EDUCATION POLICY ANALYSIS ARCHIVES

A peer-reviewed scholarly journal
Editor: Sherman Dorn
College of Education
University of South Florida

Volume 13 Number 1 **January 2, 2005** **ISSN 1068-2341**

Change and Continuity in Student Achievement from Grades 3 to 5: A Policy Dilemma

Mary McCaslin
Heidi Legg Burross
Thomas L. Good
University of Arizona

Citation: McCaslin, M., Burross, H. L., & Good, T. L. (2005, January 2). Change and continuity in student achievement from grades 3 to 5: A policy dilemma. *Education Policy Analysis Archives, 13*(1). Retrieved [date] from http://epaa.asu.edu/epaa/v13n1/.

(This article was accepted for publication by Gene V Glass, Editor 1993-2004.)

Abstract

In this article we examine student performance on mandated tests in grades 3, 4, and 5 in one state. We focus on this interval, which we term "the fourth grade window," based on our hypothesis that students in grade four are particularly vulnerable to decrements in achievement. The national focus on the third grade as *the* critical benchmark in student performance has distracted researchers and policy makers from recognition that the fourth grade transition is essential to our understanding of how to promote complex thinking and reasoning that is built upon a foundation of basic skills that may be necessary, but are not sufficient, for the more nuanced learning expected in subsequent grades. We hypothesized that the basic skills that define a successful third grade performance do not predict successful performance in subsequent years. We examined student performance over time using two measures of student success: the Arizona Instrument to Measure Standards (AIMS), a standards-based test; and the Stanford 9 (SAT9), a norm-referenced test. Three groups of schools were included in these analyses. Schools were individually matched to the original sample of interest, which were schools serving students of poverty that received state funding to implement Comprehensive School

Reform (CSR) models that emphasize continuity across grade levels. The first comparison sample includes schools that also serve students of poverty but did not receive CSR funding, "nonCSR" schools. The second comparison sample includes schools individually matched on all variables *except* economic status. These schools, which we term "low poverty" schools, are the wealthiest public schools in the state, with less than 10% of attending students receiving free or reduced lunch. Student test scores in math, reading, and writing (AIMS) or language (SAT9) were analyzed for the years 2000-2003. These intervals allowed the analysis of two cohorts of the fourth grade window. Our results suggest that the reliance on third grade performance to label students and schools is untenable.

Introduction

This investigation began with a hypothesis that the fourth grade is a critical period of schooling—especially for students of poverty. Our initial focus was on schools selected for funding by the Arizona State Department of Education to implement a Comprehensive School Reform (CSR) model. CSR models are "school-wide" reform efforts supported by Federal Title One funds that attempt to improve the educational outcomes of schools serving students of poverty by unifying curriculum, instruction, and management of that instruction across grades within a school. Several CSR models, derived from "best practices" research, are available for schools to implement (e.g., Expeditionary Learning Outward Bound, Success for All) or design (the so-called "home grown" approach). Our initial task was to assess the potential for various CSR models to promote student achievement in grades 3-5 (see Good, Burross, & McCaslin, in press). In this paper we attend to our hypothesis, that the transition between grades 3 and 5, what we term the "fourth grade window," mediates student performance in important ways.

Elementary schools implementing funded CSR models were individually matched with schools not receiving state funds for school reform (nonCSR schools) based on geography, grade composition, size, and poverty levels (defined as % of students receiving free or reduced lunch). Changes in student test performance associated with the "fourth grade window" occurred similarly in both CSR and nonCSR schools. These findings are consistent with the "cumulative deficit" attributable to poverty (Hess & Shipman, 1965; Pogrow, 1999); however, our hypothesis is that the fourth grade window is more pervasive than poverty, although it may well be exacerbated by it. To test this hypothesis we included schools individually matched to the original CSR schools using the same criteria for the nonCSR schools serving students of poverty—geography, grade composition, and size—but with low levels of poverty. In these low poverty comparison schools, less than 10% of the students received free or reduced lunch. Thus, the analyses we focus on involve comparisons among three groups of schools, two matched groups of poverty schools in Arizona, one group receiving state funding to implement comprehensive school reform models and the other not, and one group of schools matched on all criteria except poverty rates of its students. The poverty schools are not the most impoverished public schools in the state; however, the low poverty comparison schools are the wealthiest public schools in the state. Student test performances on the Arizona Instrument to Measure Standards (AIMS) and the Stanford-9 (SAT9) are tracked for four years, 2000-2003. These multi-year performances allow two

replications of longitudinal analyses of the fourth grade window, that is, two cohorts of students moving from grade 3 to grade 5. These comparisons inform the: 1) viability and robustness of the fourth grade window in student performance, 2) function of student socio-economic status (and school resources) in this phenomenon, and 3) representation of student knowledge as a function of test used (criterion- or norm-referenced) and the policy implications that emerge.

Related Literature

The economics of student performance

Ample evidence suggests that poverty interferes with student performance (Ladd & Hansen, 1999). The number of children living in poverty is increasing rapidly (e.g., National School Boards Association, 1999; US Government Printing Office, 1999, Forum on Child and Family Statistics, 1999). Additionally, states and school districts have unequal resources for schooling. Generally, schools that serve low-income students receive fewer funds than do schools serving more affluent communities; unequal resources have been distributed within a school district as well as among them (Stiefel, Rubenstein, & Berne, 1998; Ladd & Hansen, 1999). Schools whose students bring fewer home resources to the classroom also are comparatively under-resourced; thus, typically the children of poverty attend schools with fewer financial resources.

Some researchers argue that these are not troublesome relationships. Earlier Coleman (Coleman, Campbell, Hobson, McPartland, Meade, Weinfeld, & York, 1966) and more recently Hanushek (1997) make the argument that school expenditures are largely unrelated to student performance. One difference between then and now is that the "genetics of home" reason for ignoring differences in school funding (e.g., Jensen, 1973) has been replaced with an "economics of home" rationale. Others have argued for a guarded optimism that underfinanced schools can use increased funding wisely and impact student performance (Hill, Cohen, & Moffitt, 1999; Ladd & Hansen, 1999). One manifestation of "funding wisely" is the comprehensive school reform initiative. Good, Burross, and McCaslin (in press) analyzed the effects of CSR programs in Arizona on reducing the differences in student test performance as a function of home or school poverty. Results suggest that money may be a necessary condition, but it may not be sufficient to increase student performance in schools serving students of poverty. In this paper we broaden the discussion of school funding and student performance by 1) considering the effects associated with the saturation level of poverty (CSR: M= 80%; nonCSR: M=71%) and 2) including schools that serve students of relative affluence (non-poverty: M= 5%). We examine the coincidence of student home economics and school resources and its relation to changes in student performance across grades 3-5.

Critical periods in student learning

It has been argued since the 1970s that student performance in the third grade (especially reading performance) predicts student performance in high school and beyond (e.g., Klaus, 1973). This reasoning is evident in the current Federal school reform initiative, No Child Left Behind. Third grade is considered a pivotal benchmark in students learning to

read. High-stakes testing (that is, tests associated with high-stakes consequences for students and/or their schools) often begin at the third grade. In some states third grade students are automatically retained if they fail to achieve a set testing standard (e.g., Florida); in others, failures in third graders' test performance yield failing labels for schools with conditional threats of state take-over (e.g., Arizona). Third grade has become the grade at which serious decisions are made about students and schools.

Pogrow (1999), argued that 3rd grade test performance *overpredicts* the achievement of students of poverty and that the apparent gains in poverty students' performance—or at least apparent decreases in the difference between students of poverty and privilege—dissipate by the time the students leave elementary school. Pogrow casts the problem as a "cognitive wall" that results from an increasingly complex curriculum for which the student of poverty is ill-prepared. Similarly, McNeil (2000) argued that school reform efforts in Texas, and the use of the high-stakes Texas Assessment of Academic Skills test, causes poverty students to receive a curriculum that is focused primarily on drill and practice of low-level reading and math skills. She notes that these students lose in two ways. First, they do not have the opportunity to engage higher-level math and reading concepts; second, they are not getting exposure to the fullness of what we consider an education (e.g., science, social studies) because time is spent on priority test areas. McNeil also described affluent school districts that argue that the mandated tests work to lower their standards—their own assessments expect more thinking and advanced knowledge than the "new" school reforms. It appears that mandated tests may restrict the opportunities for students of poverty to be exposed to higher-order learning while they restrict the opportunities for students of privilege to display their higher-order learning. If this is the case, then the apparent gaps between students of poverty and wealth are *more* disparate than they appear on mandated tests. At minimum, they appear to reify a basic level curriculum for students of poverty.

Others point to fourth grade as a particularly susceptible time for learners. Students are transitioning into more complex cognitive mechanisms (Case & Okamoto, 1996; Piaget, 1983) that can challenge their "simple and sure" (Hofer & Pintrich, 1997) knowledge base at the same time they confront more complex learning formats (McCaslin, et al., 1994) and tasks (Chall, 1996). For example, the pattern of declining scores from third to fourth grade was observed on a standardized mathematical instrument in 26 nations (Wang, 2003). In this study, the *exact same* 20 test items were given to third and fourth grade students. The third graders outperformed the fourth graders on an average of 5.7 of the examined items and up to 16 of the items in one country.

It may well be that the "simple and sure" curriculum and test representation of knowledge and knowing at the third grade does not serve subsequent learning as expected. This could be a due to a straightforward disconnect between the curricula and instructional strategies of the third and fourth grades, but it is also possible that the mechanization procedures that result in a "successful third grader" obviate the enhancement of subsequent thinking and learning of the fourth grader. The *Einstellung* of Luchins and Luchins (1950) may apply to more than immediate problem solving. Consider the difficulty in getting students who have learned how to do long division—with remainders!—to keep their pencils on their desks as they mentally estimate how many of one unit is found in another. Do the learning habits and beliefs about knowledge instilled in the early grades and reified in high-stakes testing interfere with the struggle to understand complexity and probabilistic reasoning that are the hallmarks of what we consider an educated learner?

We study students in grades 3-5, the period that we term the "fourth-grade window," because we suspect there is too much attention to the predictive power of grade 3

and not enough attention to the subsequent 2 years of schooling and their relationship with earlier learning opportunities and ultimate educational attainment—especially for students of poverty. We want students to succeed in the long-term and the current focus on 3rd grade as *the* critical period in student performance seems ill-advised.

The measurement of student performance

Students can fail test items for many different reasons. We typically think that a failure suggests that material was too difficult for students; however, students may not have had an opportunity to learn material that is *not* too difficult for them, it is simply *unknown* to them. Opportunity to learn is a basic tenet for interpretation of student performance, both theoretically (Carroll, 1963) and practically (e.g., Berliner & Biddle, 1995; Good & Grouws, 1979). Students also can make simple material problematic and fail items that under-represent their understanding. As we have noted, this is especially the case when students are progressing into a more sophisticated level of thinking about content (Case & Okamoto, 1996; Piaget, 1983) as higher levels of thinking and understanding are not always represented by the "right" answer.

Successful test taking often is quite different from successful classroom learning. When learning, students complete assignments that show their work and thinking. In math, the problems are worked out and teachers want to see the process students used to solve the problem, and in writing the revisions count. Directions are supposed to be clear and the objectives known: students know what to do and why they are doing it. Students believe their teachers want them to succeed. Not so, the test makers. Taking mandated tests is another story. When taking tests, classroom bulletin boards, student work samples, and decorative posters are removed or covered for fear students might "see" something that helps them remember or answer an item correctly. Students show their knowledge in formats that require eye-hand coordination to stay on the right bubble. Successful test-taking is all about reading directions that can (and do) change unexpectedly, resisting the lure of the first familiar and *intentionally* seductive answer, moving on when confronted by difficulty, not wasting time working the problems through to completion, and keeping one eye on the clock. It is a considerable leap from student test performance to student learning. Even among those who agree about the use of testing, there are disagreements about the type of test, time of administration, and stakes involved with successes and failures.

One consideration at any level of testing involves the method for reporting results. Specifically, norm-referenced and standards-based reporting provides different information. Norm-referenced tests describe the individual's (i.e., student, class, school) performance in terms of how s/he did in relation to others who took the same test (e.g., percentiles). Standards-based performances are reported based on the individual's performance in relation to a standard of excellence (e.g., percentage correct). Both methods of reporting results have advantages and problems. Norm-referenced methods allow the user to determine the individual's relative standing, but do not provide general performance information. Standards-based methods depict the level of the individual's performance, but do not provide details about how others performed, and the standard and the cut-score for success or failure may, at times, be arbitrary.

Arizona Instrument to Measure Standards. The Arizona Instrument to Measure Standards (AIMS) was born out of the Arizona Student Assessment Program (ASAP) test, both of which were designed by the Arizona Department of Education to measure state standards for students. Students take the AIMS test in grades 3, 5, 8, and 10

through 12 in math, reading, and writing. These tests were developed in response to nationwide calls for stricter high school graduation requirements (Jorgensen, 1999). Both have reported reliability and validity problems since their inceptions (Smith, Heinecke, & Noble, 1999). Plans to make the AIMS test a requirement for high school graduation are in place despite many revisions of the test and delays in the implementation of the graduation requirement. This year's sophomore students took their first crack at the AIMS test in February 2004; the current plan is to allow up to 4 retakes by the end of senior year to achieve graduation. One wonders what incentives to complete high school remain for a successful sophomore, but the focus of criticisms of the test has largely been on the lack of time provided between the introduction of the AIMS test in 1998 and related standards and the passing requirement for graduation originally proposed for the 1999-2000 school year. This narrow time frame gave teachers little time to enact the standards within the classrooms and prevented revision and review to determine whether the standards were appropriately set (Jorgensen, 1999). Critics also claim that with standards set at college-entrance levels and the lack of appeal process, special education and non-native English speakers are unfairly denied graduation rights. At last report, surveys were being conducted across the state to gather public opinion about the timing of the graduation requirement and stringency of the standards (WestEd, 2001). The recommendations by the board that conducted this survey included waiting another three to four years for graduation requirement implementation, review and implement individual sections of the test in stages, and review current results to set transitory standards.

It is useful to consider the standards represented in the AIMS test in relation to the National Assessment of Educational Progress (NAEP). In 2003, only 25% of Arizona fourth graders scored at the "proficient" level in math and 23% scored proficient in reading on the NAEP (Gassen, 2003). Both of these performances are at least 7% below the national average. The state superintendent of education, Tom Horne, has noted that the state of Arizona's standards tend to be lower than the nation's standards (in Gassen, 2003).

Stanford-9. Arizona started using the SAT9 during the 1996-1997 school year. It was administered in grades 2-11 to students across the state. This standardized measure is given nationally and results are reported in terms of national percentile rankings. SAT9 results are used for ranking high schools. This method of reporting results has been criticized by some for lacking information about comparison to an "absolute standard" (www.sandiegodialoggue.org/pdfs/sddr_feb_mar02.pdf). Also, some states use the same form of the test year after year because of the costs associated with buying newer forms (http://www.ppic.org/main/commentary.asp?i=225). Another common criticism with this and any standardized measure (especially those with rankings and finances hinged on students success) is teaching to the test.

Method

Two measures of academic standards were used in this state: Arizona Instrument to Measure Standards (AIMS) and Stanford-9 (SAT9). The AIMS test was administered in grades 3, 5, 8, and 10 through 12. The SAT9 was administered to grades 2 through 12 and included reading, language, and math performance areas.

For this research, three samples of schools were used: CSR-funded schools ("CSR schools", n = 21); schools individually matched to the CSR schools based on geography, grade composition, size, and poverty level ("non-CSR schools", n = 23); and schools individually

matched to the CSR schools based on geography, grade composition, and size, but have low poverty levels, defined as less than 10% of attending students received free or reduced lunch ("low-poverty schools", n = 21). There were originally 27 CSR schools, but only SAT9 scores for grades three through five and AIMS scores for grades three and five were included in this study. There were more of the non-CSR schools with grades 3 and 5 than the CSR schools with grades 3 and 5 because one criterion for matching with the CSR schools was that the non-CSR and low poverty schools had *at least* the same grades as the CSR schools, two non-CSR kindergarten through grade eight schools were matched to CSR grades six through eight schools.

Poverty level was defined as percentage of students receiving free or reduced lunches. The percentage of students receiving free and reduced lunches was presented on the state web site (http://www.ade.az.gov/health-safety/cnp/frpercentages.asp). This information was broken into frequencies of students receiving reduced-price lunches, free lunches, and those who paid full price. The free/reduced lunch percentage was calculated by adding all of these frequencies and dividing that into the sum of those receiving free and those receiving reduced-price lunches. Poverty matches were conservative: non-CSR matching schools were selected at the same poverty level or less so that CSR schools as a group have the highest saturation of poverty in the study.

The free/reduced lunch rates on the three types of schools were subjected to ANOVA procedures ($F(2, 346) = 719.77$, $p < .001$). Scheffé post hoc analyses revealed that each group significantly differed from the others with $p < .001$: CSR: $M = 80.74$, $s = 14.86$; nonCSR: $M = 69.87$, $s = 18.58$; and low-poverty: $M = 4.99$, $s = 2.20$. The nonCSR schools had lower poverty—less saturation—levels than the CSR schools (see Table 1).

Table 1
Free and reduced lunch percentages by school type

School type		2000	2001	2002	2003	Overall Means
CSR	Mean	79.81	79.29	79.41	80.76	79.82
	N	25	25	24	24	24.5
	SD	15.77	15.24	15.15	11.49	14.41
	Minimum	29.00	38.00	29.09	51.61	36.93
	Maximum	95.24	100.00	95.01	95.11	96.45
Non-CSR	Mean	69.72	68.61	72.80	71.81	70.74
	N	23	23	23	23	23
	SD	18.99	17.01	15.59	17.28	17.22
	Minimum	32.90	37.22	40.78	34.92	36.46
	Maximum	97.33	94.96	95.03	97.06	96.10
LowPov	Mean	4.55	4.82	5.28	5.54	5.06
	N	19	20	20	20	19.75
	SD	2.25	2.11	1.95	2.25	2.13
	Minimum	1.10	1.16	1.41	.95	1.16
	Maximum	8.84	8.66	8.72	9.20	8.86

Variables

AIMS results were reported in terms of percentage of students by grade and school who fell into the following categories: "Exceeds the Standard", "Meets the Standard", "Approaches the Standard", and "Falls Far Below the Standard" (http://www.ade.az.gov/standards/aims/PerformanceStandards/performancelevels.asp). By both state standards and for use in this report, students who "Exceed" and "Meet" the standard were considered to have "passed" the AIMS test. AIMS results for third and fifth grade students were used in analyses. Percentages were reported only when at least 10 students had taken the exam within each category.

Third through fifth grade results for SAT9 also were used in these analyses (http://www.ade.az.gov/ResearchPolicy/SAT9Results/2003/default.asp). SAT9 results were reported as norm-referenced national percentile ranks by grade and performance area. These data were transformed to normal curve equivalence scores and missing data imputed using regression analyses.

Both tests have math and reading subtests. The AIMS test has a writing section and the SAT9 has language. The AIMS test was not administered to fourth graders, but the SAT9 test was. The tests were similar in many ways; however, the methods of reporting results, subtests, and grade compositions of each test differ. These similarities and differences will be described in more detail as the results of the analyses are presented subsequently.

The correlations between AIMS and SAT9 overall mean scores were all significant (all above $r = .86$, $p < .01$), across and between years. Schools maintained relative standings on these two measures every year. Table 2 contains the correlations between AIMS and SAT9 for grades 3 and 5 for each year of the study. The correlations remain strong and relatively constant in each instance.

Table 2
Correlations between AIMS and SAT9 scores by grade, 2000-2003.

		Third grade				Fifth grade			
		AIMS 2000	**AIMS 2001**	**AIMS 2002**	**AIMS 2003**	**AIMS 2000**	**AIMS 2001**	**AIMS 2002**	**AIMS 2003**
SAT9 2000	r	.868	.842	.838	.799	.837	.858	.905	.868
	N*	159	171	160	146	171	171	177	156
SAT9 2001	r	.862	.911	.865	.804	.795	.889	.878	.837
	N*	159	174	163	149	171	174	180	159
SAT9 2002	r	.864	.868	.899	.828	.806	.872	.908	.867
	N*	159	174	163	149	171	174	180	159
SAT9 2003	r	.827	.823	.848	.901	.771	.826	.856	.907
	N*	159	174	163	149	171	174	180	159

Note. All are significant at the 0.01 level.
* N is the number of grades for each year.

Results

Free and reduced lunch percentages

Because of the manner in which schools were selected, there are three distinct distributions of free and reduced lunch percentages over the four years. Table 3 displays the correlations for each year (2000 through 2003) between free/reduced lunch percentages and the AIMS and SAT9 scores for all schools.

Table 3
Correlations between free/reduced lunch percentages and AIMS and SAT9 scores, 2000-2003

		AIMS 2000	AIMS 2001	AIMS 2002	AIMS 2003	SAT9 2000	SAT9 2001	SAT9 2002	SAT9 2003
POV 2000	r	.826	-.810	-.845	-.778	-.770	-.727	-.758	-.726
	N*	59	62	61	55	62	63	63	63
POV 2001	r	-.945	-.922	-.947	-.892	-.879	-.841	-.864	-.828
	N*	57	59	58	52	60	61	61	61
POV 2002	r	-.921	-.900	-.942	-.880	-.885	-.842	-.871	-.834
	N*	56	58	57	51	57	58	58	58
POV 2003	r	-.910	-.904	-.934	-.877	-.863	-.835	-.855	-.831
	N*	57	59	58	52	58	59	59	59

Note. All significant at the 0.01 level (2-tailed).
* N is the number of schools since poverty data is available at the school level only.

Correlations between free/reduced lunch percentages and AIMS and SAT9 scores were all below $r = -.72$ ($p < .001$) across and between years. That is, the higher the test scores, the lower the percentage of students receiving free and reduced lunches. This finding also was obtained when the low-poverty schools were removed from the analysis and just the CSR and non-CSR schools (which both served students of poverty yet differed in saturation of poverty) were analyzed. The correlations between free/reduced lunch percentages and AIMS scores for these two poverty groups were below $r = -.46$, $p < .01$. The relationships between free and reduced lunch percentages and SAT9 scores in these poverty schools were in the same direction, between $r = -.56$ and $r = 0$, and many were non-significant. The relationship between saturation of poverty (the percentage of students receiving free and reduced lunches) and performance on the AIMS was stronger than the relationship between the saturation of poverty and SAT9 scores.

Differences among school types

Low-poverty schools had higher mean scores than the CSR and non-CSR schools on the AIMS performance areas, with overall mean scores 40-50 points higher in all cases (Table 4). The lowest percentage of students in low-poverty schools who passed in any year and performance area was 40% of fifth grade students in math in 2002 at one school, and there were schools with 100% passing in third grade writing in 2000, 2002, and 2003. At least six CSR and non-CSR schools had no students pass math in third or fifth grade one or more years.

Table 4
AIMS passing percentage means by year, performance area, grade and school type

Area	Grade	School type	Year			
			2000	2001	2002	2003
Math	Third	CSR (N = 14)	23.57	34.36	36.50	37.21
		Non-CSR (N = 14)	34.14	40.86	46.14	49.14
		Low-Pov (N = 19)	83.42	84.53	85.37	83.53
	Fifth	CSR (N = 15)	10.87	15.33	18.60	22.60
		Non-CSR (N = 17)	19.12	31.06	25.88	29.00
		Low-Pov (N = 19)	66.68	72.68	75.32	74.32
Reading	Third	CSR (N = 15)	46.40	50.80	50.87	52.73
		Non-CSR (N = 14)	55.29	56.93	59.64	61.64
		Low-Pov (N = 19)	92.00	91.21	91.95	89.26
	Fifth	CSR (N = 15)	38.07	29.00	29.40	32.40
		Non-CSR (N = 17)	46.71	38.29	38.59	41.18
		Low-Pov (N = 19)	89.05	82.21	83.37	79.47
Writing	Third	CSR (N = 15)	54.00	52.93	53.00	54.14
		Non-CSR (N = 17)	68.77	63.15	66.08	67.62
		Low-Pov (N = 19)	96.21	92.95	93.95	90.79
	Fifth	CSR (N = 15)	26.73	30.00	25.53	32.67
		Non-CSR (N = 17)	36.06	41.29	36.35	39.76
		Low-Pov (N = 19)	79.05	80.00	83.47	80.05

Note. Includes only those schools with reported scores for all years within grade by each year and performance area.
* N is the number of schools with reported passing percentages within each grade and performance area.

Student performance on the SAT9 show similar, although weaker, trends (Table 5). The low-poverty schools consistently outperformed the CSR and non-CSR schools both within and between grades across years.

Table 5
SAT9 NCE score means by year, performance area, grade and school type

Area	Grade	School type*	Year			
			2000	2001	2002	2003
Math	Third	CSR (N = 21)	35.24	39.70	41.22	39.86
		Non-CSR (N = 21)	39.21	39.39	42.11	43.34
		Low-Pov (N = 19)	65.42	66.89	66.89	65.95
	Fourth	CSR (N = 22)	40.14	40.46	42.22	41.69
		Non-CSR (N = 22)	42.61	44.42	45.24	46.00
		Low-Pov (N = 19)	67.89	68.53	68.68	70.26
	Fifth	CSR (N = 22)	36.86	40.58	41.34	42.28
		Non-CSR (N = 22)	43.75	44.65	46.07	46.05
		Low-Pov (N = 19)	68.89	69.53	70.63	70.16
Reading	Third	CSR (N = 21)	33.90	37.66	38.85	38.00
		Non-CSR (N = 21)	39.55	39.53	41.49	42.10
		Low-Pov (N = 19)	62.32	62.16	62.58	62.84
	Fourth	CSR (N = 22)	39.05	39.69	40.27	41.19
		Non-CSR (N = 22)	41.93	43.96	43.69	43.50
		Low-Pov (N = 19)	65.05	67.37	66.74	68.00
	Fifth	CSR (N = 22)	37.23	38.54	37.93	40.00
		Non-CSR (N = 22)	41.61	41.87	43.25	44.78
		Low-Pov (N = 19)	63.84	63.53	64.89	64.11
Language	Third	CSR (N = 21)	39.00	41.47	43.22	41.96
		Non-CSR (N = 21)	42.26	42.82	44.68	44.96
		Low-Pov (N = 19)	64.89	66.53	66.26	65.00
	Fourth	CSR (N = 22)	38.91	38.85	40.84	40.32
		Non-CSR (N = 22)	42.16	42.46	42.78	42.91
		Low-Pov (N = 19)	60.16	62.21	62.26	62.84
	Fifth	CSR (N = 22)	35.64	36.76	36.75	37.87
		Non-CSR (N = 22)	40.66	40.19	41.80	42.46
		Low-Pov (N = 19)	59.74	59.47	60.53	59.89

Note. Includes only those schools with reported scores for all years within grade by each year and performance area.
* N is the number of schools with reported NCE percentile scores within each grade and performance area.

Longitudinal analyses

Repeated-measures analyses of variance (RMANOVA) were performed on AIMS scores from third to fifth grades with a two-year lapse (third in 2000 to fifth in 2002, "cohort 1"; third in 2001 to fifth in 2003, "cohort 2"). In all cases, the low poverty schools outperformed the CSR and non-CSR schools ($p < .001$). There were ordinal interaction effects for school type (CSR, non-CSR, and low poverty) over time for reading and for writing in cohort 1, 2000 third graders to 2002 fifth graders, with less of a decrease in scores in the low poverty schools than the CSR or non-CSR schools.

There were decreases in scores for all AIMS performance areas and school types for cohort 1, third grade in 2000 to fifth grade in 2002, and cohort 2, third grade in 2001 to fifth grade in 2003 ($p < .001$; Table 6). A comparison between cohorts shows that although a decrement in their own performance trajectory, fifth grade students in poverty schools (CSR

and nonCSR) in 2003 scored higher in math than the fifth grade students in these schools in 2002. Further, the variation in student performance in third grade differed as a function of school type (p = .01) such that AIMS scores in CSR and nonCSR schools were more varied than in nonpoverty schools. This difference in dispersion as a function of school type was not found in the fifth grade.

Table 6
AIMS performance area percentage passing means by school type and year/grade

Area	School type	Year/grade			
		2000 3rd grade	**2002 5th grade**	**2001 3rd grade**	**2003 5th grade**
Math	CSR*	20.89 (10.82)	16.67 (11.22)	34.12 (19.57)	25.13 (14.24)
	Non-CSR (N = 18)	35.11 (22.06)	26.17 (15.69)	39.83 (21.63)	30.50 (19.19)
	Low Poverty**	83.42 (7.96)	75.32 (12.18)	84.30 (6.22)	74.20 (12.17)
Reading	CSR*	44.06 (13.22)	28.00 (12.65)	51.56 (17.31)	34.31 (15.60)
	Non-CSR (N = 18)	55.44 (17.59)	40.44 (14.90)	55.67 (17.53)	41.39 (17.49)
	Low Poverty	92.00 (3.94)	83.37 (8.41)	90.90 (5.73)	79.00 (9.80)
Writing	CSR*	52.17 (12.97)	24.83 (8.91)	53.25 (15.41)	34.56 (14.75)
	Non-CSR (N = 18)	67.39 (21.28)	38.44 (13.70)	62.00 (21.53)	39.72 (18.20)
	Low Poverty**	96.21 (2.42)	83.47 (7.46)	93.00 (3.89)	79.70 (9.04)

Note. Values in parentheses are standard deviations.
* N = 18 for CSR schools third grade in 2000 and fifth grade in 2002 and N = 16 for CSR schools in third grade in 2001 and fifth grade in 2003.
** N = 19 for non-CSR schools third grade in 2000 and fifth grade in 2002 and N = 20 for non-CSR schools in third grade in 2001 and fifth grade in 2003.

Student performance on the SAT9 indicated changes in performance from third to fifth grade; however, these results are not as straightforward as the AIMS test data (Tables 6 and 7). For both sets of longitudinal analyses (cohort 1, third grade in 2000 to fifth grade in 2002; cohort 2, third grade in 2001 to fifth grade in 2003), the statistical results were the same. There were no interaction effects for time by school type in any performance area. Scores changed significantly over time in all performance areas (p < .01): there was a linear drop in language, a linear *improvement* in math, and a quadratic change in reading, with an increase in fourth grade scores then slight decrease in fifth grade for almost all school types.

Table 7
SAT9 performance area means and standard deviations by school type and year / grade

Area	School type	Year					
		2000 3rd grade	2001 4th grade	2002 5th grade	2001 3rd grade	2002 4th grade	2003 5th grade
Math	CSR (N = 21)	35.24 (9.77)	40.11 (8.18)	40.88 (8.26)	39.70 (7.21)	41.76 (7.94)	41.81 (8.67)
	Non-CSR (N = 21)	39.21 (10.06)	44.11 (7.25)	45.69 (9.32)	39.39 (8.75)	44.87 (8.49)	45.53 (10.48)
	Low Poverty*	65.42 (6.55)	68.53 (4.02)	70.63 (5.00)	66.75 (6.29)	68.45 (5.82)	69.95 (5.91)
Reading	CSR (N = 21)	33.90 (8.32)	39.49 (8.86)	37.65 (7.27)	37.66 (8.22)	39.95 (11.44)	39.43 (10.51)
	Non-CSR (N = 21)	39.55 (7.56)	43.73 (11.88)	42.98 (10.53)	39.53 (9.62)	43.58 (10.38)	44.29 (10.61)
	Low Poverty*	62.32 (7.30)	67.37 (5.83)	64.89 (5.49)	61.95 (6.23)	66.50 (4.81)	63.80 (5.19)
Language	CSR (N = 21)	39.00 (8.45)	38.56 (8.14)	36.36 (13.62)	41.47 (7.29)	40.65 (9.30)	37.53 (9.41)
	Non-CSR (N = 21)	42.26 (8.32)	42.25 (12.53)	41.36 (13.27)	42.82 (7.80)	42.58 (8.27)	42.15 (8.33)
	Low Poverty*	64.89 (7.24)	62.21 (6.38)	60.53 (6.23)	66.40 (5.39)	62.15 (6.59)	59.80 (5.28)

Note. Values in parentheses are the standard deviations.
N = 19 for low-poverty schools in third 2000 to fifth 2002 and N = 20 for low-poverty schools in third 2001 to fifth 2003.

Discussion

The viability of the "fourth grade window" in student performance

Third grade scores on the AIMS test were a poor predictor of performance on the fifth grade test. The percentages of students passing the AIMS test in all performance areas decrease as the same cohort of students moves from third to fifth grade. Scores declined as predicted in both student cohorts. All schools dropped in percentage of students passing in each performance area of the AIMS test. Students in low-poverty schools, however, earned higher scores than those students in schools of both levels of poverty. Even in these schools, which had 80-90% of students passing the AIMS test, however, the "fourth grade window" is evident, indicating that greater resources alone are not the solution to declining performances in fifth grade.

The same trend is evident on the SAT9 for language, but not in math or reading. In those performance areas, the relative ranking of grades improved after third grade. Since the performance areas for reading and math on the AIMS and the SAT9 are highly correlated, the difference may be less due to content and more to the way in which test results are reported. If all students perform poorly on a norm-referenced exam, their relative ranking can remain the same and difficulties experienced by all of the students go unnoticed. The correlations

between the AIMS and the SAT9 tests in the fifth grade remain strong, suggesting that the tests continue to be aligned, thus, the drop from third to fifth grade does not appear to be a function of abnormalities in the AIMS test, although the feasibility of the cut-scores—the standard of excellence criteria—is worthy of consideration.

Policy implications

It is likely that policy makers using the results from the AIMS test would conclude that, despite several years of reform efforts, students across the board are dropping in their achievement from grades 3 to 5. This conclusion could warrant increasing sanctions to keep fourth and fifth grade teachers more squarely on a curriculum aligned with the test. This would mean a curriculum focused even more on reading and math and less time on science, physical education, music, and other non-tested content areas. It would not be surprising to find pejorative notions of "youth" (Nichols and Good, 2004) moving further into childhood as students are held accountable for their achievement decrements. School leaders may interpret the problem as reassigning effective teachers to the fifth grade (as likely has already been done with the third grade), thereby rendering fourth grade students even more vulnerable to achievement difficulties. Consider as well that there is some indication that poverty learners are becoming more similar with schooling while advantaged learners are becoming more diverse. The variation in third grade performance associated with school type dissipates at the fifth grade. Although this may be seen as a laudable achievement by some (exposure to schooling restricting the variation among poverty learners even if associated with a lower mean), others might worry that the variation among fifth graders of relative privilege is eroding earlier accomplishments. In each case, a more clear focus on the fourth grade window—rather than a policy of benign neglect—seems warranted.

In contrast, policy makers using the Stanford 9 results can maintain their current position regarding school reform as the data are essentially non-informative. We already know that poverty interferes with student performance. The "economics of home" in combination with a normal distribution of student achievement suggests that if someone is to be at the bottom it is understandably the poor. The same conclusion could support a call for increased resources for schools serving students of poverty. The notion of saturation of poverty affords a third alternative: the feasibility of designing school populations sensitive to home economics such that the saturation of poverty students attending a given school is kept below a specific ratio. Our analysis suggests that 80% level of poverty is more formidable than 70%. Research on poverty saturation thresholds and their relation to changes in student achievement seems warranted.

A major implication that emerges from both the AIMS and SAT9 results is that third grade performance is not particularly informative. The notion of third grade as *the* critical moment in learning that predicts future success is unwarranted. The fourth grade window is a compelling and understudied interval in student achievement. It is important that research examine more deeply the potential linkages between, and enactment of, curriculum and instruction expectations across the third, fourth, and fifth grades. Student mediation of these linkages seems especially promising. A better understanding of instructional dynamics in relation to the changing learning, reasoning, motivational and emotional capabilities of students is an important step toward understanding—and potentially reversing—their achievement declines.

Acknowledgment

This work was supported by a grant from the US Office of Educational Research and Improvement (OERI) Grant No. R306S000033. The authors take full responsibility for the work and no endorsement from OERI should be assumed. Thanks to Darrell Sabers for his helpful comments.

References

Berliner, D. C., & Biddle, B. (1995). *The manufactured crisis: Myths, fraud, and the attack on America's public schools.* Reading, MA: Addison-Wesley Publishing Co.

Carroll, J. B. (1963). A model for school learning. *Teachers College Record, 64*, 723-733.

Case, R., & Okamoto, Y., in collaboration with Griffin, S., McKeough, A., Bleiker, C., Henderson, B., & Stephenson, K. M. (1996). The role of central conceptual structures in the development of children's thought. *Monographs of the Society for Research in Child Development, 61*(1-2, Serial No. 246).

Chall, J.S. (1996). American Reading Achievement: Should We Worry? *Research in the Teaching of English, 30* (3), 303-310.

Coleman, J., Campbell, E., Hobson, C., McPartland, J., Meade, A., Weinfeld, F., & York, R. (1966). *Equality of educational opportunity.* Washington, DC: US Department of Health, Education and Welfare.

Gassen, S.G. (2003). Ariz. scores below U.S. average on federal test. *Arizona Daily Star,* 14 Nov., A6.

Good, T., Burross, H. L., & McCaslin, M. (in press). Comprehensive School Reform: A longitudinal study of school improvement in one state. *Teachers College Record*.

Good, T., & Grouws, D. (1979). The Missouri mathematics effectiveness project: An experimental study in fourth-grade classrooms. *Journal of Educational Psychology, 71,* 355-362.

Hanushek, E. (1997). Assessing the effects of school resources on student performance: An update. *Educational Evaluation and Policy Analysis, 19*(2), 141-164.

Hess, R., & Shipman, V. C. (1965). Early experience and the socialization of cognitive modes in children. *Child Development, 36*, 869-886.

Hill, H., Cohen, D., & Moffitt, S. (1999). Instruction, poverty, and performance. In G.Orfield & E. Debray (Eds.), *Hard work for good schools: Facts not fads in Title I reform.* The civil rights project: Harvard University (mimeo, pp. 55-76).

Hofer, B. K., & Pintrich, P. R. (1997). The development of epistemological theories: Beliefs about knowledge and knowing and their relation to learning. *Review of Educational Research, 67*, 88-140.

Jensen, A. R. (1973). *Educability and group differences.* New York: Harper & Row.

Jorgensen, O. (1999). Arizona AIMS for success. Graduation Standards put learning – and diplomas – on the line. *Clearing House, 73(1),* 23-25.

Klaus, P. E. (1973). *Yesterday's children: A longitudinal study of children from Kindergarten into the Adult years.* A Wiley-Interscience Publication. New York: John Wiley & Sons.

Ladd, H., & Hansen, J. (Eds.). (1999). *Making money matter: Financing America's schools.* Washington, DC: National Academy Press.

Luchins, A.S., & Luchins, E. H. (1950). New experimental attempts at preventing mechanization in problem solving. *Journal of Genetic Psychology, 42,* pp. 279-297.

McCaslin, M., Tuck, D., Wiard, A., Brown, B., LaPage, J., & Pyle, J. (1994). Gender composition and small-group learning in fourth-grade mathematics. *Elementary School Journal, 94,* pp. 467-482.

McNeil, L. (2000). Creating new inequalities: Contradictions of reform. *Phi Delta Kappan, 81*(10), 728-734.

National School Boards Association (1999, November 29). *Ten critical threats to America's children: Warning signs for the next millennium.* A report to the nation presented by The Hospital, Youth Crime Watch of American [online]. Available at http://www.nsba.org/highlights/ten_threats.htm

Nichols, S., & Good, T. (2004). *America's teenagers—myths and realities: Media images, schooling, and the social costs of careless indifference.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Piaget, J. (1983). Piaget's theory. In P. Mussen (Ed.), *Handbook of child psychology: Vol. 1. W. Kessen (Ed.), History, theory, and methods* (pp. 103-128). New York: John Wiley & Sons, Inc.

Pogrow, S. (1999). Overcoming the cognitive wall: Accelerating the learning of Title I students after the third grade. In G.Orfield & E. Debray (Eds.), *Hard work for good schools: Facts not fads in Title I reform.* The civil rights project: Harvard University (mimeo).

Smith, M. L., Heinecke, W. F. & Noble, A. J. (1999). Assessment policy and political spectacle. *Teachers College Record, 101*(2), 157-191.

Stiefel, L., Rubenstein, R., & Berne, R. (1998). Intra-district equity in four large cities: Data, methods, and results. *Education Finance, 23*(4), 447-467.

US Government Printing Office. (1999, July). *America's children: Key national indicators of well being.* (NCES 1999-019). Washington DC: Federal Interagency Forum on Child and Family Statistics.

Wang, J. (2003). An analysis of item score difference between 3rd and 4th grades using the TIMSS database. Presented at the Annual Meeting of the American Educational Research Association (Chicago, IL, April 21-23, 2003).

WestEd (2001). AIMS as a high school graduation requirement: Analysis of public survey data and recommendations. http://www.ade.az.gov/standards/aims/publicinput/SurveyFinal.pdf

About the Authors

Mary McCaslin is a professor of Educational Psychology at the University of Arizona. Her research interests include the co-regulation of classroom opportunities for student learning, motivation, and identity.

Heidi Legg Burross is adjunct instructional faculty at the University of Arizona. Her research interests include student transitions and perceptions of performance and achievement.

Thomas L. Good is a professor of Educational Psychology at the University of Arizona. His research interests include the study of classrooms, the communication of expectations, and the socialization of youth.

Education Policy Analysis Archives http://**epaa.asu.edu**