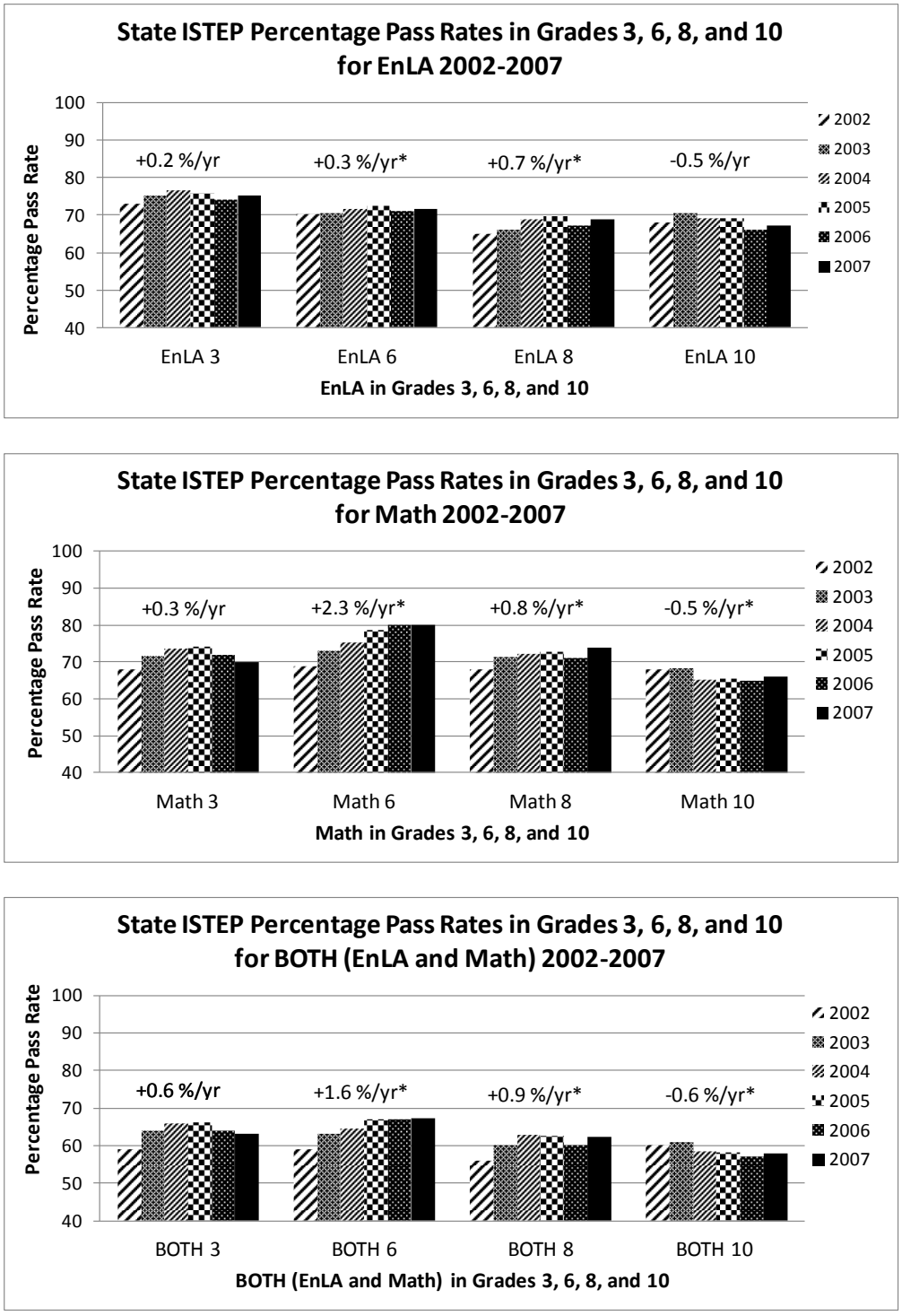

Figure 1 State ISTEP pass rates for ENLA, Math, and BOTH (ENLA and Math) are presented for grades 3, 6, 8, and 10 for the time period of data investigated in this study – fall 2002–2007. Statistically significant (95% confidence) growth/decline rates are designated with an asterisk*.

Table 2
Level of Aggregation and Source of Variables Used in the Analysis

Grade level	Grade-level Demographic Percentages	School-level Teacher Characteristics	Corp-level Indiana Records	Corp-level 2000 Census
Annual Pass Rate for ENLA, Math, and BOTH	Free-Reduced Lunch	Age Salary Experience	Per Pupil Expenditures	Parent Education Less Than HS HS Education
	Ethnicity American Indian Black Asian Hispanic White Multi-Racial		Student/Teacher Ratio Ratio of 1st to Kindergarten Ratio of First to Preschool	Some College BS Degree Household Head Married Couple Single Male Single Female
	ESL – LEP Special Education			Median Income

These rates of gain are somewhat different than the typical pattern of NAEP scores where the largest gains are for lower grades but are lower for higher grades. One should also be aware of the uncertainties associated with an average Indiana school's performance and improvement (or decline). Standard errors for a 60-percent pass rate and for average Indiana school sample sizes in Table 1 range from 3.2 percent to 5.6 percent for grade 10 and grade 3, respectively.

These uncertainties become 4.5 percent and 7.9 percent for grade 10 and grade 3, respectively, for standard errors associated with the differences in pass rates between two successive years. The smallest NCLB subgroup (30 students) will have standard errors of 8.9 percent and 12.6 percent associated with yearly pass rates and differences in pass rates between two successive years, respectively.

Table 3 provides the demographic, family and school characteristics of the top 10th percentile and bottom 10th percentile of schools ranked according to their percentage of students passing both English and Math tests. These data show the typical contrasts in achievement based on family/demographic characteristics. Schools in the top 10th percentile have pass rates of 85 percent while the bottom 10th percentile have pass rates of 36 percent. The bottom scoring schools compared to the top scoring schools have substantially higher populations of minorities, higher populations of single parent homes, and less educated parents with lower incomes. The lower scoring schools also have higher proportions of special education and ESL students, and these schools are much more likely to be in metropolitan and rural areas. However, the bottom scoring schools have somewhat higher funding per pupil and lower teacher-student ratios. Figure 2 shows pass rates for 2007 by school location; the rates follow the well-known patterns of lower scores in metropolitan areas, higher scores in suburban areas, and towns and rural areas scoring between metropolitan and suburban areas. An important question is whether the consistent patterns in pass rates by demographic characteristics and school location predict which schools are making statistically significant gains and losses. That is, will schools making the strongest (weakest) gains have family and location characteristics similar to those that predict the highest (lowest) scores?

Table 3
Average Family, Education, and School Demographic Variable Percentages, Expenditures, and Ratios Across Grade Levels (3, 6, 8, and 10) of Indiana's Lowest 10 percent and Highest 10 percent Performing Schools (N = 2029)

Variable	Percents, Ratios, and Expenditures		
	Lowest 10 percent	Highest 10 percent	Difference Low-High
ISTEP Pass Rates	36	85	-49
Free-Reduced Lunch	67	17	50
White	41	89	-48
Black	36	2	34
Hispanic	16	3	13
Multi Racial	6	3	3
Married Couple	62	82	-20
Single Female	30	13	17
Single Male	8	5	3
Less than HS Ed	23	14	9
BS Degree	12	21	-9
Median Fam Income	$44k	$62k	-$18k
Special Ed	18	12	6
ESL-LEP	10	2	8
Student Tea Ratio	17	18	-1
Expenditure/Student	$12k	$10k	$2k
Metropolitan	70	22	48
Suburban	16	46	-30
Town	6	2	4
Rural	7	30	-23
School Type Total	100	100	

Presented in Table 4 are the 2002 percentages of Indiana children at grades 3, 6, 8, and 10 receiving free-reduced lunches and the state-wide percentages of White and Hispanic children; also presented in the table are the corresponding percent per year trends of these family variables between 2002 and 2007. The annual percentages of Black children were relatively steady during this time. The state-wide percentage of free-reduced lunch children in 2002 was 35.2 percent in grade 3 with a declining percentage across grades to 22.0 percent by grade 10. These same free-reduced lunch percentages across grade levels had annual increases of 1.3 to 1.9 percentage points between 2002 and 2007.

However, when the six-year annual demographic trends of the 2029 schools studied in this investigation are examined individually, a pattern emerges across grade levels indicating annual demographic changes can show wide variations between schools. At grade 3 the 2002–2007 free-reduced lunch trends averaged 6.6 percent per year increases and -2.4 percent per year decreases for the most rapidly increasing and decreasing deciles, respectively. At grade 10, the corresponding decile increases and decreases were 5.3 and -1.4 percent per year, respectively.

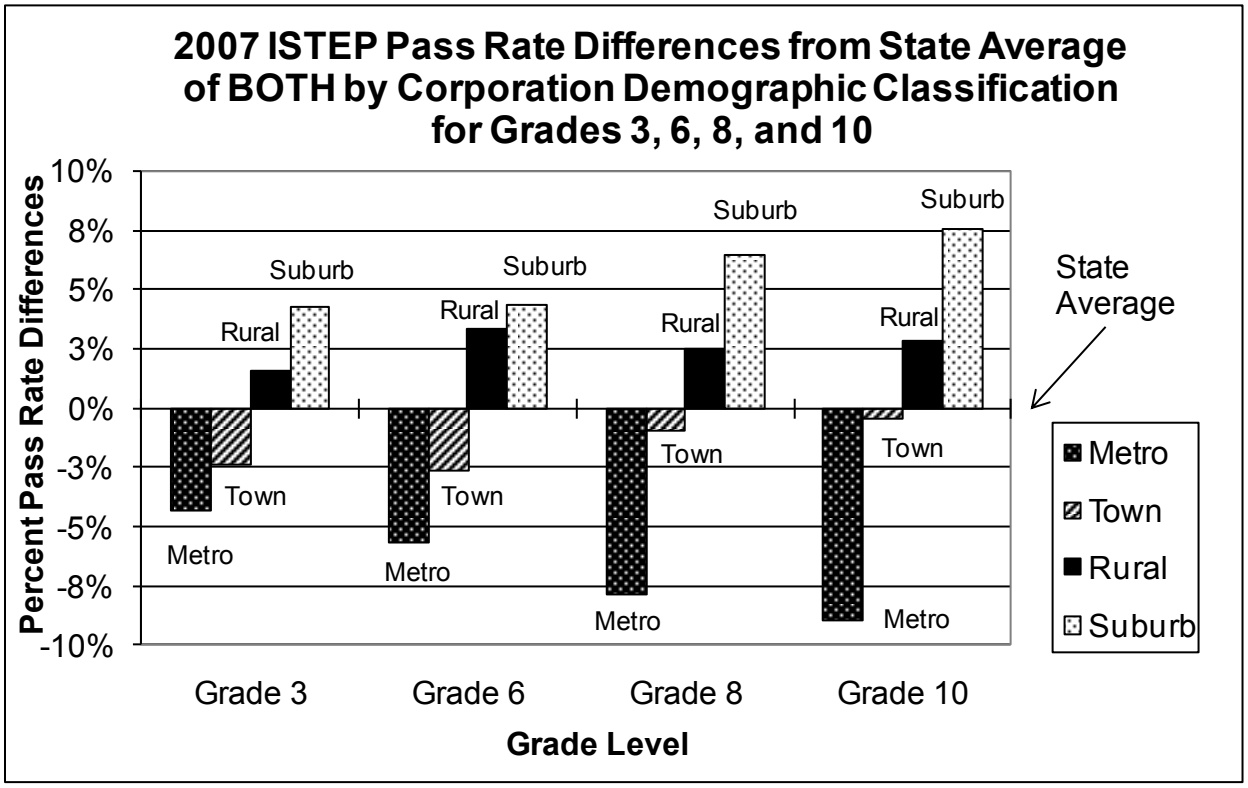

Figure 2 2007 ISTEP pass rate gap percentages of BOTH (ENLA and Math) for metropolitan, town, rural and suburban school corporations measured from the state averages for grades 3, 6, 8, and 10.

Table 4
State 2002 Percentages and Percent per Year Demographic Changes 2002–2007

Variable	Free Reduced Lunch		White		Hispanic	
Grade	2002 Percentage	Percent per Year Increase[1]	2002 Percentage	Percent per Year Decline[1]	2002 Percentage	Percent per Year Increase[1]
3	35.2	1.3	79.6	-1.1	3.9	0.7
6	33.3	1.3	80.8	-1.0	3.5	0.6
8	29.1	1.7	82.2	-1.1	3.1	0.6
10	22.0	1.9	83.1	-1.0	3.0	0.4

[1]Percent per year increases and declines had R^2 values between 0.94 and 0.99 for 2002–2007.

Ethnicity changes between grades 3 and 10 of Indiana's public school populations have also occurred between 2002 and 2007. The average school White school population has decreased from approximately 83 percent in grade 10 in 2002 to 74 percent in grade 3 in 2007, while the Hispanic population grew about four to five percentage points during this time to 7 percent in grade 3. These gains and losses showed significant variation across schools.

The wide variation in demographic shifts across schools suggests that schools may not be on a level playing field when long-term trends are used to evaluate schools, and that schools with increasing concentrations of children eligible for free-reduced lunch and minority children may have

systematic bias in their trends that can place them at a disadvantage in such rankings unless family demographic trends are controlled.

Results

Annualized State-wide Gains

Figures 3a-3d contrast the estimated state-wide annual trends using Equation 1 for estimates that include and exclude family characteristics. Appendix A has the estimations when family variables are included. For grades 3, 6 and 8 these figures indicate that demographically adjusted gains are significantly higher than gains estimated without demographic variables. For instance, gains at third grade between 2002 and 2007 are 0.21 and 0.09 standard deviation units, respectively, for adjusted and non-adjusted estimates. The differences at grades 6, 8 and 10 are 0.32 vs. 0.19, 0.26 vs. 0.17 and 0.05 vs. -0.07 standard deviation units, respectively. These differences are large from a policy perspective when evaluating the performance of schools statewide. They suggest that demographic changes are a very significant factor to take into account when assessing the long-term performance of Indiana schools. The results also suggest that demographic changes become more important as the length of the period for estimating increases. For instance, the gains between 2002 and 2003 are not affected as much by the inclusion of demographic factors compared to the difference between 2002 and 2007. Although including more years in an analysis will improve the reliability of trends, the longer period also increases the effects of demographic characteristics as long as demographic trends are steadily increasing.

School-level Trends

Equation 2 was used to compute annualized school trends at each grade level for Indiana's schools. The determination of grade-level statistically significant six-year annualized gains gi in the regression analyses and six-year slopes in OLS were computed by dividing the annualized gain and slope by its respective uncertainty, respectively. Table 5 summarizes the results. The table contrasts the percentage of schools at each grade that had statistically significant (95 percent confidence) gains or losses with family variables excluded and included. For instance, at grade 3, 17.9 percent of schools had statistically significant gains, while 8.4 percent had statistically significant losses when demographic variables are excluded. With demographic variables, the percentage with statistically significant gains increases from 17.9 to 22.7, while those with statistically significant losses changes from 8.4 to 6.0. Using either measure, it suggests that less than one-quarter of schools at third grade have statistically significant long-term gains. The demographic adjustments added 44 schools or 4.8 percent of total grade 3 schools to the category of statistically significant gains, and reduced the number of statistically significantly declining schools by 23 or 2.4 percent of total grade 3 schools. Appendix B has the regression coefficients for the family variables used in Equation 2 for computing the above estimations.

For all grades, including demographics increases the number of statistically significantly gaining schools and decreases the number of statistically significantly declining schools. Results are better for grades 6 and 8 with up to 42.4 percent of grade 6 schools showing statistically significant gains and less than 3 percent with statistically significant losses. However, at grade 10, only 8.4 percent of schools have statistically significant gains while 16.2 percent have statistically significant declines.

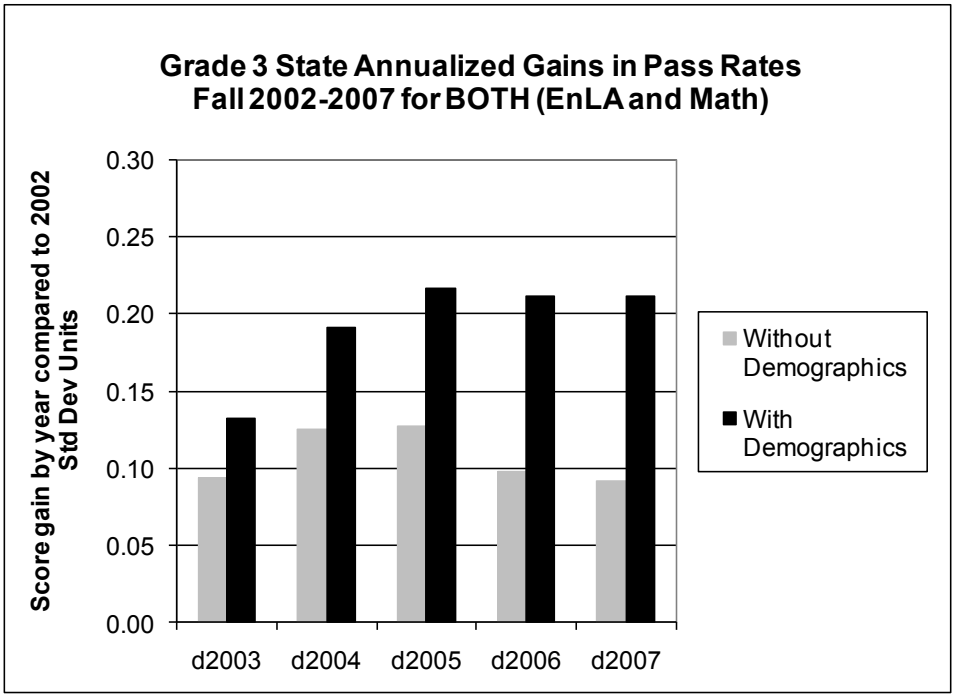

Figure 3a Comparisons of state annualized-gain coefficients of Equation (1) by regression (with demographics), and state pass rate gains (without demographics) of BOTH (ENLA and Math) in Grade 3 for fall 2002–2007

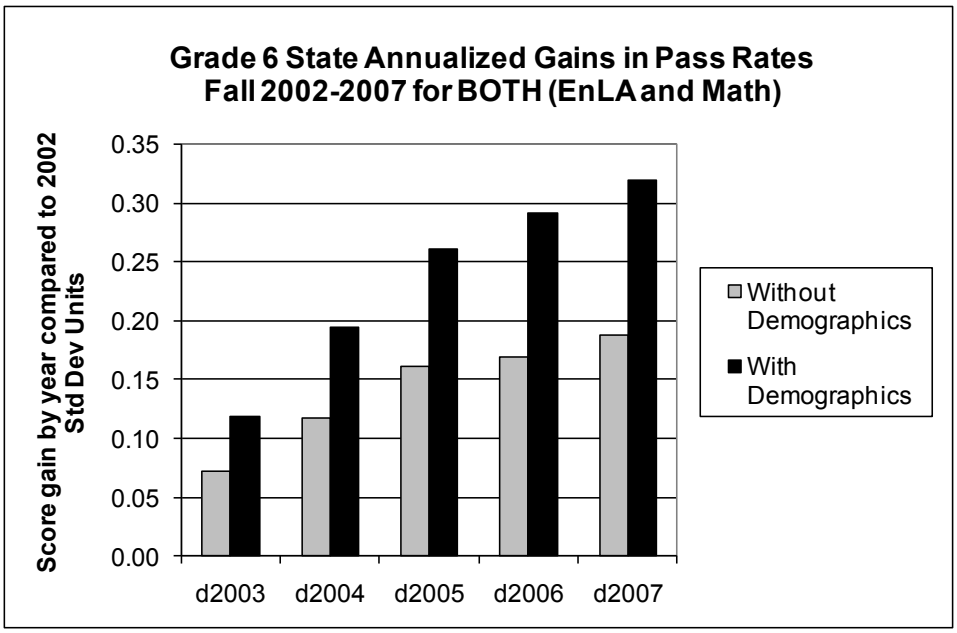

Figure 3b Comparisons of state annualized-gain coefficients of Equation (1) by regression (with demographics), and state pass rate gains (without demographics) of BOTH (ENLA and Math) in Grade 6 for fall 2002–2007

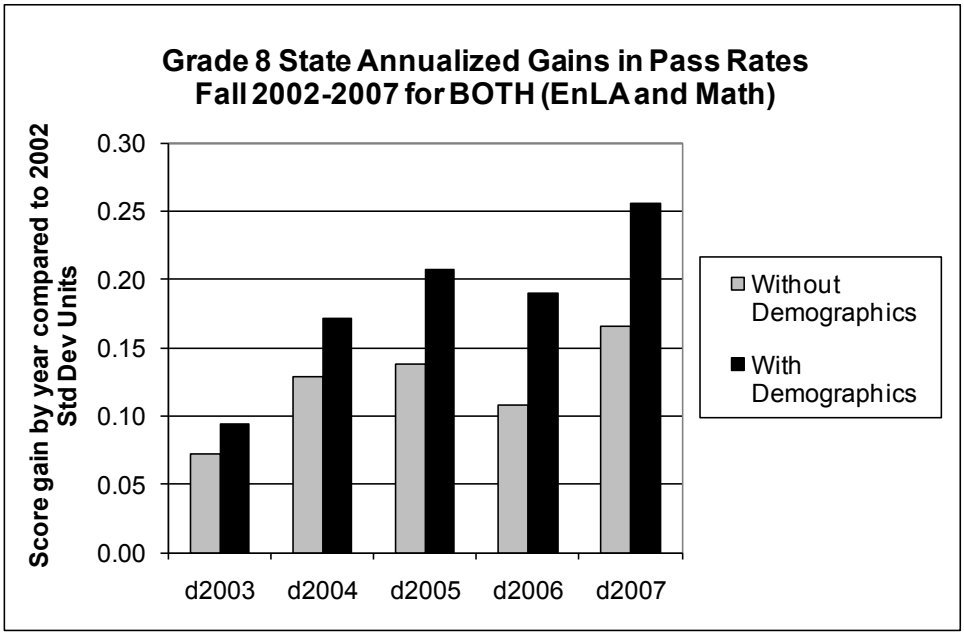

Figure 3c Comparisons of state annualized-gain coefficients of Equation (1) by regression (with demographics), and state pass rate gains (without demographics) of BOTH (ENLA and Math) in Grade 8 for fall 2002–2007

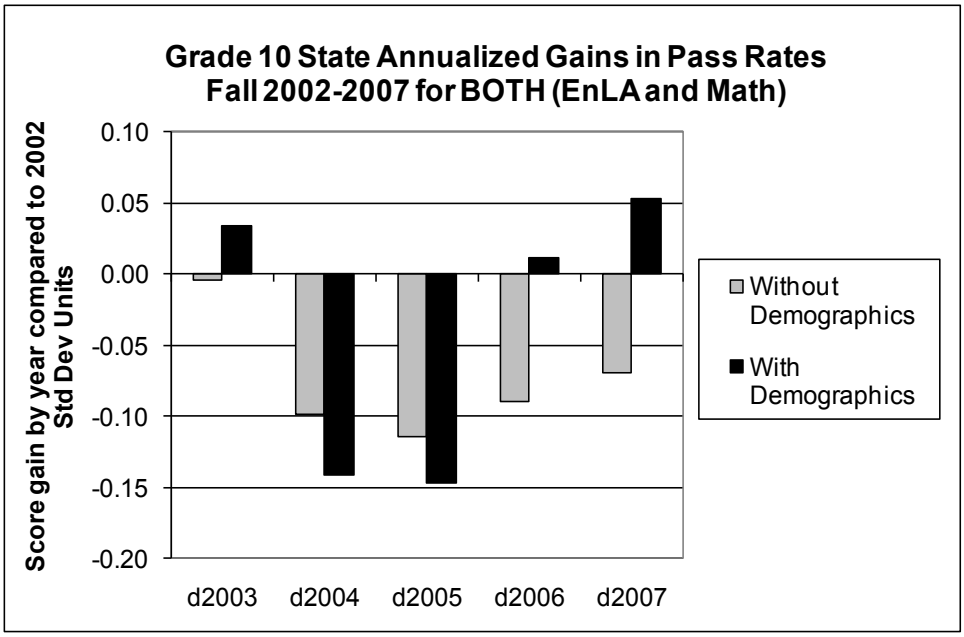

Figure 3d Comparisons of state annualized-gain coefficients of Equation (1) by regression (with demographics), and state pass rate gains (without demographics) of BOTH (ENLA and Math) in Grade 10 for fall 2002–2007

Table 5

Comparisons of Percentages of schools with Statistically Significant (95 percent confidence) Gains and Declines in Six-year (2002–2007) Pass Rates for BOTH (ENLA and Math) Estimated with and without Demographic Characteristics by Regression and Ordinary Least Squares (OLS), respectively

Grade	N Schools	OLS Trends Percent		Regression Trends Percent		Regression - OLS Percent Difference	
		Gains	Declines	Gains	Declines	Gains	Declines
3	862	17.9	8.4	22.7	6.0	4.8	-2.4
6	455	37.8	3.5	42.4	2.4	4.6	-1.1
8	378	28.6	3.2	34.4	1.3	5.8	-1.9
10	334	6.3	23.4	8.4	16.2	2.1	-7.2
Total	2029	22.4	8.8	27.0	6.0	4.6	-2.8

Table 6 shows the estimated average annual gains or losses for the four improvement categories of schools. For instance for grade 3, schools with statistically significant gains increased their pass rates by 3.33 percentage points a year, while those with statistically significant losses declined by 3.6 percentage points a year. A typical school with a 60-percent student pass rate in 2002 could increase their pass rate to 76 percent in the statistically significantly gaining category by 2007, while those in the statistically significant loss category would have a rate of 42-percent pass rate by 2007. These are large differences, although the differences decline in higher grades.

Indiana's K-12 accountability system became law (PL 221) in 1999 and was enacted to serve as a basis for evaluating schools. As with NCLB, performance and improvement are measured with the state's ISTEP exams in English-Language Arts and Mathematics in grades 3–10. Currently, PL 221 incorporates the AYP criteria of NCLB. The five measures of PL 221 are as follows: A - Exemplary Progress, B - Commendable Progress, C - Academic Progress, D - Academic Watch (priority), and F - Academic Probation (high priority).

Annual improvement is computed yearly for each school and each school corporation as a whole. Improvement is based on the pass rates on the sum totals of students across grade levels passing ENLA and Math in Elementary Schools (grades 3–5), Middle Schools (grades 6–8) and High Schools (grades 9–10). Improvement is then computed from one year to the next for non-mobile cohorts; a three-year average of improvement is then computed and compared to the improvement of the most recent year with the higher percentage being used as that year's improvement.

Table 7 shows the Indiana Public Law 221 average category placement percentages of improvement for schools using data from 2006, 2007, and 2008. We have grouped the top two categories of Exemplary and Commendable to make comparisons with our results in Table 5. Table 7 shows that grade 3 has 46.5 percent of schools that are Exemplary or Commendable, while grades 6, 8 and 10 have much lower percentages around 16 percent. Grade 3 has only 5 percent of schools on Academic Probation, compared to around 12 percent for grades 6, 8 and 10.

These rankings show a substantially different pattern than Table 5 based on estimations of trends. The trend estimation shows grades 6 and 8 to have markedly higher percentages of schools with significant gains with grade 10 having by far the lowest percentage with significant gains. The Indiana evaluations show the opposite trends at grades 3, 6, and 8 with grade 3 having the highest percentage of Exemplary or Commendable schools with grades 6 and 8 showing substantially smaller percentages than at grade 3. The trend estimates show grade 10 to have the highest percentage of statistically significantly declining schools, while the Indiana evaluations show similar percentages of probationary schools at grades 6, 8, and 10. It is important to reliably identify what parts of the school system are under or over performing. The Indiana evaluations would show

elementary grades performing better than middle or high schools, while the trend system would identify middle schools as the top performers and high schools as the lowest performers.

Table 6

Average Annualized Gains and Declines (percent/yr) After Controlling for Family Demographics of Schools with Improving, Declining and Statistically Significantly (95 percent confidence) Improving and Declining Pass Rates of BOTH (ENLA and Math) in Grades 3, 6, 8, and 10 between Fall 2002 and Fall 2007

Improvement Category	Average Percent/yr Pass Rate Changes of Students Passing BOTH (ENLA and Math)			
Grade	3	6	8	10
Statistically Significant Improve	3.33	2.68	2.33	1.55
Improving	1.03	0.90	0.80	0.52
Declining	-1.52	-1.03	-0.89	-1.15
Statistically Significant Decline	-3.64	-2.03	-2.33	-2.39
	Percent of Schools in Each Improvement Category			
Statistically Significant Improve	22.7	42.4	34.4	8.4
Improving	44.4	42.0	47.6	30.8
Declining	26.8	13.2	16.7	44.6
Statistically Significant Decline	6.0	2.4	1.3	16.2
N - Schools	862	455	378	334

Table 7

Indiana Public Law 221 Average Category Placement Percentages of Elementary (Grades 3–5), Middle Schools (Grades 6–8), and High Schools (Grades 9–10) for the Three-year Period Fall 2006 through Fall 2008

State Improvement Category	Percentages of Schools in Each Improvement Category by Grade		
	3	6 and 8	10
Exemplary plus Commendable	46.5	16.2	15.8
Academic Progress	22.4	18.1	14.4
Academic Watch	26.6	54.4	57.1
Academic Probation	4.5	11.7	13.1
N - Schools	1169	317	372

Characteristics of Schools with Improving and Declining Performance

Presented in Table 8 are the grade 3, 6, 8, and 10 comparisons of the 2007 characteristics of the schools with (1) statistically significant (95 percent confidence) improving, (2) improving, (3) declining, and (4) statistically significant (95 percent confidence) declining pass rates for estimates with and without demographic characteristics. At each grade level, the unadjusted (without family-demographic characteristics) data show that schools with statistically significant gains compared to statistically significant losses have much lower percentages of free-reduced lunch students and lower percentages of minority students. These differences narrow considerably if the comparison is between statistically significantly gaining schools vs. either declining schools (column 3) or improving schools (column 2). For instance at grade 3, the free-reduced lunch percentage of statistically significantly gaining schools is 43.5 percent compared to 42.4, 46.0, and 54.5, respectively, for improving, declining, and statistically significantly declining schools. The schools in

the statistically significantly improving, improving and declining categories tend not to have large differences in free-reduced lunch or minority diversity. This indicates that the top scoring schools do not dominate the schools having statistically significant gains, but rather schools with statistically significant gains are closer to schools with more typical characteristics. It is only the statistically significantly declining schools that show markedly different characteristics with much higher free-reduced lunch and minority populations.

The characteristics of statistically significantly gaining schools shift if demographic controls are included in the regression. Generally, the characteristics of the statistically significantly gaining schools shifts to be higher free-reduced lunch and lower minority indicating that schools with more racial/ethnic diversity have a higher probability of being in the statistically significantly gaining category with demographic controls. The characteristics of statistically significantly declining schools shifts with the inclusion of demographic controls toward being a much lower percentage of minority and similar or lower percentage of free-reduced lunch students. Thus, schools with more diversity are less likely to have statistically significant losses if demographic controls are included. Overall, the inclusion of demographic controls tends to include more diverse schools in the statistically significantly gaining category and fewer in the statistically significantly declining category.

Estimating the Percentage of Mislabeled Schools with Improving or Declining Pass Rates

A final question is how many of the schools are likely mislabeled, that is, how many of the schools that were estimated to have statistically significant gains or losses are in that category by luck. Another way to state the problem is how much difference would there be between the number of estimated schools with statistically significant gains or losses using the actual data versus how many would there be if the scores were known exactly–with no error. The presence of uncertainty means that some schools with estimated statistically significant results (95 percent confidence) would not be statistically significant if scores were known without error, and some schools that were insignificant would have been significant if scores were known without error. Estimating the percentage of schools that are mislabeled is an important policy parameter, especially if rewards or sanctions are applied to schools.

Knowing an estimate of the percentage of schools that may be mislabeled should modify rewards or sanctions because it provides a measure of the confidence with which rewards or sanctions are being applied to the appropriate schools. If the percentage of mislabeled schools is a substantial percentage of those with statistically significant results, it should help determine whether applying sanctions and rewards would provide the planned incentives and/or should temper the size and severity of such rewards and sanctions. Each school properly labeled would receive the right signals and contribute to making a policy effective if the reward or sanction was effective in supporting teachers and administrators producing the gains, or in helping teachers and administrators who have negative trends to reverse such trends. However, for every mislabeled school, teachers and administrators would receive a wrong signal and result in changing effective policies and/or continuing ineffective policies. Much of the critique of NCLB revolved around the question how much confidence there was in the labels applied to schools, and whether rewards and sanctions were being applied fairly to the schools. Using longer-term trends for accountability will not remove the problem of mislabeling schools, but estimates of the degree of mislabeling can help determine how many years of data are needed to provide adequate reliability when applying rewards and sanctions and how large such rewards and sanctions might be.

Table 8
Grades 3, 6 ,8, and 10 Characteristic Demographic Percentages (fall2007) and 2002–2007 Pass-Rate Percentages of BOTH (ENLA and Math) in Schools with Improvements, Declines, and Statistically Significant (95 Percent Confidence) Improvements and Declines

Grade and Demographic		OLS - Exclude Demographics Schools with				Regression - Include Demographics Schools with			
		SS[1] Improve	Improve	Decline	SS[1] Decline	SS[1] Improve	Improve	Decline	SS[1] Decline
		Demographic Percentages				Demographic Percentages			
3	Free Reduced	43.5	42.4	46.0	54.5	43.8	42.7	46.4	55.6
6		38.8	40.2	39.7	49.9	39.8	40.7	37.5	40.1
8		37.7	35.4	36.8	40.8	38.7	35.3	35.3	42.7
10		25.8	26.2	29.6	35.4	26.5	26.8	31.6	32.4
3	Black	12.2	10.3	13.2	17.2	11.6	11.0	12.4	20.3
6		7.2	6.2	6.9	10.2	7.6	6.8	5.0	6.1
8		7.8	8.4	11.2	19.4	9.9	7.9	11.4	11.0
10		0.7	5.3	4.9	15.3	3.6	5.1	8.9	8.3
3	Hispanic	6.8	6.3	7.9	11.4	8.1	6.2	7.9	9.6
6		5.3	5.2	8.5	9.8	5.9	5.7	7.6	6.9
8		4.8	3.8	7.3	6.5	5.3	4.4	5.4	5.6
10		3.5	4.2	2.9	5.1	5.5	3.9	3.8	2.4
3	White	75.5	77.7	72.4	62.9	74.5	76.8	73.2	61.9
6		83.1	84.3	79.8	72.7	81.8	83.5	82.7	78.8
8		83.6	84.2	77.7	68.0	80.8	84.1	79.4	78.4
10		93.4	87.2	88.3	74.9	87.0	87.6	83.2	85.2
3	Pass Rate	70.8	65.8	58.1	51.4	70.2	64.8	57.6	49.3
6		71.7	68.2	65.3	57.3	71.3	67.3	66.0	57.8
8		66.2	62.6	57.9	58.2	65.3	62.3	57.8	53.5
10		69.2	64.0	58.6	50.0	69.2	64.0	58.6	50.0
3	Percent of Schools	70.8	65.8	58.1	51.4	22.7	44.4	26.8	6.0
6		71.7	68.2	65.3	57.3	42.4	42.0	13.2	2.4
8		66.2	62.6	57.9	58.2	34.4	47.6	16.7	1.3
10		69.2	64.0	58.6	50.0	8.4	30.8	44.6	16.2
Total 2029		451	818	583	177	547	857	503	122
Percent		22.2	40.3	28.7	8.7	27.0	42.2	24.8	6.0

[1] SS - Statistically Significant (95 Percent Confidence).

Monte Carlo Simulations

Developed during the Manhattan Project of World War II, Monte Carlo simulation is used by actuaries, education researchers, medical researchers, military planners, physicists, and others. It allows such persons to simulate some random event a large number of times, calculate the resulting percentage of "success", sample means and variances, and other functions of the outcomes. Actuaries can use the technique in setting premiums for health insurance, retirement pensions, life insurance, and property damage-liability insurance. Educational researchers can use Monte Carlo simulations to estimate the number of schools that may be mislabeled as having statistically significant gains or losses on state exams. Researchers of surgical techniques, internal medicine topics, neonatal subjects, military plans, and other subjects can economically view the random

outcomes of hundreds of repetitions of some event. They can use the total set of outcomes in their business, education plans, medical research, military planning, and other area decisions involving life changing events.

We estimate the percentage of schools mislabeled by using a Monte Carlo simulation that uses actual standard errors at each grade and for each school to randomly generate annual gains and losses for each school and grade over a six-year period assuming no systematic trend. We fit these estimates with trend lines and determine the number of schools that would have statistically significant gains and losses (95 percent confidence). We then compare the percentage of these randomly significant schools to the actual estimates to determine what percentages of schools are likely mislabeled.

Table 9 shows the estimates for the number of schools estimated to have statistically significant gains and losses (95 percent confidence) randomly. The estimates show that approximately 7–10 percent of schools at each grade would have had statistically significant gains or losses given the errors in scores and assuming no systematic trend. Table 10 compares these estimates of random statistically significant gains and losses to our actual estimates. The results show that the percentage of schools who had an estimated statistically significant decline in grades 3, 6 and 8 or a significant gain in grade 10 are as large or smaller than what would be expected randomly. These estimates should induce extreme caution in labeling any school in Indiana as having statistical significant losses because those labeled as such are highly likely to be in that category due to poor luck. For those that are estimated to have statistically significant gains, about 48 percent of grade 3 schools are estimated to be mislabeled, while only about 23–30 percent of schools at grade 6 and grade 8 are estimated to be mislabeled (see Table 10). Mislabeling occurs more frequently where trends are weaker such as at 3rd and 10th grade compared to 6th and 8th grade.

Table 9

Estimates of Percentages of Schools Chosen with Statistically Significant (95 Percent Confidence) Six- year Gains and Losses by Monte Carlo Simulation

Schools	Grade	Percentages 3	6	8	10	Totals
Statistically Significant Improving		8.6	8.8	8.5	6.9	8.3
Statistically Significant Declining		10.1	8.4	8.2	7.5	8.9
N - Schools		862	455	378	334	

Table 10

Comparison of the Percentages of Schools Estimated to Have Statistically Significant (95 Percent Confidence) Six-year Gains and Losses Compared to Percentages Generated Randomly

		Actual Percentages		Random Percentages		Percent Mislabeled (Random/Actual)	
Grade	N	Gains	Declines	Gains	Declines	Gains	Declines
3	862	17.9	8.4	8.6	10.1	48.0	>100
6	455	37.8	3.5	8.8	8.4	23.3	>100
8	378	28.6	3.2	8.5	8.2	29.6	>100
10	334	6.3	23.4	6.9	7.5	>100	32.0
All Schools	2029	22.4	8.8	8.3	8.9		

The mislabeling of schools is partly due to the level of statistical significance assumed necessary to be placed in the significant categories. Setting a more stringent significance level would lower the amount of mislabeling, as would the inclusion of more years of data. These results would suggest the use of some caution in even using six years of data as a basis for rewards, and it suggests that rewards not be uniform across any category, as schools that are more statistically significant are less likely to be mislabeled.

It is important to note that these results do not imply that the number of schools with statistically significant gains is, for instance, for 3rd grade the percentage actually estimated (17.9) minus the percentage estimated randomly (8.6) or 9.3 % (see Table 10). There would be an approximately equal number of schools that were mislabeled as statistically insignificant, but should have been labeled statistically significant. The good and bad luck schools are approximately equal and similarly mislabeled in both the significantly positive and insignificant positive category. There is a correct percentage of schools with statistically significant gains, but we cannot find it exactly, because there is a percentage of schools which are not correctly classified.

Discussion

This paper utilized 6 years (2002–2007) of achievement data from Indiana to answer three questions that should be of primary interest to educational school reformers, educators, and the public. These questions focus on long-term gains, whereas almost all current state and federal policy has focused almost exclusively on short-term gains. The research literature has provided evidence that the uncertainty in short-term gains makes their use for policymaking problematic. Such a short-term focus has delayed analysis of long-term gains, yet it is long-term gains that should be the primary focus of policymakers.

As stated earlier the three questions we addressed in this paper are important to evaluating and identifying schools that produce statistically significant long-term gains and declines. First, are the same schools identified when controls for socioeconomic status and ethnicity are incorporated to assess long-term gains and declines? Second, in long-term determinations, what percentages of schools register trends that are statistically significant (gains and declines) due to inherent randomness? Third and most important, are schools that have statistically significant long-term trends of improvement (and decline) being properly identified with current short-term annual measures associated with NCLB and PL 221 in Indiana?

Inherent Uncertainties and Changing Demographics Mask Short-term Gain/Loss Measures

It is the uncertainty inherent in annual individual achievement scores combined with the relatively small sample sizes at the school level that causes annual school gains by grade to be problematical. Another problem that can make annual gains problematical is the migration of students into and out of schools that can often change demographic characteristics and scores in significant ways. These sources of uncertainty must be small compared to the expected size of annual gains for short-term gains to be meaningful. Unfortunately, the uncertainty in annual gains in scores can be of the same magnitude as expected gains.

Even if short-term gains could be made more reliable, the policies that might flow based on short-term gains would not necessarily lead to long-term gains. Educational policies and pedagogical practices that produce short-term gains may be much different than policies required to produce sustained long-term gains. Many educational and early childhood interventions have produced achievement gains in the short term, but such gains often decline when longer-term measurements are made.

Policies that produce long-term sustained gains must not only show such gains at a given grade, but must insure that gains at one grade carry over to the next grade, and become cumulatively enhanced such that each cohort shows cumulative growth over grades. Producing such gains may require a much greater coordination of teaching and curriculum across grades for gains to accumulate. For instance, large gains at one grade may require changing the curriculum at the next grade so that excess repetition is avoided and additional new and challenging material is covered in the next grade.

Research can contribute to improved policymaking partly by focusing policymakers on those questions that are central to our objectives of obtaining sustained improvement in student proficiency in math and reading such that U.S. scores on international exams are more competitive, and achievement gaps are narrowed or eliminated in the U.S. Even if historically high rates of annual gains of 1–2 percentile points a year could be sustained, it would take 10–20 years to make substantial progress on closing international and national achievement gaps. Policies that could sustain such long-term gains are likely to be different than policies that can produce short-term gains.

Regression Model Predictions of Annualized State Gains and School-Level Trends

We analyzed data at four grades: 3rd, 6th, 8th, and 10th using the Indiana state tests. We utilized the percentage of students attaining proficiency in both English and Math as our dependent variable in our analyses. We estimated trends using two methods: school fixed-effect methods including trends and school fixed-effect with family/demographic characteristics and trends. The latter measure takes account of the changing demographics of students in schools and may provide a fairer measure when evaluating and comparing schools since teachers and principals have no control over demographics. Essentially Indiana schools are compared without any consideration of the demographic characteristics of the schools.

At the state level, the estimated annual gains unadjusted by demographic variables were largest at 6th grade (1.6 percentage points per year), followed by 8th grade (0.9 percentage points per year), 3rd grade (0.6 percentage point per year), and 10th grade (-0.6 percentage points per year). These gains show somewhat different patterns from national trends measured by NAEP scores where the largest gains are at 4th grade, with somewhat slower gains at 8th grade and very small gains at 12th grade. In Indiana, gains at 6th and 8th grade are much larger than 3rd grade gains, while both national and Indiana gains during high school are small.

The percentages of schools with long-term statistically significant gains (95 percent confidence) using unadjusted trends were 17.9, 37.8, 28.6, and 6.3, respectively, for grades 3, 6, 8, and 10 (see Table 5). The smaller percentage of schools with gains at 3rd grade is not only caused by smaller statewide gains at 3rd grade than at 6th and 8th grade, but is partially due to the smaller number of students per school at 3rd grade (77) compared to 6th grade (135), 8th grade (192), and 10th grade (232). The smaller samples would have increased standard errors making statistically significant trends less likely.

The importance of demographic adjustments is illustrated by comparing the results to the unadjusted trends. The percentage of schools with statistically significant gains increased significantly when adjusted by demographics. The percentages of schools with statistically significant gains were 22.7, 42.4, 34.4, and 8.4, respectively, for grades 3, 6, 8, and 10 (see Table 5). Therefore, the number of schools with statistically significant gains increased by 27, 12, 20, and 33 percent, respectively, at grades 3, 6, 8, and 10 illustrating the importance of making demographic adjustments. The demographic trends in Indiana changed slightly faster than in most states. For instance, Indiana's percentage of White students declined 6.1 percentage points from 83.0 to 76.9 percent at 8th grade from 2002–2007 compared to a 5.2 percentage-point drop from 61.1 to 55.9

percent nationally. For states with a more rapidly changing population, the importance of making demographic adjustments increases.

These results suggest that a much lower percentage of schools are making long-term statistically significant gains than suggested by annual state evaluations. In Indiana, the average percentages of schools (2006–2008) that make acceptable annual gains as measured by a formula incorporating AYP are 68.9, 34.3, 34.3, and 30.2 (summing Exemplary, Commendable, and Academic Progress in Table 7) at grades 3, 6, 8, and 10, respectively. Clearly many of these schools rated as making annual gains do not show long-term gains.

A small percentage of Indiana schools show statistically significant (95 percent confidence) declining trends. The unadjusted trends are 8.4, 3.5, 3.2, and 23.4 at grades 3, 6, 8, and 10, respectively, compared to the adjusted estimates of 6.0, 2.4, 1.3, and 16.2 (see Table 5). The demographically adjusted trends place fewer schools in the statistically significantly declining category.

We have compared the pattern of gains and losses by grade to evaluations done by Indiana based on shorter-term measures. Our estimates show distinctly contrasting patterns compared to Indiana evaluations. While Indiana evaluations show the highest percentage of schools at 3rd grade that are Commendable or Exemplary with much smaller percentages at grades 6, 8, and 10, our estimates show grades 6 and 8 with much higher percentages of statistically significant gains than grade 3. Our estimates show grade 10 with the smallest percentage of statistically significantly gaining schools, while Indiana evaluations show little difference between grades 6, 8, and 10.

Characteristics of Improving Schools with and without Controlling for Demographics

Indiana students and schools show the typical patterns of significantly higher scores for students that are White, are ineligible for free-reduced lunches, have higher family income and better educated parents, and live in two-parent families. However, the demographic differences between schools that are making statistically significant gains compared to the remaining schools show a much smaller or little difference in demographic characteristics. For instance, the fall 2007 average free-reduced lunch percentages of schools with statistically significant gains without demographic adjustments are 43.5, 38.8, 37.7, and 25.8, respectively, at grades 3, 6, 8, and 10 (see Table 8) compared to the remaining schools (improving, declining, and statistically significant declining) of 45.0, 40.6, 36.1, and 30.1 at grades 3, 6, 8, and 10. The percentages of White students in statistically significantly gaining schools without adjustments is 75.5, 83.1, 83.6, and 93.4, respectively, (see Table 8) compared to 74.2, 82.3, 81.5, and 84.7 in remaining schools (improving, declining, and statistically significantly declining) at grades 3, 6, 8, and 10.

The characteristics of the schools with statistically significant gains change when demographic controls are included. The characteristics of schools with statistically significant gains with demographic adjustments generally have increased percentages of free-reduced lunch students and lower proportions of White students compared to unadjusted results. This indicates that including demographic adjustments increases the number of schools with statistically significant gains having more minority and free-reduced lunch populations. Thus, unadjusted trends provide less chance for more demographically diverse and poorer schools to be selected as having statistically significant gains, and provide an unfair advantage to schools with higher-income White students. The demographics of schools have little predictive power when identifying schools that will have statistically significant gains, and even less power when demographic adjustments are made.

Only a small percentage of schools in each grade had statistically significant losses (95 percent confidence). Using unadjusted trends, the characteristics of these schools showed markedly different characteristics than remaining schools. The significantly declining schools had much higher

percentages of free-reduced lunch populations and lower percentages of White students. For instance, the free-reduced lunch percentages of statistically significantly declining schools were 54.5,49.9, 40.8, and 35.4, respectively, at grades 3, 6, 8, and 10 (see Table 8) compared to 43.8, 39.6, 36.4, and 28.2 for remaining schools (statistically significant improving, improving, and declining) at grades 3, 6, 8, and 10. The percentages of White students for statistically significantly declining schools were 62.9, 72.7, 68.0, and 74.9, respectively at grades3, 6, 8, and 10 (see Table 8) compared to 75.5, 82.9, 82.6, and 88.3 for remaining schools (statistically significant improving, improving, and declining) at grades 3, 6, 8, and 10.

The characteristics of statistically significantly declining schools change when demographic adjustments are included. Except for grade 3, the percentage of White students increased markedly compared to unadjusted results. For instance, the percentages of White students were 61.9, 78.8, 78.4, and 85.2, respectively, for grades 3, 6, 8, and 10 for adjusted losses compared to 62.9, 72.7, 68.0 and 74.9 for unadjusted losses (see Table 8). The changes in the free-reduced lunch percentages between adjusted and unadjusted were grade dependent with small increases in free-reduced lunch percentages for losses at grades 3 and 8; however, there are decreases in free-reduced lunch percentages at grades 6 and 10. The inconsistencies may be due to the small numbers of schools with statistically significant losses, especially at grades 6 and 8. However, the adjusted loss results remove schools with higher percentages of minority populations from the statistically significant loss categories. Therefore, schools with greater percentages of minority students may not be fairly ranked with unadjusted results.

Monte Carlo Simulations: Statistically Significant Long-term Gains/Losses due to Chance

We made estimates by Monte Carlo simulations of the expected percentages of schools at each grade that would have had statistically significant gains or losses (95 percent confidence) assuming no overall trend present. The estimated percentages of "lucky" and "unlucky" schools were 8.6, 8.8, 8.5, and 6.9, respectively, for grades 3, 6, 8, and 10 (see Table 9). For grades 3, 6, and 8 the percentages of schools with statistically significant losses estimated with Monte Carlo simulations were similar to or greater than the estimated numbers of schools with losses using unadjusted results. This suggests that there is little reliability in identifying schools with statistically significant losses at these grades. The unadjusted results also suggest that about 48, 23, and 30 percents of the schools having statistically significant gains, respectively, at grades 3, 6, and 8 may be mislabeled. These results would suggest that applying sanctions to schools with estimated statistically significant losses using only six years of data would be unlikely to work since bad luck is the major reason that separates them from somewhat higher performing schools. Similar caution is warranted for schools at grade 3 that have statistically significant gains since about one-half are in that category due to good luck. Grades that have stronger upward trends like grades 6 and 8 are less vulnerable to mislabeling. Using more years of data in the analysis would reduce the percentage of schools that are mislabeled at each grade.

Severity of Increased Mislabeling when Using Four or Five Years of Data to Obtain Trends

All previously described school-level trends (based on six years of data) were compared to trends obtained when using four years of data (2004–2007) and five years of data (2003–2007). Estimates were again made by using Equation 2 to obtain annualized gains for all regression analyses when controlling for family variables; the OLS analyses and the Monte Carlo simulations were obtained as previously described for the six-year analyses.

Results are presented in Table C1 of Appendix C which compares the percentages of schools at grades 3, 6, 8, and 10 with statistically significant gains and declines for four, five and six years when using OLS and regression; these results are comparable to Table 5 (the six-year results). Similarly the Monte Carlo simulations for four, five and six years are summarized in Table C2 of Appendix C; these results are comparable to those presented in Table 9 (the six-year results).

Grades 3, 6, and 8 all experienced similar changes in trends where both OLS and regression gains decreased as much as a factor of two in going from six years to four years of analyzed data. The statistically significant declines for both the OLS and regression analysis increased when going from six years to five years to four years of analyzed data (see Table C1). Both of these changes are favorable to schools–more statistically significant gains for schools and fewer statistically significant declines when using six years of data. The authors do not have a simple explanation for why the trend from four years to five years to six years at grade 10 is just the opposite of what occurs at grades three, six, and eight.

The Monte Carlo simulation results follow a similar behavior at all grade levels as one compares the schools with statistically significant gains and declines. As one would expect, the Total Schools percentages for six years of data at 8.3 and 8.9 for gaining schools and declining schools, respectively, increase to 10.3 and 13.2 from 8.3 percent, and to 9.0 and 12.6 percent from 8.9 percent for gains and declines, respectively, when using only five and four years of data, respectively (see Table C2). This demonstrates the increased mislabeling that occurs when basing improvement on short-term measures.

Conclusions

Results from this paper illustrate that using six years of data to measure whether statistically significant gains or losses by grades in schools are occurring must be done with some caution. Using six years of data provides a more reliable basis for categorizing school improvement than the use of two years of data which is widely used currently to evaluate schools across the U.S. The results suggest that methods that provide recognition to schools based on short-term measures likely produce overly optimistic evaluations of schools and extensive mislabeling of schools. The results of this study also suggest that the proportions of schools making statistically significant gains can vary markedly by grade where gains were highest at 6th and 8th grades, smaller at 3rd grade and absent at 10th grade (see Table 5). This pattern contrasted with evaluations by Indiana which are based on shorter-term gains; the short-term gains determined showed grade 3 significantly outperforming grades 6 and 8, with grades 6, 8, and 10 being given similar evaluations (see Table 7).

This research suggests that controlling for demographic changes in schools increases the proportion of schools with statistically significant gains at each grade, and decreases the proportion of schools with statistically significant declines. The research also suggests that schools with higher proportions of minority students are more represented in the schools with statistically significant gains and generally less represented in schools with statistically significant losses when demographics are included in the analysis. Such schools may be unfairly classified by methods that do not incorporate demographic characteristics; these methods fail to recognize schools and teachers who are performing well with difficult populations. Finally, our analysis suggests that even six years of data cannot eliminate the role of chance in categorizing schools. Chance can become a major factor in mislabeling schools when less years of data are used and when overall trends are not robust.

The continuing emphasis on categorizing schools based on short-term comparisons like AYP or on categorizing teachers based on a single year of gains leaves policymakers highly vulnerable to the mislabeling of performance categories by schools and teachers. Sanctions and rewards can be applied to the wrong schools and teachers. Any evaluation system will have some

flaws and can still provide appropriate incentives as long as the participants feel fairly treated and rewards and sanctions have moderately high probabilities of properly identifying and classifying organizations and personnel. However, when an evaluation policy misidentifies organizations or personnel systematically due to demographic characteristics or based on statistical procedures that cannot reliably separate those organizations or personnel that are performing much better or worse than the average, such a policy cannot be expected to provide appropriate incentives and sanctions. Unfair evaluation policies may even do damage when organizational or individual morale declines or people, especially high performers, leave such organizations. In order to reduce bad policy decisions made on short-term measures, immediate policy changes should require that state Departments of Education provide standard errors when releasing pass-rate data, growth model measures, and associated year-to-year percentage changes.

The recent central focus of policymakers on annual score improvements engendered by both state accountability systems and national policies arising from No Child Left Behind may have caused large misallocations of effort by researchers and policymakers as well as misallocations of resources that have flowed based on this short-term focus. But research has been slow to provide alternatives to short-term analysis. More recent research that focuses on estimating annual teacher value-added measures has only increased the focus on short-term gains. Interestingly, Ballou, Sanders, and Wright (2004), have suggested that estimating accurate value-added measures requires observation of growth for several years for a given teacher. Ewing (2011) summarizes the issues with short-term value-added measures that have been raised by other researchers.

A shift of research focus toward explaining long-term gains may require more emphasis on the use of longitudinal data, especially data beginning collection prior to school, and a shift of emphasis toward formation of early developmental and academic skills prior to school entry in preschool and at home. Recent research has suggested that achievement at 8th grade in math, reading, and science is mainly accounted for by three early developing skills that may correspond to the early formation and use of neural networks that are used for executive function and procedural and declarative learning (Grissmer et al., 2010). Ironically, sustained and cumulative gains may require emphasis on skills learned mainly outside of schools and far removed in time from when they are used.

References

Ballou, D., Sanders, W., & Wright, P. (2004). Controlling for student background in value-added assessment of teachers. *Journal of Educational and Behavioral Statistics*, 21(1), 37–66. http://dx.doi.org/10.3102/10769986029001037

Brewer, D.J., Krop, C., Gill, B.P. & Reichardt, R. (1999). Estimating the cost of national class size reductions under different policy alternatives. *Educational Evaluation and Policy Analysis*, 21(2), 179–192. http://dx.doi.org/10.3102/01623737021002179

Brown, K.T. (2008). Testing the testing: Validity of a state growth model. *International Journal of Education Policy and Leadership*, 3(6), 1–14. Retrieved from http://journals.sfu.ca/ijepl/index.php/ijepl/article/view/106/51

Coleman, J.S., Campbell, E.Q., Hobson, C. J., McPartland, J., Mood, A.M., Weinfeld, F.D., & York, R.L. (1966). *Equality of educational opportunity*. Washington, DC: U.S. Government Printing Office.

Ewing, J. (2011). Mathematical Intimidation: Driven by the Data. *Notices of the American Mathematical Society, 58, 667-673*. Retrieved from http://www.ams.org/notices/201105/rtx110500667p.pdf

Finn, J.D., & Achilles, C.M. (1999). Tennessee's class size study: Findings, implications, misconceptions. *Educational Evaluation and Policy Analysis*, 21(2), 97–109. http://dx.doi.org/10.3102/01623737021002097

Grissmer, D., Kawata, J. & Williamson, S. (1998). Why Did Black Test Scores Rise Rapidly in the 1970s and 1980s, in *The Black-White Test Score Gap*, Christopher Jencks and Meredith Phillips (Eds.). Brookings Institution, Washington, D.C.

Grissmer, D., Flanagan, A., Kawata, J.H., & Williamson, S. (2000). *Improving Student Achievement: What Do State NAEP Scores Tell Us*. Santa Monica, Calif.: RAND Corporation, MR-924-EDU, 2000. Retrieved from http://www.rand.org/pubs/monograph_reports/MR924

Grissmer, D., & Flanagan, A., (2006). *Improving the Achievement of Tennessee Students: Analysis of the National Assessment of Educational Progress*. TR-381-EDU, 2006. Retrieved from http://www.rand.org/pubs/technical_reports/TR381.html

Grissmer, D., Grimm, K.J., Aiyer, S.M., Murrah, W.M., & Steele, J.S. (2010). Fine Motor Skills and Attention: Primary Developmental Predictors of Later Achievement. *Developmental Psychology*, 46(5): 1008–1017. http://dx.doi.org/10.1037/a0020104

Hamilton, L. & Stecher, B. (2006). *Measuring Instructional Responses to Standards-Based Accountability*. Santa Monica, CA: Rand. Retrieved from http://130.154.3.14/content/dam/rand/pubs/working_papers/2006/RAND_WR373.pdf

Indiana Department of Education [DOE] (2011). *A-F Accountability*. Retrieved June 5, 2013, from http://www.doe.in.gov/improvement/accountability/f-accountability

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* (pp. 235–269). Washington, DC: Brookings Institution.

Kim, J.S., & Sunderman, G.L. (2005). Measuring academic proficiency under the No Child Left Behind Act: Implications for educational equity. *Educational Researcher*, 34 (8), 3–13. http://dx.doi.org/10.3102/0013189X034008003

Krueger, A. B. (1999). Experimental estimates of education production functions. *Quarterly Journal of Economics*, 114(2), 497–532. http://dx.doi.org/10.1162/003355399556052

Krueger, A.B., & Whitmore, D.M. (2001). The effect of attending a small class in the early grades on college-test taking and middle school test results: Evidence from project star. *The Economic Journal*, 111(468), 1–28. http://dx.doi.org/10.1111/1468-0297.00586

Linn, R.L., & Haug, C. (2002). Stability of school-building accountability scores and gains. *Educational Evaluation and Policy Analysis*, 24(1), 29–36. http://dx.doi.org/10.3102/01623737024001029

Linn, R.L., Baker, E.L., & Herman, J.L. (2002, Fall). Minimum group size for measuring adequate yearly progress. The CRESST Line, 1, 4–5.

Marion, S.T., White, C., Carlson, D., Erpenbach, W.J., Rabinowitz, A., & Sheinker, J. (2002). *Making valid and reliable decisions in determining adequate yearly progress*. Washington, DC: Council of Chief State School Officers. Retrieved from http://programs.ccsso.org/content/pdfs/AYPpaper.pdf

Metropolis, N. & Ulam, S. (1949). The Monte Carlo method, *Journal of the American Statistical Association*, 44 (247), 335–341. http://dx.doi.org/10.1080/01621459.1949.10483310

Mintrop, H. & Trujillo, T.M. (2005). Corrective action in low performing schools: Lessons for NCLB implementation from first-generation accountability systems. *Education Policy Analysis Archives,13*(48). Retrieved from http://epaa.asu.edu/epaa/v13n48/

Raudenbush, S.W. (2004). Schooling, statistics, and poverty: Can we measure school improvement? Princeton, NJ, Educational Testing Service. Retrieved from http://www.ets.org/Media/Education_Topics/pdf/angoff9.pdf

Rogosa, D. (2005), Statistical misunderstandings of the properties of school scores and school accountability. *Yearbook of the National Society for the Study of Education*, 104: 147–174. http://dx.doi.org/10.1111/j.1744-7984.2005.00029.x

Rothstein, R., (2008), Grading education: Getting accountability right. Economic Policy Institute, Washington, D.C.

Stecher, B., & Hamilton. L. (2002). Putting Theory To The Test: Systems of "Educational Accountability" Should Be Held Accountable, *RAND Review*, 26(1), 16–23. Retrieved from http://www.rand.org/publications/randreview/issues/rr-04-02/theory.html

Stecher, B., Hamilton. L., & Gonzalez, G. (2003). *Working Smarter to Leave No Child Behind Practical, Insights for School Leaders*. Santa Monica, CA: Rand. Retrieved from http://www.rand.org/pubs/white_papers/WP138/WP138.pdf

Thompson, B.R. (2004), Equitable measurement of school effectiveness. *Urban Education*, 39 (2), 200–229 Retrieved from http://people.msoe.edu/~thompson/Papers/Article-Model.pdf

Usher, A., (2012). AYP Results for 2010–11 - November 2012 Update. Center on Education Policy. (2012), Washington, D.C. 20037. Retrieved from http://www.cep-dc.org/index.cfm?DocumentSubSubTopicID=8

Wiley, E., Mathis, W., & Garcia, D. (2005). *The Impact of the Adequate Yearly Progress Requirement of the Federal No Child Left Behind Act on Schools in the Great Lakes Region.* East Lansing, MI: Great Lakes Center for Education Research and Practice. Retrieved from http://greatlakescenter.org/docs/early_research/g_l_new_doc/EPSL-0505-109-EPRU.Great_lakes.pdf

Winston, W. (2004). Adapted from *Microsoft Excel Data Analysis and Business Modeling.* Retrieved from http://office.microsoft.com/en-us/excel-help/introduction-to-monte-carlo-simulation-HA001111893.aspx

Yeh, S.S. (2005). Limiting the unintended consequences of high-stakes testing. *Education Policy Analysis Archives, 13(43).* Retrieved from http://epaa.asu.edu/ojs/article/view/148

Appendix A

Table

Regression Coefficients in Grades 3, 6, 8, and 10 for Estimating Annualized State Gains from ISTEP Pass Rates of BOTH (ENLA and Math) Using Family Variables in Equation 1

Variable	Beta (Standard Deviation Units)			
Grade	3	6	8	10
d_{2003}	0.132***	0.118***	0.094***	0.034**
d_{2004}	0.191***	0.194***	0.171***	-0.142***
d_{2005}	0.217***	0.261***	0.207***	-0.147***
d_{2006}	0.212***	0.292***	0.190***	0.012
d_{2007}	0.212***	0.320***	0.256***	0.053***
Free Reduced Lunch	-0.399***	-0.413***	-0.349***	-0.330***
Black	-0.225***	-0.213***	-0.306***	-0.388***
Special Education	-0.103***	-0.190***	-0.174***	-0.199***
BS Degree	0.122***	0.122***	0.182***	0.176***
Single Female	0.120***	-0.031	-0.060***	-0.034*
Hispanic	-0.111***	-0.082***	-0.109***	-0.028**
Less than HS Education	-0.019	0.084***	0.085***	0.077***
Single Male	0.023*	-0.020*	-0.053***	-0.109***
Median Family Income	0.090***	0.025	0.104***	0.136***
Asian	0.031***	0.065***	0.029**	0.006
Multiracial	-0.084***	-0.064***	0.003	-0.025**
ESL-LEP	-0.071***	-0.039***	-0.021**	0.008
College Plus	-0.013	0.063***	0.048***	0.024
American Indian	-0.018**	-0.033***	-0.016*	-0.009
Adjusted R^2	0.467	0.583	0.754	0.730
N-Schools	862	455	378	334

*= $p < 0.10$, **= $p < 0.05$, ***= $p < 0.01$

Appendix B

Table

Regression Coefficients in Grades 3, 6, 8, and 10 for Estimating Six-year Annualized School Gains from ISTEP Pass Rates of BOTH (ENLA and Math) Using Family Variables in Equation 2

Variable	Beta (Standard Deviation Units)			
Grade	3	6	8	10
Free Reduced Lunch	-0.124***	-0.154***	-0.078***	-0.066***
Black	-0.325***	-0.244***	-0.379***	-0.384***
Hispanic	-0.195***	-0.102**	-0.142***	-0.110***
Median Family Income	0.224***	0.251**	0.236***	0.328***
Asian	-0.019	-0.008	-0.020	0.025
American Indian	-0.017**	-0.015*	-0.007	0.030***
Multiracial	-0.022*	-0.002	-0.009	-0.009
Adjusted R^2	0.712	0.770	0.859	0.869
N-Schools	862	455	378	334

*= $p < 0.10$, **= $p < 0.05$, ***= $p < 0.01$

Note The following comparison category variables were excluded from the regressions: White (Ethnicity), HS Education (Parent Education Level), and Married Status (Household Head). The following variables were Excluded (census) variables in each of the above sets of analysis: BS Degree, Single Male, College Plus, Less HS Ed, and Single Female.

Appendix C

Table C1

Comparisons of percentages of schools with statistically significant (95 percent confidence) gains and declines in four-, five-, and six-year pass rates (2004-2007, 2003-2007, and 2002-2007, respectively) for BOTH (ENLA and Math) estimated with and without demographic characteristics

Years of Trend	Grade	N Schools	OLS Trends Percent		Regression Trends Percent		Regression-OLS Percent Difference	
			Gains	Declines	Gains	Declines	Gains	Declines
4	3	862	11.8	18.8	13.2	16.5	1.4	-2.3
5			10.8	14.4	13.1	12.1	2.3	-2.3
6			17.9	8.4	22.7	6.0	4.8	-2.4
4	6	455	18.2	7.7	19.3	6.6	1.1	-1.1
5			22.9	4.0	25.9	2.9	3.0	-1.1
6			37.8	3.5	42.4	2.4	4.6	-1.1
4	8	378	15.6	11.9	18.8	10.8	3.2	-1.1
5			19.3	6.1	23.3	5.0	4.0	-1.1
6			28.6	3.2	34.4	1.3	5.8	-1.9
4	10	334	15.0	9.9	18.8	9.9	3.8	0.0
5			7.5	14.4	10.8	11.4	3.3	-3.0
6			6.3	23.4	8.4	16.2	2.1	-7.2
4	Total Schools	2029	14.5	13.6	16.7	12.1	2.2	-1.5
5			14.5	10.5	17.5	8.6	3.0	-1.9
6			22.4	8.8	27.0	6.0	4.6	-2.8

Table C2

Comparisons of percentages of schools with statistically significant (95 percent confidence) gains and declines in four-, five-, and six-year pass rates for BOTH (ENLA and Math) estimated with demographic characteristics and compared to Monte Carlo Simulations of random data

Years of Trend	Grade	N Schools	Regression Trends Percent		Monte Carlo OLS Trends Percent	
			Gains	Declines	Gains	Declines
4	3	862	13.2	16.5	13.6	12.4
5			13.1	12.1	10.1	9.0
6			22.7	6.0	8.6	10.1
4	6	455	19.3	6.6	11.6	15.2
5			25.9	2.9	9.7	9.0
6			42.4	2.4	8.8	8.4
4	8	378	18.8	10.8	16.4	11.4
5			23.3	5.0	12.7	9.3
6			34.4	1.3	8.5	8.2
4	10	334	18.8	9.9	10.8	10.8
5			10.8	11.4	9.3	8.7
6			8.4	16.2	6.9	7.5
4	Total Schools	2029	16.7	12.1	13.2	12.6
5			17.5	8.6	10.3	9.0
6			27.0	6.0	8.3	8.9

About the Authors

David W. Grissmer
University of Virginia
dwg7u@virginia.edu
Dr. Grissmer is a Research Professor in the Center for the Advanced Study of Teaching and Learning at the University of Virginia, Charlottesville, Virginia. Prior to 2006 he was a Senior Management Scientist in the Washington, DC, Office of Rand Corporation. His current research interests are directed toward understanding the origin of the gaps in achievement between Black, Hispanic, and White students, and between advantaged and disadvantaged students. He also has investigated the developmental origins of these cognitive gaps prior to school entry using the ECLS-K and ECLS-B data bases. He is currently Co-PI on two major random controlled trials involving evaluations of an after school socio-emotional program and CORE Knowledge charter schools.

David R. Ober
Ball State University
dober@bsu.edu
Dr. Ober is Department Chairperson and Professor of Physics and Astronomy Emeritus at Ball State University. He has served as PI and Co-PI of summer instruction programs for high ability high school students and physics teacher preparation and retention initiatives. He directed a university sponsored updating/retaining program for physics and physical science teachers. He also served for four years as a Co-PI to the university's Physics Teacher Education Coalition (PhysTEC) project that was initiated in 2001 by the American Physical Society and the American Association of Physics Teachers to address the critical shortage of qualified physics and physical science teachers.

John A. Beekman
Ball State University
jjbeekman@comcast.net
Dr. Beekman is Distinguished Professor Emeritus of Mathematics and Actuarial Science at Ball State University. His research and writing have been in the areas of stochastic processes, mathematical statistics, partial differential equations, actuarial models, actuarial projections for the U.S. Social Security, mathematical demography, and educational statistics. He also has actuarial experience in the life insurance industry. Since establishing the university's graduate and undergraduate Actuarial Science programs in 1970, over 200 Actuarial Science students at Ball State University have received a part of their actuarial training from Dr. Beekman.

Acknowledgements

The authors acknowledge with much gratitude the helpful comments and suggestions made by the editor and the reviewers to improve this manuscript.

education policy analysis archives

Volume 22 Number 5 February 3rd, 2014 ISSN 1068-2341

education policy analysis archives
editorial board