JOURNAL OF RESEARCH IN SCIENCE TEACHING VOL. 49, NO. 6, PP. 804–830 (2012)

Research Article

The Consequences of "School Improvement": Examining the Association Between Two Standardized Assessments Measuring School Improvement and Student Science Achievement

Adam V. Maltese[1] and Craig D. Hochbein[2]

[1]*School of Education, Indiana University, 201 North Rose Avenue, Bloomington, Indiana 47404*
[2]*College of Education & Human Development, University of Louisville, Louisville, Kentucky*

Received 14 January 2012; Accepted 4 June 2012

Abstract: For more than half a century concerns about the ability of American students to compete in a global workplace focused policymakers' attention on improving school performance generally, and student achievement in science, technology, engineering, and mathematics (STEM) specifically. In its most recent form—No Child Left Behind—there is evidence this focus led to a repurposing of instructional time to dedicate more attention to tested subjects. While this meant a narrowing of the curriculum to focus on English and mathematics at the elementary level, the effects on high school curricula have been less clear and generally absent from the research literature. In this study, we sought to explore the relationship between school improvement efforts and student achievement in science and thus explore the intersection of school reform and STEM policies. We used school-level data on state standardized test scores in English and math to identify schools as either improving or declining over three consecutive years. We then compared the science achievement of students from these schools as measured by the ACT Science exams. Our findings from three consecutive cohorts, including thousands of high school students who attended 12th grade in 2008, 2009, and 2010 indicate that students attending improving schools identified by state administered standardized tests generally performed no better on a widely administered college entrance exam with tests in science, math and English. In 2010, students from schools identified as improving in English scored nearly one-half of a point lower than their peers from declining schools on both the ACT Science and Math exams. We discuss various interpretations and implications of these results and suggest areas for future research. © 2012 Wiley Periodicals, Inc. J Res Sci Teach 49: 804–830, 2012
Keywords: accountability; science education; mathematics; achievement

For decades Americans have expected their public schools to accomplish a multitude of objectives, benefitting both the individual and collective good. As Labaree (2010) observed, "We want schools to provide us with good citizens and productive workers; to give us opportunity and reduce inequality; to improve our health, reduce crime and protect the environment" (p. 1). The prioritization of this throng of school objectives varies by person, place, and time (Payne, 2008; Ravitch, 2000; Tyack & Cuban, 1995). However, public concerns resulting from high-profile events and inauspicious publications about American student performance, specifically in the areas of science, technology, engineering, and mathematics (STEM), have galvanized school improvement and STEM initiatives as priorities on the agendas of educational reformers.

Additional Supporting Information may be found in the online version of this article.
Correspondence to: A.V. Maltese; E-mail: amaltese@indiana.edu
DOI 10.1002/tea.21027
Published online 11 July 2012 in Wiley Online Library (wileyonlinelibrary.com).

At least since scientists in the former Soviet Union launched *Sputnik* in 1957, educational, corporate, and political leaders have voiced concerns about the number of American students adequately prepared to enter STEM careers (Daniels, 2006; Obama, 2009, 2011; President's Council of Advisors on Science and Technology (PCAST), 2010, 2012). More recently, on assessments of scientific and mathematics literacy American students scored below many of their international peers (Organization for Economic Cooperation and Development (OECD), 2007, 2010; Provasnik, Gonzales, & Miller, 2009). In addition, international students have comprised an increasing number of post-secondary graduates in STEM fields (National Science Foundation, 2011). Policy experts warned that continued decline in the preparation of American students for careers in STEM fields would weaken the US economy and harm the competitiveness of American corporations in the increasingly global marketplace (National Academy of Sciences (NAS), 2005; PCAST, 2010, 2012).

Similarly, after the *Brown v. the Board of Education of Topeka* ruling by the Supreme Court in 1954 and the publication of the Coleman Report in 1966, educators, policymakers, and researchers searched for methods to improve the educational opportunities for all students through effective and improved schools (Purkey & Smith, 1983; Sammons, 2007; Teddlie & Reynolds, 2000; Wimpelberg, Teddlie, & Stringfield, 1989). To stem the rising tide of mediocrity noted in *A Nation at Risk* (1983) and reduce the savage inequalities observed by Kozol (1991), federal legislators required public reporting of school performance in the areas of literacy and numeracy under the No Child Left Behind Act of 2002 (NCLB). More recently, US Secretary of Education Arne Duncan challenged educators, policymakers, and researchers to turn around the lowest achieving schools in the country (Duncan, 2009), and supported his challenge with federal grant initiatives, such as Race to the Top and School Improvement Grants (Maxwell, 2009).

The purpose of this analysis was to explore the association between educational policies intended to improve school performance and those seeking to bolster student achievement in science. Specifically, we used hierarchical linear modeling (HLM) to determine if students who attended schools identified as improving fared better on a science section of a common college admissions exam than their peers from declining schools. Advocates of accountability reforms assume that higher performing schools better prepare students to successfully enter post-secondary institutions and compete in the global workplace. However, some authors have suggested that the limited metrics utilized by school accountability agencies inherently narrow the curriculum of schools (Ravitch, 2010; Rothstein, 2009; Rothstein, Jacobsen, & Wilder, 2009), and a limited body of research has supported such hypotheses (Brown & Clift, 2010; Spillane, Diamond, Walker, Halverson, & Jita, 2001). To achieve our purpose, we addressed two research questions:

(1) How is school improvement in English achievement associated with student achievement on the science section of a college admissions exam?
(2) How is school improvement in mathematics achievement associated with student achievement on the science section of a college admissions exam?

In the text that follows we present background information on extant research related to these questions and the analysis we completed. First we discuss policy objectives and research related to improvement of STEM throughout the US. We follow this with a more general discussion of policy and research in the area of school improvement. The review ends with a discussion of the role assessment plays in evaluating the research and the effectiveness of existing and recommended policy measures.

STEM Improvement

In much of the educational research literature authors synonymously use the terms science and STEM. Sometimes STEM will be used to discuss work involving both science and math, but rarely do papers explicitly include discussion of all four STEM content areas. In this paper, we are making a distinction that although science and math are anchoring members of the STEM policy efforts, educators, and policymakers treat the subjects quite differently. Math functions as a required component of school improvement efforts and therefore has been tested since initial implementation of NCLB. In contrast, state departments of education might administer standardized science tests, but not include the results as part of school accountability systems. Finally, the new Framework for K-12 Science Education stated the clear linkage in the STE components of STEM; however the science we write of in this paper does not incorporate the technology or the engineering components (National Research Council (NRC), 2011).

Keeping the US at the forefront of research and innovation is a common talking point at the highest levels of government (Obama, 2011). *A Blueprint for Reform: The Reauthorization of the Elementary and Secondary Education Act* released by the United States Department of Education (USDOE) stated, ''America needs to increase the number of students pursuing STEM fields in their academic studies and careers, and improve preparation for the next generation of engineers, scientists, mathematicians, and technicians'' (USDOE, 2010a, p. 1). Preceding many of the recent calls-to-action was the NRC report *Rising Above the Gathering Storm* (NAS, 2005), which discussed the condition of STEM fields in the US and stressed the importance of STEM education for maintaining our technical workforce. Based on such reports policy initiatives have extolled the need to increase the rigor of math and science preparation in US schools in order to bolster the STEM workforce (Bill, H. R. 5116, 2010; Singer, Hilton, & Schweingruber, 2006). Unfortunately, the federal, state, district, and school level interventions required to increase the number of students pursuing degrees in STEM remain unclear.

In 2009, *Educate to Innovate* (Obama, 2009) was initiated by the White House to address poor results on both national (NAEP) and international (PISA, TIMSS) assessments in math and science. This initiative espoused the tripartite goals of increasing STEM literacy for all students, moving the US from the middle to the front of the pack on international assessments, and expanding the educational and career goals for traditionally underrepresented groups in STEM fields. To accomplish these objectives supporters stressed the need to form public–private partnerships, increase the use of technology and hands-on instruction in classrooms, and increase the visibility of STEM initiatives and efforts to increase student knowledge and interest (Obama, 2009, 2010).

In addition to President Obama's stated goal of moving our students to the elite cadre of nations leading the world in science and math achievement, the USDOE *Blueprint* espoused a clear goal to make students in America college- and career-ready. As part of the assessment plan, the *Blueprint* suggested:

> States and districts will collect and make public data relating to student academic achievement and growth in English language arts and mathematics, student academic achievement in science, and, if states choose, student academic achievement and growth in other subjects, such as history (USDOE, 2010b, p. 8).

Although such a statement enabled the possibility of including student achievement in science as an integral part of the assessment scheme, no official mandate has yet required the inclusion of science.

Despite these calls for improvement across STEM, Spillane et al. (2001) indicated that school administrators often devalued science in light of reforms to improve instruction and student achievement in literacy and math. Spillane et al. found that this effect was particularly strong in elementary schools, and especially in urban districts where educators focus instruction on the basics required to pass standardized reading and math tests. Recent data collected by the National Science Teachers Association (NSTA) from elementary and middle school teachers supported these results. The NSTA (2011) data indicated that 45% of respondents reported a decrease in instructional time for science in the 2010–2011 school year. For those citing this decrease, most attributed the decline to a repurposing of the time to literacy and math instruction.

Similar findings from a report by the Center on Education Policy (McMurrer, 2008) indicated that since the implementation of NCLB an average decline in weekly science instructional time of 33% occurred in elementary schools that reported increases in time spent teaching English and math. Analysis of nationally representative data gathered by the National Center for Education Statistics in the Schools and Staffing Survey (Jacob & Dee, 2010) and an independent study by RAND (Hamilton et al., 2007), supported the notion that elementary and middle school teachers and administrators made shifts in instructional time to bolster instruction in English Language Arts and math as a consequence of NCLB. These findings seem in direct opposition to the recent recommendations by the NRC (2011) for the inclusion of ''adequate instructional time and resources'' when teaching science in elementary grades (p. 27).

Although we searched for literature that evaluated changes in instructional time in high school as the result of NCLB, we did not find any studies that specifically presented data to support or refute such changes. It may be that curriculum narrowing is less common in high school settings because core classes are treated as distinct units of credit that are generally fixed within a student's schedule. Yet we believe that it is well within reason that high school administrators and faculty will do whatever they can to shift instructional resources to areas of critical need when facing the threat of sanctions. This notion is supported by a summary of research on the impacts of an accountability system in Texas predating NCLB. McNeil and Valenzuela (2001) found high school administrators required teachers to squeeze test preparation and extra literacy or math instruction into non-tested classes, such as science and social studies, a practice that was most common in the lowest performing schools.

However, the recent PCAST (2010) report stated that the focus should not just be on remediation of the lowest-performing students, ''Even as the United States focuses on low-performing students, we must devote considerable attention and resources to all of our most high-achieving students from across all groups'' (p. viii). This sentiment is echoed by the NRC (2011) report as a way to better prepare all citizens for life in the 21st century. When school administrators and teachers feel pressured from policies to focus efforts on literacy and math performance for all students, while simultaneously asked to improve student performance in science, such policies and reforms may operate at cross purposes.

School Effectiveness and School Improvement

In response to research findings that questioned the capability of schools to overcome inherent disparities imposed by student background characteristics (Coleman et al., 1966; Jencks, 1972), educators, researchers, and policymakers designed, implemented, and tested a multitude of reforms intended to improve school performance (Borman, Hewes, Overman, & Brown, 2003; Edmonds, 1977; Heck & Mayor, 1993; Klitgaard & Hall, 1975; Teddlie & Reynolds, 2000; Tyack & Cuban, 1995). Schools demonstrating effectiveness in teaching all

students, including those from disadvantaged backgrounds, as well as schools achieving dramatic improvement exist (Austin, 1979; Bryk, Sebring, Allensworth, Luppescu, & Easton, 2010; Duke & Jacobson, 2011; Johnson & Asera, 1999; Teddlie & Stringfield, 1993). Unfortunately, researchers have been unable to discern a universal set of methods or characteristics that reliably produces, replicates, or sustains such school achievement (Payne, 2008; Ravitch, 2000). Reviews of research such as Sammons (2007) and Teddlie and Reynolds (2000) highlight common discoveries and lessons derived from effective and improved schools, but findings like Stuit (2010) and Balfanz, Bridgeland, Moore, and Fox (2010) demonstrated the need for enhanced school improvement techniques.

From the continued existence of persistently low-achieving schools, as well as critiques of prior research and initiatives (Purkey & Smith, 1983; Rowan, Bossert, & Dwyer, 1983; Wimpelberg, Teddlie, & Stringfield, 1989) developed a multitude of reforms attempting to improve school performance (Datnow, Borman, Stringfield, Overman, & Castellano, 2003). Among these reforms, which primarily defined school performance using math and literacy metrics, school accountability, and school turnaround became federal strategies to improve persistently low-achieving schools. Accountability efforts in both the US and the United Kingdom publicly reported school performance results based upon student achievement and implemented sanctions for schools identified with substandard performance (Matthews & Sammons, 2004, 2005; Stringfield & Yakimowski-Srebnick, 2005). Advocates of accountability policies assumed that public reporting and impending sanctions would spur improvement in under- and low-achieving schools, as well as close the achievement gaps that exist between different racial and socioeconomic groups of students (Forte, 2010; Kane & Staiger, 2002). In the US, schools received state sanctions and interventions when they failed to demonstrate AYP which was based upon student achievement on standardized tests of reading and math (Commission on NCLB, 2007; Hemelt, 2010).

A second school improvement derivation, school turnaround, evolved from strategies and operations that business leaders utilized to turnaround bankrupt or declining organizations (Duke et al., 2005; Murphy & Meyers, 2008). Adapting techniques from for-profit organizations, some educational leaders achieved dramatic improvement in a condensed period of time (Duke, 2007; Duke & Jacobson, 2011; Duke & Salmonowicz, 2010; Picucci, Brownson, Kahlert, & Sobel, 2002). Attempting to discern effective characteristics and operations of turnarounds, Herman et al. (2008) conducted a review of school turnaround research that restricted findings to examples of turnaround exhibiting a 10% increase in reading or math improvement within a 3-year period of time. Similar to prior work assessing the effectiveness or improvement of schools, the researchers operationally defined school performance using student achievement in only literacy or math.

Throughout the history of school reform, educators, policymakers, and researchers have relied most heavily upon literacy and numeracy metrics to identify and evaluate school performance (Brookover & Lezotte, 1979; Bryk et al., 2010; Duke, 2007; Edmonds, 1979; Gray, Goldstein, & Thomas, 2001; Hemelt, 2010; Opdenakker & Van Damme, 2006; Palardy, 2008; Purkey & Smith, 1983; Teddlie & Stringfield, 1993; Wimpelberg et al., 1989). Results from the testing of these two outcome measures often compelled educational leaders to tinker with operations and policies (Tyack & Cuban, 1995). Proponents of such data-driven decision-making assumed that improved literacy and math scores would produce a trickledown effect, increasing other desirable outcome measures, such as graduation rates, college and career readiness, and also science preparation. However, because of policies that prescribe severe sanctions for schools with poor performance in literacy and math (Dillon, 2011; Maxwell, 2009), educators might focus their resources on these subject areas at the expense of content

in non-tested subject areas, including science (Brown & Clift, 2010; McNeil & Valenzuela, 2001; NSTA, 2011; Spillane et al., 2001).

The Critical Role of Assessment

NCLB and other federal reform initiatives have created an unprecedented level of focus on student and school assessment. In 2001, NCLB directly tied student performance on standardized exams to a school's eligibility for federal funding. More recently, some state educational leaders have considered the use of student performance on state administered assessment to evaluate and compensate teachers (Cavanaugh, 2011). Some policymakers have even suggested using such assessments to determine the effectiveness of teacher preparation programs (USDOE, 2011).

With such important outcomes in the balance, one might expect that the ''high stakes'' standardized assessments at the core of these educational issues would be stringently evaluated. Although state administered assessments might satisfy psychometric scrutiny for content validity and internal reliability, a lack of information exists about the external validity of student and school level results with other desirable outcomes, such as graduation rates, discipline referrals, or college and career readiness. For instance, consider the implementation of state administered assessments in the state of Indiana, which is the focus of our analysis. Students across primary and secondary education take the Indiana Statewide Testing for Educational Progress (ISTEP) to not only measure individual academic progress, but also to evaluate Indiana teachers, schools and districts. Ruiz-Primo, Shavelson, Hamilton, and Klein (2002) would consider the ISTEP a ''distal'' assessment because it is a state-level test aligned with the published academic standards in various content areas.

Within the 2010–2011 ISTEP program manual the Indiana Department of Education (IDOE) presented statistical evidence from the 2009 ISTEP administration to support the notion that the assessments provide data that can lead to both valid and reliable conclusions about students' achievement levels from various racial/ethnic subgroups and across multiple content domains (IDOE, 2010). The presented data support claims of reliability including: item level reliability (point biserial correlation, differential item functioning), test level reliability (Cronbach's alpha, standard error of measurement), classification consistency and classification accuracy. The manual also discussed attempts to establish validity for the assessments including evidence from an exploratory/confirmatory factor analysis that produced a single factor for each of the content assessments. These statistical results generally support the notions of validity and reliability of the data produced by the ISTEP exams; however it is not clear if the data presented in the manual[1] included the End of Course Assessments (ECAs) taken by high school students in English, math (Algebra), and science (Biology). Additionally, little information describes how student performance on the ISTEP exams may generalize to other recognized indicators of knowledge, beyond the Indiana Academic Standards.

Although many of the principles undergirding school accountability seem reasonable, the poor execution of accountability measures and sanctions has raised concerns about the current system (Kahle, 2004; Stringfield & Yakimowski-Srebnick, 2005; Wood, Lawrenz, Huffman, & Schultz, 2006). From a technical standpoint, Penfield and Lee (2010) discussed how these high stakes measures, which are meant to ensure that no child is left behind, are biased against many students who take these state administered tests. The authors raised concerns that the linguistic complexity and context of items negatively impact non-majority students and lead the users of test results to make invalid conclusions based on scores from minority

students. Kane and Staiger (2002) questioned the sanctions resulting from such testing results and concluded, ''The problem resides not with the measures themselves, but with the way that these measures are often used,'' (p. 100).

Although NCLB did not mandate the inclusion of science assessments in the determination of AYP, states can include science assessments to provide extra evaluative data or as one of an additional set of academic indicators. While it is not clear how many states currently include science assessments in their AYP calculations, there are indications that most or all states now include some form of science assessments as part of their testing regimen (Penfield & Lee, 2010). Similar to their counterparts across the US, the annually administered ISTEP tests are a fixture in the academic experiences of Indiana students and teachers (Geier et al., 2008). These assessments operate in multiple capacities to provide both an indicator of necessary reform within schools and districts, as well as a measure of change within these organizations over time.

Methods

Research Design

Proponents of school accountability reforms assume public reporting of school performance in the subjects of English and math will provide sufficient incentives for educators in low-achieving schools to initiate and sustain necessary improvements to raise levels of student achievement. However, Campbell's Law asserts, ''The more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor'' (Campbell, 1976, p. 49). Invoking Campbell's Law Rothstein (2009) questioned the utility of school accountability measures, ''Attempts to hold schools accountable for math and reading test scores have corrupted education by reducing the attention paid to other important curricular goals'' (p. 23). To explore such assumptions and assertions, we used longitudinal student and school level data to examine the science achievement of students, as measured by a college admissions test, from schools with improving and declining performance, as measured by performance on standardized English and math assessments.

Data Source

The dataset for this analysis came from two separate data streams within IDOE. At the student level we compiled data records for sets of students who were 12th grade students in Indiana high schools in 2008, 2009, or 2010. The student data included school attended and demographic variables for age, gender, and race. These data included students' test scores on standardized assessments including ISTEP (10th grade scores), SAT and ACT exams. The data files also had information pertaining to socioeconomic status, required education services, language proficiency, and the number of AP exams. Although we present some tables comparing groups that include these data, their inclusion in student records seemed neither consistent (i.e., similar variables from different sources lacked agreement) nor complete for all students. Therefore, rather than impute missing values, we excluded these variables from the models. At the school level, we created longitudinal data files for public high schools in the state. These data included annual values related to enrollment, attendance rate and size of graduating class, demographic information (% minority, % Free, and Reduced Meals) as well as mean achievement (ISTEP, SAT, ACT, and % taking ACT). We then linked the student and schools files using the high school that students attended.

Analytical Approach and Rationale

To examine the association between school improvement and student science achievement, we used both school- and student-level results in an HLM analysis. Application of HLM enabled us to better model the inherently nested structure of student achievement within school performance (Raudenbush & Bryk, 2002). We used longitudinal state test results in English/language arts (''English'') and mathematics (''Math'') to identify parallel samples of schools that were improving or declining for each subject area. We made these classifications for three separate cohorts of students who were in residence at the schools during the period of interest. We then applied HLM analysis to school- and student-level measures to assess the relationship between school improvement and individual science test scores on the ACT exam completed during their 12th grade year.

Our analysis focused on students from high schools in Indiana who took the ACT for four reasons. First, we limited the study to the state of Indiana because of unique data access that enabled us to connect students' performances on state administered and college placement assessments with the performance of the school each attended. Second, we analyzed ACT scores because these entrance exams include a science section, are independent of the state administered exams, and essentially available to all students unlike Advanced Placement testing. Third, analysis of the high school level fulfilled a current gap in the school effectiveness and STEM literature because extant school level studies most often utilized elementary schools (Brown & Clift, 2010; McNeil & Valenzuela, 2001; NSTA, 2011; Spillane et al., 2001). Fourth, unlike elementary schools, high school schedules allot specific times and Carnegie units to subject areas. This type of dedicated scheduling reduced the likelihood that differences between schools resulted from limited science instructional time as found by Spillane et al. (2001).

Population and Sample of Schools

All public high schools and their students monitored by the IDOE between 2005 and 2010 comprised the population of subjects considered for inclusion in the analysis. To define the sample of schools, we pared down the population of high schools based on the availability of data elements for each of the separate cohorts. Exclusion of schools with missing annual data eliminated the inclusion of newly opened or closed schools in the analysis. Although recently opened schools might demonstrate improvement and decline, the initiation of operations likely poses a unique set of circumstances unlike those found in more established schools. Similarly, school closure might represent the final phase of a decline cycle, but many school closings result from chronic low-performance or consolidation, rather than a brief period of decline.

To operationally define improving and declining schools, we used yearly school performance on Indiana's ISTEP English and Math tests, which measured English skills and algebra content, respectively. The IDOE administered the ISTEP English and Math tests annually to 10th grade students through fall 2009. Using schools with valid ISTEP results, we created a measure of continual school performance change over each of three consecutive years using mean school ISTEP test scores. For example, for the 2010 cohort of 12th grade students we evaluated school test scores from 2007, 2008, and 2009 such that if scores from ISTEP 2009 $>$ ISTEP 2008 $>$ ISTEP 2007, that school was labeled as ''improving'' across that time period. Conversely, schools with test scores matching the inverse pattern, ISTEP 2009 $<$ ISTEP 2008 $<$ ISTEP 2007, were labeled as ''declining.''

Although our operational definitions' reliance on just three time points and no specified magnitude might appear simplistic, both prior research and pragmatic considerations informed our decisions. The 3-year time period aligns with other examinations of school improvement and decline (Gray et al., 2001; Hochbein, 2012a,b), including the guidelines from the Institute of Education Science for school turnaround (Herman et al., 2008). As Thomas, Peng, and Gray (2007) observed, ''For the majority of schools 3 years of upward movement seems to have been the typical limit'' (p. 280). Furthermore, to curb the potential influence of ceiling and floor effects, we did not specify a minimum magnitude of change in our operational definitions. A magnitude requirement would have excluded high-performing schools demonstrating improvement, as well as low-performing schools enduring decline from the examined sample of schools.

Each of the individual students included in the analysis satisfied three basic criteria. First, they attended a school identified as improving/declining in one of the three cohorts. Second, students must have completed the ACT exam[2] with results included in their 12th grade data file. Third, all students included in the study had complete data records on key demographic variables (gender, race, birth year, etc.). In Supporting Information Appendices A and B we have provided additional data regarding the comparability of schools and students included in our analyses with those excluded from the analyses.

The combination of the longitudinal nature of the data and the extensive grouping utilized in the research design provided practical challenges to the concise nomenclature of subjects included in the analyses. Table 1 provides details about the number of subjects included in the analyses and labels we created for the specific comparison groups. Cohorts referred to the temporal grouping of students and schools. We named the three cohorts by the students' 12th grade year ''2008'' (includes school data 2005–2008), ''2009'' (2006–2009), and ''2010'' (2007–2010). Samples referred to the grouping of schools based on their improvement status (Improve or Decline) for the two subject area ISTEP exams (English or Math) over a 3-year period of time. Therefore, within each of the three 12th grade cohorts we examined four sub-samples of schools.

HLM Models

To model the relationships between school and student performance, we used the *HLM 7* (Raudenbush, Bryk, & Congdon, 2011) software package so that we could account for the hierarchical nature of the students nested within schools. Our purpose was to focus precisely on this relationship, the association between schools with improving or declining performance over a 3-year period and the science achievement of students who attended the school during

Table 1
Count of students and schools included in each of the cohorts and samples

	Cohorts					
	2008		2009		2010	
Samples	Students	Schools	Students	Schools	Students	Schools
English decline	2,181	46	2,207	58	2,902	59
English improve	1,180	41	1,325	38	2,300	55
Math decline	1,644	38	1,371	28	2,263	69
Math improve	1,921	65	3,516	86	1,306	39

the same period of measurement. To build models with parallel samples for this exploration we purposefully sought to keep the models simple. At the student level we included background variables commonly included when modeling achievement including: gender, race/ethnicity, birth year, and 10th grade ISTEP scores in English and math, where the range of scores generally falls between 300 and 850 (IDOE, 2010). The gender (Female = 1, Male = 0) and race/ethnicity variables were entered as dummy variables. Birth year and ISTEP scores were centered around the grand mean for all students from each cohort. We included these variables in the model based on previous research indicating that they often have a significant relationship with student achievement, although sometimes with mixed results (Holme, Richards, Jimerson, & Cohen, 2010; Maerten-Rivera, Myers, Lee, & Penfield, 2010).

At the school level we included the indicator of improvement or decline (Improving schools = 1, Declining schools = 0), the proportion of minority students within the school during each graduating cohort's 10th grade year, and the proportion of students who completed the ACT exam during each cohort's 12th grade year. To account for variations in the demographic composition of schools, we entered the 10th grade minority proportion rather than Free and Reduced-price meals (FARM). Although FARM constitutes the most common metric for capturing student SES in educational literature, the high correlation between FARM and minority proportion ($r \sim 0.68$) led us to leave this variable out of the model based on concerns of multicollinearity. In addition, Harwell and LeBeau (2010) questioned FARM as a valid measure of socioeconomic status, and Pogash (2008) reported that in one large urban district, only 37% of eligible high school students participated in the FARM program. We included the proportion of ACT test-takers as a general proxy measure of the proportion of college-bound students within each school being assessed. Both the proportion minority variable and the proportion of ACT test-takers were centered around the grand mean for all schools in the models. An example of our model predicting ACT Science scores of students from the English sample of the 2010 cohort is shown in Figure 1.

Level-1 Model

$$12^{th}\text{ Grade ACT - SCIENCE} = \beta_0 + \beta_1*(\text{Female}) + \beta_2*(\text{Native American}) + \beta_3*(\text{Black}) + \beta_4*(\text{Asian}) + \beta_5*(\text{Hispanic}) + \beta_6*(\text{Multi-racial}) + \beta_7*(\text{Birth year}) + \beta_8*(10^{th}\text{ Grade ISTEP Language Score (2008)}) + r$$

Level-2 Model

$$\beta_0 = \gamma_{00} + \gamma_{01}*(\text{Percentage taking ACT (2010)}) + \gamma_{02}*(\text{Proportion Minority Students} - 2008) + \gamma_{03}*(\text{Improvement Status 2007-2009 (based on ISTEP Language Scores)}) + u_0$$

$$\beta_1 = \gamma_{10}$$
$$\beta_2 = \gamma_{20}$$
$$\beta_3 = \gamma_{30}$$
$$\beta_4 = \gamma_{40}$$
$$\beta_5 = \gamma_{50}$$
$$\beta_6 = \gamma_{60}$$
$$\beta_7 = \gamma_{70}$$
$$\beta_8 = \gamma_{80}$$

Figure 1. Two-level HLM model predicting ACT Science scores of students from the English sample of the 2010 cohort.

We conducted the analyses using parallel models of cohorts of 12th grade students from 2008, 2009, and 2010 who completed the ACT during that same year. We evaluated school performance using ISTEP English and Math school-level test results separately. As outcomes for each of these models, we assessed the association of student- and school-level factors on ACT scores in science. As a comparison for the ACT Science results, we also ran models with ACT English and Math results as dependent variables. Model results for each factor indicated the differences in ACT performance while holding all other factors constant.

Analysis and Results

Descriptive Analysis

The extensive amount of descriptive data calculated from the three cohorts revealed few systematic patterns of similarity between the academic and non-academic measures of the improving and declining schools. For instance, we found no discernible pattern among the percentage of students taking college entrance exams (Table 2). The percentage of students taking the ACT or SAT varied inconsistently between improving and declining schools, between the English and math groups, and longitudinally among the cohorts. In addition, aggregate school-level ACT performance in English, math, and science demonstrated no consistent patterns between or within the 12 sub-samples of schools.

Despite the lack of trends among the ACT performance data, the operational definitions of school improvement and decline identified samples of schools with definitive patterns among other academic and non-academic measures. Comparison of the cohorts revealed that the declining and improving samples manifested with inverted longitudinal achievement based on ISTEP performance (Figure 2). For each of the improving-declining paired samples the initial mean achievement of the improving schools was below the mean achievement of the declining schools. However, three years later the mean achievement of five of the six improving school samples exceeded the mean achievement of the declining counterpart. Interestingly, each of the improving samples also demonstrated a continuous longitudinal decrease in their corresponding SAT subject area performance (Table 2). The improving samples exhibited no less than a five-point cumulative decrease on the mean SAT verbal score and two points on the SAT math score. Declining schools showed no consistent pattern across the cohorts or subject areas.

Additional trends appeared among the school composition factors. In each cohort, for example, declining schools tended to enroll more students. All 12 samples demonstrated a continual increase in both the percentage of FARM and minority students. Both the 2008 and 2009 cohorts exhibited larger percentages of minority students in the English and math declining samples, but that trend reversed for the 2010 cohort. A similar pattern manifested within the percentage of FARM students, although it was not as substantial.

The improving and declining sample sizes ranged from 1,180 to 3,516 students, demonstrating similar student characteristics across all samples. Female students represented the majority of the ACT test-takers, consistently representing approximately 57% of both the declining and improving samples (Table 3). White students comprised the majority of each sample, ranging between 67% and 92%, with black students constituting the largest single minority group. In five of the six paired samples, the declining sample exhibited a greater proportion of minority students than the improving sample, with 2010 Math providing the lone exception. All 12 samples demonstrated similar percentages of students with limited English proficiency, as well as those who required special education services. The percentage

Table 2
Comparison of school samples demographics by ISTEP assessment, cohort, and improvement status

	English						Mathematics					
	2008		2009		2010		2008		2009		2010	
	Decline ($n = 46$)	Improve ($n = 41$)	Decline ($n = 58$)	Improve ($n = 65$)	Decline ($n = 59$)	Improve ($n = 65$)	Decline ($n = 65$)	Improve ($n = 65$)	Decline ($n = 28$)	Improve ($n = 86$)	Decline ($n = 69$)	Improve ($n = 39$)
ISTEP												
04-05	575 (14)	569 (10)					612 (23)	606 (19)				
05-06	571 (14)	572 (10)	575 (17)	568 (11)			606 (23)	612 (18)	612 (19)	601 (24)		
06-07	567 (14)	577 (10)	571 (17)	573 (12)	578 (11)	573 (12)	590 (33)	619 (18)	605 (17)	609 (24)	622 (18)	590 (33)
07-08			566 (16)	577 (13)	574 (11)	569 (13)			599 (18)	616 (23)	616 (18)	596 (32)
08-09					570 (12)	573 (13)					609 (17)	601 (31)
ACT Science												
04-05	NA	NA					NA	NA				
05-06	21.0 (1.7)	21.5 (1.4)	20.9 (2.2)	21.1 (1.6)			21.2 (1.7)	21.3 (1.5)	21.0 (1.7)	21.1 (1.6)		
06-07	21.4 (1.8)	21.3 (1.9)	21.0 (2.1)	21.3 (2.1)	21.3 (1.7)	20.9 (2.2)	21.7 (1.8)	21.4 (14)	21.1 (1.6)	21.3 (1.6)	21.5 (1.5)	20.3 (2.6)
07-08			21.1 (2.1)	21.3 (1.5)	21.3 (1.9)	20.9 (1.5)			21.4 (2.0)	21.2 (1.7)	21.8 (1.5)	20.3 (1.9)
08-09					21.5 (1.7)	20.7 (1.9)					21.8 (1.4)	20.5 (2.0)
ACT English												
04-05	NA	NA					NA	NA				
05-06	20.7 (2.2)	20.7 (1.7)	20.6 (2.8)	20.4 (2.0)			20.7 (2.5)	20.7 (1.7)	20.6 (2.5)	20.6 (2.0)		
06-07	21.3 (2.1)	20.5 (2.0)	20.5 (2.7)	20.8 (3.1)	20.8 (2.2)	21.0 (1.9)	21 4 (2 3)	21 0 (1 9)	20 5 (2 1)	20 8 (2 1)	21 0 (2.1)	20 0 (3 2)
07-08			20.8 (2.7)	20.8 (2.1)	20.5 (2.3)	20.3 (2.0)			20.9 (2.6)	20.6 (2.1)	21.4 (2.0)	20.0 (2.7)
08-09					21.4 (2.1)	20.0 (2.0)					21.5 (1.9)	19.9 (2.4)
ACT Math												
04-05	NA	NA					NA	NA				
05-06	21.3 (2.0)	21.8 (1.7)	21.2 (2.3)	21.3 (1.9)			21.2 (2.0)	21.5 (2.0)	21.4 (1.8)	21.2 (2.0)		
06-07	21.6 (2.0)	21.6 (2.0)	21.3 (2.4)	21.7 (2.4)	21.5 (2.3)	21.3 (2.6)	21.7 (1.9)	21.8 (1.8)	21.3 (1.6)	21.6 (1.9)	22.0 (2.0)	20.6 (2.8)
07-08			21.7 (2.4)	22.1 (1.9)	21.9 (2.2)	21.6 (2.0)			22.1 (2.3)	21.8 (2.1)	22.6 (1.7)	20.9 (2.4)
08-09					22.3 (2.0)	21.3 (2.1)					22.6 (1.8)	21.1 (2.3)

(*Continued*)

Table 2
(*Continued*)

	English						Mathematics					
	2008		2009		2010		2008		2009		2010	
	Decline ($n = 46$)	Improve ($n = 41$)	Decline ($n = 58$)	Improve ($n = 65$)	Decline ($n = 59$)	Improve ($n = 65$)	Decline ($n = 65$)	Improve ($n = 65$)	Decline ($n = 28$)	Improve ($n = 86$)	Decline ($n = 69$)	Improve ($n = 39$)
ACT% taking												
04-05	NA	NA					NA	NA				
05-06	17.5 (13.6)	14.2 (8.3)	19.7 (18.3)	17.1 (11.6)			24.9 (22.7)	15.1 (9.9)	17.3 (12.8)	17.0 (13.0)		
06-07	16.5 (11.5)	13.5 (6.9)	18.4 (13.3)	16.0 (11.3)	17.4 (13.0)	18.4 (13.5)	20.2 (13.8)	15.5 (10.7)	17.4 (11.5)	16.1 (12.7)	14.7 (9.3)	21.9 (16.3)
07-08			17.5 (11.6)	16.2 (10.4)	17.3 (11.0)	19.8 (12.9)			16.5 (10.2)	16.6 (11.5)	15.4 (9.9)	20.6 (13.0)
08-09					20.9 (15.4)	22.0 (16.0)					18.2 (11.9)	23.0 (16.5)
SAT												
04-05	494 (30)	489 (18)					502 (20)	499 (25)				
05-06	489 (27)	487 (17)	482 (30)	480 (22)			501 (24)	498 (31)	497 (27)	495 (31)		
06-07	489 (26)	481 (18)	475 (34)	479 (24)	485 (29)	480 (24)	502 (24)	495 (30)	491 (28)	494 (30)	502 (27)	481 (45)
07-08			483 (30)	472 (25)	485 (23)	479 (26)			500 (28)	493 (32)	505 (27)	481 (50)
08-09					489 (23)	475 (26)					507 (28)	478 (49)
SAT% taking												
04-05	54 (13)	54 (9)					56 (13)	55 (12)				
05-06	52 (14)	52 (9)	52 (9)	53 (13)			52 (14)	52 (14)	52 (14)	48 (14)		
06-07	59 (12)	59 (9)	58 (15)	61 (11)	58 (14)	53 (15)	58 (11)	61 (10)	56 (13)	58 (15)	62 (10)	52 (17)
07-08			55 (13)	54 (13)	56 (12)	52 (14)			55 (10)	54 (13)	58 (10)	51 (17)
08-09					54 (12)	47 (13)					55 (10)	48 (16)
% FARM												
04-05	20 (13)	20 (11)					20 (15)	18 (13)				
05-06	21 (14)	18 (11)	22 (16)	18 (11)			22 (16)	17 (12)	20 (11)	19 (13)		
06-07	22 (13)	20 (1)	24 (17)	24 (17)	19 (12)	24 (15)	21 (16)	19 (12)	21 (12)	20 (13)	18 (10)	29 (18)
07-08			25 (17)	21 (11)	21 (13)	26 (16)			24 (11)	21 (13)	20 (11)	30 (19)
08-09					23 (14)	28 (16)					22 (12)	33 (20)

(*Continued*)

Table 2
(*Continued*)

	English						Mathematics					
	2008		2009		2010		2008		2009		2010	
	Decline ($n = 46$)	Improve ($n = 41$)	Decline ($n = 58$)	Improve ($n = 65$)	Decline ($n = 59$)	Improve ($n = 65$)	Decline ($n = 65$)	Improve ($n = 65$)	Decline ($n = 28$)	Improve ($n = 86$)	Decline ($n = 69$)	Improve ($n = 39$)
% Minority												
04-05	17 (23)	9 (12)					15 (21)	10 (16)				
05-06	18 (23)	9 (12)	17 (26)	9 (12)			16 (22)	16 (24)	16 (24)	11 (17)		
06-07	19 (24)	10 (13)	18 (27)	0 (13)	17 (22)	15 (22)	15 (22)	18 (27)	17 (25)	12 (18)	10 (14)	24 (32)
07-08			19 (27)	10 (13)	14 (21)	16 (22)			18 (26)	18 (26)	11 (15)	25 (33)
08-09					15 (21)	16 (23)					11 (15)	25 (33)
Total enrollment												
04-05	1,123 (633)	904 (518)					1,096 (614)	899 (516)				
05-06	1,155 (665)	926 (534)	958 (608)	821 (731)			1,121 (640)	968 (744)	1,089 (702)	968 (744)		
06-07	1,167 (683)	946 (531)	961 (618)	833 (752)	1,022 (800)	928 (536)	1,149 (710)	948 (561)	1,139 (778)	977 (764)	871 (523)	803 (447)
07-08			961 (619)	826 (755)	1,024 (805)	920 (543)			1,136 (778)	975 (775)	883 (531)	799 (460)
08-09					1,022 (808)	918 (564)					879 (530)	790 (471)

Values presented are mean values or mean percents, with standard deviations in parentheses.

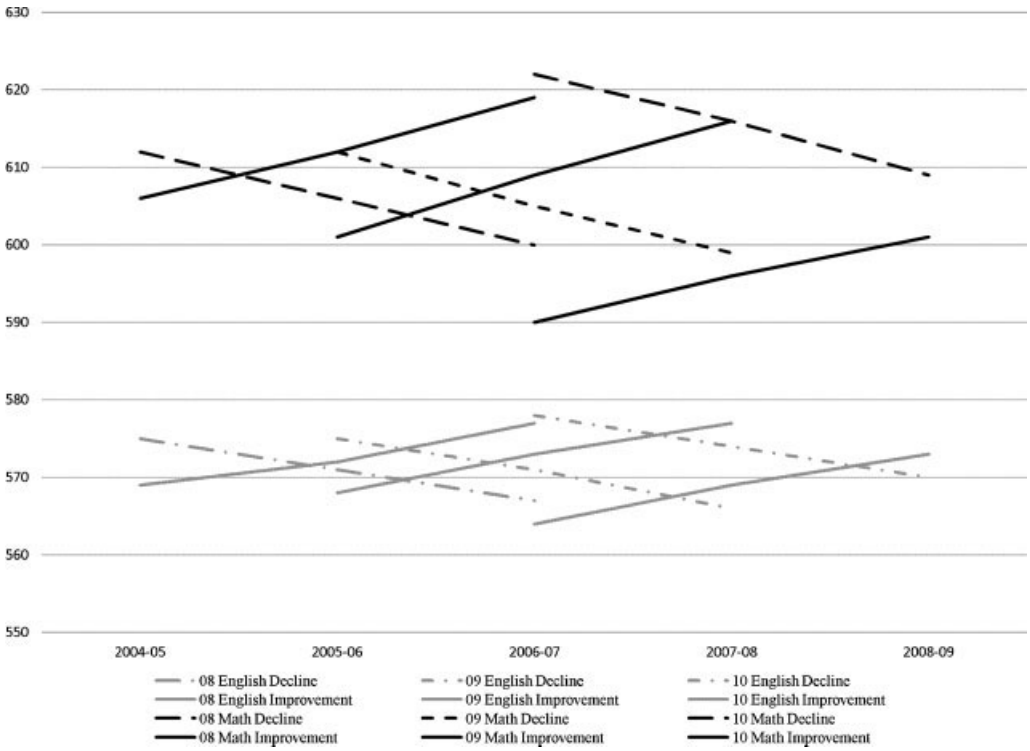

Figure 2. Longitudinal comparison of ISTEP Math and English performance in improving and declining cohort samples from 2005 through 2009.

of FARM students showed modest variations between improving and declining schools, but did not manifest consistent patterns.

HLM Results

Student Level. Individual student results maintained consistency across the three cohorts for both the English and Math samples (Tables 4 and 5). Controlling for all other student- and school-level factors, female students generally attained lower scores than their male counterparts on both the ACT Math and Science sections. However, for female students from the English sample, their ACT Science scores improved by .35 points in comparison to males across time from 2008 to 2010. Females either demonstrated no statistical difference or outperformed males on the English section of the ACT. With regard to race, Black, Hispanic, and Multiracial students scored approximately 1–2 ACT points lower than their white peers. For the samples defined by school performance on ISTEP English tests, Asian students consistently earned scores 1–2 points higher than their white peers on all three ACT sections. Results also indicated that younger students in all of the cohorts earned higher scores on all three sections of the ACT test. Finally, students' performance on their 10th grade ISTEP English and Math exams were consistently and positively associated with scores on the ACT Science, Math, and English tests.

School Level. Unlike the individual factors, our models indicated that school-level factors exhibited little to no association with individual student ACT performance. Although individual student racial identification was significantly and consistently associated with ACT

Table 3
Comparison of individual samples' demographics by operational definition, cohort, and improvement status (Student values are counts, other values are percentages)

	2008		2009		2010	
	Decline	Improve	Decline	Improve	Decline	Improve
ISTEP English cohorts						
Students	2,181	1,180	2,207	1,325	2,902	2,300
Sex						
Male	44	44	43	43	43	43
Female	56	56	57	57	57	57
Ethnicity						
Nat. Am.	<1	<1	<1	<1	<1	<1
Black	11	9	15	5	17	12
Asian	3	2	1	2	3	2
Hispanic	4	3	3	2	3	3
White	80	85	78	89	75	80
Multiracial	2	1	2	2	2	3
FARM						
Free	6	4	8	5	7	9
Reduced	4	5	6	2	5	6
Language proficiency						
Non-LEP	99	99	97	97	97	97
LEP	1	1	3	3	3	3
Education services						
GenEd	97	97	96	97	96	96
SPED	3	3	4	3	4	4
ISTEP Math cohorts						
Students	1,644	1,921	1,371	3,516	2,263	1,306
Sex						
Male	42	43	43	42	44	44
Female	58	57	57	58	56	56
Ethnicity						
Nat. Am.	<1	<1	<1	<1	<1	<1
Black	18	3	27	8	5	13
Asian	3	1	2	2	2	2
Hispanic	3	2	2	2	2	5
White	74	92	67	85	88	78
Multiracial	2	1	3	2	3	2
FARM						
Free	10	4	9	6	5	12
Reduced	5	4	6	5	4	5
Language proficiency						
Non-LEP	99	99	98	97	97	96
LEP	1	1	2	3	3	4
Education services						
GenEd	96	97	95	97	96	97
SPED	4	3	5	3	5	3

performance across the cohorts, the proportion of minority students within the schools did not reveal a similar relationship. Not only did the coefficient demonstrate substantial fluctuation between cohorts and ACT sections, the factor was most often not statistically significant. Likewise, in all but a few instances the percentage of students within each school taking the ACT was not statistically significant and manifested with similarly small coefficients across cohorts and samples.

Table 4
Results of HLM models evaluating the association of school improvement status (ISTEP English) with student ACT English, Math, and Science scores

	ACT English Scores						ACT Math Scores						ACT Science Scores					
	2008		2009		2010		2008		2009		2010		2008		2009		2010	
	Coeff.	SE	Coeff.	SE	Coeff.	SE	Coeff.	SE	Coeff.	SE	Coeff.	SE	Coeff.	SE	Coeff.	SE	Coeff.	SE
Intercept	21.31***	0.20	21.28***	0.17	21.35***	0.14	23.41***	0.20	23.34***	0.21	23.58***	0.18	22.73***	0.20	22.37***	0.17	22.77***	0.15
School-level factors																		
Percentage taking ACT	0.02*	0.01	0.00	0.01	0.01	0.01	0.00	0.01	0.00	0.01	0.01	0.01	0.01	0.01	0.02*	0.01	0.01	0.01
Proportion minority students	1.31*	0.55	0.40	0.38	0.55	0.51	0.60	0.86	−0.72	0.42	−0.99**	0.37	0.17	0.50	−0.59*	0.30	0.08	0.34
Improvement status	**−0.15**	**0.21**	**−0.12**	**0.23**	**−0.22**	**0.18**	**0.21**	**0.25**	**−0.31**	**0.29**	**−0.50***	**0.22**	**0.01**	**0.20**	**−0.18**	**0.21**	**−0.41***	**0.17**
Student-level factors																		
Female	−0.28	0.16	−0.16	0.10	−0.22*	0.10	−2.22***	0.13	−1.95***	0.15	−2.12***	0.10	−2.01***	0.15	−1.70***	0.16	−1.66***	0.11
RACE																		
Native American	−0.19	0.88	0.37	1.68	−0.44	1.03	−0.20	1.48	−0.82	0.43	−0.99	1.19	−0.12	1.00	−3.37	1.95	0.33	1.08
Black	−2.00***	0.38	−1.55***	0.21	−1.68***	0.23	−2.23***	0.32	−1.47***	0.25	−1.56***	0.19	−1.61***	0.28	−1.58***	0.36	−1.56***	0.16
Asian	1.30***	0.37	2.39***	0.51	0.84	0.49	2.82***	0.43	2.96***	0.48	2.17***	0.36	1.70***	0.44	2.29***	0.40	0.73*	0.31
Hispanic	−0.55	0.50	−0.12	0.40	−0.87*	0.35	−1.01**	0.29	0.06	0.37	−0.65*	0.27	−0.79*	0.39	−0.50	0.45	−1.05***	0.28
Multiracial	−0.36	0.49	−0.59	0.36	−1.31***	0.31	0.12	0.43	−1.42***	0.29	−0.89**	0.32	−1.01	0.54	−1.36***	0.35	−0.99***	0.28
Birth year	0.75***	0.12	0.72***	0.10	0.61***	0.09	0.62***	0.13	0.37**	0.12	0.48***	0.08	0.62***	0.10	0.50***	0.10	0.62***	0.10
ISTEP Scores—English (10th grade)	0.10***	0.00	0.10***	0.00	0.11***	0.00	0.08***	0.00	0.07***	0.00	0.07***	0.00	0.07***	0.00	0.06***	0.00	0.07***	0.00
Intraclass correlation coefficient	0.04		0.04		0.03		0.05		0.07		0.07		0.03		0.02		0.03	

*$p < 0.05$.
**$p < 0.01$.
***$p < 0.001$.

Table 5
Results of HLM models evaluating the association of school improvement status (ISTEP Math) with student ACT English, Math, and Science scores

	ACT English scores						ACT Math scores						ACT Science scores					
	2008		2009		2010		2008		2009		2010		2008		2009		2010	
	Coeff.	SE	Coeff.	SE	Coeff.	SE	Coeff.	SE	Coeff.	SE	Coeff.	SE	Coeff.	SE	Coeff.	SE	Coeff.	SE
Intercept	19.94***	0.19	20.61***	0.22	20.29***	0.17	22.17***	0.15	22.60***	0.15	22.44***	0.15	21.51***	0.15	21.80***	0.18	21.80***	0.14
School-level factors																		
Percentage taking ACT	0.03	0.02	0.01	0.01	0.01	0.01	−0.01	0.01	−0.01	0.00	0.01	0.01	0.00	0.01	0.01	0.01	0.01	0.01
Proportion minority students	0.29	0.77	−0.02	0.46	1.39*	0.64	0.05	0.46	0.17	0.40	0.31	0.48	−0.28	0.48	−0.47	0.39	0.57	0.56
Improvement status	**−0.04**	**0.23**	**−0.07**	**0.22**	**0.02**	**0.29**	**0.09**	**0.17**	**0.10**	**0.16**	**−0.09**	**0.19**	**−0.10**	**0.18**	**−0.05**	**0.18**	**−0.09**	**0.24**
Student-level factors																		
Female	1.81***	0.13	1.23***	0.12	1.58***	0.15	−0.36***	0.10	−0.56***	0.08	−0.56***	0.10	−0.19	0.12	−0.49***	0.10	−0.41**	0.13
RACE																		
Native American	−1.21	1.41	−1.04	1.30	−0.79	0.97	0.15	0.78	−1.34	0.87	−2.03	1.35	−0.36	0.84	−0.13	0.80	−0.53	0.80
Black	−1.81***	0.42	−1.53***	0.25	−2.26***	0.35	−0.19	0.26	−0.72***	0.19	−0.50	0.29	−0.72**	0.27	−1.22***	0.24	−0.98**	0.28
Asian	−0.07	0.60	0.00	0.55	−0.10	0.85	1.20**	0.38	0.97***	0.28	1.13*	0.45	0.03	0.39	0.53	0.45	0.12	0.48
Hispanic	−1.62***	0.44	−1.30***	0.32	−1.76**	0.60	−0.10	0.35	−0.47*	0.22	−0.24	0–0	−0.46	0.30	−0.71**	0.25	−0.81	0.44
Multiracial	−0.63	0.46	0.15	0.37	−1.65***	0.46	−0.09	0.31	−0.50*	0.25	−0.53*	0.26	−0.09	0.48	−0.43	0.30	−0.52	0.28
Birth year	0.56***	0.14	0.77***	0.10	0.72***	0.12	0.33***	0.09	0.43***	0.06	0.46***	0.08	0.48***	0.11	0.49***	0.08	0.54***	0.11
ISTEP Scores—Math (10th grade)	0.07***	0.00	0.06***	0.00	0.07***	0.00	0.07***	0.00	0.07***	0.00	0.07***	0.00	0.05***	0.00	0.05***	0.00	0.05***	0.00
Intraclass correlation coefficient	0.05		0.02		0.06		0.05		0.04		0.07		0.03		0.02		0.06	

*$p < 0.05$.
**$p < 0.01$.
***$p < 0.001$.

The improvement status of the school, in either English or math, did not generally demonstrate significant association with individual student ACT scores. On the ACT Science exam, schools identified as improving based on ISTEP English performance demonstrated a downward trend with increasingly negative coefficients. In 2010, students from these improving schools averaged 0.4 ACT points [$p = 0.02$; 95% confidence interval (CI) = -0.08, -0.74], lower than their peers from schools with declining ISTEP English performance. Student ACT Math scores, revealed a similar pattern across the cohorts, with the 2010 cohort exhibiting that students from declining schools outperformed those in improving schools on average by 0.5 points ($p = 0.03$; 95% CI = -0.07, -0.93). Among the three cohorts identified by ISTEP English performance, none of the cohorts demonstrated a statistically significant difference in student ACT English scores between improving and declining schools.

Shifting attention to the samples of schools identified by their performance on ISTEP Math assessments, we see no statistical difference between students in improving schools and declining schools on any of the three ACT sections. While not achieving statistical significance, negative ACT Science coefficients manifested. The negative coefficients suggested students from schools with three consecutive years of improving ISTEP Math scores consistently scored below those from declining schools. Interestingly, ACT Math results, which were also statistically non-significant, revealed an increasingly negative trend from the 2008 to the 2010 cohort.

To summarize these results, students who attended schools with three years of consecutive performance gains on ISTEP English and ISTEP Math exams scored lower on ACT Science exams than their peers who attended schools with three years of consecutive ISTEP decreases. In addition, results indicated that the score gap increased over the three cohorts. This trend was repeated on the ACT Math exams, with students from improving schools performing higher than their peers in 2008, but progressively declining so that a similarly identified cohort in 2010 performed lower than students from declining schools. However, in both cases, the only statistically significant difference in ACT performance between students from improving and declining schools was based on ISTEP English performance for the 2010 cohort. There was no consistent pattern of performance on the ACT English exam for any students based on improvement status across the cohorts.

Synthesis and Discussion

Application of the operational definitions of school improvement and decline based on ISTEP performance identified approximately 25–30% of the population of public high schools per cohort. Despite the differences in school performance trajectories across the cohorts and between improving and declining samples, few differences manifested among the school- and student-level demographic factors. However, the definitions did not identify samples overrepresented by particularly affluent or disadvantaged students, nor select schools marked by particularly high or low achievement. School-level composition measures suggested the analysis examined similar samples that lacked any extraordinary characteristics. Nonetheless, the restriction of our individual samples to students who completed the ACT exam within improving and declining schools limits the generalizability of our findings. Given that participation in the ACT exams is not required of all students in Indiana, insights from this analysis are likely limited to only students preparing for college; however, within this sample it is a logical conclusion that a proportion of these students will pursue degrees in STEM fields.

Results of the HLM analyses indicated that school-level performance trajectories measured by state administered standardized tests in English and math demonstrated little association with individual student performance on a widely used college entrance examination

(see bolded Improvement Status coefficients in Tables 4 & 5). Students from schools identified as improving on ISTEP English exam performance showed progressive declines in their ACT Science and Math exam scores from 2008 to 2010, with overall drops of −.42 and −.71 ACT points, respectively. A slight demographic shift occurred (increased % minority, increased % FARM) for improving schools within the English sample across the cohorts that could have accounted for the perforamance decline. However, a more significant change in the demographics of the Math sample occurred during the same period and a parallel evaluation of improvement status based on ISTEP Math performance yielded no statistically significant differences between improving and declining schools on any of the ACT tests. Given that ISTEP assessments are used as accountability measures and the ACT exams are not, it appears possible that the hard-earned gains by educators to satisfy accountability mandates are not strongly related to the performance of students who seek to attend college and potentially enter the STEM career pipeline. Additionally, if one assumes that the ACT can provide a form of external validity check of the ISTEP exams, and that consistent improvement in English or math on the ISTEP should be matched with improvement on the ACT exams, the results indicated that there is little connection. At the student-level, the HLM models provided evidence that significant score gaps on standardized math and science assessments still remain between males and females and across the racial and ethnic groups traditionally underrepresented in STEM.

Although we caution against over-interpretation of these findings, we next offer a spectrum of potential explanations and implications. A positive interpretation of these results champions the similarity of student ACT performance regardless of the overall improvement trajectory of the school. Because the samples of schools do not necessarily represent chronically high- or low-performing schools, the generally indistinguishable achievement implies a certain amount of resiliency among the students aspiring to attend college. Within a moderate range of school competency and effectiveness, college-bound students appear to be insulated from large school-level effects (both positive and negative) on their performance in science, math, and English college admissions exams.

A related explanation would assert that improving schools dedicated efforts to increasing the quality of education for students ''on the bubble'' (i.e., the middle of the pack) or those who score lowest on standardized assessments. This type of improvement initiative might increase the standardized state test scores of students who once lagged behind, but not necessarily impact high-performing students. While this strategy is not universally beneficial, it would likely produce overall improvement in school testing results.

A more pessimistic interpretation suggests that reforms and policies focused educators' attention and finite school resources on basic literacy and numeracy skills of the lowest-performing students, neglecting the continued development of higher-performing students. The pressures and incentives to demonstrate that all children receive an acceptable level of education may divert attention and efforts from students with demonstrated talent, capabilities, or competencies. In addition, as the authors of a 2011 NRC report suggested, because literacy and numeracy metrics provide the basis for district, school, teacher, and student performance, students possibly receive lower quality and/or less instructional time in the ''non-tested'' subjects like science and history. While our initial hypothesis was that this would not be the case in high schools, the results indicated that such curricular narrowing may have existed in the secondary schools within our samples. Although school-level strategies produced gains that might help satisfy expectations of school reform efforts like NCLB, Race to the Top, and School Improvement Grants, the results from our analyses suggested that strategies leading to improved test scores were not positively associated with gains in

performance for aspiring college students, including the sub-group of these students who will pursue STEM majors and careers.

One limitation of this study is that our analyses cannot connect the particular types of instruction or reform efforts enacted within schools with the results of those efforts, as presented elsewhere (Geier et al., 2008; Schroeder, Scott, Tolson, Huang, & Lee, 2007). Additionally, given that we only use data from a single state, we strongly recommend that other researchers extend these analyses to other states and assessment contexts.

Returning to the core purpose of this study, we reiterate two critical findings. First, we did not find a consistent direct relationship between performance on standardized assessment metrics utilized to evaluate school improvement in literacy and numeracy and student performance on college entrance examinations in science and math. Specifically, our models indicated that in most cases students from schools labeled as improving on ISTEP English and Math exams had lower ACT Science scores than those from declining schools. We suggest these non-significant or negative relationships are an indication that school improvement and STEM reforms are possibly working at cross purposes. Second, school improvement, as typically measured through standardized assessments, does not seem associated with improved achievement for all students, nor in all subject areas. Although commendable and hard-earned efforts have surely improved literacy and numeracy for many students, it appears that not all students benefitted from these efforts.

Conclusions and Recommendations

Results from this analysis suggest that school reform efforts evaluated through assessments in literacy and math may be negatively associated with gains in science achievement and therefore contradict related STEM reform initiatives to improve the achievement of American students in science and math. Although we make the assumption that designers of school reforms intended to improve the educational outcomes across a broad range of intellectual areas of all students, this study offers some evidence of counterproductive outcomes. Model results, consistent across three cohorts of students, indicated that improved school-level performance in English and math, as measured by state administered standardized tests, was generally associated with lower individual achievement scores on the ACT Science exam. For the 2010 cohort, students from schools with improving ISTEP English scores over the previous three-year period produced statistically significant and lower mean ACT Science scores than their peers in schools exhibiting declines on the ISTEP exams. A nearly identical result was found for ACT Math scores, where scores across the three cohorts declined and 'improving' schools from the 2010 cohort had students with significantly lower mean ACT scores than their peers. In more general terms, our models showed no significant and positive relationships between schools marked as 'improving' in math or English and student performance on ACT exams in science, math or English. Based on these results, we provide recommendations for educators, policymakers, and researchers.

As Duke's (2007) study of principals attempting to turn around chronically low-performing schools suggested, literacy might be the cornerstone of effective school reform. If students cannot proficiently read and comprehend material, then all other efforts toward improvement might prove futile. However, increasing literacy instruction only at the most basic levels might produce gains on state-mandated measures of school achievement, but not necessarily enhance the educational experience of students desiring to be college-bound. As both Holland (2010) and Suskind (1998) demonstrated, even within chronically low-performing schools, capable and competent students aspire to attend prestigious universities. When educators design and implement initiatives intended to improve the literacy of low-performing

students, we believe they should simultaneously consider ways to improve literacy instruction for all students, including those who are high-performing. Rather than focusing on leaving no child behind, we should reframe the policies to focus on helping all students achieve their academic potential. More directly, these initiatives should be seen as ways to simultaneously deepen the content knowledge of students in the areas of science, math and other disciplines.

When policymakers devise statutes intended to assess school performance, we recommend development of a more balanced approach across the disciplines. Current federal and state laws focus almost exclusively on student literacy and numeracy. This narrow definition of school effectiveness and school improvement does not address the multitude of purposes that stakeholders have come to expect from public schools (Labaree, 1997). In addition, as Balfanz, Legters, West, and Weber (2007) and Stringfield and Yakimowski-Srebnick (2005) have indicated, the NCLB metric of Adequate Yearly Progress in reading and math performance often produces confusing and counterproductive results. The current analysis suggests that current methods of school improvement, as measured by student English and math scores, likely may not enhance the science content preparedness of students intending to attend college and potentially enter the STEM pipeline. Echoing this finding, the NRC (2011) suggested, ''To make progress in improving STEM education for all students, policy makers at the national, state, and local levels should elevate science to the same level of importance as reading and mathematics,'' (p. 28). We feel it is illogical to think that elevating the testing status or increasing instructional time for science would solve these problems. Given that the school day and year are generally bounded quantities of time, increasing instructional focus on science as well as in English and math, would require less instructional time for other subjects and leave our peers from those disciplines in a similar situation as we are today. Based on this, we suggest a more balanced approach that utilizes the assessment of cross-disciplinary subjects like science and history to test both content knowledge in tested disciplines and critical reading, writing and numeracy skills. Using ''applied'' discipline areas for assessment could encourage greater school-level collaboration of teachers and significant integration of instruction across class content boundaries. Improving student literacy and numeracy then becomes spread across a diverse curricula rather than just the sole responsibility of English and math teachers.

Finally, we recommend researchers from multiple, often disparate, academic disciplines collaborate to examine the intersection of their interests. As previously discussed, research from discipline-based fields and school administration rarely intersect. The lack of overlap between the two research areas facilitates the production and existence of gaps in educator preparation and policy implementation. An increase in multidisciplinary collaboration may lead toward a more holistic and effective approach to improving the education of all students across subject areas.

We sincerely thank Gary Crowe, Meredith Park Rogers, Sam Stringfield, Tom Tretter and the dedicated reviewers whose feedback improved this paper.

Notes

[1]The IDOE appendix discussing reliability and validity mentions many data tables, but none of these are found within the actual report or on the IDOE website and so could not be evaluated further. The original technical evaluation report from 2003 could not be located via an internet search.

[2]Unlike many states (e.g., Illinois, Kentucky) Indiana students are not required to take ACT exams during high school.

References

America COMPETES Reauthorization Act of 2010. (2010). H. R. 5116, 111th Congress, 2nd Session.

Austin, G. R. (1979). Exemplary schools and the search for effectiveness. Educational Leadership, 37(1), 10–12.

Balfanz, R., Bridgeland, J. M., Moore, L. A., & Fox, J. H. (2010). Building a grad nation: Progress and challenge in ending the high school dropout epidemic. Baltimore, MD: Civic Enterprises Everyone Graduates Center at Johns Hopkins University.

Balfanz, R., Legters, N., West, T. C., & Weber, L. M. (2007). Are NCLB's measures, incentives, and improvement strategies the right ones for the nation's low-performing high schools? American Educational Research Journal, 44(3), 559–593. DOI: 10.3102/0002831207306768

Borman, G. D., Hewes, G. M., Overman, L. T., & Brown, S. (2003). Comprehensive school reform and achievement: A meta-analysis. Review of Educational Research, 73, 125–230.

Brookover, W. B., & Lezotte, L. W. (1979). Changes in school characteristics coincident with changes in student achievement. East Lansing, MI: Institute for Research on Teaching, Michigan State University.

Brown, A. B., & Clift, J. W. (2010). The unequal effect of Adequate Yearly Progress: Evidence from school visits. American Educational Research Journal, 47(4), 774–798.

Bryk, A. S., Sebring, P. B., Allensworth, E., Luppescu, S., & Easton, J. Q. (2010). Organizing schools for improvement: Lessons from Chicago. Chicago, IL: The University of Chicago Press.

Campbell, D. T. (1976). *Assessing the impact of planned social change*. Occasional paper, Kalamazoo, MI: The Evaluation Center, Western Michigan University.

Cavanaugh, S. (2011, May 25). State legislatures notch major K-12 policy changes. Education Week, 30(32), p1. Available from: http://www.edweek.org/ew/articles/2011/05/25/32legisoverview_ep.h30.html

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.

Commission on No Child Left Behind. (2007). Beyond NCLB. Washington, DC: Aspen Institute.

Daniels, M. (2006, December 12). Press release downloaded on January 25, 2010. Available from: http://www.in.gov/apps/utils/calendar/presscal?PF=gov2&Clist=196&Elist=87828

Datnow, A., Borman, G. D., Stringfield, S., Overman, T., & Castellano, M. (2003). Comprehensive school reform in culturally and linguistically diverse contexts: Implementation and outcomes from a four-year study. Educational Evaluation and Policy Analysis, 25, 143–170.

Dillon, S. (2011, February 8). *U.S. plan to replace principals hits snag: Who will step in*? The New York Times, p. A15.

Duke, D. L. (2007). Keys to sustaining successful school turnaround. Charlottesville, VA: Darden-Curry Partnership for Leaders in Education. Unpublished manuscript.

Duke, D. L., & Jacobson, M. (2011). Tackling the toughest turnarounds: Low-performing high schools. Phi Delta Kappan, 92(5), 34–38.

Duke, D. L., & Salmonowicz, M. (2010). Key decisions of first-year 'turnaround' principal. Educational Management Administration & Leadership, 38(1), 33–58.

Duke, D. L., Tucker, P. D., Belcher, M., Crews, D., Coleman-Harrison, J., Higgins, J., & West, J. (2005). Lift-off: Launching the school turnaround process in 10 Virginia schools. Charlottesville: Darden/Curry Partnership for Leadership in Education.

Duncan, A. (2009, June 17). Start over. Education Week, 28(35), 36.

Edmonds, R. (1977). Search for effective schools: The identification and analysis of city schools that are instructionally effective for poor children. Washington, DC: United States Department of Health, Education & Welfare, National Institute of Education.

Edmonds, R. (1979). Effective schools for the urban poor. Educational Leadership, 37(1), 15–18, 20–24.

Forte, E. (2010). Examining the assumptions underlying the NCLB federal accountability policy on school improvement. Educational Psychologist, 45(2), 76–88.

Geier, R., Blumenfeld, P. C., Marx, R. W., Krajcik, J. S., Fishman, B., Soloway, E., & Clay-Chambers, J. (2008). Standardized test outcomes for students engaged in inquiry-based science curricula in the context of urban reform. Journal of Research in Science Teaching, 45(8), 922–939.

Gray, J., Goldstein, H., & Thomas, S. (2001). Predicting the future: The role of past performance in determining trends in institutional effectiveness at A level. British Educational Research Journal, 27, 391–405.

Hamilton, L. S., Stecher, B. M., Marsh, J. A., McCombs, J. S., Robyn, A., Russell, J. L., ... Barney, H. (2007). Standards-based accountability under No Child Left Behind: Experiences of teachers and administrators in three states. Santa Monica, CA: RAND Corporation.

Harwell, M., & LeBeau, B. (2010). Student eligibility for a free lunch as a SES measure in education research. Educational Researcher, 39(2), 120–131.

Heck, R. H., & Mayor, R. A. (1993). School characteristics, school academic indicators and student outcomes: Implications for policies to improve schools. Journal of Education Policy, 8, 143–154.

Hemelt, S. W. (2010). Performance effects of failure to make Adequate Yearly Progress (AYP): Evidence from a regression discontinuity framework. Economics of Education Review, 30(4), 702–723. DOI: 10.1016/j.econedurev.2011.02.009

Herman, R., Dawson, P., Dee, T., Greene, J., Maynard, R., & Redding, S. (2008). Turning around chronically low performing schools. Washington, DC: United States Department of Education.

Hochbein, C. (2012a). Downward spirals, boiled frogs, and catastrophes: Examining the rate of school decline. Leadership and Policy in Schools, 11(1), 66–91.

Hochbein, C. (2012b). Relegation and reversion: A longitudinal examination of school turnaround and downfall. Journal of Education for Students Placed At-Risk, 17(1/2), 92–107.

Holland, W. R. (2010). A school in trouble: A personal story of Central Falls High School. Lanham, MD: Rowman & Littlefield Education.

Holme, J. J., Richards, M. P., Jimerson, J. B., & Cohen, R. W. (2010). Assessing the effects of high school exit examinations. Review of Educational Research, 80(4), 476–526.

Indiana Department of Education. (2010). *2010–2011 ISTEP+ Program Manual: Policies and Procedures for Indiana's Assessment System*. Originally retrieved from: http://www.doe.in.gov; but archived manuals appear to no longer be available. The 2011–2012 manual is available and discusses the same results.

Jacob, B. A., & Dee, T. S. (2010). The impact of no child left behind on students, teachers, and schools. Brookings Papers on Economic Activity, 2010(2), 149–194.

Jencks, C. (1972). Inequality: A reassessment of the effect of family and schooling in America. New York, NY: Harper & Row.

Johnson, J. F., & Asera, R. (Eds.). (1999). Hope for urban education: A study of nine high-performing, high-poverty, urban elementary schools. Austin, TX: The Charles A. Dana Center, The University of Texas.

Kahle, J. B. (2004). Will girls be left behind? Gender differences and accountability. Journal of Research in Science Teaching, 41(10), 961–969.

Kane, T. J., & Staiger, D. O. (2002). The promise and pitfalls of using imprecise school accountability measures. The Journal of Economic Perspectives, 16(4), 91–114.

Klitgaard, R. E., & Hall, G. R. (1975). Are there unusually effective schools? Journal of Human Resources, 10, 90–106.

Kozol, J. (1991). Savage inequalities: Children in America's schools. New York, NY: Crown.

Labaree, D. F. (1997). Public goods, private goods: The American struggle over educational goals. American Educational Research Journal, 34(1), 39–81.

Labaree, D. F. (2010). Someone has to fail: The zero sum-game of public schooling. Cambridge, MA: Harvard University Press.

Maerten-Rivera, J., Myers, N., Lee, O., & Penfield, R. (2010). Student and school predictors of high-stakes assessment in science. Science Education, 94(6), 937–962. DOI: 10.1002/sce.20408

Matthews, P., & Sammons, P. (2004). Improvement through inspection: An evaluation of the impact of Ofsted's work. London: The Office of Standards in Education/Institute of Education.

Matthews, P., & Sammons, P. (2005). Survival of the weakest: The differential improvement of schools causing concern in England. London Review of Education, 3(2), 159–176.

Maxwell, L. A. (2009, December 2). Stimulus rules on turnarounds shift: Stimulus guidelines changed for turning around schools. Education Week, 29(13), 1, 19.

McMurrer, J. (2008). Instructional time in elementary schools: A closer look at changes for specific subjects (pp. 8). Washington, DC: Center on Education Policy.

McNeil, L., & Valenzuela, A. (2001). The harmful impact of the TAAS system of testing in Texas: Beneath the accountability rhetoric. In M. Kornhaber & G. Orfield (Eds.), Raising standards or raising barriers? Inequality and high stakes testing in public education (pp. 127–150). New York: Century Foundation.

Murphy, J., & Meyers, C. V. (2008). Turning around failing schools: Leadership lessons from the organizational sciences. Thousand Oaks, CA: Corwin.

National Academy of Sciences. (2005). Rising above the gathering storm: Energizing and employing America for a brighter economic future. Washington, DC: National Academies Press.

National Commission on Excellence in Education. (1983). A nation at risk: The imperative for educational reform. An open letter to the American people. A report to the Nation and the Secretary of Education. Washington, DC: Superintendent of Documents, Government Printing Office (ERIC Document Reproduction Service No. ED 226 006).

National Research Council. (2011). Successful K-12 STEM Education: Identifying Effective Approaches in Science, Technology, Engineering, and Mathematics. Committee on Highly Successful Science Programs for K-12 Science Education. Board on Science Education and Board on Testing and Assessment, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

National Science Foundation. (2011). *WebCASPAR: Data from IPEDS Completions Publication*, retrieved on November 12, 2011 from National Science Foundation: http://webcaspar.nsf.gov/

National Science Teachers Association. (Summer 2011). *Elementary teachers getting less time for science*. NSTA Reports, p. 17.

Obama, B. (2009, Nov. 23). *Remarks by the President on the "Education To Innovate" Campaign*. Copy of text downloaded on July 5, 2011 from: http://www.whitehouse.gov/issues/education/educate-innovate

Obama, B. (2010, Jan. 6). *President Obama Expands "Educate to Innovate" Campaign for Excellence in Science, Technology, Engineering, and Mathematics (STEM) Education*. Copy of text downloaded on April 15, 2012 from: http://www.whitehouse.gov/the-press-office/president-obama-expands-educate-innovate-campaign-excellence-science-technology-eng

Obama, B. (2011, Jan. 25). *State of the Union Address*. Copy of text downloaded on July 5, 2011 from: http://www.whitehouse.gov/the-press-office/2011/01/25/remarks-president-state-union-address

Opdenakker, M.-C., & Van Damme, J. (2006). Differences between secondary schools: A study about school context, group composition, school practice, and school effects with special attention to public and Catholic schools and types of schools. School Effectiveness and School Improvement, 17, 87–117.

Organization for Economic Cooperation and Development (OECD), Program for International Student Assessment (PISA). (2007). *PISA2006: Science Competencies for Tomorrow's World (Vol. 1)*. Copy of text downloaded on July 14, 2010 from: http://www.oecd.org

Organization for Economic Cooperation Development (OECD), Program for International Student Assessment (PISA). (2010). *PISA 2009 at a glance*. Copy of text downloaded on July 10, 2011 from http://dx.doi.org/10.1787/9789264095298-en

Palardy, G. J. (2008). Differential school effects among low, middle, and high social class composition schools: A multiple group, multilevel latent growth curve analysis. School Effectiveness and School Improvement, 19, 21–49.

Payne, C. M. (2008). So much reform so little change: The persistence of failure in urban schools. Cambridge, MA: Harvard Press.

Penfield, R. D., & Lee, O. (2010). Test-based accountability: Potential benefits and pitfalls of science assessment with student diversity. Journal of Research in Science Teaching, 47(1), 6–24.

Picucci, A. C., Brownson, A., Kahlert, R., & Sobel, A. (2002). Driven to succeed: High-performing, high-poverty, turnaround middle schools. Austin, TX: The Charles A. Dana Center, The University of Texas.

Pogash, C. (2008, March 1). *Free lunch isn't cool, so some students go hungry.* The New York Times. Retrieved December 16, 2011 from: http://www.nytimes.com/2008/03/01/education/01lunch.html?pagewanted=all

President's Council of Advisors on Science and Technology. (2010). *Report to the President. Prepare and Inspire: K-12 Education in Science, Technology, Engineering, and Math (STEM) for America's Future: Executive Office of the President*. Retrieved July 14, 2011 from: http://www.whitehouse.gov/sites/default/files/microsites/ostp/pcast-stem-ed-final.pdf

President's Council of Advisors on Science and Technology. (2012). *Report to the President. Engage to Excel: Producing One Million Additional College Graduates with Degrees in Science, Technology, Engineering, and Mathematics: Executive Office of the President*. Retrieved March 1, 2012 from: http://www.whitehouse.gov/administration/eop/ostp/pcast/docsreports

Provasnik, S., Gonzales, P., & Miller, D. (2009). U.S. Performance Across International Assessments of Student Achievement: Special Supplement to the Condition of Education 2009 (NCES 2009-083). Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education.

Purkey, S. C., & Smith, M. S. (1983). Effective schools: A review. The Elementary School Journal, 83, 426–452.

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

Raudenbush, S. W., Bryk, A. S., & Congdon, R. (2011). HLM 7 for Windows [Computer software]. Lincolnwood, IL: Scientific Software International, Inc.

Ravitch, D. (2000). Left back: A century of failed school reforms. New York, NY: Simon & Schuster.

Ravitch, D. (2010). The death and life of the great American school system: How testing and choice are undermining education. New York, NY: Basic Books.

Rothstein, R. (2009). What's wrong with accountability by the numbers? American Educator, 33(1), 20–23.

Rothstein, R., Jacobsen, R., & Wilder, T. (2009). Grading education. American Educator, 33(1), 24–32.

Rowan, B., Bossert, S. T., & Dwyer, D. C. (1983). Research on effective schools: A cautionary note. Educational Researcher, 12(4), 24–31.

Ruiz-Primo, M. A., Shavelson, R. J., Hamilton, L., & Klein, S. (2002). On the evaluation of systemic science education reform: Searching for instructional sensitivity. Journal of Research in Science Teaching, 39(5), 369–393.

Sammons, P. (2007). School effectiveness and equity: Making connections: A review of school effectiveness and improvement research—Its implications for practitioners and policy makers. Berkshire, UK: CfBT Education Trust.

Schroeder, C. M., Scott, T. P., Tolson, H., Huang, T.-Y., & Lee, Y.-H. (2007). A meta-analysis of national research: Effects of teaching strategies on student achievement in science in the United States. Journal of Research in Science Teaching, 44(10), 1436–1460.

Singer, S. R., Hilton, M. L., & Schweingruber, H. A. (Eds.). (2006). America's lab report: Investigations in high school science. Washington, DC: National Academy Press.

Spillane, J. P., Diamond, J. B., Walker, L. J., Halverson, R. & Jita, L. (2001). Urban school leadership for elementary science instruction: Identifying and activating resources in an undervalued school subject. Journal of Research in Science Teaching, 38(8), 918–940.

Stringfield, S. C., & Yakimowski-Srebnick, M. E. (2005). Promise, progress, problems, and paradoxes of the three phases of accountability: A longitudinal case study of the Baltimore City Public Schools. American Educational Research Journal, 42(1), 43–75. DOI: 10.3102/00028312042001043

Stuit, D. A. (2010). Are bad schools immortal? The scarcity of turnaround and shutdowns in both charter and district sectors. Washington, DC: Thomas B. Fordham Institute.

Suskind, R. (1998). A hope in the unseen: An American odyssey from the inner city to the Ivy League. New York, NY: Broadway Books.

Teddlie, C., & Reynolds, D. (2000). The international handbook of school effectiveness research. London, UK: Falmer Press.

Teddlie, C., & Stringfield, S. (1993). Schools make a difference. Lessons learned from a 10-year study of school effects. New York, NY: Teachers College Press.

Thomas, S., Peng, W. J., & Gray, J. (2007). Modelling patterns of improvement over time: Value added trends in English secondary school performance across ten cohorts. Oxford Review of Education, 33, 261–295.

Tyack, D. B., & Cuban, L. (1995). Tinkering toward utopia: A century of public school reform. Cambridge, MA: Harvard University Press.

US Department of Education. (March, 2010a). *A Blueprint for Reform: The Reauthorization of the Elementary and Secondary Education Act: Supporting Science, Technology, Engineering and Mathematics Education*. 4 pp. Retrieved from: http://www2.ed.gov/policy/elsec/leg/blueprint/index.html

US Department of Education. (March, 2010b). *A Blueprint for Reform: The Reauthorization of the Elementary and Secondary Education Act*. 45 pp. Retrieved from: http://www2.ed.gov/policy/elsec/leg/blueprint/index.html

US Department of Education. (September, 2011). *Our Future, Our Teachers: The Obama Administration's Plan for Teacher Education Reform and Improvement*. 18 pp. Retrieved from: http://www.ed.gov/teaching/our-future-our-teachers

Wimpelberg, R. K., Teddlie, C., & Stringfield, S. (1989). Sensitivity to context: The past and future of effective schools research. Educational Administration Quarterly, 25, 82–107.

Wood, N. B., Lawrenz, F., Huffman, D., & Schultz, M. (2006). Viewing the school environment through multiple lenses: In search of school-level variables tied to student achievement. Journal of Research in Science Teaching, 43(3), 237–254.