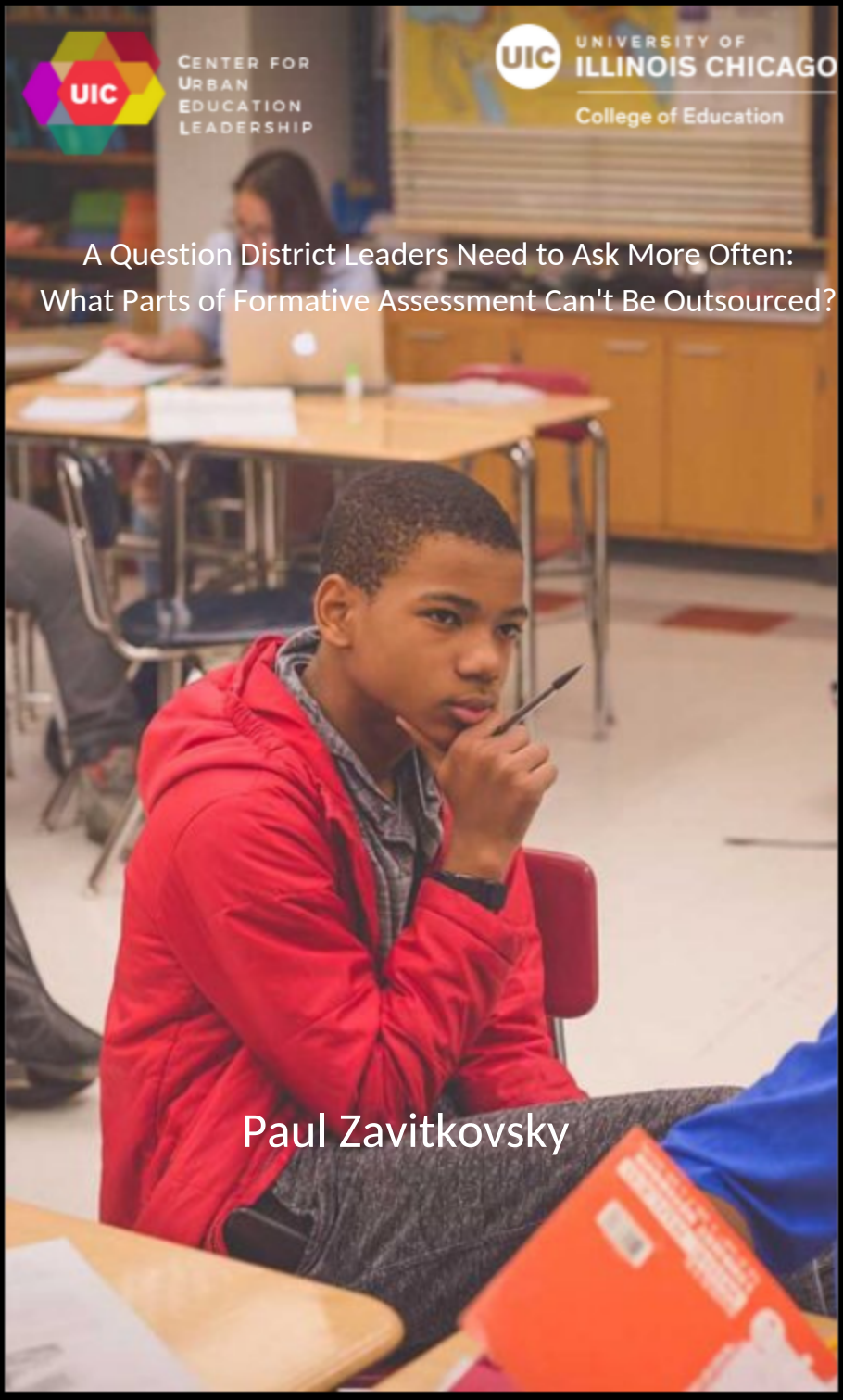

CENTER FOR
URBAN
EDUCATION
LEADERSHIP
UIC
UNIVERSITY OF
ILLINOIS CHICAGO
College of Education
A Question District Leaders Need to Ask More Often:
What Parts of Formative Assessment Can't Be Outsourced?
Paul Zavitkovsky

UIC Center for Urban Education Leadership

Publication Date: December 2022

Design: Angela Staples, CUEL Marketing Intern

Zavitkovsky, P. (2022). *A Question District Leaders Need to Ask More Often: What Parts of Formative Assessment Can't Be Outsourced?* Chicago, IL: Center for Urban Education Leadership. Retrieved from http://www.urbanedleadership.org

The Center for Urban Education Leadership (CUEL) is a research and development center housed in the College of Education at the University of Illinois at Chicago in Chicago IL, USA. The center is directed by Dr. Shelby Cosner. The center includes researchers, developers, and policy advocates with expertise in educational leadership, organizational development, continuous improvement, and equity/social justice. The center is driven to use its expertise and passion to IMPACT the lives of PK-12 urban students locally and throughout the world. Independently and in collaboration with other research/development organizations, CUEL has secured over $16 million to fuel a broad assortment of research and development projects. Learn more about our work at: https://urbanedleadership.org

Table of Contents

Introduction

"For a while in my life, I was senior research director at Educational Testing Service in Princeton, New Jersey. We had three different teams working on trying to get good diagnostic information from high-stakes accountability tests . . . Nothing came out of that work. It was simply, in most practical situations, impossible to get any insights into what you might do with this information apart from just looking at the total score and seeing 'this kid's struggling, this kid's OK."

. . . I would love it if we could derive diagnostic information from [large-scale standardized] tests. Everybody wants things to do double duty. But what we find in assessment is that generally, when we try to make things do double duty, they don't do either of them very well."

Dylan Wiliam, Learning Sciences International Conference on Formative Assessment, (April 2022)[1]

Back in 1998, Paul Black and Dylan Wiliam created big excitement in the education world when they published compelling evidence that teachers who regularly engage in formative assessment increase average student achievement by a grade level or more beyond expected gains. Their core finding was that, unlike conventional grading, the practices they identified as "formative"[2] increase shared ownership of learning between teachers and students and help teachers adjust instruction in ways that respond more effectively to specific student needs.

After the passage of *No Child Left Behind*, and especially after Obama-era Race to the Top initiatives, rising accountability for improved achievement and widespread dissatisfaction with the instructional value of statewide test reports led many school districts to purchase computer-adaptive "interim" testing systems. A common rationale for making this investment has been that computer-adaptive systems help educators scale up more formative approaches to teaching and learning.

For the most part, testing organizations have responded to consumer demand by reporting results immediately using two types of scoring information that state accountability tests do not normally provide:

- Detailed diagnostics about individual student learning that provide content-specific advice for day-to-day instruction
- National achievement and growth percentiles that offer richer normative information than the vaguely-defined proficiency categories that state tests report[3]

Marketing materials from the organizations that produce interim systems often make strong claims that this kind of reportage can support improved teaching and learning.[4] And rapid growth in the popularity of interim systems suggests that many educators agree. But independent assessment professionals have been warning for decades that large-scale standardized assessments are structurally incapable of delivering on these claims.[5] And now, growing empirical evidence is amplifying those warnings.

The Not-so-Common Sense of Standardized Assessment

As the opening quote from Dylan Wiliam suggests, a longstanding principle of modern psychometrics is that no single assessment can serve all purposes well.[6] For example, standardized test scales make it possible to track general academic progress over time with high levels of validity and reliability. But equating, the technical process used to produce

standardized test scales, makes it impossible for those same tests to generate valid, diagnostic information about specific curricular content.[7] The reverse is also true for assessments designed to produce small-grain diagnostic information. A rule of thumb among assessment professionals is that the better a test does serving one purpose, the worse it will do with a different purpose.

Nevertheless, rising accountability for improved teaching and learning, a dearth of useful reportage from statewide tests, and what James Popham calls "abysmal assessment literacy"[8] have led large numbers of educators and policy makers to think otherwise. In my own state of Illinois, for example, 70% of districts now spend $50 million annually on standardized interim testing.[9] This level of popular appeal suggests that large numbers of educators and policy makers have come to believe that advances in computer-adaptive technology now make it possible to do what Dylan Wiliam and independent assessment experts everywhere continue to reject out of hand.[10]

In recent years, changes in federal law have encouraged several testing organizations to develop even more audacious "through-year" systems that eliminate year-end accountability testing altogether by "rolling up" interim test results into a single composite accountability score for each student tested.[11] So doing, they reduce overall testing time in most districts and let states pick up the tab for interim testing that districts currently pay for on their own. In 2022, despite ongoing warnings from independent assessment professionals, close to a dozen states were poised to adopt some form of through-year testing.[12] Many other states are actively exploring the idea.

Where's the Beef?

"Extraordinary claims require extraordinary evidence." Carl Sagan Cosmos—Encyclopedia Galactica (1980)[13]

In 2019, Robert Slavin was the Director of the Center for Research and Reform in Education at Johns Hopkins University.[14] That year, he published an updated summary of independent research that examined the relationship between use of large-scale interim testing and changes in standardized achievement. Slavin concluded, "Benchmark assessments fall into the enormous category of educational solutions that are simple, compelling and wrong . . . average outcomes are not just small; they are zero."[15]

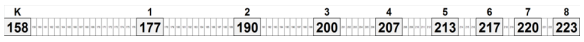
Summary of Independent Research on the Relationship Between Use of Large-Scale Interim Assessments and Standardized Achievement Scores

	Number of Studies	Mean Effect Size
Elementary Reading	6	-0.02
Elementary Math	4	0.00
Study-weighted mean	10	-0.01

SOURCE:
https://robertslavinsblog.wordpress.com/2019/04/11/benchmark-assessments-weighing-the-pig-more-often/

Data on the relationship between use of large-scale interims and achievement growth over time has been equally disappointing. As illustrated below, growth norms for all of America's most widely used standardized tests have long shown that the kind of new learning standardized tests value most declines progressively as grade levels rise.[16] This makes year-to-year growth after grade three a particularly fertile target for improvement initiatives.

Median Scale Scores by Grade on the NWEA MAP Reading Exam (2011 norms)



A common claim of large-scale interim systems is that they help educators increase students' year-to-year growth. But comparisons of grade-level norms from 2011 and 2020 for just those students tested by NWEA's MAP Growth

system show no statistically meaningful changes in growth norms over time at the 25th, 50th and 75th percentile of NWEA's national scoring distributions. This comparison suggests that, on average, information from America's most widely used interim system had no demonstrable impact on year-to-year-growth in those districts that used MAP Growth during the decade between 2011 and 2020.

MAP READING Scale Score Growth: 2011 Norms vs. 2020 Norms

		Grade 4		Grade 5		Grade 6		Grade 7		Grade 8		5-year Δ
25th Percentile	2011	189	+8	197	+5	202	+4	206	+4	210	+2	212	+23
	2020	186	+8	194	+6	200	+5	205	+2	207	+3	210	+24
50th Percentile	2011	199	+8	207	+5	212	+4	216	+4	220	+2	222	+23
	2020	197	+8	205	+6	211	+4	215	+3	218	+4	222	+25
75th Percentile	2011	209	+7	216	+6	222	+4	226	+3	229	+3	232	+23
	2020	208	+8	216	+6	222	+4	226	+3	229	+4	233	+25

MAP MATH Scale Score Growth: 2011 Norms vs. 2020 Norms

		Grade 4		Grade 5		Grade 6		Grade 7		Grade 8		5-year Δ
25th Percentile	2011	194	+9	203	+8	211	+4	215	+3	218	+4	222	+28
	2020	192	+8	200	+7	207	+4	211	+3	214	+3	217	+25
50th Percentile	2011	203	+9	212	+9	221	+5	226	+4	230	+4	234	+31
	2020	201	+10	211	+8	219	+4	223	+4	227	+3	230	+29
75th Percentile	2011	212	+10	222	+9	231	+5	236	+6	242	+4	246	+34
	2020	211	+10	221	+9	230	+5	235	+4	239	+5	244	+33

SOURCE: *RIT Scale Norms Study* (2011) Northwest Evaluation Association; *NWEA MAP Growth: Achievement Status and Growth Norms Tables for Students and Schools (2020).* Northwest Evaluation Association

Pre-pandemic evidence from the Stanford Education Data Archive (SEDA 4.1)[17] tells a similar story for all but the wealthiest of the PK-12 school districts that make up Illinois' Large Unit District Association (LUDA).[18] Ninety-five percent of LUDA's 52 districts used commercial interim systems in grades three through eight during all or most of the years between 2009 and 2018.[19] But during this same period, only 11% (3/27) of LUDA districts with average or below-average family income levels achieved above-average, cohort growth rates from grades three through eight, and only 19% (5/27) achieved average growth rates[20] (see green shading below).

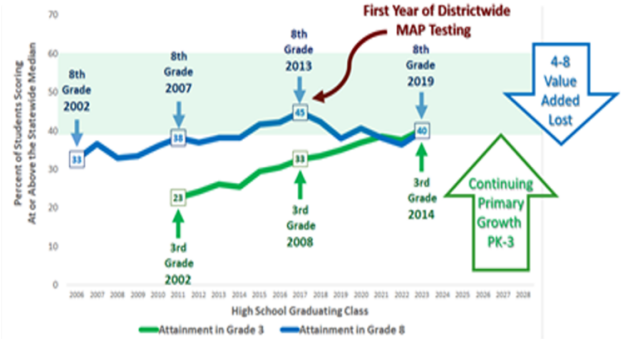
Cohort Growth Rates from Grade 3 through Grade 8 in 52 Large Unit Districts: 2009-2018

		Socio-economic Status of a Typical District Family				
		Below US Average		US Average	Above US Average	
		> -1.50 StDev	-0.50 to -1.49 St Dev	-0.49 to 0.49 St Dev	0.50 to 1.49 St Dev	>1.50 St Dev
Higher Growth >>>	≥ 15.0% Above US Growth Average 6 Districts	0	0	1	2	3
	5.0% to 14.9% Above US Growth Average 9 Districts	0	1	1	5	2
	15 Districts	7%		13%	80%	
	Average US Growth +/- 4.9% 15 Districts	0	0	5	3	7
	15 Districts	0%		33%	67%	
<<< Lower Growth	-5.0% to -14.9% Below US Growth Average 14 Districts	0	6	5	3	0
	≤ -15.0% Below US Growth Average 8 Districts	1	2	5	0	0
	22 Districts	41%		45%	14%	

SOURCE: Stanford Education Data Archive Version 4.1 https://edopportunity.org/

Recent pre-pandemic evidence from Chicago is even more sobering. In 2017, Stanford researcher Sean Reardon made national headlines by showing that Chicago achieved a full extra year of learning-beyond-expected-gains between third and eighth grade during the years between 2009 and 2014. That rate of growth was higher than 96% of all district growth rates nationwide.[21] But from 2013 through 2019, when the country's most widely used interim system (MAP Growth) was adopted district wide for grades three through eight, five-year cohort growth plummeted to a little below the national average.[22] After the fact, there's no way to know how much MAP Growth systems may have contributed to this decline. But they clearly didn't prevent it.

Growth Beyond Expected Gains from Grade 3 through Grade 8 Evaporated after MAP Growth Testing Began in 2013



SOURCES: Illinois State Board of Education Student-Level Media Files—2002-2019: ftp://ftp.isbe.net/schoolreportcard/ Zavitkovsky, Paul (November 2021) "MAP to Nowhere" Consortium for Research on Educational Assessment and Teaching Effectiveness

Explaining the Dissonance: Seeing the System that Produces Current Outcomes

"A close look at state test items shows me that both test-prep 'teaching' and test bashing get it wrong. The test items that our students do most poorly on demand interpretation and transfer, not rote learning and recall. Better teaching and (especially) better local testing would raise state test scores. Teaching for greater understanding would improve results, not threaten them—as both common sense and research indicate." Grant Wiggins
Why We Should Stop Bashing State Tests (2010)[23]

The dissonance between growing support for standardized interims and growing evidence that they aren't working in many districts is loud and getting louder. This is especially true for schools and districts where large percentages of students come from low-income households. How long it will take to resolve this dissonance is anybody's guess. But one promising way to understand it is to look more closely at how well current forms of interim reportage match up with the kinds of learning that standardized tests value the most.

In 2001, the University of Chicago's Consortium on School Research released a ground-breaking study called, *Authentic Intellectual Work and Standardized Tests: Conflict or Co-existence?*[24] Over a three year period, this study assessed the connection between standardized achievement and the intellectual demand of classroom assignments. A key feature of the study was that it took place in schools serving high percentages of Black and Latinx students from low-income households; 89% of students studied were eligible for free or reduced lunch; 53% were Black and 39% were Latinx.

Authentic Intellectual Work found that students at all achievement levels who were regularly challenged by intellectually demanding assignments scored about a grade level higher on standardized tests than students who were not. This finding flew in the face of conventional notions that only higher-achieving students can handle and benefit from intellectually challenging tasks. It also highlighted what test designers have always known.[25] Depth of

knowledge (DOK) is a significant contributor to higher scores on standardized tests and accounts for why their power to predict early college success closely rivals that of a student's high school GPA.[26]

But as obvious as these findings might sound to some readers . . . and as often as they have been replicated in other studies[27]. . . they are actually at odds with the reporting practices of most interim testing systems and with teaching practices in most American classrooms. With deep roots in 20th century behaviorism,[28] these practices continue to reflect what Harvard professor Jal Mehta calls "the pernicious myth" that mastery of basic skills must necessarily precede deeper learning.[29]

"If there is one prevalent assumption that stands in the way of deeper learning, it is that you have to do 'the basics' before you can engage in deeper learning . . . You can see the appeal of this idea. Learn the notes before you play the concerto . . . [but] this line of thinking tends to foreground students' deficits over their assets.

For disadvantaged and lower-track students, it serves to justify teaching as transmission and what Freire called the 'banking model' of education . . . No matter how well-intentioned, the result in practice is that, yet again, the most privileged students are being taught how to think, whereas less advantaged students, who are often students of color, are being taught how to follow the directions of authorities . ."[30]

A decade of international research and analysis on classroom teaching from the Third International Math and Science and Study (TIMSS) illustrates how powerfully the pernicious myth still guides the prevailing norms of American teaching culture.[31] Unlike teaching practices in virtually all higher-achieving countries worldwide, TIMSS video studies showed that most American teachers break down curricular content into easily digestible pieces that do most of the thinking for students. Connection-making challenges that require sustained intellectual struggle are almost completely absent from most American classrooms . . . even when curriculum materials have been explicitly designed to engage students in problem solving and productive struggle.[32] In more recent studies of students entering community college, TIMSS researcher James Stigler has described the devastating impact that this kind of teaching has on student understanding and post-secondary success many years after it occurs.[33]

Even a cursory analysis of reporting practices used by most large-scale interim testing systems shows how much these practices tacitly endorse atomized concepts of curriculum that are at odds with what standardized tests value most. For example, the Renaissance Star 360 system does not explicitly declare that curriculum can be best understood as a large collection of discrete standards that computer-adaptive technology can assess one standard at a time. But that is the message its reportage consistently communicates.[34] And color-coded sub-score reports[35] close the loop by tacitly encouraging teachers to design grouping and instruction accordingly in order to "meet students where they are." Test reports from NWEA's MAP Growth system are even more fine-grained. They match up long lists of discrete skills from the MAP Learning Continuum with specific, 10-point scoring ranges on MAP's overall achievement scale.[36]

Tim Shanahan is a past president of the International Literacy Association and the former co-chair of the National Reading Panel.

In multiple summaries of research about teaching that improves reading comprehension, Shanhan has criticized the impact that one-skill-at-a-time test reportage can have on policy and practice in American schools.[37]

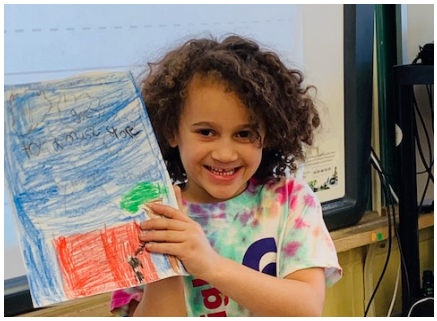
"A common error in reading education is to treat reading comprehension as if it were a skill or a collection of discrete skills. Skills tend to be highly repeatable things… Many of the items listed as comprehension skills are not particularly repeatable. All these standards or question types aimed at main idea, central message, key details, supporting details, inferencing, application, tone, comparison, purpose, etc. are fine, but none is repeatable in real reading situations.

"Studies have repeatedly shown that standardized reading comprehension tests measure a single factor—not a list of skills represented by the various types of question asked."[38]

To be clear, the core critique here is not that mastery of discrete skills and standards is somehow unimportant. The critique is that:

- large-scale standardized tests, both fixed form and computer-adaptive, are structurally incapable of producing valid, small-grain diagnostic information;[39] apples and oranges may both be fruits, but you can't get orange juice from apples no matter how cleverly you squeeze them;
- reporting systems that focus almost entirely on discrete standards and skills tacitly legitimate over-dependence on one-skill-at-a-time teaching that has long been identified as a major driver of race/class opportunity gaps;[40]
- reporting systems that translate test results into discrete standards and skills filter out information about interpretation, application and transfer that is central to deep learning[41] and is closely associated with higher scores on standardized tests.[42]

Why these problems seem to be impacting poorer districts more negatively than wealthier ones is an open question. But one promising explanation is that districts where large percentages of students are scoring well below grade level are more likely to use interim systems to drive instruction for all students rather than restricting their use to preliminary screening for the 15-20% of students who are most in need of supplemental support.



Sense-Making: The Part of Formative Assessment that Can't be Out-Sourced

"In August of 2008, the chairman of the Federal Reserve Bank called an emergency meeting with then-President George W. Bush to inform him that the entire financial system was melting down. Bush, shocked, responded by asking, 'How did we get here?' . . . Part of the answer is that the system was designed to fail. Naturally, the banks did not want to fail. They did not want the economy to fall apart. But these results were nevertheless natural outgrowths of the choices they made about measuring and rewarding performance. Investment banks failed to hold their employees accountable for key decisions that were well within their control." Douglas Harris *Value-Added Measures in Education (2011)*[43]

There's plenty of blame to go around for relying too heavily on large-scale standardized testing to define curricular priorities and support day-to-day teaching in American classrooms. For example:

- At the federal level, NCLB and all subsequent legislation has codified assessment illiteracy

into law by requiring large-scale standardized tests to do more than their psychometric DNA allows;[44]

- Leaders in commercial testing organizations have responded to consumer demand by fudging on the technical limitations of large-scale testing and by using marketing and reporting practices that minimize or obscure those limitations;[45]
- Limited resources and limited internal capacity have led many state education leaders to devote far more attention to efficient compliance with federal regulations than to using their bully pulpit to deepen assessment literacy and encourage investment in richer assessment practices at the school and district level; for similar reasons, leadership in many statewide teachers unions and other educational advocacy groups have made their peace with specious reportage from large-scale testing systems rather than confronting it head-on;[46]
- Locally, limited resources and limited internal capacity have led many school and district leaders to accept the claims of commercial testing organizations more or less uncritically, and to build elaborate cultures of assessment around them at the school and district level.[47]

The list goes on. One way or another, it now includes virtually all of the major stakeholders in the American educational system. The question, of course, is how could this happen?

In December 2004, a disgruntled American soldier challenged Secretary of Defense Donald Rumsfeld to explain why his unit had to rummage through trash heaps to find scrap metal they could use to strengthen the armor of their Humvees. Rumsfeld famously responded, "You go to war with the army you have . . . not the army you might want or wish to have at a later time."

When *No Child Left Behind* became law, the army we had for revolutionizing large-scale assessment design was big banks of norm-referenced test items and close to a century of experience building tests that compared students with each other. Since that time, it is mostly resources of this kind that the testing industry has relied on to build large-scale, "standards-based" assessments. As a result, most of what we now routinely call standards-based assessment is really just norm-referenced testing dressed up in standards-based clothing.[48]

Twenty years in, we're still jerry-rigging tools for the assessment army we have because the assessment army we wish to have requires way more organizational support and political will than most states, districts and testing organizations have been able to muster.[49] In a 2018 article for the National Council on Measurement in Education, Scott Marion described the situation most districts face like this, *"Districts do not implement [interim testing systems] in an effort to waste money. They do so because they think that such assessments are a critical component of an assessment system, and they are struggling to find any ray of hope to improve performance in situations with scarce resources. There is also a belief that test results from an external entity are 'official' . . ."*[50]

An encouraging response to the ongoing disruptions created by COVID 19 has been a

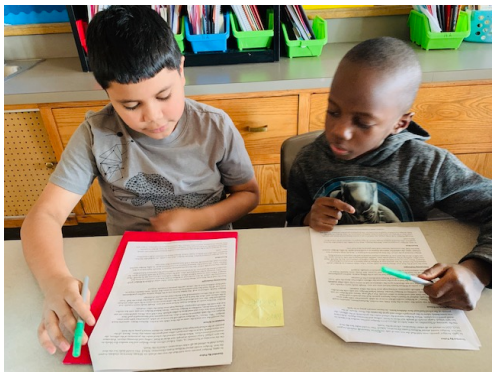
renewed call for fundamental reform in the way we do school. For example, growing numbers of educators and policy makers are talking about replacing over-reliance on remediation with "acceleration of deeper learning" and "just in time skills instruction."[51] But a century of earlier attempts to close opportunity gaps and take deep learning to scale in American schools offers a sobering lesson. The acceleration army we wish to have will need more than souped up Humvees to meet its goals. It will need new kinds of equipment to help it *make better sense* of the instructional battles it needs to fight, and to *make better choices* about the tools and tactics it needs to be successful.

Early in the NCLB era, Richard Elmore wrote,

"[After a century of large-scale efforts to alter the fundamental processes of schooling], basic conceptions of knowledge, of teacher's and student's roles in constructing knowledge, and of the role of classroom- and school-level structures in enabling student learning remain relatively static. The problem lies not in the supply of new ideas, but in the demand . . . The primary problem of scale is understanding the conditions under which people working in schools seek new knowledge and actively use it to change the fundamental processes of schooling."[52]

Two high-profile conditions under which people working in American schools now operate are high-stakes accountability and widespread use of standardized testing. These are the tools that we're currently relying on to "change the fundamental processes of schooling." But after two decades of disappointing results from state accountability testing and large-scale interim testing, an obvious question is whether standardized tests can play any role at all in producing formative information that educators can "actively use to change the fundamental processes of schooling."

The good news is that better reportage of standardized test items actually can provide powerful, context-specific models of what challenging standards require. And better reportage actually can help educators gain novel insights about where their students are getting stuck and why. We've known this since at least 2003 when the National Research Council brought together educators and independent assessment professionals from all over the country to spell out more clearly what large-scale standardized assessments can and can't do.[53]



The bad news is that support systems for extracting formative insights from released test items have been too labor-intensive, and depend too heavily on interpretation and collaborative sense-making to fit easily into the limited time that most American teachers have to do collaborative work. In most elementary and middle schools, for example, the time available for collaborative work is rarely more than one or two 45-minute periods per week.

In 2019, Charles DePascale wrote a series of articles on assessment innovation for the National Center for Improvement of Educational Assessment. In the first of these articles,[54] he argued that a key characteristic of any true innovation is that it adds value to a product by making it more useful to its consumers. In the 2010s, enthusiasm for large-scale interim assessments grew quickly because they delivered more useful reportage than almost all statewide

accountability exams. But useful for what? Quick, user-friendly access to reportage that supported one-skill-at-a-time grouping and remediation? For sure. Better access to information that supported acceleration and deep learning for all. Not so much, especially in schools and districts where the majority of students were scoring one or more grade-equivalents below national norms.

Getting More Serious about Equity-Driven Test Reportage

"Your system . . . any system . . . is perfectly designed to produce the results you're getting." [Systems maxim—original source unknown]

"*One cannot understand the history of education in the United States during the twentieth century unless one realizes that Edward L. Thorndike won and John Dewey lost." Ellen Condliffe Lagemann An Elusive Science: The Troubling History of Education Research (2002)*[55]

"To understand structural racialization . . . we have to entertain the idea that a series of seemingly benign or supposedly well-intended policies actually create a negative, cumulative and reinforcing effect that supports rather than dismantles the status quo within institutions." Zaretta Hammond *Culturally Responsive Teaching and the Brain* (2015)[56]

As long as standardized tests continue to play a major role in the American policy environment, there is still a strong case for getting as much valid information from them as possible to help "change the fundamental processes of schooling."[57] But the evidence is strong that most current forms of large-scale interim reportage do not meet conventional standards of validity with the "formative" information they provide.[58] And as earlier illustrations have shown, many of them filter out higher-order learning information and tacitly encourage the very beliefs and practices that we most need to change in order to address chronic race/class opportunity gaps.

Overly didactic, one-skill-at-a-time teaching practices have persisted for over a century in American classrooms despite repeated, high-profile efforts to engage both teachers and students in deeper forms of learning. We have ample evidence that instructional practices of this kind exacerbate race/class opportunity gaps, and ample evidence that these gaps have been magnified substantially by pandemic disruptions.[59] But when officially-sanctioned test reportage describes curriculum as a large collection of discrete skills and standards, it reifies the "pernicious myth" and builds deficit-based pedagogy into the woodwork of school and district policy. So doing, it further postpones the long-overdue work of closing opportunity gaps . . . work that cannot happen without making deep learning much more accessible to all of our students regardless of the way they currently score on standardized tests.

It's not like we don't have good prototypes for reporting rich, formative information from standardized tests. The National Assessment of Educational Progress (NAEP) has spent years building user-friendly item maps and released items that demystify what scale scores and proficiency levels mean, and provide concrete examples of how students respond to different levels of academic demand,[60] The College Board released items and response frequencies from its PSAT exams with the same purpose in mind.

The Smarter Balanced Assessment Consortium (SBAC) has gone one step further by developing a comprehensive system of fixed-form, curriculum-specific interims that is "loosely-coupled"[61] with SBAC's year-end summative test. These systems are explicitly designed to help grade/departmental teams use item analysis to make better sense of how

deeply students have mastered state learning standards, where they're getting stuck, and why.[62] And at least in the State of Connecticut, there's good evidence that their use produces better-than-expected gains across achievement levels on year-end accountability exams.[63]

To be clear, there is nothing bullet-proof about deeper analysis of test items to improve school and district assessment practices. For example, without clear guidance, practitioners can easily misunderstand the representative nature of individual test items. This misunderstanding leads them to:

- look too literally for specific skills and content contained in items that have low, correct-response frequencies;
- create what Lorrie Shepard describes as "1,000 mini-lessons"[64] that are narrowly designed to improve mastery of discrete skills and content.

In the process, they can overlook the broader requirements for application, transfer and other context-specific understandings that the items they selected were actually designed to assess.

Better reportage of large-scale test information won't in and of itself solve the challenge of better formative sense-making. The solution for that lies where it always has: in thoughtful leadership and in ongoing collaboration and experimentation by teams of teachers who commit to helping each other build greater rigor and depth of knowledge into the work they do with all their students. Supporting those teams at scale is the job of states and districts. That job begins by equipping teachers with the best information we have about the challenges we expect them to address.

Long before the disruptions created by COVID-19, there was clear evidence that return on investment from state accountability exams and most large-scale interim systems was inadequate at best and harmful at worst.[65] The price we've paid for failing to act on that evidence is widespread illiteracy about what large-scale standardized tests can and can't do, and a generation of lost capacity building for improving teaching in general and for closing chronic, race/class opportunity gaps in particular.

Like the launch of Sputnik in the late 1950s, the publication of *A Nation at Risk* in the 1980s, the passage of *No Child Left Behind* in the 2000s, and *Race to the Top* in the 2010s, pandemic disruptions have created new conditions of possibility for concerted action that can "change the fundamental processes of schooling." What's still unclear is whether state and local leaders will be any better able than leaders in past eras to confront the ghosts of 20th century behaviorism and our competing commitments[66] to outmoded views of learning and intelligence.[67] These commitments continue to speak loudly through policies, systems, beliefs and practices in schools, districts and states throughout America. A small but crucial step toward something better will come from providing more adequate sense-making tools to the teachers, parents and administrators who do the daily work of schooling.

Paul Zavitkovsky

CUEL Faculty Affiliate

REFERENCES

1. Wiliam, D. (2022, April 8). A Fireside Chat with Dylan Wiliam. Dylan Wiliam International Conference on Formative Assessment. *Learning Sciences International.* https://vimeo.com/699790414

2. For a more detailed description of the practices Black and Wiliam identified as formative, see Black, P. and Wiliam, D. (October 1998). Inside the Black Box. *Phi Delta Kappa. 80* (2), 144, 146-148. https://kappanonline.org/inside-the-black-box-raising-standards-through-classroom-assessment/

3. For example, unreported State of Illinois percentile ranges for proficiency levels on Illinois' 2015 eighth grade PARCC ELA exam were:

   · Level 1—Did Not Meet Expectations (0-13th percentile)

   · Level 2—Partially Met Expectations (14th to 33rd percentile—a range of 1.5 grade equivalents)

   · Level 3—Approached Expectations (34th to 60th percentile—a range of 1.5 grade equivalents)

   · Level 4—Met Expectations (61st to 93rd percentile—a range of 2.5 grade equivalents)

   · Level 5—Exceeded Expectations (94th to 99th percentile)

   For at least the last two decades, State of Illinois percentile ranges have consistently been within a standard error of NAEP national percentile ranges for all U.S schools

4. For example, online marketing materials for NWEA's MAP Growth system say, *"Easy-to-use, standards-aligned reports put the information teachers need at their fingertips. Reliable insights make it simple for teachers to find common areas of need among their students, identify students who could benefit from intervention, and determine which instructional strategies are generating the most academic growth. Higher-level reports provide administrators with the context to drive improvement across entire schools and educational systems."* https://www.nwea.org/map-growth/

5. See, for example, Shirley, D. and Hargreaves, A. (October 3, 2006). "Data Driven to Distraction." *Education Week*. https://www.edweek.org/technology/opinion-data-driven-to-distraction/2006/10 , and Shepard, L. "Formative assessment: Caveat Emptor" *ETS Invitational Conference 2005, The Future of Assessment: Shaping Teaching and Learning* https://csaa.wested.org/wp-content/uploads/2019/11/shepard-formative-assessment-caveat-emptor.pdf

6. For example, in a 2003 report from the National Research Council's workshop on bridging the gap between large-scale and classroom assessment (see endnote #52), Lorrie Shepard wrote, *"While the value of large-scale assessment for [some] purposes is clear, it is equally clear that they are not useful for many other educational purposes, particularly that of providing detailed understanding of individual students' performance. Professional standards are firm on the point that it is not a test itself that can be established as valid, but particular inferences that may be made from test data (see National Science Education Standards (NSES) Standard 13.2, NRC. 1996)"* [emphasis added]

7. Charles DePascale (November 19, 2019) offers a straightforward explanation of the limitations that the equating process place on information that can be derived from large-scale tests in a short article called "How We Can Generate Actionable, Student-Level from the Summative State Test While Honoring Their Design," *Center for Assessment Centerline*, https://www.nciea.org/blog/making-the-most-of-the-summative-state-assessment/

8. Popham, W. James. (2009). *Unlearned Lessons: Six Stumbling Blocks to Our Schools' Success*. (Chapter 6—Abysmal Assessment Literacy). *Cambridge: Harvard Education Press.*

9. See hour 2:52:00 of a presentation by State Superintendent Carmen Ayala to the Illinois State Board of Education at their regular monthly meeting on May 18, 2021, available at https://register.gotowebinar.com/recording/724526900778132481

10. See, for example, Marion, S. (April 29, 2021). "It Might Just Be a Pile of Bricks! The Challenges of Creating Balanced Assessment Systems." *Center for Assessment Centerline*. https://www.nciea.org/blog/it-might-just-be-a-pile-of-bricks/. For a more detailed analysis, see Marion, S. et. al. (April 26, 2019). "A Tricky Balance: The Challenges and Opportunities of Balanced Systems of Assessment, Annual Meeting of the National Council on Measurement in Education." https://www.nciea.org/wp-content/uploads/2021/11/A-Tricky-Balance_031319.pdf

11. See, for example, the landing page of the Northwest Evaluation Association's promotional website for its through-year assessment model https://www.nwea.org/through-year-assessment/

12. A thorough discussion of through-year systems that includes a wide range of independent and vendor-based assessment professionals is available in recordings and writings from a national convening on through-year testing that was sponsored by the National Center for Improvement of Educational Assessment on November 15 and November 16, 2021. https://www.nciea.org/blog/collaboratively-learning-about-through-year-assessments/

13. https://effectiviology.com/sagan-standard-extraordinary-claims-require-extraordinary-evidence/

14. The link to the Johns Hopkins Center for Research and Reform in Education website is https://education.jhu.edu/crre/

15. Robert Slavin's research summary is available at https://robertslavinsblog.wordpress.com/2019/04/11/benchmark-assessments-weighing-the-pig-more-often/

16. Growth patterns from grades K through 12 on seven of the country's most widely used standardized tests are reported in Lipsey, M. et. al. (2012) "Translating the Statistical Representation of Effects of Education Interventions into More Readily Interpretable Forms." *NC SER 2013-3000, Institute for Education Sciences*, US Dept. of Education

17. The Stanford Education Data Archive (SEDA) houses normalized achievement and growth information for cohorts of students in grades 3-8 in over 11,000 U.S. school districts. A unique feature of SEDA data displays is that they use a single equated scale based on national norms from the National Assessment of Educational Progress (NAEP) to report five-year cohort achievement and growth patterns that are derived from 50 separate state accountability tests. The link to the SEDA website is https://edopportunity.org/ Additional descriptive information about SEDA that is designed for general audiences is available at https://www.nytimes.com/interactive/2016/04/29/upshot/money-race-and-success-how-your-school-district-compares.html?action=click&contentCollection=upshot&region=rank&module=package&version=highlights&contentPlacement=1&pgtype=sectionfront&smid=tw-upshotnyt&smtyp=cur&_r=2 and at https://www.nytimes.com/interactive/2017/12/05/upshot/a-better-way-to-compare-public-schools.html?hp&action=click&pgtype=Homepage&clickSource=story-heading&module=second-column-region&region=top-news&WT.nav=top-news&_r=0

18. The link to LUDA's website is https://www.ludaillinois.org/ After data analysis for this study was completed, LUDA district membership increased to 53 when the Chicago Public Schools rejoined the organization With Chicago included, LUDA districts represent close to 50% of students enrolled in all Illinois public schools. The demographics of student enrollments in LUDA districts are highly representative of the state as a whole, and public school enrollments in the State of Illinois are more demographically representative of the nation as a whole than those of any other American state.

19. In the 2019-2020 school year, 62% of LUDA school districts used MAP Growth, 21% used Fast Bridge and 12% used Renaissance Star 360 as their core interim assessment system for grades three through eight https://www.isbe.net/Pages/AssessmentSurveyResults.aspx

20. An important finding from analysis of school- and district-level SEDA data nationwide is that the correlation between student/family SES and cohort growth from end of grade three through end of grade eight is much lower than the familiar correlation between SES and normative achievement status. At the school level, this correlation virtually disappears.

21. See "Chicago Schools Lead Country in Academic Growth, Study Finds" *Education Week,* November 9, 2017 https://www.edweek.org/leadership/chicago-schools-lead-country-in-academic-growth-study-finds/2017/11, and the research brief by Sean Reardon and Rebecca Hinze-Pifer titled, "Test Score Growth among Chicago Public School Students: 2009-2014." https://cepa.stanford.edu/sites/default/files/chicago%20public%20school%20test%20scores%202009-2014.pdf

22. This analysis uses normalized State of Illinois distributions to determine the percentage of Chicago students in each cohort who scored at or above statewide median scores in the years shown. A more detailed analysis of cohort growth over time in Chicago is scheduled to be posted on the Center for Urban Education Leadership website in November 2022. This analysis illustrates the close correspondence of trends based on State of Illinois norms with scoring trends derived from the most recent data that are publicly available from SEDA version 4.1

23. Wiggins, G. (2010). "Why We Should Stop Bashing State Tests". *Education Leadership,* 67:6, pages 48-52 https://www.ascd.org/el/articles/why-we-should-stop-bashing-state-tests

24. Newmann, F. Bryk, A. and Nagaoka, J. (2001). "Authentic Intellectual Work and Standardized Tests: Conflict or Coexistence?" *UChicago Consortium on School Research, https://consortium.uchicago.edu/publications/authentic-intellectual-work-and-standardized-tests-conflict-or-coexistence*

25. For a description of the role played by DOK in the construction of items used in the pre-Common Core, Illinois Standards Achievement Test (2006-2014) see Appendix D "Webb Alignment Analysis of Reading, Mathematics and Science Standards and Assessments in the 2007 ISAT Technical Manual https://www.isbe.net/Documents/isat_tech_2007.pdf

26. For a discussion of the predictive power of high school GPA versus standardized test scores, see Watkins, S. (2020). "GPA or SAT—Two Measures are Better than One." *James G. Martin Center for Academic Renewal*, https://www.jamesgmartin.center/2020/02/gpa-or-sat-two-measures-are-better-than-one/#:~:text=There%20is%20strong%20evidence%20that%20GPA%20and%20test,together%20is%20a%20better%20predictor%20than%20GPA%20alone .

27. See, for example, Allensworth, E. et. al. (May 2022). "Standards Driven Instructional Improvement" *UChicago Consortium on School Research*, https://consortium.uchicago.edu/publications/standards-driven-instructional-improvement

28. Mehta, J. (January 4, 2018). "A Pernicious Myth: Basics Before Deep Learning." *Education Week*. https://www.edweek.org/teaching-learning/opinion-a-pernicious-myth-basics-before-deeper-learning/2018/01 A detailed description of the behaviorist roots of Mehta's pernicious myth can be found in Labaree, D. "How Dewey Lost." *School of Education*, Stanford University. https://web.stanford.edu/~dlabaree/publications/How_Dewey_Lost.pdf

29. As recently as the 1980s and 1990s, Madeline Hunter's Science of Teaching dominated the American professional landscape and was the backbone of many district observations and evaluation protocols. For an illustration of how clearly Hunter's approach reflects Mehta's pernicious myth, see Minute 11:15 through Minute 14:00 in the video of an April 2022 presentation by Laquita Louie and Paul Zavitkovsky called "Getting Smarter about Equity Metrics" at Learning Science International's 2022 Dylan Wiliam International Conference on Formative Assessment. https://vimeo.com/699775172

30. A more detailed discussion of one-skill-at-a-time teaching in schools that serve mostly low-income students of color can be found in Haberman, M. (December 1991). "The

Pedagogy of Poverty" *Phi Delta Kappan,* 73:4, pages 290-294. https://www.pblworks.org/blog/pedagogy-poverty-and-its-antidote-pbl

31. The findings of TIMSS studies are summarized in a short 2009 *Phi Delta Kappan* article by Stigler and Hiebert called "Closing the Teaching Gap." https://www.researchgate.net/profile/James-Stigler/publication/285720440_Closing_the_Teaching_Gap/links/56659c9b08ae192bbf925443/Closing-the-Teaching-Gap.pdf?_sg%5B0%5D=_ejGHjyxt7nPj-pTV0hdn29LQoPyM59_wgsk6bPFKiCAOw0IytJ7OBJEaYe0OSPKcMQaMelLzq2HZao-nO7CUA.1rcuUmKFcaYm8Ys76tunae6-6LOwZnyNsHMnxyExMj2x77FpTlQj-P8IJRmuZPUd5ah1f97PVzeYAC75jAZ6nA&_sg%5B1%5D=v5VNORmuaECuQw-93icK-UkCJoy9JxGJUrBZ_GY599h2Q6RYK6zHNAYDuqRKRgnMXx9PTIOmygwkLtZ_R9N_maeZ941YIqOQbIyuHFsvcOFD.1rcuUmKFcaYm8Ys76tunae6-6LOwZnyNsHMnxyExMj2x77FpTlQj-P8IJRmuZPUd5ah1f97PVzeYAC75jAZ6nA&_sg%5B2%5D=Dgzam4hvi8XtkXAbA2oeg4by88fTBwOI5B6kP7gtnMtg2fh9zHhOqU1dh4_tXtVMxSFyey7L5P90m50.Kp7xZO-d6B3SFGw5KyHf1lm6mMAfma2oiK1O8ssiX2XoV2WDFNyO8gk3SctxHfm3F0Nc9bbl4ZIQnBH2l02uYA&_iepl=

32. See Minute 18:40 through Minute 20:15 in "Getting Smarter about Equity Metrics" https://vimeo.com/699775172

33. See Minute 5:10 through Minute 14:45 of James Stigler's address to the Precision Institute of National University called "Teaching for Understanding: What Will It Take?" https://vimeo.com/691127511

34. For a more extended analysis, see Zavitkovsky, P. (November 2021). "MAP to Nowhere" presented to the *2021 Conference of the Consortium for Research on Educational Assessment and Teaching Effectiveness (CREATE).* https://vimeo.com/673344384

35. Sample Renaissance STAR 360 sub-score reports can be found at https://cdn5-ss12.sharpschool.com/UserFiles/Servers/Server_538005/File/Curriculum/Teachers/Assessment/R0053249615EE616.pdf

36. A sample of NWEA Classroom Breakdown Reports can be found at https://teach.mapnwea.org/impl/maphelp/Content/Data/SampleReports/ClassBreakdownbyRIT_byGoal.htm?Highlight=class%20report%20by%20RIT

37. Two blog posts by Tim Shanahan that describe misguided offshoots of teaching reading comprehension primarily through the mastery of discrete skills are, "Should We Grade Students on the Individual Reading Standards?" (September 17, 2019). https://www.shanahanonliteracy.com/blog/should-we-grade-students-on-the-individual-reading-standards?utm_source=Shanahan+on+Literacy&utm_campaign=9995eee1a2-EMAIL_CAMPAIGN_2019_08_31_10_22&utm_medium=email&utm_term=0_95269a2ffa-9995eee1a2-252666097#sthash.rQsCKnk5.dpbs and "Should We Administer Weekly Tests Linked to Standards?" (February 23, 2019) https://www.shanahanonliteracy.com/blog/should-we-administer-weekly-tests-linked-to-standards#sthash.3PGWL9AG.dpbs

38. Shanahan, Tim. (November 9, 2019). "How to Analyze or Assess Reading Comprehension" https://www.shanahanonliteracy.com/blog/how-to-analyze-or-assess-reading-comprehension?utm_source=Shanahan+on+Literacy&utm_campaign=d185450e39-EMAIL_CAMPAIGN_20

19_11_08_06_22&utm_medium=email&utm_term=0_95269a2ffa-d185450e39-252666097#sthash.7DIprU3F.dpbs

39. See DePascale, C. (November 8, 2019). "Can We Generate Actionable Student-Level Information from the Summative State Assessment While Honoring Their Design?" *Center for Assessment Centerline.* https://www.nciea.org/blog/making-the-most-of-the-summative-state-assessment/

40. For extended discussion and analysis of this connection see Hammond, Z. (2015). Culturally Responsive Teaching and the Brain. *Thousand Oaks: Corwin Press.*

41. Foundational texts on the nature and measurement of deep learning published by the National Research Council include: Pellegrino, J. Chudowsky, N. and Glaser, R. (Eds.) (2001). Knowing What Students Know: The Science and Design of Educational Assessment. See also Pellegrino, J. and Hilton, M. (2011). Education for Life and Work: Developing Transferable Knowledge and Skills in the 21st Century. *Washington D.C: National Academies Press.*

42. See, for example, ACT Inc.'s 2006 Study called "Reading between the Lines" that illustrated the relationship between overall scale scores on the ACT and test items that were coded separately by the skill/standard measured and by the item's level of complexity/cognitive demand. https://achievethecore.org/content/upload/act_reading_between_the_lines_research_ela.pdf -

43. Harris, D. (2011) Value-Added Measures in Education. *Cambridge: Harvard Education Press.*

44. For a more detailed description of how federal law has codified assessment illiteracy in statutory language, see Marion, S. and Briggs, D. (July 13, 2022). "Just Give Us A Little." *Center for Assessment Centerline*.

45. For specific examples both quantitative and qualitative, see Zavitkovsky, P. (November 2021). "MAP to Nowhere" *CREATE Annual Conference.* https://vimeo.com/673344384

46. For example, AFT and NEA affiliates enthusiastically endorsed Florida Governor Ron DeSantis' proposal to replace year-end accountability exams with a through-year model called FASTER, ostensibly because it would reduce overall testing time. For more details about FASTER, see https://www.washingtonpost.com/education/2021/09/18/florida-desantis-standardized-testing-overhaul/

47. See Marion, S. (October 30, 2019). "Do Interim Assessments Have a Role in Balanced Assessment?" *Center for Assessment Centerline*. https://www.nciea.org/blog/do-interim-assessments-have-a-role-in-balanced-systems-of-assessment/

48. For more extended analysis of how conventional norm-referenced testing results have been dressed up in standards-based clothing, see Section 2 of Zavitkovsky, P. Roarty, D. and Swanson, J. (2016). Taking Stock: How Standardized Test Reports Let Us Down Under No Child Left Behind . . . And How We Can Fix What's Wrong. *Center for Urban Education*

Leadership.
https://urbanedleadership.org/wp-content/uploads/2020/02/UPDATED-FILE-for-TAKING-STOCK-06.06.17-2.pdf

49. For a more detailed description of the tension between validity and efficiency in decision making about assessment policy, see DePascale, C. (October 14, 2019). "The Reality Faced by Innovators of Educational Assessment" *Center for Assessment Centerline.* https://www.nciea.org/blog/the-reality-faced-by-innovators-of-educational-assessments-part-1-what-does-innovative-assessment-really-mean-and-what-has-recent-assessment-innovation-looked-like/

50. Marion, S. (March 25, 2018). "The Opportunities and Challenges of a Systems Approach to Assessment," *National Council on Measurement in Education.* https://www.researchgate.net/profile/Scott-Marion/publication/324214256_The_Opportunities_and_Challenges_of_a_Systems_Approach_to_Assessment/links/5bd891bf4585150b2b91fcb6/The-Opportunities-and-Challenges-of-a-Systems-Approach-to-Assessment.pdf?_sg%5B0%5D=aabDMtEMHSnFKYqvryw5GoeuGli5xxeoXyWRkM6R2-6PFfHlDxUbp_SX9Yznj05A89ho636592viQC6R8267UA.aqbUFIeDxOBPklvPcVwGEW1qNSgKraLimlfAtehIc-4DQynwa3hA4PridelxfGJBD5Nsk6SxRkp4ekk9emKYXg&_sg%5B1%5D=-Myu7EL7vCI6AnWQbDe1Ndew9bPNAuZxReQzMh29uTgi6hSvYNAL_uOehlKAM2-iuMR6kj_4gaVoG1XEUJ4E3wk6hcpn_IdBzYdD03-OhkMR.aqbUFIeDxOBPklvPcVwGEW1qNSgKraLimlfAtehIc-4DQynwa3hA4PridelxfGJBD5Nsk6SxRkp4ekk9emKYXg&_iepl=

51. See, for example, "Accelerate, Don't Remediate" https://tntp.org/publications/view/accelerate-dont-remediate and "Unlocking Acceleration: How Below Grade Level Work is Holding Students Back in Literacy" https://tntp.org/publications/view/teacher-training-and-classroom-practice/unlocking-acceleration, both published by TNTP.

52. Elmore, R. (2004). School Reform from the Inside Out. *Cambridge: Harvard Education Press.* (page 12).

53. "Assessment in Support of Instruction and Learning: Bridging the Gap Between Large-Scale and Classroom Assessment – Workshop Report." (2003). *National Research Council.* https://nap.nationalacademies.org/cart/download.cgi?record_id=10802

54. DePascale, C. (October 14, 2019). "The Reality Faced by Innovators of Educational Assessment" *Center for Assessment Centerline,* https://www.nciea.org/blog/the-reality-faced-by-innovators-of-educational-assessments-part-1-what-does-innovative-assessment-really-mean-and-what-has-recent-assessment-innovation-looked-like/

55. Lagemann, Ellen Condliffe. (2002). An Elusive Science: The Troubling History of Educational Research. *Chicago: University of Chicago Press.*

56. See endnote #39

57. See DePascale, C. (November 8, 2019). "Can We Generate Actionable Student-Level Information from the Summative State Assessment While Honoring Their Design?" *Center*

for Assessment Centerline.
https://www.nciea.org/blog/making-the-most-of-the-summative-state-assessment/

58. See, for example, Marion, S. (April 27, 2022). "Advancing Contemporary Validity Theory and Practice" *Center for Assessment Centerline*.
https://www.nciea.org/blog/advancing-contemporary-validity-theory-and-practice/

59. See 2022 Long-Term Trend Reading and Math Age 9 Highlights Report (September 9, 2022) *National Center for Educational Statistics*
https://www.nationsreportcard.gov/highlights/ltt/2022/

60. The introductory paragraph on the NAEP landing page for item maps and released items reads, "*NAEP item maps are tools that help readers understand student performance. For each assessment, example questions are "mapped" onto the NAEP scale for that subject—with more difficult questions at the top of the map and easier questions at the lower part of the map. The map provides a description of the knowledge or skill needed to answer each question. The location of the questions on the map indicates that students with that score had a high probability of answering the question correctly."* Interactive exploration of NAEP item maps and released items for multiple years, subjects and grade levels can be done at https://www.nationsreportcard.gov/itemmaps/?subj=MAT&grade=4&year=2017

61. In the article cited in endnote #49, Scott Marion describes why "loose coupling" of large-grain standardized assessment and small-grain curriculum-based assessment is critical for preserving the validity of information produced at different levels a balanced assessment system.

62. The Connecticut State Board of Education has summarized SBAC's fixed-form interim assessment system in a reader-friendly online document called "Sensible Assessment Practices for 2020-2021 and Beyond," available at
https://www.google.com/url?client=internal-element-cse&cx=004354853108091474482:sp6telu2lxi&q=https://portal.ct.gov/-/media/SDE/COVID-19/SensibleAssessmentPractices.pdf&sa=U&ved=2ahUKEwj9nNHR5tv5AhUzFlkFHcDxAm8QFnoECAEQAQ&usg=AOvVaw3kPijx3_8HFP4taaJIQjyr

63. See "The Relationship between Student Participation on the Smarter Balanced Interim Assessment Blocks and Student Growth on the Smarter Balanced Summative Assessment – Phase I." (2020). *Connecticut State Department of Education.*
https://portal.ct.gov/-/media/SDE/Performance/Research-Library/RelationBetweenIABParticipationAndSummativePhase1Final.pdf

64. See the Lorrie Shepard article cited in endnote #5.

65. See, for example, A Public Policy Statement. (2013). *The Gordon Commission on the Future of Assessment in Education.*
https://origin-www.ets.org/Media/Research/pdf/gordon_commission_public_policy_report.pdf

66. For more on how tacit competing commitments based on cultural norms and prior personal experiences create powerful obstacles to growth and development, see Kegan and

Lahey's 2001 Harvard Business Review article titled "The Real Reason People Won't Change." https://hbr.org/2001/11/the-real-reason-people-wont-change For a more in-depth discussion, see their book-length text titled, *Immunity to Change* (2009).
For detailed descriptions of Kegan and Lahey's developmental principles are being used to support adult learning and development in schools, see Drago-Severson, E. (2009). *Leading Adult Learning.*

67. In the modern era, these views were made explicit and transparent in a popular text by Harvard behavioral psychologist Richard Herrnstein and political scientist Charles Murray called *The Bell Curve* (1994). Widely rejected by contemporary learning scientists in books like T*he Bell Curve Wars* (1995) and *Outsmarting IQ* (1995), the idea that intelligence is a fixed characteristic that is closely associated with race and class is still quietly preserved in the thinking of many otherwise well-informed people, including many educators. As recently as 2021, Murray continued to make the case for intelligence as a fixed characteristic in a text called, *Facing Reality: Two Truths about Race in America.*
In 2011, Alix Spiegel hosted an 8-minute National Public Radio podcast called "Struggle for Smarts." In this podcast, developmental psychologists Jin Li and Jim Stigler offer compelling evidence for a more culturally determined perspective on intelligence that has very direct implications for one-skill-at-a-time teaching in American classrooms.
https://vimeo.com/738770508