



RESEARCH REPORT

Breaking the Curve

Promises and Pitfalls of Using NAEP Data to Assess the State Role in Student Achievement

Matthew M. Chingos

October 2015



ABOUT THE URBAN INSTITUTE

The nonprofit Urban Institute is dedicated to elevating the debate on social and economic policy. For nearly five decades, Urban scholars have conducted research and offered evidence-based solutions that improve lives and strengthen communities across a rapidly urbanizing world. Their objective research helps expand opportunities for all, reduce hardship among the most vulnerable, and strengthen the effectiveness of the public sector.

Contents

Acknowledgments	iv
Executive Summary	v
Breaking the Curve	1
Do States Matter?	3
In Which States Do Students Break the Curve?	7
How Has Performance Changed over Time?	9
Conclusions	13
Appendix A	16
Appendix B	20
Data	20
Methods	22
Notes	24
References	26
About the Author	27
Statement of Independence	28

Acknowledgments

This report was funded by an anonymous donor. We are grateful to them and to all our funders, who make it possible for Urban to advance its mission. Funders do not, however, determine our research findings or the insights and recommendations of our experts. The views expressed are those of the authors and should not be attributed to the Urban Institute, its trustees, or its funders.

The author thanks Katharine Lindquist and Kristin Blagg for outstanding research assistance and Greg Acs, Matthew Di Carlo, Tom Loveless, and Morgan Polikoff for helpful comments on drafts of this report.

Executive Summary

States are increasingly at the center of education policymaking in the United States. Shifts in governance in recent decades have increased the role of state governments at the expense of local school districts. The federal government's role in education operates largely through state policy, and current trends suggest that Congress is likely to return even greater autonomy to the states.

The biennial release of results from the National Assessment of Educational Progress (NAEP) highlights the perceived state role in educational policy. Politicians and advocates from across the political spectrum point to high or improved NAEP scores as evidence supporting their preferred policies, ignoring researchers who argue that these scores rarely provide credible evidence that policies are working or failing.

This report provides a systematic framework for assessing how much student achievement varies across observationally similar states and the extent to which changes in state performance on the NAEP are accounted for by changes in the demographics of the state. Specifically, I measure the average NAEP performance of students in each state relative to similar students in other states. In addition, I assess how the performance of each state's students changed over 10 years relative to what might have been expected based on changes in demographics.

The analysis of student-level, restricted-use NAEP data from the US Department of Education yields two sets of findings. First, depending on the state they live in, similar students vary significantly in their test performance. But the states where students "break the curve" (i.e., perform better than their demographic peers) are often not the states with high scores overall. This means that any comparison of NAEP scores across states must consider each state's student population.

Second, NAEP scores in all 50 states have increased more than would be expected based on demographic shifts between 2003 and 2013. This means that analyses of trends over time in state performance on the NAEP need to consider changes in student demographics.

The NAEP is an incredibly important and useful tool for researchers and policymakers as the only set of exams that can be used to compare student performance across all 50 states. The federal entities responsible for NAEP should more proactively encourage the responsible use of this vital resource by enabling the public to make demographic adjustments across states and by enhancing the underlying database that makes these adjustments possible.

Breaking the Curve

Public education policy in the United States is made largely at the state and local level. Most of the shifts in governance in recent decades have increased the role of the state government at the expense of local school district control. In the 1970s and 1980s, many states adopted school finance equalization policies aimed at decreasing disparities in funding across school districts or, at least, increasing funding at the lowest-spending districts (Hoxby 2001). In 1987, New Jersey was the first state to adopt a law allowing the state government to take over troubled school districts. Since then, at least 28 states have adopted similar laws.¹

The federal government's role in education operates predominantly through state policy. Federal education spending often flows through states, and federal policy largely works through mandates implemented at the state level. The passage of the No Child Left Behind Act (NCLB) of 2001 represented the most significant change in federal education policy since the adoption of the original Elementary and Secondary Education Act in 1965. But NCLB did not create a federal system of accountability—it mandated that states design and implement testing and accountability systems that meet certain federal guidelines. The next reauthorization of NCLB is likely to return even greater autonomy to the states.

The perceived importance of the state role in education policy is most visible every two years when scores from the National Assessment of Educational Progress (NAEP) are released. Since 2003, every state has participated in the main NAEP exams in math and reading, so it has become increasingly common to use states' scores, and changes in those scores over time, to evaluate both the overall quality of states' education systems and the impact of recent education reforms.

Politicians and advocates from across the political spectrum have pointed to high or improved NAEP scores as evidence supporting their preferred policies. On the Republican side, Tennessee governor Bill Haslam proudly proclaimed that the state had the largest increase in NAEP scores in the country.² Former Florida governor Jeb Bush has also trumpeted his state's NAEP scores, especially the average reading score of fourth-grade Hispanic students in Florida—ranked highest in 2013.³ And the American Legislative Exchange Council has pointed to increases in NAEP scores in states that have adopted school choice policies: “[S]tates that have adopted school choice policies have seen an increase in test scores, including Indiana, which leads the states in education reform policy.”⁴

On the Democratic side, US Secretary of Education Arne Duncan joined local officials to announce Washington, DC's, significant gains in NAEP scores in 2013, claiming “The reforms D.C. has put in place

are working.”⁵ American Federation of Teachers president Randi Weingarten has used NAEP data to argue in favor of strong teachers’ unions: “[I]n states like Massachusetts and Minnesota, where public schools are heavily unionized, students earn the highest scores on the National Assessment of Educational Progress....In contrast, students in states such as Mississippi, Louisiana and Arkansas, which have few if any teachers union members and virtually no union contracts, have the lowest NAEP scores.”⁶

This biannual exercise is predictably accompanied by a chorus of researchers who warn against interpreting NAEP scores, or changes in NAEP scores, as evidence that policies are working or failing (Loveless 2013).⁷ A leading concern is that NAEP punditry usually ignores differences across states in demographics and other student characteristics that affect test performance. Not all analyses of NAEP scores ignore the role of student demographics in test-score performance, but what is missing from this discussion is a systematic framework for assessing how much student achievement varies across observationally similar states and the extent to which changes in state performance on NAEP are accounted for by changes in the demographics of the state (Loveless 2011).

This report begins to fill this gap with a detailed analysis of the most recent (2013) NAEP data available and of changes over the previous decade (2003–13). First, I estimate how much states matter for student achievement by examining how much test performance varies across states once student characteristics are taken into account. Second, I measure the average performance of students in each of the 50 states relative to similar students in other states. In other words, which states “break the curve” of student achievement? Finally, I examine how the performance of each state’s students changed from 2003 to 2013 relative to what might have been expected based on changes in demographics.

The NAEP data reveal significant variation in average test scores across states, even after taking student characteristics into account. But the relative ranking of states changes substantially when each state’s students are compared with similar students in other states, as opposed to a simple comparison of average scores. This means that interstate comparisons of NAEP scores should not be done without considering the characteristics of each state’s student population. And though the adjusted measures allow for more meaningful comparisons across states, they are still unlikely to reveal the causal impact of particular education policies. There remain a number of unmeasured student characteristics and differences in non-education policies that could shed further light on these data.

Accounting for demographics also affects the size and direction of changes in state performance on the NAEP. In particular, NAEP scores in all states have increased more than would be expected based

on demographic shifts between 2003 and 2013. But the ranking of states based on *changes* in their NAEP scores is relatively unaffected by the demographic adjustments. In other words, NAEP scores are likely to be much more useful for evaluating the impact of policy changes within a state than for comparing the effects of policies across different states.

Do States Matter?

Existing research indicates that levels of education service delivery that are further removed from students tend to have weaker associations with student achievement. A recent study used student-level data from two states' tests to partition the variation in student achievement into the components associated with students, classrooms, schools, and districts (Chingos, Whitehurst, and Gallaher 2015). The analysis revealed that, in general, more variation was associated with teachers than with schools and with schools than with districts. For example, data on fourth- and fifth-grade math scores in North Carolina indicated that 63 percent of the variance in achievement was at the student level, 5 percent at the teacher level, 3 percent at the school level, and 2 percent at the district level.

This logic suggests that states might matter little for student achievement, but there are several reasons why they may have a significant impact on students. First, even small shares of variance can still translate into significant differences in student achievement. The aforementioned study found that a one standard deviation difference in district effectiveness translated into 0.11 standard deviations in student achievement (about 2–3 months of schooling), despite the fact that districts only accounted for 1–2 percent of the total variation in test scores.

Second, variation in education policy at the state level may be more substantial and meaningful than variation in policy at the district level. This is especially likely to be the case in states where many district policies—such as funding, class size, teacher evaluation, and employment rules—are set or constrained by state policy. State policy may shift the distribution of student achievement across all districts, schools, and classrooms. And though state policy has the clearest implications for public schools, it also has the potential to affect enrollment in and the operation of private schools (e.g., through regulation and school choice policies).

Finally, we already know that NAEP scores vary significantly across states. It is less clear how much differences in student characteristics drive differences in test scores across states. For example, Massachusetts students post among the highest scores on the NAEP exams, whereas Mississippi is among the lowest-scoring states. But how much of that difference can be attributed to demographic

differences between the two states, such as the fact that Mississippi's poverty rate was twice that of Massachusetts (24 versus 12 percent) in 2013 (Bishaw and Fontenot 2014, table 1)?

I address this question using restricted-use, student-level NAEP data from the US Department of Education. The data cover four exams: math and reading in the fourth and eighth grades. I include students in both public and private schools in all 50 states.⁸ I exclude the District of Columbia because it is a large city (not a state).

I calculate the relative performance of each state, adjusting for the following rich set of student-level factors:

- gender
- race and ethnicity
- eligibility for free or reduced-price lunch
- limited English proficient
- special education
- age
- whether the student was given an accommodation on the NAEP exam (e.g., extra time or a separate room)
- whether the student has various amenities in their home (e.g., computer, Internet, own room, dishwasher, and clothes dryer)
- the number of books in the home
- the language spoken at home
- the family structure (e.g., two-parent, single-parent, foster)

The statistical adjustment for student demographics is made using student-level data, as opposed to state-level data (see appendix B for more detail of the data and methodology). This means that states are judged by how well their students do relative to students with similar characteristics across the country. This methodology is not perfect, as the NAEP data do not measure every possible student characteristic that could affect test performance. But with more than 160,000 student observations per test, the relationship between a substantial number of student characteristics and test scores can be

estimated much more accurately than with state-level data, which has only 50 observations. In appendix B, I show that my results are much less sensitive to the choice of control variables than a state-level approach.

There is also an important difference in interpretation between state- and student-level analyses. The student-level methodology I use compares how students in a given state score relative to similar students around the country, regardless of where they live. A methodology that uses state-level data would only compare states with other states with similar student populations. For example, a state-level analysis would compare a state with 30 percent of students eligible for free or reduced-price lunch to states with similar proportions of eligible students, whereas my analysis compares how those 30 percent of eligible students score compared with all eligible students, and how the 70 percent of ineligible students compare with all ineligible students.

Table 1 reports the standard deviation across states in NAEP performance. One standard deviation across states is roughly the difference between the state ranked 35th and the state ranked 16th in performance. The first two columns are reported in terms of student-level standard deviations on the test. To facilitate interpretation, the table also converts the student standard deviations into months of learning, reported in the final two columns.⁹

If states matter little for student achievement, then the variation in average NAEP scores across states would be greatly diminished by accounting for the rich set of student characteristics available in these data. But that is not the case—student demographics only explain about one third of the variation in NAEP scores across states. The remaining variation is substantial. Averaged across the four tests, a one-standard-deviation increase in state performance is associated with about four more months of learning (0.12 score standard deviations).

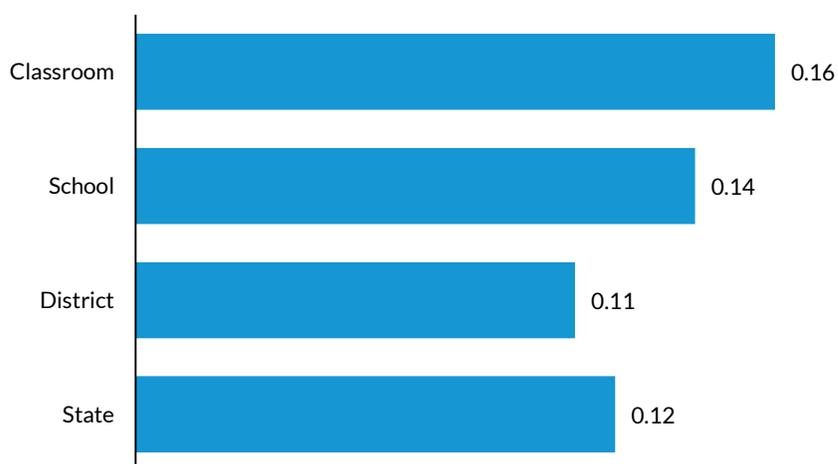
TABLE 1
State Variation in NAEP Scores, 2013

	Student Standard Dev		Months of Learning	
	Unadjusted	Adjusted	Unadjusted	Adjusted
4th-grade math	0.18	0.12	3.2	2.1
4th-grade reading	0.18	0.12	4.5	3.0
8th-grade math	0.20	0.14	9.0	6.4
8th-grade reading	0.17	0.10	7.2	4.4
Average	0.18	0.12	6.0	4.0

How does variation across states compare with other institutional levels, such as districts and schools? Figure 1 shows that states appear to vary about as much in their performance as districts do, somewhat less than schools, and significantly less than teachers (Whitehurst, Chingos, and Gallaher 2013, figure 4). Further, the greatest variation in test scores occurs within classrooms, and this comparison understates the relative importance of teachers because their effects are measured over a single year whereas the effects at the other levels are measured over all of the years the student spends in the school, district, or state.

FIGURE 1

Comparison of One Standard Deviation of Classroom, School, District, and State Differences in Student Achievement



Sources: Result for states is from table 1 of this report; result for other levels is from Whitehurst, Chingos, and Gallaher (2013, figure 4).

Notes: 0.10 student standard deviations is equivalent to roughly two months of learning in fourth grade. Classroom effects condition on prior-year test scores whereas the school, district, and state effects do not condition on prior-year test scores.

The bottom line is that states vary in terms of their students' NAEP performance, and within-state variation across students, classrooms, schools, and districts is much greater. In the words of education researcher Tom Loveless, "[A]nyone who follows NAEP scores knows that the difference between Massachusetts and Mississippi is quite large. What is often overlooked is that every state has a mini-Massachusetts and Mississippi contrast within its own borders" (2013).¹⁰

In Which States Do Students Break the Curve?

Which states come out on top in terms of how well their students perform compared with similar students in other states? I address this question using a similar methodology, in which the NAEP scores of students in each state are compared with the scores of similar students in other states.¹¹ Using the rich set of control variables described above, I calculate a predicted score for each student and then I calculate how much better or worse the student did compared with this prediction. Each state's performance is measured as the average of the differences between actual and expected test scores.

Figure 2 ranks states by the demographically adjusted performance of their students, averaged across the four tests and reported in months of learning.¹² It is important to bear in mind that this is only a relative measure (with the average state arbitrarily assigned a value of zero), not an absolute measure of performance. Examining both the adjusted and unadjusted measures shows that the two measures diverge significantly. The demographic adjustment leads to an average change of nine positions in the 50-state ranking (see table A.1 for both rankings).

The top two states in terms of adjusted student performance are the same two states with the top unadjusted scores: Massachusetts and New Jersey. To be sure, the adjusted performance measures for these states are smaller than their unadjusted measures—my analysis indicates that their students are 6–9 months ahead of their peers in the average state, as compared with the 10–12 months suggested by the raw data.

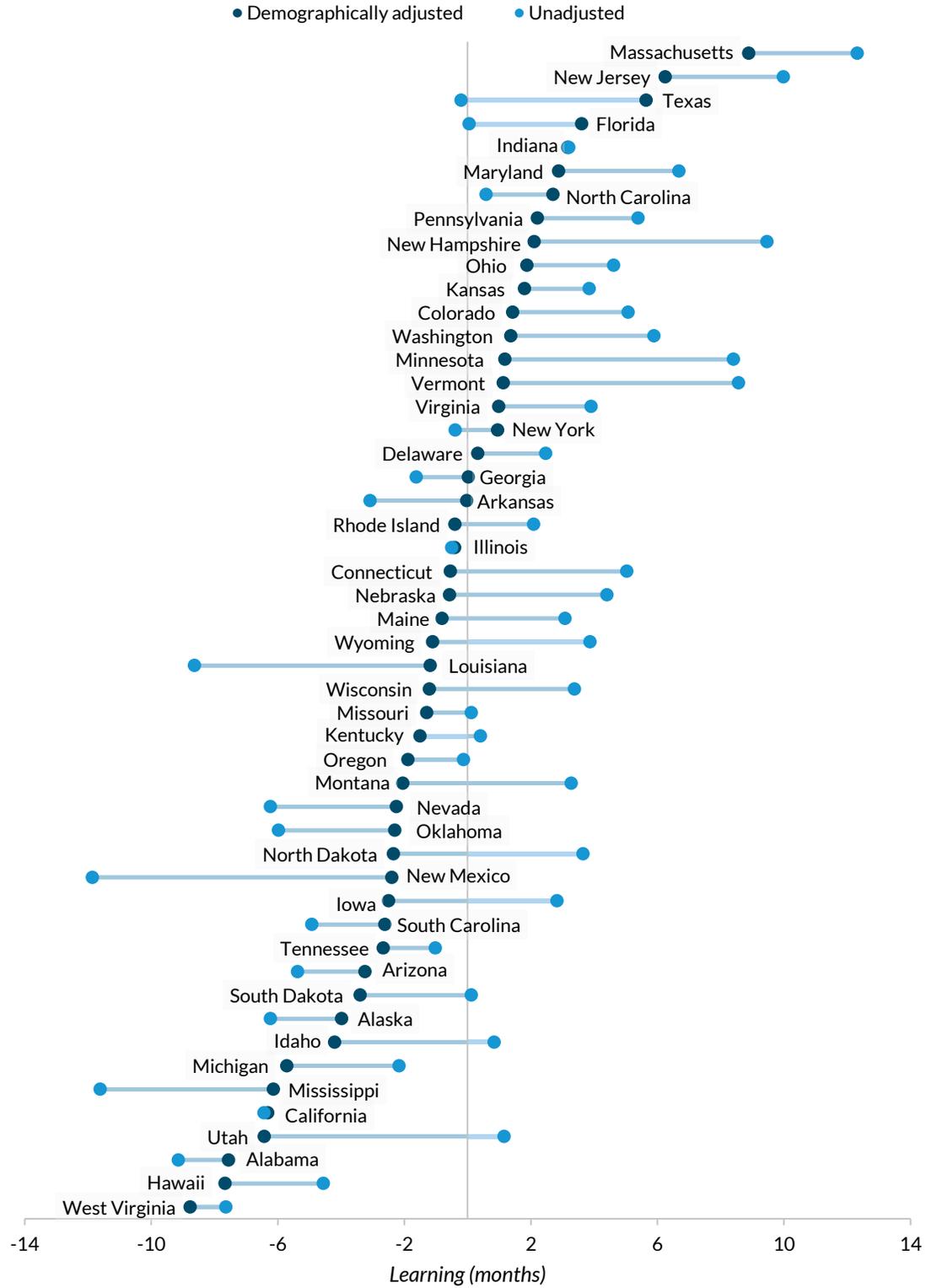
But the next two spots go to Texas and Florida, respectively, which have average overall performance, but where student performance significantly exceeds the test scores of similar students in other states. A relatively large share of students in these states come from demographic groups that tend to score less well on NAEP. But, as the results show, these students score better than similar students in other states.

This is illustrated by a simple comparison of fourth-grade reading scores in Oregon and Texas. Texas limited English proficient (LEP) students perform much better than LEP students in Oregon, and non-LEP students perform about the same in both states. But 21 percent of Texas students are LEP, as compared with 13 percent of Oregon students, so the overall average NAEP score is higher in Oregon than in Texas.¹³

Figure 2 also shows the substantial overall variation in state NAEP performance averaged across the four exams, which differs between the top- and bottom-ranked states—Massachusetts and West Virginia—by almost 18 months of learning.

FIGURE 2

State Performance on 2013 NAEP



How Has Performance Changed over Time?

Punditry using NAEP data often focuses more on changes in NAEP performance than on the absolute levels of performance. This is especially true for those seeking to associate a change in NAEP performance with the recent implementation of a policy change. Setting aside the issue of associating a change in performance with any single policy change or set of changes, examining changes in performance is potentially informative especially because it takes into account any unmeasured characteristics of each state that are constant over time. Whereas we might worry that two states like Oregon and Texas differ in hard-to-measure ways, we are probably more willing to believe that Texas in 2003 is not that different from Texas in 2013.

The data reported above show that states vary markedly in their overall performance on NAEP, with differences at any given point in time partly reflecting differences in student populations. Such differences in characteristics can also appear in data points for the same state at different points in time. For example, the share of fourth-grade Texas students classified as LEP increased from 13 to 22 percent from 2003 to 2013.¹⁴

Shifts in demographics can affect trends in performance. For example, scores for all racial and ethnic groups of students could increase in a state, but the average score could remain flat or even decline if the number of students in the lower-performing groups increases faster than the number from the higher-performing groups.

I address this question empirically by calculating how much NAEP scores increased between 2003 and 2013, relative to what might have been expected based on changes in student demographics. I use 2003 as the starting year because it was the first year that all states were required to participate in NAEP. I use 2013 as the ending year because it is the most recent year for which NAEP data are available.

I first measure the relationship between scores in 2003 and a set of student factors that are available in both the 2003 and 2013 data: gender, race and ethnicity, eligibility for free or reduced-price lunch, LEP, special education, age, and exam accommodation. I then predict what each student's score in 2013 would have been if the relationships between demographics and scores were the same in 2013 as they were in 2003. Finally, I measure how much better or worse students in 2013 did relative to the prediction.

The results of this analysis for all 50 states combined are shown in table 2. The first row shows that fourth-grade math scores were predicted to fall 0.10 standard deviations (1.8 months of learning) over

this decade. This means that there was greater growth in the population of students who tended to score lower in 2003, on average, than students who tended to score higher in 2003.

But despite this predicted fall in scores, fourth-grade math scores actually increased 0.24 standard deviations (4.3 months of learning). In other words, scores increased 0.34 standard deviations (6.2 months) more than would have been expected based on changes in demographics. A similar pattern holds for the other three tests, with the average across all tests indicating a demographically adjusted increase in scores of 0.28 standard deviations (8.9 months).

TABLE 2

Predicted and Actual Changes in NAEP Scores

All 50 states, 2003–13

	Student Standard Deviations			Months of Learning		
	Predicted change	Actual change	Difference	Predicted change	Actual change	Difference
4th-grade math	-0.10	0.24	0.34	-1.8	4.3	6.2
4th-grade reading	-0.13	0.10	0.23	-3.2	2.6	5.8
8th-grade math	-0.10	0.20	0.30	-4.5	9.2	13.7
8th-grade reading	-0.10	0.13	0.24	-4.4	5.5	9.8
Average	-0.11	0.17	0.28	-3.5	5.4	8.9

I repeat this analysis for each of the 50 states. The results, reported in figure 3, show that scores in every single state increased more than would have been expected based on demographics. But adjusted score increases varied widely, from a 2-month increase of learning in South Dakota to a 16-month increase in Nevada.

In general, the demographic adjustment makes less of a difference to the ranking of state changes in scores over time than it did for the relative ranking of 2013 scores. The average state moves five positions as a result of the adjustment (see table A.2 for both rankings), and the correlation between adjusted and unadjusted changes in scores is 0.90 (as compared with 0.69 for 2013 scores). States with large increases in raw scores tend to also have large increases relative to expectations based on demographic adjustments. This is unsurprising given that demographic differences between two years in a single state are likely to be smaller than differences across states at a single point in time.¹⁵

But there are some notable exceptions, with Indiana moving 20 positions up in the ranking (from 38th to 18th) and Alabama falling 21 positions (from 17th to 38th). Other notable shifts include North

Carolina and Oregon moving up in the rankings and Idaho, Mississippi, and Wyoming moving down in the rankings.

And even if the adjustment does not move the relative ranking of many states, it still leads to different conclusions than a simple examination of average scores in 2003 and 2013. The adjustment shows how much more scores have increased than would have been expected based on changes in demographics (figure 3).

The results of the two analyses are summarized in figure 4, which shows demographically adjusted 2013 scores and 2003–13 changes for each state (averaged across in the four tests). The blue lines are the average of each measure. States such as Florida, Maryland, Massachusetts, New Jersey, and Texas have students who outperform their peers in other states and where performance increased more than in the average state. Other states, such as Michigan, Mississippi, and West Virginia, have not increased their performance as rapidly and are below average in terms of demographically adjusted performance levels (although, as noted above, all states have made gains after accounting for demographic shifts).

There are states with above-average performance but below-average growth (e.g., New York and North Carolina) as well as states with below-average performance but above-average growth (e.g., Hawaii and Nevada). Finally, there is a large but diverse group of states who are about average on both measures, including Kentucky, Maine, and Nebraska.

FIGURE 3

Change in State NAEP Performance

2003-13

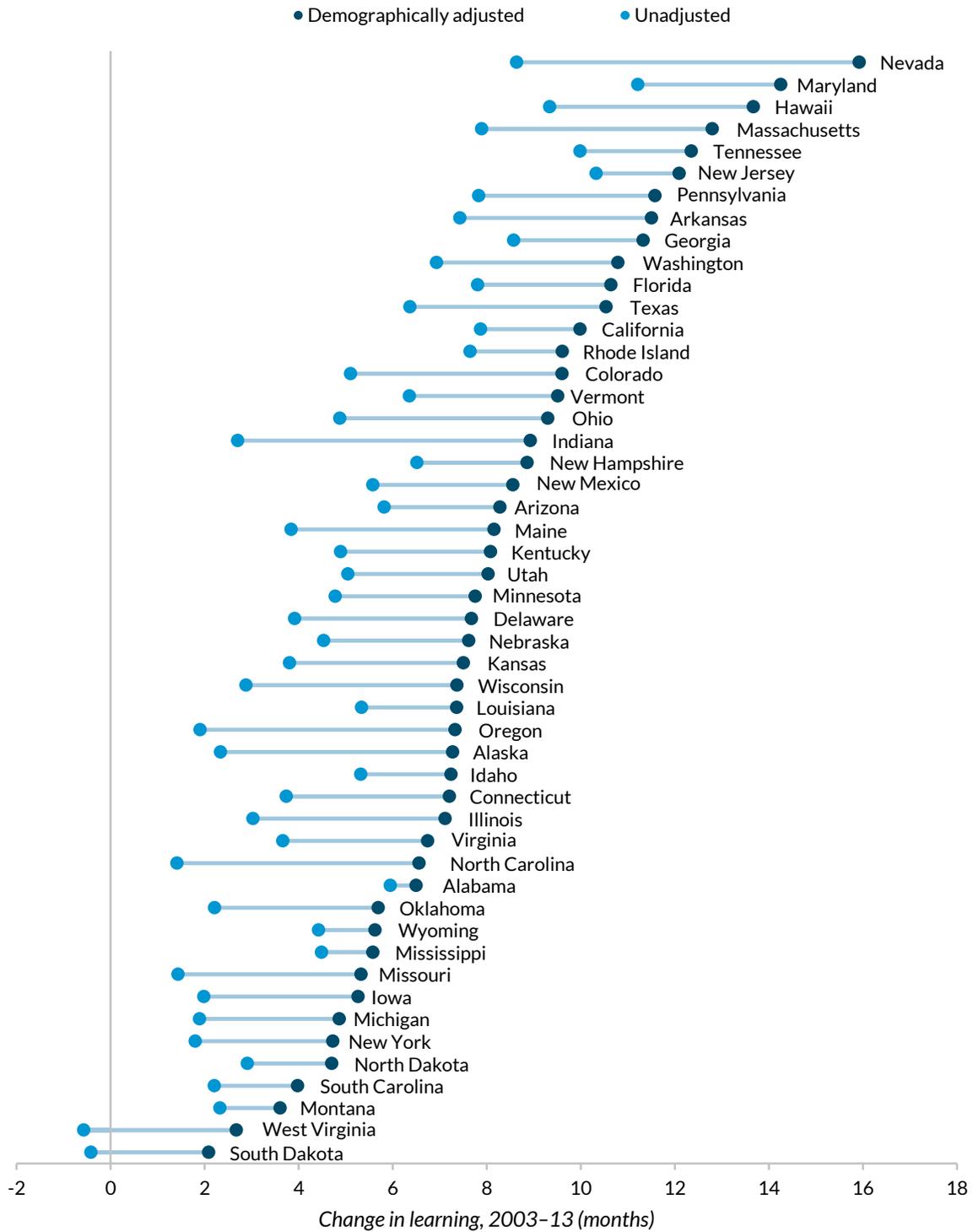


FIGURE 4

State NAEP Performance (2013) and Changes in Performance (2003-13)

Adjusted for demographics, in months of learning



Conclusions

NAEP scores have been misused as long as the test has been given. They are cited as evidence our education system is broken, as cause for optimism that things are getting better, as evidence that pet policies are working, and as proof that disfavored reforms have failed. NAEP scores are used to point to exemplars of perceived success backed up with a convenient narrative. High-scoring Massachusetts is a frequently cited example, which some say has everything to do with its test-based accountability system, while others credit its strong teachers' unions.

All uses of NAEP scores have important limitations, and the demographically adjusted scores presented in this report are no exception. Chief among these limitations is the underlying assumption that the student factors mean the same thing across and states and over time. For example, accounting

for free and reduce-price lunch eligibility in the analysis assumes that the typical eligible student in New York is similar to the one in Florida, and that the average socioeconomic background of eligible students (relative to ineligible students) was the same in 2003 and 2013. Of course, this may not be the case; for example, students eligible for free lunch may be more disadvantaged in some states than in others, or the relative disadvantage of eligible and ineligible students may change over time (e.g., because of broader economic changes).¹⁶

Another limitation is that the NAEP data, like any dataset, do not capture every student and family factor that could possibly affect student achievement. But if these limitations are kept in mind, the adjustments that are possible with the student-level NAEP data represent a clear improvement over just eyeballing the raw data. The LEP population in Texas likely differs in important unmeasured ways from the LEP students in Oregon, but these two groups of students likely share more in common than a pair of randomly selected students from the two states.

Finally, it is important to emphasize that differences across states in the NAEP performance of similar students reflect both differences in education policy and variation in all state policies that affect children, not to mention differences in history and culture. This is also potentially the case when tracking student performance in the same state over time. A score increase (adjusted for demographics) could reflect changes in the schools, but it might also be the result of changes in the economy or in the availability of social safety nets.

The NAEP is an incredibly important and useful tool for researchers and policymakers if it is used in ways that are mindful of its limitations. The analyses presented in this report have at least two sets of implications for the future use of NAEP data to evaluate states' education policies.

First, comparisons of student performance across states are likely to be much more informative if they adjust for differences in student demographics. The striking differences in the average test performance of similar students who reside in different states can be used to generate hypotheses for testing in future research. For example, demographically adjusted test performance could be used to identify state education policies (e.g., regarding funding or accountability) that coincide with strong performance and thus merit more careful study.

Second, the results in this report make clear that tracking student achievement over time requires accounting for population shifts in our ever-changing nation. The 2003 and 2013 data show that recent gains in NAEP scores have been partially concealed by increased enrollments of lower-scoring demographic groups in all 50 states. Careful comparisons of changes in NAEP performance within states over time also have the greatest potential to reveal the impact of state policy changes, as

exemplified by previous high-quality studies on subjects ranging from universal prekindergarten to school accountability (Fitzpatrick 2008; Dee and Jacob 2011).

Better use of NAEP data will ultimately come down to the users themselves, but the federal entities responsible for NAEP can help. The US Department of Education and the National Assessment Governing Board (which oversees NAEP) should proactively encourage the responsible use of NAEP data, for example by releasing demographically adjusted NAEP data alongside the raw scores for each state. There is not a single preferred way to make these adjustments, but the National Center for Education Statistics could develop an online tool that would allow users to calculate demographically adjusted scores (and trends over time) based on a user-selected set of student demographic factors.

NAEP scores will become more useful for making comparisons across states if the underlying data include more detailed information on the characteristics of students and their families. NAEP data currently rely on administrative data from schools and survey responses from students, teachers, and administrators. These data could be augmented by links to other federally held data, such as detailed earnings records on students' parents maintained by the Internal Revenue Service. Such a linkage would permit significant improvements over the current use of eligibility for the federal free and reduced-price lunch program as the main proxy for socioeconomic status.

Education researchers, policymakers, and practitioners are hungry for information on what works in education. The NAEP truly is the "nation's report card," as it is the only set of exams that can be used to compare student performance across all 50 states. Coupling principles of sound research design with the use of NAEP data is critical to maximizing the positive impact of this vital resource.

Appendix A

TABLE A.1

State Performance on NAEP, 2013

State	Adjusted		Unadjusted		Rank difference
	Learning (months)	Rank	Learning (months)	Rank	
Massachusetts	8.9	1	12.3	1	0
New Jersey	6.3	2	10.0	2	0
Texas	5.6	3	-0.2	32	29
Florida	3.6	4	0.0	30	26
Indiana	3.2	5	3.2	19	14
Maryland	2.9	6	6.7	6	0
North Carolina	2.7	7	0.6	26	19
Pennsylvania	2.2	8	5.4	8	0
New Hampshire	2.1	9	9.5	3	-6
Ohio	1.9	10	4.6	11	1
Kansas	1.8	11	3.8	15	4
Colorado	1.4	12	5.1	9	-3
Washington	1.4	13	5.9	7	-6
Minnesota	1.2	14	8.4	5	-9
Vermont	1.1	15	8.6	4	-11
Virginia	1.0	16	3.9	13	-3
New York	0.9	17	-0.4	33	16
Delaware	0.3	18	2.5	22	4
Georgia	0.0	19	-1.6	36	17
Arkansas	0.0	20	-3.1	38	18
Rhode Island	-0.4	21	2.1	23	2
Illinois	-0.4	22	-0.5	34	12
Connecticut	-0.5	23	5.0	10	-13
Nebraska	-0.6	24	4.4	12	-12
Maine	-0.8	25	3.1	20	-5
Wyoming	-1.1	26	3.9	14	-12
Louisiana	-1.2	27	-8.6	47	20

State	Adjusted		Unadjusted		Rank difference
	Learning (months)	Rank	Learning (months)	Rank	
Wisconsin	-1.2	28	3.4	17	-11
Missouri	-1.3	29	0.1	29	0
Kentucky	-1.5	30	0.4	27	-3
Oregon	-1.9	31	-0.1	31	0
Montana	-2.0	32	3.3	18	-14
Nevada	-2.2	33	-6.2	43	10
Oklahoma	-2.3	34	-6.0	42	8
North Dakota	-2.3	35	3.6	16	-19
New Mexico	-2.4	36	-11.9	50	14
Iowa	-2.5	37	2.8	21	-16
South Carolina	-2.6	38	-4.9	40	2
Tennessee	-2.7	39	-1.0	35	-4
Arizona	-3.2	40	-5.4	41	1
South Dakota	-3.4	41	0.1	28	-13
Alaska	-4.0	42	-6.2	44	2
Idaho	-4.2	43	0.8	25	-18
Michigan	-5.7	44	-2.2	37	-7
Mississippi	-6.1	45	-11.6	49	4
California	-6.3	46	-6.4	45	-1
Utah	-6.4	47	1.2	24	-23
Alabama	-7.6	48	-9.1	48	0
Hawaii	-7.7	49	-4.6	39	-10
West Virginia	-8.8	50	-7.6	46	-4

TABLE A.2

Change in State NAEP Performance, 2003–13

State	Adjusted		Unadjusted		Rank difference
	Learning (months)	Rank	Learning (months)	Rank	
Nevada	15.9	1	8.6	5	4
Maryland	14.3	2	11.2	1	-1
Hawaii	13.7	3	9.3	4	1
Massachusetts	12.8	4	7.9	7	3
Tennessee	12.3	5	10.0	3	-2
New Jersey	12.1	6	10.3	2	-4
Pennsylvania	11.6	7	7.8	9	2
Arkansas	11.5	8	7.4	12	4
Georgia	11.3	9	8.6	6	-3
Washington	10.8	10	6.9	13	3
Florida	10.6	11	7.8	10	-1
Texas	10.5	12	6.4	15	3
California	10.0	13	7.9	8	-5
Rhode Island	9.6	14	7.7	11	-3
Colorado	9.6	15	5.1	22	7
Vermont	9.5	16	6.4	16	0
Ohio	9.3	17	4.9	25	8
Indiana	8.9	18	2.7	38	20
New Hampshire	8.9	19	6.5	14	-5
New Mexico	8.6	20	5.6	19	-1
Arizona	8.3	21	5.8	18	-3
Maine	8.2	22	3.8	31	9
Kentucky	8.1	23	4.9	24	1
Utah	8.0	24	5.0	23	-1
Minnesota	7.8	25	4.8	26	1
Delaware	7.7	26	3.9	30	4
Nebraska	7.6	27	4.5	27	0
Kansas	7.5	28	3.8	32	4
Wisconsin	7.4	29	2.9	37	8
Louisiana	7.4	30	5.3	20	-10
Oregon	7.3	31	1.9	44	13
Alaska	7.3	32	2.3	39	7
Idaho	7.2	33	5.3	21	-12

State	Adjusted		Unadjusted		Rank difference
	Learning (months)	Rank	Learning (months)	Rank	
Connecticut	7.2	34	3.7	33	-1
Illinois	7.1	35	3.0	35	0
Virginia	6.7	36	3.7	34	-2
North Carolina	6.6	37	1.4	48	11
Alabama	6.5	38	5.9	17	-21
Oklahoma	5.7	39	2.2	41	2
Wyoming	5.6	40	4.4	29	-11
Mississippi	5.6	41	4.5	28	-13
Missouri	5.3	42	1.4	47	5
Iowa	5.3	43	2.0	43	0
Michigan	4.9	44	1.9	45	1
New York	4.7	45	1.8	46	1
North Dakota	4.7	46	2.9	36	-10
South Carolina	4.0	47	2.2	42	-5
Montana	3.6	48	2.3	40	-8
West Virginia	2.7	49	-0.6	50	1
South Dakota	2.1	50	-0.4	49	-1

Appendix B

Data

This report draws on restricted-use, student-level NAEP data on the 2003 and 2013 administrations of fourth- and eighth-grade reading and math tests. These tests are given every two years to a nationally representative sample of US students. Since 2003, every state has been required to participate so that state-level results can be calculated.¹⁷ I use data on students in the national reporting sample (RPTSAMP = 1) who reside in one of the 50 states (I exclude DC and other jurisdictions that are not states).

To minimize the time that students sit for the tests, no student takes an entire test. For the analysis, I use a statistical estimate of what each student's score on the test would have been had he or she taken the test in its entirety. For 2013, using the procedures described in the documentation provided with the restricted-use data, this estimate is based on 20 plausible test-score values and 62 replicate weights (in prior years, including 2003, there were 5 plausible values instead of 20).

For the analysis of the 2013 data, I use a rich set of student-level control variables that are drawn from administrative records and a student survey.¹⁸ The variables used and the coding of them is as follows:

- SEX: gender (male or female)
- DRACEM: race and ethnicity, as reported by the student (white, black, Hispanic, Asian American or Pacific Islander, American Indian or Alaska Native, or multiple)
- SLUNCH: eligibility for federal free and reduced-price lunch program (not eligible, eligible for reduced-price lunch, eligible for free lunch, or other or missing)
- LEP: student classified as an English language learner (yes or no)
- IEP: student classified as having a disability (yes or no)
- Age on February 1 of testing year, using date of birth estimated as 15th day of birth month [BMONTH] in birth year [BYEAR], with ages more than two years from the mean weighted national age recorded to the mean

- ACCOMCD: whether the student received an accommodation (no accommodation, accommodation in regular testing session, or accommodation in separate testing session)
- B017101: whether student reported having a computer in the home (yes or no)
- B0267A1: whether student reported having Internet access in the home (yes or no)
- B0267B1: whether student reported having a clothes dryer in the home (yes or no)
- B0267C1: whether student reported having a dishwasher in the home (yes or no)
- B0267E1: whether student reported having his/her own bedroom (yes or no)
- B013801: number of books in the home, as reported by the student (0–10, 11–25, 26–100, or more than 100)
- B018201: language other than English spoken at home (never, once in a while, about half of the time, or all or most of the time)
- B0268A1 through B0268F1: family structure, measured as which parents child lives with (mother and father, mother only, father only, mother and other parent or guardian, father and other parent or guardian, foster parent(s), or other or missing).

Student survey data are missing for nearly all students in Alaska, so performance metrics for Alaska are calculated using a more restricted set of control variables (gender, race and ethnicity, free and reduced-price lunch eligibility, LEP, special education, age, and test accommodations). This is unlikely to be a significant limitation given that scores for the other 49 states based on this limited set of control variables are highly correlated with the metrics based on the full set of controls (coefficients range from 0.95 to 0.97 depending on subject or grade).

Methods

All analyses are run separately by grade and subject and are weighted to be nationally representative (using weight variable ORIGWT).

Measuring Variation across States

I measure variation across states using a fixed-effects analysis. Specifically, I regress the test score of each student (standardized based on the weighted national distribution in 2013) on the set of control variables described above and a set of state dummies. All control variables are included in the regression using dummy variables identifying each of the groups of students for each construct, with the exception of the one arbitrarily chosen group that is the omitted category (except for age, which is included as a continuous variable). I measure variation across states as the standard deviation of the coefficients on the state dummies.

Measuring State Performance Adjusted for Student Characteristics

For each state, I regress student test scores in all other states on the control variables described above. I then calculate predicted scores for each student in that state, based on the estimated relationship in the other states, and the difference between the predicted scores and the actual scores of students in that state. I aggregate these estimated residuals to the state level as our measure of state NAEP performance adjusted for student characteristics.

In practice, this measure of state performance is almost perfectly correlated with a measure that runs a single regression for all states (i.e., does not leave out each state from the prediction for its students) and with the fixed-effects estimates used in the previous analysis ($r > 0.99$ in both cases for all four subject-grade combinations). Leaving a state's students out of the prediction regression only makes a noticeable (but still small) difference for large states, such as California and Texas.

As discussed above, the student-level regression approach is significantly different from a regression approach based on state-level aggregate data. The most important difference is the counterfactual. In the student-level analysis, the counterfactual against which each student is measured is the performance of similar students in other states. In a state-level analysis, each state would be measured against the average performance of states with similar average student characteristics.

Using the same set of control variables, the correlation between state-level measures based on student- and state-level regressions range from 0.28 in fourth-grade reading to 0.43 in eighth-grade math. This low correlation likely results partly from the number of control variables included in the analysis (32), which is small relative to the number of students (more than 160,000) but large relative to the number of states (50). The correlations are noticeably higher when using a more parsimonious model (with only 14 control variables), ranging from 0.51 to 0.60.

This implies that the state-level regression approach is more sensitive to the inclusion of control variables, with correlations between the results of the full and parsimonious state-level models in the 0.43–0.68 range, as compared with 0.95–0.97 for the student-level models. The ability of the student-level model to accommodate a greater number of control variables is a clear advantage.

Measuring Changes in State Performance over Time Adjusted for Student Characteristics

Separately for each state, I regress student test scores (standardized based on the weighted national distribution in 2003) on a more limited set of control variables that are available in both the 2003 and 2013 data. These include gender, race and ethnicity, eligibility for free and reduced-price lunch, LEP, special education, age, and test accommodation. All are coded as above except for test accommodation, which in this analysis is a single binary variable (received accommodation or did not, with no distinction by type of accommodation).

For students in each state, I use the estimated relationship between the test scores and predictors in their state in 2003 to calculate a predicted score for the students in the same state in 2013. I calculate each state's adjusted change in performance over time as the difference between its average actual score in 2013 and its average predicted score in 2013. The predicted score in 2013 is what the state's score would have been had the relationship between student characteristics and test scores in that state in 2003 remained the same in 2013.

Notes

1. Jaclyn Zubrzycki, 2013 “N.J. Moves to Take Over another District.” *Education Week*, May 31, 2013. <http://www.edweek.org/ew/articles/2013/06/05/33nj.h32.html>.
2. Tennessee State Government, “Tennessee Students the Fastest Improving in the Nation,” news release, November 7, 2013, <https://news.tn.gov/node/11644>.
3. Amy Sherman, “Jeb Bush says Florida Hispanic students perform ‘the best’ of any Hispanic group in the US.” *PolitiFact*, February 11, 2015. <http://www.politifact.com/florida/statements/2015/feb/11/jeb-bush/jeb-bush-says-florida-hispanic-students-perform-be/>.
4. American Legislative Exchange Council, “New Report Reveals Ties between School Choice Policies and Student Achievement,” news release, April 1, 2013. <http://www.alec.org/new-report-reveals-ties-between-school-choice-policies-and-student-achievement/>.
5. DC Office of the State of Superintendent of Education, “Mayor Gray and U.S. Secretary of Education Duncan Hail Continued Improvement in District’s Test Scores,” news release, November 7, 2013. <http://osse.dc.gov/release/mayor-gray-and-us-secretary-education-duncan-hail-continued-improvement-district%E2%80%99s-test>.
6. Pedro Noguera and Randi Weingarten, “Beyond Silver Bullets for American Education,” *The Nation*, December 22, 2010, <http://www.thenation.com/article/beyond-silver-bullets-american-education/>.
7. See also Matthew Di Carlo, “When You Hear Claims That Policies Are Working, Read the Fine Print,” (blog), Albert Shanker Institute, November 19, 2012, <http://www.shankerinstitute.org/blog/when-you-hear-claims-policies-are-working-read-fine-print>.
8. In practice, the results are not sensitive to the inclusion of private school students. For example, the standard deviations in table 1 are similar if private school students are excluded, and the correlation between state-fixed-effects estimates with and without private school students are in the 0.97–0.99 range.
9. Standard deviations are converted to months of schooling by assuming a 10-month school year and using the average annual learning gains for fourth and eighth grades reported by Hill and colleagues 2008. Specifically, I use yearly score gains of 0.40 in fourth-grade reading, 0.56 in fourth-grade math, 0.24 in eighth-grade reading, and 0.22 in eighth-grade math.
10. Loveless reports that the within-state standard deviation in NAEP scores was 4–5 times that between-state standard deviation in 2011. The adjustments for student demographics made in my analysis do not alter that basic conclusion, as the adjustments reduce both the within- and between-state standard deviations. For example, in fourth-grade math, the demographic adjustment changes the within/between ratio from $0.97/0.18 = 5.4$ to $0.74/0.11 = 6.7$.
11. In other words, I ensure that students are not compared with students in their own state. In practice, this makes almost no difference for the results for most states and only a small difference in a handful of large states. See the appendix B for details.
12. Results in student-level standard deviations for each of the four tests are published in a separate data file on Urban’s publication page (appendix C). The same-subject and same-grade correlations of the demographically adjusted measures are in the 0.7–0.8 range (the cross-subject/grade correlations are lower).
13. This phenomenon is sometimes called Simpson’s paradox.
14. These statistics are averaged across the NAEP math and reading tests.
15. In addition, the reliability (i.e., stability) of changes/differences in scores is almost always lower than the reliability of raw scores.
16. Dropping free and reduced-price lunch eligibility from the demographic adjustment in the change analysis reduces the predicted drop in scores from 0.10 to 0.05 standard deviations, but the correlation between state-level adjusted changes with and without this control variable is high ($r = 0.98$).

17. Detailed information on the NAEP, including on its history and design, see “NAEP: Measuring Student Progress since 1964,” National Center for Education Statistics, last modified November 27, 2012, accessed October 12, 2015, <http://nces.ed.gov/nationsreportcard/about/naephistory.aspx>.
18. Copies of the student questionnaires are available at <http://nces.ed.gov/nationsreportcard/bgquest.asp>.

References

- Bishaw, Alemayehu, and Kayla Fontenot. 2014. "Poverty: 2012 and 2013." Washington, DC: US Census Bureau. <https://www.census.gov/content/dam/Census/library/publications/2014/acs/acsbr13-01.pdf>.
- Chingos, Matthew M., Grover J. Whitehurst, and Michael R. Gallaher. 2015. "School Districts and Student Achievement." *Education Finance and Policy* 10 (3): 378–98.
- Dee, Thomas, and Brian Jacob. 2011. "The Impact of No Child Left Behind on Student Achievement." *Journal of Policy Analysis and Management* 30 (3): 418–46.
- Fitzpatrick, Maria Donovan. 2008. "Starting School at Four: The Effect of Universal Pre-Kindergarten on Children's Academic Achievement." SIEPR Discussion Paper 08-05. CA: Stanford Institute for Economic Policy Research. <http://www-siepr.stanford.edu/Papers/pdf/08-05.pdf>.
- Hill, Carolyn J., Howard S. Bloom, Alison R. Black, and Mark W. Lipsey. 2008. "Empirical Benchmarks for Interpreting Effect Sizes in Research." *Child Development Perspectives* 2 (3): 172–77.
- Hoxby, Caroline M. 2001. "All School Finance Equalizations Are Not Created Equal." *The Quarterly Journal of Economics* 116 (4): 1189–231.
- Loveless, Tom. 2011. "Who's Winning the Real Race to the Top?" In vol. 2, no. 5 of *The 2010 Brown Center Report on American Education: How Well Are American Students Learning?*, 13–19. http://www.brookings.edu/~media/research/files/reports/2011/2/07-education-loveless/0207_education_loveless.pdf.
- Loveless, Tom. 2013. "Be Wary of Ranking NAEP Gains," *Brown Center Chalkboard* no. 45. Washington, DC: Brookings Institution. <http://www.brookings.edu/research/papers/2013/11/13-interpreting-naep-gains-loveless>.
- Whitehurst, Grover J., Matthew M. Chingos, and Michael R. Gallaher. 2013. *Do School Districts Matter?* Washington, DC: Brookings Institution. http://www.brookings.edu/~media/research/files/papers/2013/3/27%20school%20district%20impacts%20whitehurst/districts_report_03252013_web.

About the Author



Matthew M. Chingos is a senior fellow at the Urban Institute, where he studies education-related topics at both the K-12 and postsecondary levels. Chingos's areas of expertise include class-size reduction, standardized testing, teacher quality, student loan debt, and college graduation rates. Chingos received a BA in government and economics and a PhD in government from Harvard University.

STATEMENT OF INDEPENDENCE

The Urban Institute strives to meet the highest standards of integrity and quality in its research and analyses and in the evidence-based policy recommendations offered by its researchers and experts. We believe that operating consistent with the values of independence, rigor, and transparency is essential to maintaining those standards. As an organization, the Urban Institute does not take positions on issues, but it does empower and support its experts in sharing their own evidence-based views and policy recommendations that have been shaped by scholarship. Funders do not determine our research findings or the insights and recommendations of our experts. Urban scholars and experts are expected to be objective and follow the evidence wherever it may lead.



2100 M Street NW
Washington, DC 20037

www.urban.org