How Much Should We Rely on Student Test Achievement as a Measure of Success?

Dan Goldhaber[1] and Umut Özek[1]

The use of test scores as a performance measure in high-stakes educational accountability has become increasingly popular since the enactment of the No Child Left Behind Act of 2001 (NCLB), which imposed sanctions such as the threat of losing federal funds unless a state implemented a school accountability system that measures student progress continuously. Since then, many in the education community have questioned whether differences in student test scores reflect actual discrepancies in the long-term well-being of individuals. In this review, we try to address this question in the light of the extant literature that examines the relationship between test scores and later life outcomes. We show that while there are certainly studies that contradict the causality of this relationship, there is also abundant evidence suggesting a causal link between test scores and later life outcomes. We conclude that any debate about the use of test scores in educational accountability (1) should be framed by use of all relevant empirical evidence, (2) should also consider the predictive validity of nontest measures of student success, and (3) should keep in mind that the predictive validity of test scores could be stronger in some contexts than others.

Keywords: accountability; achievement; educational policy; high-stakes testing; meta-analysis; school/teacher effectiveness

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 but became more widespread after the No Child Left Behind Act of 2001 (NCLB), which mandated the use of test scores as a measure of school quality in state accountability systems (Vinovskis, 2008).[1] The 2009 Race to the Top (RttT) federal grant program further expanded the use of standardized tests in educational accountability by promoting teacher evaluation reforms that included test scores as a component of a teacher's evaluation (Goldhaber, 2015).[2]

But there has been pushback against the use of tests in educational accountability. 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 both school and teacher accountability purposes. One strand of criticism in this context relates to the psychometric properties of measures derived from standardized tests: Recent examples of this strand include the statements by the American Statistical Association (ASA) and the American Educational Research Association (AERA) cautioning against the use of value-added scores as the main factor in high-stakes decisions regarding educators (AERA, 2015; ASA, 2014). Another concern is that overreliance on test scores could corrupt the educational process and be harmful to student learning (Koretz, 2017).

While there is academic and policy disagreement about the efficacy of using test scores for accountability purposes,[3] there is no doubt that policymakers are placing less weight on student test scores in high-stakes decisions. As a chief example, the 2015 passage of the Every Student Succeeds Act (ESSA) continued NCLB's requirement that students be tested annually in Grades 3 through 8 but introduced a number of new measures to be used in school accountability systems, lessening the role that tests play overall.[4]

More recently, policy scholars have even begun to question whether we should use test scores as a measure of success at all—a

[1]American Institutes for Research, Washington, DC

Educational Researcher, Vol. 48 No. 7, pp. 479–483
DOI: 10.3102/0013189X19874061

question that is gaining broad public attention.[5] Much of this argument is based on a claim that test score gains are not always associated with changes in other schooling outcomes. One recent example of this argument is a recent report by Hitt, McShane, and Wolf (2018) that examines the use of test scores to evaluate school choice programs. In particular, 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 suggest there is 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).[6]

One might 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.[7] In other words, it is reasonable to argue that we should hold schools/teachers accountable for the test performance of their students but only if test scores at least partially reflect their causal impact on the underlying learning that is important for later life success. In this review, we examine the evidence on the relationship between test scores and later life outcomes using Hitt et al. (2018) as an illustrative example, especially the extent to which we can infer causality in the relationship, and then discuss what this might imply for the use of test scores as a component in educational accountability systems.

Test Scores and Later Life Outcomes

There is a vast literature linking test scores and later life outcomes, such as educational attainment, health, and earnings. Hanushek (2009) provides a 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.[8] 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, and choice of occupation) but also with risky behavior (teenage pregnancy, smoking, and participation in illegal activities). However, as Hanushek (2009) notes, *these observed correlations do not necessarily reflect causal effects of schools on later life outcomes*.

For example, the observed differences in later life outcomes between students with higher and lower test scores could be driven by differences in unobservable attributes of students such as their levels of grit. Test achievement is also likely to be significantly influenced by learning opportunities outside of school—the supportiveness of families or the communities in which students live. This is an important reason why some scholars doubt that static measures of test performance alone are reflective of contributions schools or teachers make toward student learning (Tienken, 2017).[9]

Establishing causal links between test scores and adult outcomes is challenging; 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. Thus, 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. This is also the approach utilized by Hitt et al. (2018) in the context of school choice program evaluations.

So what does the broader literature (beyond school choice) say about whether there is a causal link? While there are studies that find test-score and long-term outcome effects that are not in the same direction (such as the ones cited in cited in Hitt et al., 2018), there is also abundant evidence suggesting a causal link between test scores and later life outcomes. Perhaps the most influential study connecting schooling, test scores, and later life outcomes was conducted by Chetty, Friedman, and Rockoff (2014a). 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.[10]

Another study by Raj Chetty et al. (2011) examines the long-term effects of peer quality in kindergarten 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, & Schanzenbach, 2013) looks at the effects of smaller classes in primary school and finds 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 that show certain school choice programs having positive effects on both test scores and later life outcomes. For example, Angrist, Cohodes, Dynarski, Pathak, and Walters (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.

Accountability Without Test Scores

Overall, all of these studies suggest that interventions that move the needle on test scores also improve later life outcomes, which lend support to the argument for using test scores as a measure of success in education systems.[11] This does not mean that test

score effects of educational interventions will always align with their effects on adult outcomes.[12] It is easy to make the case that interventions can and do improve later life outcomes without affecting the cognitive skills of children. For example, choice schools may have stronger pipelines into college, leading to better college-going results while not affecting test results. In short, test scores will not encompass the full impact of schools and teachers on students, and hence we should not expect them to fully capture all the contributions that schools and teachers make toward influencing long-term student outcomes.[13]

But we need to think carefully about what abandoning the use of test scores altogether might mean for education policy and practice.[14] From a practical perspective, we cannot 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 educational 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.

ESSA, for instance, encourages states to rely more on nontest outcomes (e.g., high school graduation rates, kindergarten readiness, college readiness, and chronic absenteeism) to assess school performance. But there are increasing concerns that these measures are "gameable." For example, as of this writing, the District of Columbia Public Schools (DCPS) has been under investigation by the U.S. Department of Education and the FBI for awarding high school diplomas to hundreds of students who failed to meet the high school graduation requirements.[15] Similarly, while some studies find that high school GPA is a better predictor of college success than standardized test scores (e.g., Geiser & Santelices, 2007), there is recent evidence suggesting grade inflation in high schools, especially in wealthier settings, which casts doubt on the reliability of high school GPA in school accountability (Gershenson, 2018).

Perhaps more importantly, we know less empirically about the causal connections between some of these new ESSA measures or other means of school or teacher accountability and long-term student prospects. Are students assigned to teachers who get good classroom observation ratings likely to have better future prospects? Perhaps, but there is less evidence about this type of measure than there is about test-based measures.

In the end, where one lands on the use of test scores to measure student or school success is a matter of subjective judgment. But the debate about this (1) should be framed by use of all relevant empirical evidence, (2) should also consider the predictive validity of nontest measures of student success, and (3) should keep in mind that the predictive validity of test scores could be stronger in some contexts than others.

NOTES

We appreciate feedback from Collin Hitt, Patrick J. Wolf, and Jay P. Greene. Views expressed here are those of the authors and do not necessarily reflect those of the institutions to which the authors are affiliated. This research was also supported by the National Center for 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.

[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]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 such as measurement error (Goldhaber, 2015).

[4]See, for example, https://www.k12insight.com/trusted/standardized-testing-essa/, accessed on March 29, 2019.

[5]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)—this debate about tests as a measure of success is now reaching a much broader audience.

[6]There are reasons to question the specific conclusions of Hitt et al. (2018). 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, Wolf (2018) responds to these criticisms.

[7]It is also reasonable to argue that education is an inherently important outcome itself, though it is important to recognize that students 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).

[8]For example, in the U.S. 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 (Lazear, 2003; Mulligan, 1999; Murnane, Willett, Duhaldeborde, & Tyler, 2000).

[9]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 (2018) letter to the California State Board of Education.

[10]Chetty et al. (2014a) and Chetty, Friedman, and Rockoff (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. (2014b), they exploit the plausibly exogenous variation in teacher quality across subsequent student cohorts in a school driven by teacher mobility.

[11]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).

[12]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.

[13]Indeed, 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 nontest outcomes such as absences, suspensions, and grade progression.

[14]Indeed, we need to be thinking about this now as the ESSA, which replaced NCLB, encourages states to rely more on nontest outcomes (Woods, 2018). The trick is finding reliable nontest measures to use.

[15]See https://www.washingtonpost.com/local/dc-politics/dc-public-schools-were-once-a-success-story-are-they-now-an-embarrassment/2018/02/01/fb15dd4c-069d-11e8-b48c-b07fea957bd5_story.html?noredirect=on&utm_term=.9a89e3e6d5dc, accessed on March 18, 2019.

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). Retrieved from https://www.brookings.edu/wp-content/uploads/2017/11/claro-and-loeb-report.pdf.

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

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.

Duncan, G., & Magnuson, K. (2013). Investing in preschool programs. *Journal of Economic Perspectives*, *27*(2), 109–132. https://doi.org/10.1257/jep.27.2.109

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.

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.

Geiser, S., & Santelices, M. V. (2007). *Validity of high-school grades in predicting student success beyond the freshman year: High-school record vs. standardized tests as indicators of four-year college outcomes*. Retrieved from http://www.escholarship.org/uc/item/7306z0zf

Gershenson, S. (2018). *Grade inflation in high schools (2005-2016). Thomas B. Fordham Institute Report*. Retrieved from http://edex.s3-us-west-2.amazonaws.com/publication/pdfs/(2018.09.19)%20Grade%20Inflation%20in%20High%20Schools%20(2005-2016).pdf

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, 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* (pp. 39–56). 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. *The 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

Vinovskis, M. (2008). *From a Nation at Risk to No Child Left Behind: National education goals and the creation of federal education policy*. New York, NY: Teachers College Press.

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/

Woods, J. R. (2018). *50-state comparison: States' school accountability systems.* Retrieved from https://www.ecs.org/50-state-comparison-states-school-accountability-systems/.

AUTHORS

DAN GOLDHABER, PhD, is the director of the Center for Analysis of Longitudinal Data in Education Research (CALDER) at the American Institutes for Research, 1000 Thomas Jefferson Street, NW, Washington, DC 20007, and the director of the Center for Education Data & Research (CEDR) at the University of Washington; *dgoldhaber@air.org*. His research focuses on issues of educational productivity and reform at the K–12 level; the broad array of human capital policies that influence the composition, distribution, and quality of teachers in the workforce; and connections between students' K–12 experiences and postsecondary outcomes.

UMUT ÖZEK, PhD, is a principal researcher at the American Institutes for Research, 1000 Thomas Jefferson Street, NW, Washington, DC 20007; *uozek@air.org*. His research focuses on immigrant students, implementation and consequences of educational accountability, design and effects of school choice programs, and value-added measurement.

Manuscript received December 17, 2018
Revisions received April 2, 2019; June 25, 2019
Accepted July 30, 2019