CALDER Policymakers Council

Opinion Brief

HOW MUCH SHOULD WE RELY ON STUDENT TEST ACHIEVEMENT AS A MEASURE OF SUCCESS?

Dan Goldhaber & Umut Özek
American Institutes for Research/CALDER

Suggested citation:
Goldhaber, D., & Özek, U. (2018). *How much should we rely on student test achievement as a measure of success?* (CALDER Policy Brief 12-1118-1). Washington, DC: National Center for Analysis of Longitudinal Data in Education Research.

The crafting and dissemination of this opinion brief was supported by the National Center for the Analysis of Longitudinal Data in Education Research (CALDER), which is funded by a consortium of foundations. For more information about CALDER funders, see www.caldercenter.org/about-calder. We appreciate thoughtful comments by Jack Buckley, Mike Petrilli, Collin Hitt, and Jay Greene on an earlier draft of this brief.

WARNING: This brief contains the authors' unmoderated opinions about controversial issues, which may cause dizziness, nausea, and/or seizures. Note that the views expressed are those of the authors and do not necessarily reflect those of the funders or the institutions with which the authors are affiliated. Also note that Goldhaber and Özek are employed by the American Institutes for Research (AIR), a division of which develops state assessments (see https://www.air.org/page/air-assessment). While this may appear to be a conflict of interest, Goldhaber and Özek swear AIR's assessment work does not factor into their thinking on the topic of this opinion brief.

Questioning Standardized Tests as a Measure of Success

The use of standardized tests as a measure of student success and progress in school goes back decades. And the 2001 passage of the No Child Left Behind Act (NCLB) established the broader use of test scores as a measure of school quality, used for accountability, nationwide.[1] The 2009 Race to the Top (RttT) federal grant program promoted teacher evaluation reforms that also included the use of standardized tests as a component of a teacher's evaluation (Goldhaber, 2015).[2]

But there has been pushback against the use of tests. Academics and advocates, prominently including the teachers' unions (Taylor & Rich, 2015), have raised various concerns about the consequences of reliance (or overreliance) on test scores for school and teacher accountability purposes.[3] And while there is certainly academic and policy disagreement about the efficacy of using test scores for accountability purposes,[4] there is no doubt that policymakers are scaling back use of tests. The 2015 passage of the Every Student Succeeds Act (ESSA), for instance, continues NCLB's requirement that students be tested annually in grades 3 through 8, but eliminates much of the federal role in enforcing test-based accountability.[5]

More recently, however, policy scholars have begun to question whether test scores are a metric that we should really care about, pointing out that test score gains are not always associated with changes in other schooling outcomes (for example, see Greene, 2016). But this critique of tests as a measure has gone beyond a narrow academic policy debate. As recent headlines in *Forbes* show—for instance, "How Much Do Rising Test Scores Tell Us About a School?" (Hess, 2018) and "Is the Big Standardized Test a Big Standardized Flop?" (Greene, 2018)[6]—this debate about tests as a measure of success is now reaching a much broader audience.

A recent report by Collin Hitt, Michael McShane, and Patrick Wolf (the "HMW report"; Hitt, McShane, & Wolf, 2018) is the latest warning from a number of prominent education policy

[1] Note that while NCLB included requirements for the use of standardized tests, a number of states had been using tests for accountability purposes since the late 1980s/early 1990s (Coley & Goertz, 1990).
[2] RttT also incentivized states to reform principal evaluation systems. For more details, see Dragoset et al. (2016).
[3] For instance, see policy positions taken by the teachers' unions (Taylor & Rich, 2015); issues raised about the psychometric properties of measures derived from standardized tests—for example, in the American Educational Research Association's research note (AERA, 2015), the American Statistical Association's statement (ASA, 2014), and Darling-Hammond, Amrein-Beardsley, Haertel, and Rothstein (2012); and other general concerns about how overreliance on tests may corrupt the educational process and be harmful to student learning (Koretz, 2017).
[4] On the use of test scores for various accountability purposes, see, for instance, Dee and Jacob (2010); Koedel, Leatherman, and Parsons (2012); and Claro and Loeb (2017). Note also that critics of using standardized tests often suggest alternative means of judging school and teacher performance that have some of the same statistical limitations (e.g., measurement error) that are cited in criticizing test-based measures of performance (Goldhaber, 2015).
[5] For example, under ESSA, each state is allowed to set its own goals for student achievement within the federal framework; there was far less flexibility for states to set their own goals under NCLB. Similarly, NCLB focused mainly on student academic achievement and used reading and math scores to assess schools. In contrast, states must consider more than just test scores under ESSA, including high school graduation rates, kindergarten readiness, college readiness, chronic absenteeism, etc.
[6] To be clear, the author of this article (Peter Greene) is different than Jay Greene, a professor of education reform at the University of Arkansas. Who knew that there are two Greenes in the education policy world who feel the same way about testing!

researchers about using test scores as a measure of student success.[7] The report focuses on a number of studies that examine the effects of different school choice programs on both student test scores and long-term outcomes (such as high school graduation and college enrollment), and examines how well test score impacts in these studies align with the attainment impacts. The authors find very little correlation between the two and conclude that "test scores should be put in context and should not automatically occupy a privileged place over parental demand and satisfaction as short-term measures of school choice success or failure" (p.20).

The measured statement that test scores should not automatically occupy a privileged place as a measure of success is at odds with the much stronger claim by (Greene, 2018) that "the test scores do not tell you what they claim they tell you. They are less like actionable data and more like really expensive noise". It is important to recognize that a lack of alignment in the choice literature (or in other education research findings) between test scores and later life outcomes is not in itself evidence that tests do not matter - educational interventions could improve later life outcomes without affecting the test scores of students. Choice schools may, for instance, have stronger pipelines into college, leading to better college-going results while not affecting test results. Certainly, a finding like this shouldn't be ignored, but we also should not conclude that tests do not matter.

There are other reasons to question the specific conclusions of the HMW report. Many of these are spelled out in a series of articles by Michael Petrilli, in which he critiques: (1) how the authors identified the studies on school choice programs (Petrilli, 2018a); (2) how the authors tested the alignment of test score and attainment impacts in these studies (Petrilli, 2018a); and (3) how the authors extrapolated their findings on school choice programs and applied them to schools (Petrilli, 2018b). And in a separate EdNext piece, Patrick Wolf responds to these criticisms (Wolf, 2018). Our objective in this opinion brief is not to delve into the specifics of the HMW report (although we revisit it to point out a few additional concerns). Instead, we take a step back and try to assess what the broader literature suggests about using test scores as a metric for student success.[8]

Test Scores and Later Life Outcomes

One might reasonably argue that test scores are only an intermediate measure of what we really care about: the extent to which students are gaining knowledge in school that enhances their later life prospects.[9] Thus, it is certainly reasonable to question whether this intermediate measure does a good job of capturing what we really care about: the underlying learning that really is important for later life success.

[7] For other examples, see Greene (2016) and Gill (2016).

[8] Keep in mind that the HMW report discusses the use of test scores in a specific context, namely the evaluation of school choice programs. In contrast, we are focusing on the general use of test scores as a measure of student success and school performance.

[9] It is also reasonable to argue that education is an inherently important outcome itself, though it is important to recognize that students clearly learn skills in school that are unlikely to be accurately measured by tests alone, whether these skills are valued for their own sake (e.g., respect for differences of opinion) or because they also translate into better future outcomes (Claro & Loeb, 2017).

There is a vast literature linking test scores and later life outcomes, such as educational attainment, health, and earnings. Hanushek (2009) provides an excellent review of the extant literature on the relationship between cognitive skills, as proxied by test scores and individual incomes in developed and developing countries, and concludes that there is considerable evidence that test scores are directly related to later life outcomes. For example, in the US context, students who score one standard deviation higher on math tests at the end of high school have been shown to earn 12% more annually, or $3,600 for each year of work life in 2001 (Mulligan, 1999; Murnane, Willett, Duhaldeborde, & Tyler, 2000; Lazear, 2003). Similarly, Heckman, Stixrud, and Urzua (2006) find that test scores are significantly correlated not only with educational attainment and labor market outcomes (employment, work experience, choice of occupation), but also with risky behavior (teenage pregnancy, smoking, participation in illegal activities).

The idea that students who learn more in school, and hence perform better on tests, have better later outcomes because of that learning has a good deal of face validity. It is no great leap to imagine that students who score in the 90th percentile on a science test are more likely to be successful scientists than those who score in the 10th percentile, and that this score reflects a better understanding of science. But we might be less sure that smaller differences in student test achievement are meaningful; and as Hanushek (2009) notes, these observed *correlations do not necessarily reflect causal effects of schools on later life outcomes*.

Maybe students who do well on tests are the same students who wake up in the morning, go to work on time, and work hard. Test achievement is also likely to largely reflect learning opportunities outside of school—the supportiveness of families or the communities in which students live. This is why scholars doubt that static measures of test performance alone are reflective of contributions that schools or teachers make toward student learning[10]—a popular critique of those who doubt the use of tests for school accountability purposes (Tienken, 2017).

Both of the above arguments highlight why it is difficult to attribute the observed associations between test scores and later life outcomes to the causal effect of schools. The key question is whether interventions that boost students' test scores are also likely to lead to better future outcomes for students. Certainly, a lack of an underlying causal link between test scores and long-term outcomes should lead us to consider downplaying (or perhaps eliminating) tests as a measure of student success. Unfortunately, definitively establishing such a causal link is challenging, given that it would be unethical to design an experiment where we randomly provide better education to some students, measure their test scores, and assess whether improvements in test scores lead to better life outcomes. Therefore, what we know about the causality of this relationship comes from a limited number of studies that examine the causal effects of different educational inputs (e.g., schools, teachers, classroom peers) on both student test scores and later life outcomes. If a study finds test score impacts and adult outcome impacts that are not in the same direction, this might be regarded as evidence that test scores do not affect the later life outcomes we care about.

[10] Figlio and Loeb (2011) present an excellent discussion about the pros and cons of using test score levels in school accountability systems. Out-of-school learning is also why some scholars have urged policymakers to link performance measures to student test growth instead. See, for example, Morgan Polikoff's letter to the California State Board of Education (Polikoff, 2018).

So, what does the literature say about whether there is a causal link? While there are certainly studies that find test-score and long-term-outcome effects that are not in the same direction (as cited in the HMW report), our reading of the broader literature in this context seems to indicate that they are outnumbered by the studies finding evidence of a strong causal link between test scores and later life outcomes. Perhaps the most influential study of all in this strand was conducted by Chetty, Friedman, and Rockoff (2014). Examining the long-term effects of teacher quality (assessed based on their effect on student test scores), the authors find that students who are assigned to highly effective teachers in elementary school are more likely to attend college and earn higher salaries.[11]

Another study by Raj Chetty and co-authors (Chetty et al., 2011) examines the long-term effects of peer quality in kindergarten (once again proxied by test scores) using the Tennessee Student Teacher Achievement Ratio (STAR) experiment, and finds that students who are assigned to classrooms with higher quality peers have higher college attendance rates and adult earnings. Similarly, using the Tennessee STAR experiment, a recent study by Susan Dynarski and colleagues (Dynarski, Hyman, and Schanzenbach, 2013) looks at the effects of smaller classes in primary school and find that the test score effects at the time of the experiment are an excellent predictor of long-term improvements in postsecondary outcomes. Lafortune, Rothstein, and Schanzenbach (2018) and Jackson, Johnson, and Persico (2016) investigate the effects of school finance reform on test scores, educational attainment, and earnings, and find significant benefits of an increase in school spending on both test scores and adult outcomes.

Finally, there are a number of studies in the school choice context (cited in the HMW report) that show certain school choice programs having positive effects on both test scores and later life outcomes. For example, Angrist and colleagues (2016) examine the effects of Boston's charter high schools and conclude that charter effects on college-related outcomes are strongly correlated with gains on earlier tests. Dobbie and Fryer (2015) find that attending a high-performing charter school not only increases test scores, but also significantly reduces the likelihood of engaging in risky behavior.

These two studies also highlight an important concern regarding the HMW report, which presents these studies as evidence of misalignment between test scores and long-term effects. While both studies find statistically insignificant, albeit positive, attainment effects (presented as evidence in the HMW report), they both find significant effects on other long-term outcomes that we care about (a shift from 2-year to 4-year college enrollment in the former study, and a significant effect on risky behavior in the latter).

Overall, all of these studies suggest that interventions that move the needle on test scores also improve later life outcomes. Thus, we contend that the weight of empirical evidence lends

[11] Chetty et al. (2014a) and Chetty et al. (2014b) conduct several robustness checks to provide evidence that their value-added measure is not picking up the effect of confounding factors on student test scores beyond teacher effectiveness (e.g., parental resources). For example, in Chetty et al. (2014a), they exploit the plausibly exogenous variation in teacher quality across subsequent student cohorts in a school driven by teacher mobility.

support to the argument for using test scores as a measure of success in education systems.[12] This does not mean that test score effects of educational interventions will always align with their effects on adult outcomes.[13] It is easy to make the case that interventions can and do improve later life outcomes without affecting the cognitive skills of children. In short, test scores will not encompass the full impact of schools and teachers on students, and hencewe should not expect them to fully capture all the contributions that schools and teachers make toward influencing long-term student outcomes.[14]

But we need to think carefully about what that might mean for education policy and practice.[15] From a practical perspective, we can't wait many years to get long-term measures of what schools are contributing to students. This does not mean that test scores ought to be the exclusive or even primary short-term measures, but if one believes in school accountability and that test scores ought to be down-weighted, it is important to consider what alternative measures of success are out there and how reliable they are. For instance, there are concerns that non-test outcomes, such as attendance, grades, suspensions, and high school graduation rates, are arguably more "gameable" than test scores. And we certainly know less empirically about the causal connections between these types of outcomes and long-term student success.

Where one lands on the use of test scores to measure student or schooling success is clearly a matter of subjective judgement. But we strongly argue that debate about this should be framed by the right interpretation of the empirical evidence, most of which does suggest that test scores are a good intermediate measure of student success.

[12] Importantly, even if tests are a good measure, that does not imply that using the tests for school or teacher accountability purposes will lead to better schooling outcomes. See, for instance, Koretz (2017).
[13] For example, Duncan and Magnuson (2013) examine the extant literature on the effects of preschool programs and conclude that while early childhood programs appear to boost cognitive ability and early school achievement in the short run, these cognitive impacts fade out within a few years. That said, long-run follow-ups of some of these programs show lasting positive effects on educational attainment and adult earnings.
[14] Indeed, a recent study by C. Kirabo Jackson (2018) finds evidence suggesting that teachers affect later life outcomes not only through their effect on test scores but also through their effect on non-test outcomes such as absences, suspensions, and grade progression.
[15] Indeed, we need to be thinking about this now as the ESSA, which replaced NCLB, encourages states to rely more on non-test outcomes (Education Commission of the States, 2018). The trick is finding reliable non-test measures to use.

References

American Educational Research Association (AERA). (2015). *AERA statement on use of value-added models (VAM) for the evaluation of educators and educator preparation programs.* Retrieved from https://www.aera.net/Newsroom/News-Releases-and-Statements/AERA-Issues-Statement-on-the-Use-of-Value-Added-Models-in-Evaluation-of-Educators-and-Educator-Preparation-Programs

American Statistical Association (ASA). (2014). *ASA statement on using value-added models for education assessment*. Retrieved from http://www.amstat.org/asa/files/pdfs/POL-ASAVAM-Statement.pdf

Angrist, J. D., Cohodes, S. R., Dynarski, S. M., Pathak, P. A., & Walters, C. R. (2016). Stand and deliver: Effects of Boston's charter high schools on college preparation, entry, and choice. *Journal of Labor Economics*, *34*(2), 275–318.

Chetty, R., Friedman, J. N., Hilger, N., Saez, E., Schanzenbach, D., & Yagan, D. (2011). How does your kindergarten classroom affect your earnings? Evidence from Project STAR. *Quarterly Journal of Economics*, *126*(4), 1593–1660.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014a). Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood. *The American Economic Review*, *104*(9), 2633–2679.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014b). Measuring the impacts of teachers I: Evaluating Bias in Teacher Value-Added Estimates. *The American Economic Review*, *104*(9), 2593-2632.

Claro, S., & Loeb, S. (2017). New evidence that students' beliefs about their brains drive learning. *Evidence Speaks Reports*, *2*(29).

Coley, R. J., & Goertz, M. E. (1990). *Educational standards in the 50 states*. Princeton, NJ: Educational Testing Service.

Darling-Hammond, L., Amrein-Beardsley A., Haertel E., & Rothstein J. (2012, February 29). Evaluating teacher evaluation. *Education Week*. Retrieved from http://www.edweek.org/ew/articles/2012/03/01/kappan_hammond.htm

Dee, T. S, & Jacob, B. A. (2010). The impact of No Child Left Behind on students, teachers, and schools. *Brookings Papers on Economic Activity*, No 2, 149–207.

Dobbie, W., & Fryer, R. (2015). The medium-term impacts of high-achieving charter schools. *Journal of Political Economy*, *123*(5), 985–1037.

Dragoset, L., Thomas, J., Herrmann, M., Deke, J., James-Burdumy, S., Graczewski, C., … Upton, R. (2016). *Race to the Top: Implementation and relationship to student outcomes: Executive summary* (NCEE 2017-4000). Washington, DC: National Center for Education

Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Dynarski, S., Hyman, J., & Schanzenbach, D. W. (2013). Experimental evidence on the effect of childhood investments on postsecondary attainment and degree completion. Journal of Policy Analysis and Management, 32(4), 692–717.

Education Commission of the States. (2018). *50-state comparison: States' school accountability systems.* Retrieved from https://www.ecs.org/50-state-comparison-states-school-accountability-systems/

Figlio, D., & Loeb, S. (2011). School accountability. In E. A. Hanushek, S. Machin, & L. Woessmann (Eds.), *Handbooks in economics* (Vol. 3, pp. 383–421). Amsterdam, The Netherlands: North-Holland.

Gill, B. (2016, May 11). *Beyond test scores: Improving research evidence on education* [Blog post]. Retrieved from https://www.mathematica-mpr.com/commentary/beyond-test-scores-improving-research-evidence-on-education

Goldhaber, D. (2015). Exploring the potential of value-added performance measures to affect the quality of the teacher workforce. *Educational Researcher*, *44*(2), 87–95.

Greene, J. P. (2016, November 5). Evidence for the disconnect between changing test scores and changing later life outcomes [Blog post]. Retrieved from https://jaypgreene.com/2016/11/05/evidence-for-the-disconnect-between-changing-test-scores-and-changing-later-life-outcomes/

Greene, P. (2018, September 20). *Is the big standardized test a big standardized flop?* Retrieved from https://www.forbes.com/sites/petergreene/2018/09/20/is-the-big-standardized-test-a-big-standardized-flop/#7e62f3024937

Hanushek, E. (2009). The economic value of education and cognitive skills. In G. Sykes, T. Ford, D. Plank, & B. Schneider (Eds.), *Handbook of education policy research*. New York, NY: Routledge.

Heckman, J., Stixrud, J., & Urzua, S. (2006). The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior. *Journal of Labor Economics*, *24*(3), 411–482.

Hess, F. (2018, September 18). *How much do rising test scores tell us about a school?* Retrieved from https://www.forbes.com/sites/frederickhess/2018/09/18/how-much-do-rising-test-scores-tell-us-about-a-school/#1a461d2922e8

Hitt, C., McShane, M. Q., & Wolf, P. (2018). *Do impacts on test scores even matter? Lessons from long-run outcomes in school choice research*. Washington, DC: American Enterprise Institute. Retrieved from http://www.aei.org/wp-content/uploads/2018/04/Do-Impacts-on-Test-Scores-Even-Matter.pdf

Jackson, C. K. (2018). What do test scores miss? The importance of teacher effects on non–test score outcomes. *Journal of Political Economy*, *126*(5), 2072–2107.

Jackson, C. K., Johnson, R. C., & Persico, C. (2016). The effects of school spending on educational and economic outcomes: Evidence from school finance reforms. *Quarterly Journal of Economics*, *131*(1), 157–218.

Koedel, C., Leatherman, R., & Parsons, E. (2012). Test measurement error and inference from value-added models. *The B.E. Journal of Economic Analysis & Policy*, *12*(1), 1–37.

Koretz, D. (2017). *The testing charade: Pretending to make schools better*. Chicago, IL: University of Chicago Press.

Lafortune, J., Rothstein, J., & Schanzenbach, D. W. (2018). School finance reform and the distribution of student achievement. *American Economic Journal: Applied Economics*, *10*(2), 1–26.

Lazear, E. P. (2003). Teacher incentives. *Swedish Economic Policy Review*, *10*, 179–214.

Mulligan, C. B. (1999). Galton versus the human capital approach to inheritance. *Journal of Political Economy*, *107*, S184–S224.

Murnane, R. J., Willett, J. B., Duhaldeborde, Y., & Tyler, J. H. (2000). How important are the cognitive skills of teenagers in predicting subsequent earnings? *Journal of Policy Analysis and Management*, *19*, 547–568.

Petrilli, M. J. (2018a, April 20). What counts as school choice in new study of short- and long-term outcomes? *Education Next.* Retrieved from https://www.educationnext.org/counts-school-choice-new-study-short-long-term-outcomes/

Petrilli, M. J. (2018b, April 24). Are there schools of choice that hurt test scores but not long-term outcomes? *Education Next.* Retrieved from https://www.educationnext.org/schools-choice-hurt-test-scores-not-long-term-outcomes/

Polikoff, M. (2018). *Letter to the CA State Board of Education.* Retrieved from https://morganpolikoff.com/2018/07/04/letter-to-the-ca-state-board-of-education/

Taylor, K., & Rich, M. (2015, April 20). Teachers' unions fight standardized testing, and find diverse allies. *New York Times.* Retrieved from https://www.nytimes.com/2015/04/21/education/teachers-unions-reasserting-themselves-with-push-against-standardized-testing.html

Tienken, C. (2017). *Students' test scores tell us more about the community they live in than what they know*. Retrieved from http://theconversation.com/students-test-scores-tell-us-more-about-the-community-they-live-in-than-what-they-know-77934

Wolf, P. J. (2018*). A Flawed Critique of Our School Choice Achievement-Attainment Divide Study*. Retrieved from https://www.educationnext.org/flawed-critique-school-choice-achievement-attainment-divide-study/