A Reset for Assessment: Toward a Less Burdensome Accountability System

By Jack Buckley

October 2021

Key Points

- The COVID-19 pandemic provides an opportunity to replace current educational assessments with new tests that provide a net increase in the overall utility of the assessment system to stakeholders on the ground—including educators and families—while reducing stakeholders' burden.
- In the short term, federal policymakers should assist states with replacing high-burden, low-value summative assessments used to meet Every Student Succeeds Act accountability requirements with higher-value interim assessments administered throughout the school year.
- In the long term, the US Department of Education should use its resources to encourage development of a low-burden, high-value assessment system in which student performance data are collected as part of routine interactions with a digital learning platform.

Standardized educational assessments are often criticized as overly burdensome, competing for vital instructional time and narrowing the curriculum to what is tested. They are regarded as expensive, diverting scarce resources from students, teachers, and classrooms to shadowy testing companies. They are unfair, showing stubborn achievement gaps between rich and poor, Black and White, and suburban and rural students year after year.

Yet, as Lindsay Fryer's report in this series highlights, despite all these criticisms, federal policymakers on both sides of the aisle have continually returned to standardized assessment as a key component of school improvement, civil rights, and accountability.[1] The reason is simple: Despite these assessments' many real and imagined flaws, they remain the best way of measuring student achievement with reliability, fairness, and validity. Indeed, it could be said that standardized assessment is the worst way of understanding students' knowledge and capabilities—except for all the other ways that have been tried.

This is not to say that educational assessment cannot be improved. The COVID-19 pandemic has profoundly affected education and has exposed significant weaknesses in assessment that have been ignored or tolerated for too long. Faced with an enormous public health crisis and an unexpected shift to emergency remote learning, all 50 states, the District of Columbia, Puerto Rico, and the Bureau of Indian Education sought and were

granted assessment waivers from the US Department of Education (ED) for the 2019–20 school year.

At the same time, other educational assessment programs either faced chaos and cancellations, as the SAT and ACT did,[2] or launched untested and unvalidated alternative virtual models, as the Advanced Placement (AP) and International Baccalaureate programs did.[3] The fortunes of testing companies with interim or formative assessment products depended on whether they could offer remote testing.

There are many possible explanations for why educational assessment in the United States was caught flat-footed by the pandemic, including a fractured marketplace of testing providers, weak federal oversight, underinvestment in broadband and computer-based assessment technology, and insufficient incentives for research and innovation in alternative security and remote delivery models. Undoubtedly, many postmortem analyses will be written in the years to come. However, like all crises, this disruption to testing-as-usual provides an opportunity to move beyond finger-pointing and rethink the form and function of educational assessment.

As educators confront the challenge of remediating COVID-19 learning loss—and policymakers wrestle with measuring it amid widespread discontent with testing—now is an apt time to consider how much of the current accountability and assessment regime is truly necessary. In this report, I examine the state of testing post–March 2020 and explore how a minimally viable, less burdensome assessment system might look.

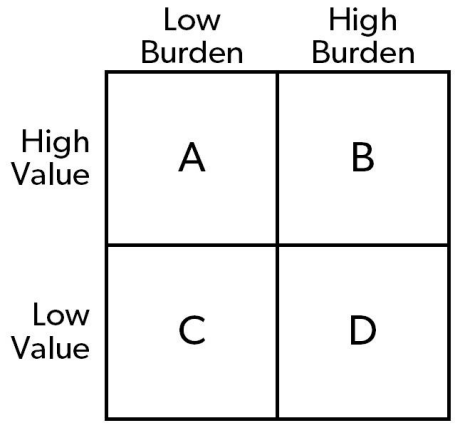
Classifying Assessments

Before examining how the future of educational assessment could look, it is useful first to clarify some concepts and introduce some terminology. Assessment programs are often described by their purpose or primary use. Summative assessments are intended to measure what students have learned or can do, in contrast to formative assessments, which are generally more diagnostic and integrated with classroom activities and are designed to assist educators in guiding instruction.

Between these lie interim assessments, such as NWEA's MAP Growth and Renaissance Learning's Star Assessments.[4] These are often similar in form and operation to summative tests but are given more often throughout the period of instruction, and they are sometimes used to determine whether students are on track to meet summative benchmarks. Historically, federal law, regulations, and rules around assessment—such as the No Child Left Behind (NCLB) Act and its successor, the Every Student Succeeds Act (ESSA),[5] along with their associated guidance—have focused on requirements for standardized summative testing at the state level.

States and school districts generally cannot satisfy all ESSA requirements, let alone meet all their educational measurement objectives, through the purchase of a single assessment program, product, or service. Instead, it is useful to think of the combination of various tests as an assessment system—one in which individual components may be provided by different vendors and designed for different purposes. As shown below, some past and potential future innovations in assessment occur at this systemic level, such as the replacement of one assessment program with another that meets multiple requirements simultaneously.

Figure 1. Value vs. Burden of Assessments



Source: Author.

Beyond the typical classification of assessment programs as summative, interim, or formative, we can consider the value they provide to students, families, educators, and policymakers and the burden they place on these stakeholders. The value of

an assessment program, from the stakeholder's point of view, might be determined by what information it provides and the timeliness of those data, whether it has additional benefits (such as college credit for scores above a certain level), and what legal, bureaucratic, or regulatory requirements it satisfies. The burden of a given assessment program is generally a function of how much testing time it requires and how much money it costs the stakeholders. Figure 1 presents a simple typology of assessment programs by these two dimensions.

The placement of various assessment programs into these categories depends a great deal on the stakeholder's point of view. For example, the AP program, from the perspective of most families and students, is a type B (high burden, high value) assessment. Although AP is relatively burdensome, requiring much in student instructional time, study and preparation, and exam fees (unless waived), it returns a lot of potential value—college credit and a strong signal of achievement and preparedness.

The National Assessment of Educational Progress (NAEP),[6] on the other hand, can be considered a type C (low burden, low value) assessment for students and families. It does not require much of the average student (who likely will not even be sampled), nor does it return much to them of use. In fact, NAEP by design cannot provide scores for individual students and thus is useful only for aggregate reporting.

For the typical student and family—and perhaps most teachers and street-level administrators—the state summative assessments used to meet ESSA requirements are type D (high burden, low value) tests. They require significant assessment time every spring for the third through eighth grades and at least once in high school (in the case of mathematics and English language arts), they cost tens of millions of dollars annually, and many stakeholders believe they shape instruction in ways that may narrow the curriculum (although this is an open research question with mixed findings).[7] In exchange for this high burden, the tests give little to parents and students; the results come at the end of the school year or even over the summer and often do not provide much in the way of actionable diagnostic data to help guide future instruction.

For education leaders and policymakers, however, the positioning of various assessment programs in this typology may be different. For example, from the perspective of a district superintendent participating in NAEP's Trial Urban District Assessment (TUDA) program, NAEP may be a type B test—reasonably high burden in administrative requirements for participation but also relatively high value, since each TUDA administration allows that district's performance to be compared to the nation's, states', and other districts' results.

Yet, even for this audience of local policymakers, in most cases the state summative assessments used to satisfy ESSA requirements are still type D: high burden, low value. They are perceived as useful for satisfying federal reporting requirements in support of civil rights monitoring and accountability objectives but not much else, at the expense of a lot of time and money that could be better used on instruction or more diagnostic assessment.

Testing in the COVID-19 Pandemic

Although the pandemic has created enormous challenges for education in general and for assessment in particular, not all assessment programs have fared equally. Regarding the typology above, stakeholders have generally pushed to preserve or even expand type B testing, ignore type C assessments, and jettison their type D tests. Arguably, type A assessment programs do not exist in the marketplace or in the portfolios of states today—more on this below.

Type B Assessments Fared Well. Business has been good for test vendors with computer-based formative and diagnostic assessments in English language arts or mathematics, as local education agencies scramble to diagnose COVID-19 learning loss and tailor instruction this fall accordingly. Perhaps more strikingly, despite concerns about fairness and an untested design-and-delivery platform,[8] the College Board had an outpouring of public support to find a technical solution that would enable it to offer AP exams after schools shut down in March 2020—and a willingness by postsecondary institutions to accept the scores. The lesson here is that some assessments are so

useful that stakeholders will take on additional risk or tolerate the bending of the usual principles of psychometrics and measurement to preserve them.

Type C Assessments Were Ignored. Even in the best of times, low-value, low-burden assessment programs are not salient to most stakeholders, at least when compared to other types of tests. Disruptions due to the pandemic have made it harder to collect data even from low-burden assessments, such as those used for research and statistical purposes and given to relatively small samples of students. The ED postponed NAEP math and reading assessments until 2022, for example, citing concerns about the pandemic.

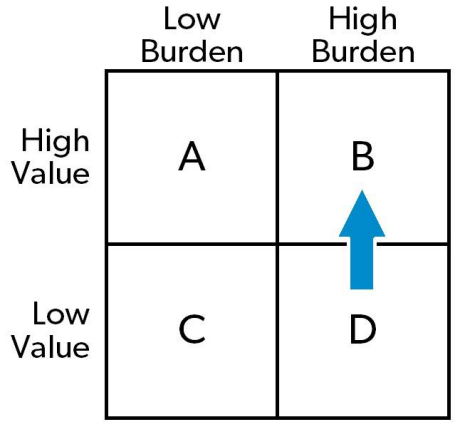
Type D Assessments Were Avoided. On the other hand, as noted above, every state and jurisdiction sought a waiver to suspend state summative assessment in the spring of 2020, and many sought waivers for 2021.[9] This is even though, in most cases, these assessments are low stakes for students, are already administered via cloud-based digital platforms, and could be transitioned to home administration with some ingenuity and resources. The lesson here is that state and district leaders have little appetite to find innovative solutions to operational challenges if there is any possibility of a waiver of state summative assessment, given the value-burden trade-off.

Toward a Minimum Viable Assessment System

Assuming there is still consensus at the federal level around the fundamental civil rights objective of state summative educational assessment as enshrined in NCLB and ESSA—ensuring that states continue to produce assessment data that can be used for subgroup accountability reporting—then the COVID-19 pandemic provides an opportunity to replace the current assessments with new tests that provide a net increase in the overall utility of the assessment system to stakeholders on the ground, including educators and families, while reducing burden. In fact, "opportunity" may be an understatement; many stakeholders will likely refuse to return to the status quo ante COVID-19.

In other words, the primary goal of educational assessment policy in the near term must be to drive the replacement of type D tests used to meet federal requirements with type B ones (Figure 2). How might this look? One idea being discussed in the assessment field is replacing the typical end-of-year state summative assessment in mathematics, English language arts, and science with measurements derived from a series of interim assessments given throughout the year.[10]

Figure 2. Increase the Value of Tests Used to Meet ESSA Requirements



Source: Author.

Such a redesigned system might indeed still be burdensome in cost and time. (Although, if a state or its districts are already using interim assessments,[11] the net effect will be a reduction in burden.) But if executed correctly, it could return more timely and actionable data to guide instruction and support improving student achievement.

Thus, in the typology presented here, this new program would be a type B assessment for educators and leaders and possibly for parents, with some additional design and outreach, since an interim assessment system could provide enhanced information on their child's academic achievement multiple times throughout the year. Also, while there are important implementation details to work out—interim assessment products can generally not be used off-the-shelf as they currently are operated but must be made suitable for accountability purposes—this shift does not require

significant investment in new technology or research and development. Digital platforms and item pools (collections of test questions) exist that are sufficient to create a minimum viable version with some modest investment, although some have questioned the item quality.[12]

How could the federal government help? A transition from status quo state summative assessments to an interim-as-summative approach, in which interim tests are rolled up into summative results for reporting, is already permissible under the ESSA. But states may need technical assistance in making the transition and developing a comprehensive assessment system using this approach.

Given that there are technical and design issues to solve, such as test security, timing of administrations, provision of accommodations, and the combination of interim data into a summative score,[13] states may also need additional resources, which will almost surely be passed through to testing vendors. Perhaps these resources can be awarded via a competitive grant program with provisions for technology and knowledge sharing to the field more broadly.

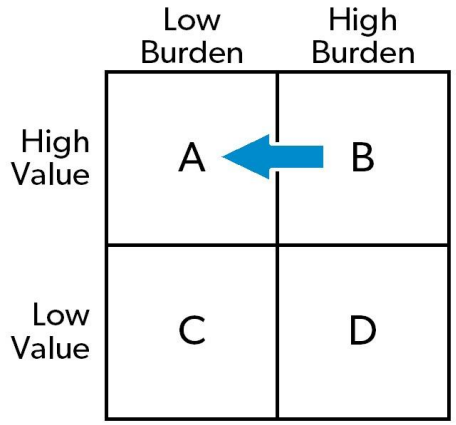
Looking Long Term

A transition to roll up interim-to-summative testing, which would increase the utility of assessments for stakeholders and possibly reduce burden across the entire assessment system, is a worthwhile first step. But in the longer term, educational assessment needs to move beyond simply shuffling the burden and increasing value and develop truly type A—high value, low burden—tests (Figure 3).

How might such an assessment program look? One idea is a shift to a more embedded assessment model, in which student performance data are collected as part of routine interactions with a digital learning platform. These measurement opportunities could include engaging, formative, computer-based enhanced performance tasks (and the development of the accommodations and accessibility technology required to deliver them for all students) that provide significantly more diagnostic assessment data while remaining valid, reliable, and fair. These (currently few) initiatives[14] all need significantly more research and development, beyond the type-D-to-B shift discussed above and resource-intensive prototyping and pilot testing.

Because of this need for significant additional research, development, and testing and because of the uncertain and risky return to providers on the large investment of development resources needed, there is likely a second federal role in supporting the infrastructure, technology, and science needed to develop this next generation of type A assessment programs.

Figure 3. Fund Research to Reduce Burden



Source: Author.

This federal support could be structured by splitting it into two separate programs: one housed in ED's Office of Innovation and Improvement and the other in the Institute of Education Sciences.

The Office of Innovation and Improvement program might be an assessment-specific innovation accelerator and venture partnership designed along the lines of the US Air Force's AFWERX[15] program—composed of several initiatives, including innovation hubs, challenge programs, and the Spark initiative, to drive grassroots innovation—and a venture arm based on more-traditional Small Business Innovation Research and Small Business Technology Transfer programs that fund small business research and development but relax some constraints to allow for adopting already-commercialized technologies.[16]

The Institute of Education Sciences program should build on that agency's experience with sci-

entific grant making and direct supervision of contracted research and assessment vendors. In particular, there are likely synergies with the NAEP program, which is considering a radical transformation from an outdated digital platform to a next-generation assessment design. By wisely using federal NAEP contracting dollars to not only improve that program but also create technology (not assessment content) that could be shared and disseminated for use in state assessment by vendors, research centers, and state agencies, the Institute of Education Sciences could be the perfect laboratory for research and development and has the capacity to manage this activity.

Conclusion

State summative assessment is at a crossroads. Dissatisfaction with current assessments coupled with an unprecedented disruption in all aspects of education have created significant pressure for an end to these assessments. However, despite its flaws and challenges, valid, fair, and reliable standardized testing remains an important component of consensus goals in education policy at the federal level.

Indeed, those who ignore assessment will be condemned to reinvent it. Stakeholders—the federal government, states, districts, and the assessment industry—should make the necessary investments to fix testing now, as measurement of what students know, and don't know, is more crucial than ever.

About the Author

Jack Buckley is the head of assessment and learning sciences at Roblox. Previously, he served as the senior vice president of research at the College Board and the commissioner of the US Department of Education's National Center for Education Statistics.

Notes

1. Lindsay Fryer, "Beyond Reading and Math Scores: Flexibility in Federal K–12 Accountability Law," American Enterprise Institute, September 29, 2021, https://www.aei.org/research-products/report/beyond-reading-and-math-scores-flexibility-in-federal-k-12-accountability-law/.

2. Nick Anderson, "One Million-Plus Juniors Will Miss Out on SATs and ACTs This Spring Because of Coronavirus," *Washington Post*, April 13, 2020, https://www.washingtonpost.com/local/education/one-million-plus-juniors-will-miss-out-on-sats-and-acts-this-spring-because-of-coronavirus/2020/04/12/4ccc827c-7a95-11ea-b6ff-597f170df8f8_story.html.

3. See College Board, "Students Take More Than 4 Million Advanced Placement Exams Online for the First Time, Working to Claim College Credit," May 22, 2020, https://newsroom.collegeboard.org/students-take-more-4-million-advanced-placement-exams-online-first-time-working-claim-college; and International Baccalaureate, "The Assessment and Awarding Model for the Diploma Programme May 2020 Session," May 13, 2020, https://www.ibo.org/news/news-about-ib-schools/the-assessment-and-awarding-model-for-the-diploma-programme-may-2020-session/.

4. For an overview of various types of assessments, see Kathy Dyer, "Understanding Formative, Interim, and Summative Assessments and Their Role in Student Learning," NWEA, July 13, 2017, https://www.nwea.org/blog/2017/understanding-formative-interim-summative-assessments-role-student-learning/.

5. See No Child Left Behind Act of 2001, Pub. L. No. 107-110, https://www.govinfo.gov/content/pkg/PLAW-107publ110/pdf/PLAW-107publ110.pdf; and Every Student Succeeds Act, Pub. L. No. 114-95, https://www.congress.gov/114/plaws/publ95/PLAW-114publ95.pdf.

6. National Assessment of Educational Progress, website, https://nces.ed.gov/nationsreportcard/.

7. For a general review of the unintended negative consequences of large-scale assessment, see Trina E. Emler et al., "Side Effects of Large-Scale Assessments in Education," *ECNU Review of Education* 2, no. 3 (2019): 279–96, https://journals.sagepub.com/doi/full/10.1177/2096531119878964. Perhaps the seminal paper introducing an empirical argument that standardized testing narrows what is taught in schools is David Berliner, "Rational Responses to High Stakes Testing: The Case of Curriculum Narrowing and the Harm That Follows," *Cambridge Journal of Education* 41, no. 3 (2011): 287–302, https://www.tandfonline.com/doi/abs/10.1080/0305764X.2011.607151. For an overview of the first wave of research after the implementation of

the No Child Left Behind (NCLB) Act, see Craig D. Jerald, "The Hidden Costs of Curriculum Narrowing," Center for Comprehensive School Reform and Improvement, 2006, https://files.eric.ed.gov/fulltext/ED494088.pdf. Jerald argues, more subtly than Emler et al., that in some cases, assessment causes the sequencing of subject matter to be changed. See also Stephen Sawchuck, "Under ESSA, an End to 'Narrowing of the Curriculum'?," *Education Week*, June 23, 2017, https://www.edweek.org/policy-politics/under-essa-an-end-to-the-narrowing-of-the-curriculum/2017/06. Sawchuck argues that, even if curricula did narrow at the elementary level under the NCLB, the accountability provisions of the Every Student Succeeds Act (ESSA) could drive a reversal of that trend, but compare that to Andrew Saultz, Jack Schneider, and Karalyn McGovern, "Why ESSA Has Been Reform Without Repair," *Phi Delta Kappan* 101, no. 2 (September 2019): 18–21, https://kappanonline.org/why-essa-has-been-reform-without-repair-saultz-schneider-mcgovern/. Most recently, Benjamin W. Arnold and M. Danish Shakeel use National Assessment of Educational Progress data to show a negative impact on instructional time and student achievement in nontested (i.e., not English language arts or math) subjects. See Benjamin W. Arnold and M. Danish Shakeel, "The Unintended Effects of the Common Core State Standards on Non-Targeted Subjects" (working paper, Annenberg Institute for School Reform at Brown University, Providence, RI, June 2021), https://www.edworkingpapers.com/sites/default/files/ai21-418.pdf.

8. These concerns led to the filing of a federal class action suit after the 2020 Advanced Placement exams were completed. See *J.P. v. Educational Testing Services*, No. 2:20-CV-04502 (C.D. Cal. May 19, 2020), https://www.fairtest.org/sites/default/files/Advanced-Placement-Lawsuit-As-Filed.pdf.
9. Ultimately, only Washington, DC, was granted a waiver from ESSA's testing requirements for the 2020–21 school year, prompting criticism from state leaders whose waivers were rejected by the Department of Education. See Perry Stein and Valerie Strauss, "D.C. Is Granted Permission to Skip National Standardized Exams," *Washington Post*, April 7, 2021, https://www.washingtonpost.com/local/education/dc-standardized-tests-waiver/2021/04/07/bd4bc928-97f3-11eb-a6d0-13d207aadb78_story.html.
10. See, for example, Bonnie O'Keefe and Brandon Lewis, *The State of Assessment: A Look Forward on Innovation in State Testing Systems*, Bellwether Education Partners, July 2019, https://files.eric.ed.gov/fulltext/ED596503.pdf.
11. As of 2016, the market for classroom assessments (i.e., formative and interim tests not mandated by the state) had surpassed that of state summative testing. See Michele Molnar, "Market Is Booming for Digital Formative Assessments," *Education Week*, May 24, 2017, https://www.edweek.org/teaching-learning/market-is-booming-for-digital-formative-assessments/2017/05.
12. See Amy Burkhardt and Derek C. Briggs, *The State of District-Level Interim Assessments*, University of Colorado Boulder, School of Education, Center for Assessment, Design, Research and Evaluation, 2018, https://www.colorado.edu/cadre/sites/default/files/attached-files/interim_assessment_report.pdf.
13. See Nathan Dadey and Brian Gong, *Using Interim Assessments in Place of Summative Assessments? Consideration of an ESSA Option*, Council of Chief State School Officers, 2017, https://ccsso.org/resource-library/using-interim-assessments-place-summative-assessments-consideration-essa-option.
14. One well-known example is the Summit Learning Platform, which contains an embedded assessment component, although the system as a whole is not without its detractors. See Summit Learning, "Summit Learning Platform Overview," https://help.summitlearning.org/hc/en-us/articles/225773488-Summit-Learning-Platform-Overview; and Tara García Mathewson, "The Overlooked Power of Zuckerberg-Backed Learning Program Lies Offline," Hechinger Report, May 2, 2020, https://hechingerreport.org/a-personalized-learning-program-with-ties-to-zuckerberg-shows-promise-despite-criticism/.
15. Details at AFWERX, website, https://www.afwerx.af.mil/.
16. Details at SBIR, "About," https://www.sbir.gov/about.