Teachers College Record
The Voice of Scholarship in Education

A History of Achievement Testing in the United States
Or: Explaining the Persistence of Inadequacy

This essay offers a historical analysis of the structural and cultural aspects of American education that help explain the durability of standardized testing in the face of more than a century of persistent criticism.

ETHAN HUTT is an assistant professor at the University of Maryland – College Park. His research focuses on the history of quantification in American schools and on the relationship between schools, law, and education policy.

JACK SCHNEIDER is Assistant Professor of Education at the College of the Holy Cross and Director of Research for the Massachusetts Consortium for Innovative Education Assessment. His latest book is *Beyond Test Scores: A Better Way to Measure School Quality* (Harvard University Press in 2017).

Background/Context: For more than a century standardized achievement tests have been a feature of American education. Throughout that time critics of standardized tests have argued that their use has detrimental effects on students, schools, and curriculum. Despite these critiques, the number and uses of standardized tests has increased steadily. Though a great deal of research has focused on the technical design of tests, the history of individual tests, and general critiques of testing, there is little research that helps explain the continued use of standardized tests in American education despite near constant criticism.

Purpose/Objective/Research Question/Focus of Study: This essay develops a framework for understanding a basic paradox in the history of standardized testing in American education: the durability of standardized testing in the face of persistent criticism. Seeking to address this paradox, the essay asks why tests have persisted and proliferated despite the fact that students dislike taking tests, educators believe that tests distort the learning process, and experts challenge the validity of test results.

Research Design: This essay involves a historical analysis of structural and cultural aspects of American education that help explain the particular uses and durability of testing.

Conclusions/Recommendations: First, we identify three master critiques of standardized tests: distortion, waste, and misclassification. We find that despite these persistent critiques, four important contextual features of the American education system help explain the continuous hold standardized tests have had on American education: the fact that the American education system is decentralized, avowedly meritocratic, publicly funded, and central to aspirations of upward mobility. These contextual factors, along with the historically contingent development of testing expertise, testing culture, and development of testing infrastructure, provide a framework for understanding the persistence of testing. Together, these factors create a dynamic system in which critiques of tests led not to the elimination of testing, but to its further elaboration and evolution.

Executive Summary

In 1922, Lewis Terman responded to critics of his efforts to quantify intelligence with the sarcastic quip that "in the interest of freedom of opinion there ought to be a law passed forbidding the encroachment of quantitative methods upon those fields which from time immemorial have been reserved for the play of sentiment and opinion." His point was simple: in the evaluation of schools, Americans would have to choose between science and anecdote. And he worried that tradition, rather than measurement, would win the day.

A century later, quantitative measurement is a standard part of life in American schools—from standardized tests of student achievement, to the use of resulting scores for the purpose of measuring the quality of schools and teachers. One might say that a new tradition has developed. Yet despite this seeming triumph, the advance of testing has never been total nor has it driven "sentiment or opinion" from schooling.

Here, however, is the dilemma. For, though critiques of standardized testing are as old as the tests themselves, and have often been relatively compelling, pushback against testing has done almost nothing to disrupt the practice. How is it, then, that testing has won the day, with regard to measurement in education, while remaining defective in the eyes of so many Americans?

Seeking to address this paradox, this essay asks why tests have persisted and proliferated, despite the fact that students dislike taking tests, educators believe that tests distort the learning process, and experts challenge the validity of test results. The essay identifies four major historical critiques of standardized tests: distortion, waste, and misclassification. We find further that despite these persistent critiques, four important contextual features of the American education system help explain the continuous hold standardized tests have had on American education: the fact that the American education system is decentralized, avowedly meritocratic, publicly funded, and central to aspirations of upward mobility.

These contextual factors, along with the historically contingent development of testing expertise, testing culture, and development of testing infrastructure, provide a framework for understanding the persistence of testing. Together, these factors create a dynamic system in which critiques of tests led not to the elimination of testing, but to its further elaboration and evolution.

Ultimately, the analysis reveals that the history of testing is not a story about bad guys (test experts and their numbers) versus good guys (teachers and parents valiantly resisting). Rather, it is a story about cross-cutting motivations and shifting alliances. Consequently, it suggests that any meaningful future change with regard to educational testing will come not from demonization or prohibitions but from reforms addressing the multiple functions served by standardized tests.

A History of Achievement Testing in the United States
Or: Explaining the Persistence of Inadequacy

Introduction

As long as we have had tests we have had critiques of tests. Writing in one of the first issues of the *Journal of Educational Research*, Professor Matthew Willing offered a critique that would fit just as comfortably in the latest edition of *Education Week*, complaining that "except in the most simple kinds of school work a mere score is not particularly illuminating, even though the unit of measure is defined" (Willing, 1920, 194). And as his contemporary, Walter S. Monroe, observed: "recent critical studies of standardized tests have proved that these new measuring instruments are far from perfect … Many of them are not 'scientific' measuring instruments unless the meaning of the term 'scientific' is materially modified" (Monroe, 1924, 255). To put it simply: Americans have never been naïve about testing and its effects.

Yet despite persistent criticism from a range of stakeholders, standardized tests have endured. In fact, as criticism has mounted and expanded over the past century, testing has continued to penetrate deeper into the fabric of American education. Today, policymakers cannot imagine a single year passing without issuing tests and circulating their results to the public.

This would seem to point to a paradox largely unaccounted for in the current literature on testing: Why is it that tests persist and proliferate when students dislike taking tests, educators believe that tests distort the learning process, and experts challenge the validity of test results?

To ask this question is not to deny that we know a considerable amount about the history of achievement testing. Scholars have written histories of city- and state-specific tests (e.g. Beadie 1999; Reese 2013). They have examined federal and international efforts—like No Child Left Behind, or the OECD's PISA test—to engineer systems for measuring student achievement (e.g. Koretz, 2002; McGuinn, 2006; Trohler et al, 2014). And they have looked beyond the state to construct histories of prominent tests like the SAT (Lemann 2000), the GED (Hutt 2014), and the AP exam (Schneider, 2009). This robust literature has provided a great deal of insight into the general allure of measurement, as well as into the specific political and cultural work performed by individual tests (e.g. Dorn, 2014).

This literature, however, has largely ignored a basic question: Why has standardized testing flourished in American education, despite unrelenting criticism?

If we eliminate explanations that are inapplicable (e.g. authoritarianism) or implausible (e.g. ignorance, misinformation), then any effort to answer this question must consider the enduring presence of testing in light of the various purposes testing has served and the varying ways it has been experienced. After all, if testing were merely a mechanism of oppression, endured unwillingly by all stakeholders, it would have long ago been jettisoned.

Before proceeding any further, it is important that we be clear at the outset about what we are

and are not attempting here. This essay looks exclusively at local and state-level achievement tests in the United States—surveying them broadly, across time and space. Insofar as that is the case, it will not offer a comprehensive account of specific achievement tests, their particular meanings, uses, or receptions by experts (e.g. Buros Institute of Mental Measurement, 1938), nor will this essay detail the work of test developers or testing advocates. Given the question at the heart of this inquiry, it is also worth stating explicitly that we are not seeking to evaluate the positions of testing advocates or opponents. Instead, our concern is with explaining why standardized testing has persisted, whatever the ultimate truth about its value.

Given that historians tend to focus on documenting and explaining change over time, it may seem odd that we are focusing on constancy rather than change. In choosing to focus on the durability of standardized testing, we are not denying its evolution or minimizing the importance of studying that evolution. Nor are we attempting to advance an ahistorical argument in which standardized tests sit as a stable, free-floating technology hovering over changing school systems below. Far from it. We are, however, trying to draw attention to what we think are striking continuities despite the clear shifts in the usage and ubiquity of standardized tests–continuities we think require both examination and explanation.

The durability of the practice of standardized testing is a remarkable thing when one considers the diversity and idiosyncrasy of the American educational landscape, making this puzzle interesting in its own right. But at a time when Americans appear to be genuinely rethinking the role that standardized testing should play in their school system, it becomes more important than ever to understand the peculiar dynamic that achievement tests have introduced into our system. This essay seeks to provide just such a framework.

Critiques of Testing

Standardized achievement tests—that is, sets of uniform examinations and their corresponding answer keys, issued across a population of students—have never been perfect. In fact, since standardized tests first appeared in 1845, their defects have been well known and frequently identified (Reese 2013). Additionally, the arguments against testing have been relatively consistent in highlighting the negative impacts of tests—offering what might seem a sufficient timespan for those arguments to take hold.

We begin this essay by highlighting three of the general criticisms of tests. The aim here is not to be comprehensive—detailing every criticism of tests ever mounted—nor is it to imply that these criticisms are indisputable, universally recognized, or mutually exclusive. Instead, it is to highlight broad categories under which critiques of testing can be organized and to illustrate that relatively similar arguments against standardized testing have been mounted for generations. For, only by understanding the nature of these criticisms can we begin to understand the surprising staying power of tests.

While not all critiques of tests will fit into this taxonomy, we think that the criticisms levied by scholars, policymakers, and the public can usefully be grouped into three general types: distortion, waste, and misclassification.

1. Distortion

The first general kind of criticism is that tests distort the educational process, particularly with regard to teaching. In 1927, for instance, a *New York Times* story relayed one educator's concern that, despite improvements in the Regents' examinations, "the pupils who cram and learn by rote a few typical 'thought' questions can still get a good mark on them" ("Blame Regents Test" 1927, 14). Another educator, in the same story, suggested that "It is quite possible to drill for an examination and to pass a large number of pupils with high ratings without giving any breadth of outlook or grasp of underlying principles." (Ibid., 14). Fifty years later, Albert Shanker, President of the American Federation of Teachers, would accuse tests of a similar distortion arguing that with everyone from the student to the school board's reputation implicated in test results "schools are devoting more and more time [to] teaching kids strategies for filling in blanks and choosing answers to multiple-choice questions. This destroys much of the value of these tests, which only tell you something if they are an independent measure of what the student knows" (1988). Now, several decades further into the high stakes accountability movement, the centrality of tests to federal education policy has turned the issue of "curriculum narrowing" into something of a battle-cry against accountability testing (e.g. Nichols & Berliner 2007).

At its core, this critique is an epistemological one. Tests, the argument runs, provide at best a partial and at worst a distorted critique of what goes on in schools. This issue is compounded when the results of testing are used to interpret the relative success of schools, teachers, and students. The higher the stakes associated with the test results, the more incentive there is for those involved to direct more of their time, attention, and effort toward catering to the demands of the test, thereby magnifying distortion effects. Over time, the number of aspects of the school system that have been subjected to evaluation by tests has increased; but the basic concern about the effects of tests on what was being evaluated—students, schools, curriculum, teachers—has endured.

2. Waste

The second enduring criticism is that testing is wasteful, particularly with regard to classroom time. In 1930, Dr. Henry Linville, president of the New York teachers union, called the Regents tests a "continuing waste of childhood that is appalling to contemplate" ("Linville Assails Test," 1939, 19). And those kinds of claims—the *New York Times* Editorial Board lamented the practice of "substituting test preparation for instruction"—have only grown stronger as tests have eaten up more of the school year ("Trouble with Testing Mania," 2013). According to one recent estimate, testing consumes 1.6 percent of all school time (ASCD, 2015).

Of course, there is also the expense of testing. A 2012 report, for instance, calculated that school systems across the country spend at least $650 million a year in contracts with test vendors (Chingos, 2012). Though different on its face, this definition of waste is rooted in the same core position—a perception that tests, to the extent that they measure anything at all, merely duplicate information already known to local stakeholders. And the cost of that duplication, in the form of

"wasted time" or "wasted dollars," has often been unacceptable to stakeholders. Thus, the waste critique provides a means of challenging the need for and value of testing without directly contesting the test results themselves.

3. (Mis)classification

The third general criticism of standardized tests does engage with the specific results of tests. Specifically, criticisms in this vein argue that tests classify, and often *mis*classify, students, schools, and, recently, teachers. The result of these classifications, many have argued, send deeply problematic messages and have profound consequences. Emblematic of this genre of complaint, one educator wrote in 1927 that test "results are used to make invidious comparisons" ("Blame Regents Test," 1927, 14). Well over a generation later, the same concern was being expressed. "The performance of students on standardized tests and their general level of academic achievement is probably determined in large measure by the general characteristics of the student body," one scholar observed, concluding that "it is obviously ridiculous to make many comparisons between schools" (Snider, 1963). At the current moment, the use of tests and test scores have produced a growing concern over the identification of students as "below grade level" in any number of skills—especially when identification leads to placement in academic tracks (Oakes & Guiton, 1995)—and the identification of teachers as ineffective (e.g. *Lederman v. King*, 2014).

Such objections challenge the meaning of the scores themselves. In this respect, the misclassification critique is similar to the distortion critique. The misclassification critique, however, distinguishes itself through attention to the comparative dimensions of standardized test scores, as well as to the faulty logic underlying such comparison. The creation of league tables based on test scores—a practice that has been part of standardized testing since the beginning (Reese 2013)—and the comparisons they invite, imply both a common game and an even playing field. Yet though such an idea meshes well with the nation's democratic ethos, the history of discrimination and inequality in American schooling and society leaves wide leeway for critique.

Persistent Core Critiques

Though we do not have the space to detail all of the ways arguments against standardized testing have been marshalled over time, this limited evidence is clearly suggestive of the durability and persistence of these core critiques. What is remarkable, then, is that standardized tests have not only endured these challenges, but that they have thrived in spite of them. Indeed, standardized tests are as central to the operation and evaluation of American schools as they have ever been.

To make this observation is to reassert the paradox of standardized testing: How have standardized tests flourished despite consistent criticism? In order to formulate an answer, we need to begin, as we do in the next section, with a consideration of structure and purpose of schools in America and what role tests serve within them.

Toward a More Coherent Theory

Our effort to develop a more robust framework for explaining the staying power of local and state-run standardized testing proceeds in two parts. The first is a consideration of the crucial contextual factors—the features inherent to American public education—that we believe must be accounted for in explaining the use of tests. These factors do not, in and of themselves, offer explanations. But they do describe an environment that is hospitable to standardized tests. In that sense, they might be understood as necessary but not sufficient conditions for the growth and persistence of standardized testing.

After establishing the context in which testing took root, we then offer three possible explanations for the enduring use of tests in the United States. Given the contingent nature of history, in which even the smallest of events cannot be completely discounted, these explanations are not by any means complete. Yet, insofar as they do collectively offer an answer to the question at the heart of this inquiry, they may prove to be sufficiently comprehensive. Additionally, given the limitations of an article-length essay, these explanations are not buttressed by mountains of evidence. Nevertheless, we believe that the evidence included can bear the weight of the analysis.

The Importance of Context

There are four crucial interrelated features of the American education system that are central to understanding the continuous hold standardized tests have had on the American school system. On their own, these features do not constitute an explanation; but they help enrich the soil in which standardized testing has taken root and grown.

1. A Decentralized System

The first of these features is the decentralized organization of American schooling. No other public school system in the world has developed with such little uniformity of purpose, organization, or curriculum.

Early on, this decentralization provided the flexibility necessary to tailor taxation levels and curricular content according to local political and cultural tastes—fostering broad-based support for public schools in the absence of large government infrastructure. However, as the public school system moved from parochial archipelagos of schools to an integrated system, the lack of uniformity created major organizational challenges. It is not surprising, then, that professional educators perceived administrative standardization as one of their chief tasks. In the absence of centralized governmental oversight, superintendents relied on their professional networks and standardized tools to bring coherence to a geographically and numerically massive school system that was growing at a rapid pace. These tools included regulations for standardized teacher qualifications and courses of study, as well as standardized school surveys in which standardized measures of student achievement became a key feature (Steffes, 2012, 37-45).

The desire to create and implement tools that could aid in system building was clearly part of the impetus behind the embrace of standardized testing in the early 20th century. Indeed, advocates of standardized testing consistently stressed the value of the common language and uniform benchmarks that tests provide. The New York Regents test, for instance, which represents the earliest statewide examination, was designed in part to prevent private academies from gaming the state funding system by advancing students without any regard for their achievement levels Beadie, 1999). Nearly a century later, not much had changed. As a 1971 *New York Times* story reported: "Defenders of the [Regents] tests say they provide an objective, uniform standard of attainment, without which it would be difficult, if not impossible, to maintain a check on the quality of education from one locality to the next" (Stevens, 1971, 41).

This basic sentiment has been be echoed by educators across the country, particularly at higher levels of the system where students from different schools are more likely to be educated together. Explaining the organizational value of tests to the *New York Times* in 1916, a college official described the challenge of integrating the mélange of entering students: "Our aim has been in part to contribute to the solution of an education problem. Students, on entering college as freshmen, are a heterogeneous lot. Some come from first-class high schools, others from poor ones. Some have already acquired in their homes habits of accuracy or of perseverance, others are not so fortunate. Some know how to study, others have been carried along without much effort of their own" (Waugh, 1916, SM12). In short, tests allowed the college—in his opinion, at least—to create apples-to-apples comparisons that otherwise would have been impossible.

Advocates of standardized testing have also made the case that the tests are essential in communicating *beyond* the school. Superintendents, particularly, tended to seize on the power of test scores as providing a common language for communicating with the public. As the author of a 1922 guide to the use of standardized tests explained: "Test scores furnish the common language, for anyone can understand what is meant by saying that our schools in Smithville are a year ahead of most schools of America in, say, arithmetic, and a year or two years behind others in music or French or manual training" (Geyer, 1922, 11-12). But tests would not only furnish common language; they would also provide explanations about school progress that were rooted in data. As the authors of another 1922 textbook wrote: "Test results constitute incontrovertible facts, so often needed by the superintendent in a campaign of education of public opinion" (Pressey & Cole, 1922, 34-35). The argument that the eclecticism of American education necessitates the production of common metrics has only accelerated in recent years with the rise federal policy more focused on standards based accountability, the adoption of Common Core State Standards, and state testing consortia (e.g. Brown, 2015; Duncan, 2009).

2. A Meritocratic System

Education is characterized by uncertainty. Its central aim—human improvement—is nebulous and context-dependent. And insofar as that is the case, it is as difficult to track as it is to prescribe. This creates a serious challenge for anyone attempting to communicate about relative success or failure in this venture—a challenge that is compounded by the considerable variation that exists in the American school system. The need for technologies that could provide a

common metric of achievement, therefore, were highly prized, particularly as historian Joseph Kett has argued, "in a republic founded on the principles of equal rights and advancement by merit" (2012, 9).

Tests did not bring the obsession with meritocracy to the schools. Instead, they enabled a particular worldview—specifically, that schools could identify and reward talent—that already existed by the mid-nineteenth century. Arguments in favor of introducing standardized tests, then, tended to focus on the objectivity, fairness, and transparency that they would ostensibly bring to the task of evaluating pupils. As one observer explained in 1916, "fairness in the award of honors, justice in determining failures and dismissals, and incitement of the student to better work can be attained only to the extent to which a common standard for the awarding of marks is understood, accepted, and acted upon" (Canning, 1916, 196). In a system saturated by the notion of meritocracy, tests were a welcome addition.

In the eyes of those committed to the notion of meritocratic achievement, the influence of tests would not merely reward the "best" students. Instead, tests would motivate *all* students to work harder. As the author of a 1922 text wrote, "The scores on standardized tests supply to the pupil's goal just this definiteness. When such scores are represented by a simple graph, say with one line showing the given pupil's attainment in these tests and another line the attainment of the average American child of this age or grade who has taken these tests, then the pupil has his strong and weak points set before him in a manner that is perfectly definite and objective" (Geyer, 1922, 8). This was true of grades doled out by teachers (Schneider & Hutt, 2014). And it soon became true of standardized tests, which were increasingly being used to affect a student's future (Lemann, 2000).

3. A Tax-Supported System

The above examples highlight the value of test scores to the school system internally. But standardized tests were equally valuable as *external* forms of communication—specifically, as a device for transparency and public accountability in a tax-supported system. Prior to the introduction of standardized tests, public accountability often took the form of public exhibitions. Yet while this could satisfy the immediate curiosity of those in attendance, it was difficult to tell whether these performances were objective measures of student success or merely carefully choreographed performances (Reese, 2013). It was similarly difficult to know whether students attending schools in other locales were demonstrating comparable levels of proficiency and achievement.

The introduction of standardized tests, then, took a big step in surmounting this difficulty, and did so almost immediately after their introduction. The results of the first standardized tests administered in the Boston area, for instance, were quickly seized upon as a definitive way to compare the relative merit and standing of the local school districts (Reese, 2013).

As school systems around the country grew and found themselves demanding larger shares of local tax revenues, superintendents found that this kind of relative comparison was invaluable. Mocking the lack of evidence for most claims of effectiveness, one author quipped: "If one could

read all the small town papers of any given state for one year, he would probably find three-fourths of them claiming that their home town had the best schools in the state." These empty assertions, the author argued, stood in contrast to the "new" method of a superintendent comparing his own school with other schools "by means of a standard test" (Alexander, 1919, 52-53).

4. A Critically Important System

The comparisons enabled by standardized tests would likely carry no more weight than so many magazine "Top 10 Lists" were it not for the centrality of schooling to American life—our fourth contextual factor. The fact that the American welfare state has been, somewhat uniquely, built with its school system at the center (Labaree, 2012; Steffes, 2013), means that test scores comparisons move out of the realm of harmless novelty and into the realm of consequential "knowledge."

The surest path to upward mobility in the United States is to succeed in school, which has always made the American system highly competitive with regard to a student's relative standing among his or her peers (Labaree, 1997). In the nineteenth century, this manifested most clearly in the struggle to gain admission to high school. By the mid-twentieth century, the competition was over spots at top colleges and universities (e.g. Karabel, 2005; Wechsler, 1977). Thus, it is easy to see how attention to test scores among those in education could radiate outward, encompassing larger groups of constituents—parents, schools, districts, and even states.

As the market for credentials becomes more competitive, parents want to live in school districts that produce the best test scores; districts justify expenditures by pointing to the achievement of students; and state policymakers justify existing policy or the need for reform by pointing out areas of deficiency. The considerable stakes involved for the actors in each of these concentric rings surrounding the school ensures that it is always in someone's best interest to put forth test score information as a definitive sign of excellence. Only those at the very top of the pyramid, whose excellence is accepted axiomatically, find themselves with a few additional degrees of freedom (Schneider, 2009).

Since the 1960s, the declining public faith in the capacity of schools to do their job effectively has only made Americans more obsessive about issues of uniformity, fairness, and transparency. Indeed, the standards movement of the past several decades has arisen in response to the belief that tests can address those concerns (Schneider, 2011). Thus, unlike the other three contextual factors, which have been relatively stable, this final factor—the centrality of schooling in American life—has grown increasingly prominent over time.

Potential Explanations

We can explain a great deal about the American inclination toward standardized testing by examining the nature of the educational system itself. Yet while contextual factors—decentralization, meritocracy, public funding, and the perceived importance of education—help

us understand the initial impetus toward standardized testing, they do not resolve the puzzle at the heart of this essay. After all, Americans have long ceased being naive about the limitations of tests.

Thus, we must still explain the *persistence* of standardized tests in the face of nearly constant, if sometimes low-level, criticism. Indeed, a purely structural explanation for the persistence of standardized testing appears particularly incomplete given the large shift in organizational structure over time (Kaestle & Lodewick, 2007). Below, then, we discuss how the American education system was populated by competing professions, shaped by the politics of knowledge, and increasingly characterized by a culture of numbers and evaluation, all of which helped make standardized achievement testing a fixture.

1. Testing Expertise/Professional Politics

Education has always been a field with relatively porous professional boundaries. The proximity and familiarity of the school system tends to mean that almost everyone has an opinion on how things are going and how things should be. Within the educational realm very few spaces have been effectively carved out and successfully defended against these types of persistent challenges to educators' professional knowledge (Abbott, 1988; Mehta, 2013).

Among the few notable exceptions to this has been testing expertise. Though assessment has always been a part of schooling, relatively early on in its history, testing became the "owned" domain of experts—evidence of a highly successful kind of professional politics (Labaree, 2004; Lagemann, 2000). Using the language of science and objective measurement, these experts—first psychologists, and later psychometricians—developed the kind of abstract knowledge that resisted easy challenge. American psychologists, in particular, were interested in availing themselves of quantitative measures that served as visible extensions of their expertise and useful bulwarks against charges of personal bias (Carson, 2007). This does not mean these testing experts were immune from criticism or buffered from sometimes intense ideological and methodological debates, but, increasingly, participation in these debates required knowledge and expertise beyond the reach of the general public (e.g. Haertel, 2013; Lindblom & Cohen, 1979).

Policymakers and administrators, too, increasingly relied on experts to help them make determinations about standardized tests. Though testing experts might disagree about particular test formats or about the uses of scores, they almost universally agreed that the tests had value. Debates over whether achievement was best understood as a function of an intelligence that was unitary or multi-faceted resulted in the development of additional tests to measure additional student abilities or achievements. Hence the creation and application of standardized tests to measure such unlikely skills as creative thinking (e.g. Torrance, 1972). Likewise, debates about the practical value of normed test score results facilitated developments in criterion referenced scoring beginning in the 1970s (Koretz, 2008). In each case, critiques originating either within the field or from outside resulted in the creation of new kinds of tests or new ways to express test results. This dynamic reflected not only the perceived value, even if imperfect, of the information provided by tests, but also the larger belief that what happened within schools could be brought into view through testing. Of course, as the century went on there was also a growing

number of testing experts capable of developing tests for these varied purposes.

All of this explains how testing experts maintained their foothold in education. But it does not explain how they initially *secured* that foothold.

As with any historical phenomenon, causality is multiple. Yet there is something of an origin moment for the testing movement in American education. This nascent field got a tremendous boost from the U.S. military's manpower placement needs in World War I. As one commentator observed, "The fact that two or three hundred young men who have for several months been working in the psychology division of the Army are now about to be discharged offers an unusual opportunity for city schools to obtain the services of competent men as directors of departments of psychology and efficiency, for such purposes as measuring the results of teaching and establishing standards to be attained in the several school studies" (Claxton, 1919, 203-204). By the early 1920s, public school officials in big cities began to consider it a functional necessity to organize departments of research and measurement (Cardozo, 1924, 797). This new set of testing professionals was well suited to help increasingly bureaucratic, hierarchically structured districts solve their chief organizational challenge: the efficient and systematic sorting of individual students. It also facilitated the increasingly common efforts to differentiate curriculum and track students based on a combination of their presumed (or measured) natural abilities and "likely" life trajectories.

The incorporation of this new brand of expert into public education became so common that complaints about their roles and distinctions about the "new" and "old" guard of testing became a topic of regular debate. As early as 1921, professional educators were complaining about the newly minted testing experts who, unlike their forerunners, lacked a basic understanding of how schools functioned. Traditionally, wrote one critic in the *Journal of Educational Research*, experts "have in the main been men and women of considerable actual experience in teaching who later added their measurement equipment. They have thus been able to make valuable interpretations and applications of their measurements. Now, however, the educational institutions are sending out many youngsters highly trained in measurement technic [sic] but sometimes woefully ignorant of the school work which they are attempting to measure" (Alexander, 1921, 354).

Testing professionals were also increasingly employed by corporate entities that channeled great energy into establishing their expertise. As one observer noted in 1924: "Since 1915 more than 300 standardized tests have been devised, and the number is constantly increasing." The reason? "The distribution of standardized tests … has become a commercial enterprise of considerable magnitude" (Monroe, 1924, 254). The viability of the testing as a commercial enterprise facilitated not only the training of more testing experts but also the search for new markets and applications of their skills. One historian has compared the testing experts of the period to "a migrating gaggle of geese, alighting on one pond before flying to another" (Kett, 2013, 159). The availability of resources to attract their skills and the hospitable organizational conditions—the demand for efficiency in identifying and sorting students—ensured that many testing experts became permanent residents of schooling environments.

Criticism of standardized testing, of course, continued. Yet that only made experts more valuable

to the organizations for which they worked. In order to remain valid, new forms had to be created and new tests had to be devised. This led one educator to observe that the "need for an increase in the number of forms of the different tests is urgent, and there is still room for contributions along this line on the part of expert educational workers. The process of constructing and standardizing the different tests is technical and laborious and the layman has neither the training nor the time for the development of first-class instruments of measurement" (Woody, 1924).

This professionalization gave rise to more rigorously constructed tests whose creators had more ambitious designs for their use. In 1923 the Stanford Achievement Test debuted as the first set of standardized tests specifically designed for a mass market. As the authors explained, "blanks and materials for testing the schools of an entire city [could] be ordered in a single letter, and they [would] come in a single shipment" (Kelley et al., 1922, 3). In 1929, E.F. Lindquist created the Iowa Testing Programs, which would subsequently produce the nationally used Iowa Tests of Educational Development and which served as the skeleton for the first Test of General Educational Development (GED) (Hutt, 2014).

This was not just a function of elementary and secondary education. In this same period, many of the most prominent colleges began to draw more heavily on the growing capacities of standardized tests. As with administrators in elementary and secondary educational settings, college administrators found standardized tests particularly adept at solving organizational challenges as the demand for college access grew. While many colleges had long relied on a system of pre-certifying individual high schools and then accepting graduates from those schools without further examination, this system proved too inconsistent and unwieldy (Schudson, 1972). In the absence of standardized courses of study, curricula, or transcripts it became increasingly difficult for colleges to determine the most qualified and capable applicants. As a result, a growing number of the country's colleges began entrusting a portion of their admissions process to the SAT—thereby intertwining their admissions process with testing experts (Lemann, 2000). In 1947, the Cooperative Testing Service and the College Entrance Examination Board (CEEB) gave way to the more commercially ambitious Education Testing Service (ETS) as well as a host of other commercial publishers including such as the World Book Company. These companies had their own symbiotic relationship with the growing number of universities training graduate students in educational testing and psychometrics (and, increasingly, with test preparation "experts" like Stanley Kaplan). In 1959, Lindquist's Iowa Testing Program at the University of Iowa, produced the American College Testing (ACT) Program as a competitor to the SAT that could be used as a tool for both admission and placement. By 1960, roughly a million students were taking the SAT or ACT annually—a sign of how central the testing industry had become to mediating this key educational transition.

As testing companies like ETS grew stronger and more consolidated, their standing in education circles grew and their influence in education policy became more pronounced. In 1968, for example, the University of California system began to require SAT scores from all applicants. Critics were quick to recognize testing companies as "major factors [that] have contributed significantly to the increased use of testing in public schools," including "a continuous growth and development of standardized testing at all grade levels made possible by the increased resources and professional competence of the major test publishers" (Snider, 1963). These

sentiments were echoed in a *New York Times* story in 1964 reported that "The publication of school exams alone has come to be an industry involving more than $25 million annually … the number of standardized grade-school and high-school tests given each year comes to 100 million … This does not include the additional two million multiple-choice tests given for college admissions and scholarship competitions" (Cook, 1964, 51).

The experts, in short, were protected not only by their expertise and their usefulness to school administrators and policymakers, but also by their centrality to a rapidly expanding industry.

Acknowledging this symbiotic relationship between a robust business in standardized test production and the value of tests in solving the organizational and administrative challenges of schools, is not intended to diminish the very real contributions tests have made to improving the quality of schools and the learning opportunities of children. As we detail more in the next section, the rigor, precision, and rhetorical power of test scores made them central to policy debates and legal disputes concerning nearly every facet of schools, from school funding to curriculum effectiveness to remedial education. The ability of school officials and researchers to utilize tests to address these issues made them an increasingly indispensable part of the school system—a development nicely captured by the increasing prevalence of the phrase "data-driven decision making" (e.g. Marsh, McCombs, & Martorell, 2010).

This indispensability often facilitated—if reluctantly—new developments and refinements in psychometrics. For instance, the passage of Title I and Congress' desire for statistics that would allow it to compare the relative effectiveness of different compensatory programs led to a demand for nationally normed standardized tests, to an increased capacity to compare norms across tests, and to new metrics for reporting student growth within this framework. This provided an important impetus for the federally funded Anchor Study (Bianchini & Loret, 1974))—designed to allow direct comparisons of scores across the most commonly used tests—and the creation of the normal-curve equivalent (NCE) as a test score metric and an evaluation tool (Tallmadge & Wood, 1976). Likewise, the decision to move from static measures of student achievement (e.g. percent proficient) to measures of student growth trajectories renewed debates among psychometricians about the best way to create a vertical test score scale to capture student growth (e.g. Briggs & Weeks, 2009; Tong & Kolen, 2007). And, the switch in policy emphasis from proficiency to growth requires tests that are designed to be equally robust at all points along the score scale not just around the proficiency cut point. The use of these kinds of test scores as inputs for attempts to evaluate specific aspects of the school like teacher quality, for example, have introduced still more considerations in the design, selection, and use of tests (e.g. Haertel, 2013; Lockwood et al., 2007; Shavelson, et al., 2010).

While these developments represent welcome innovations in the effort to provider a firmer empirical grounding for education policy and school system decision-making, one challenge is the increasing sophistication necessary to accurately interpret these metrics. Though psychometricians and researchers are usually quite circumspect about what can and cannot be claimed on the back of existing metrics and available data, the possibility for misinterpretation by the broader public remains. The more statistical calculation involved in the production of a test score or an associated metric, the more difficult it becomes for laypeople to understand the measures at anything but the broadest conceptual level. While few parents in the nineteenth

century likely needed an explanation to understand percent correct or class rank, entire books are now written to explain to educators and the public what they need to know and understand about tools like value-added measures (e.g. Harris, 2011). The more the public is reliant on experts to explain and interpret the meaning of the measures used to guide school decision making and inform public policy, the more integral these experts become to our schools.

The overarching point is that, throughout this history, the locus of control over standardized testing has moved increasingly away from teachers and into the hands of experts whose job it has become to design, implement, and control the use and interpretation of tests. It is difficult to understand the foothold and influence tests have had without accounting for the politics of professional expertise and increasingly entrenched interests of those who controlled the technology.

2. Testing Culture

Another important reason why tests have maintained a foothold in education is because testing, in addition to being a technical and informational exercise, became a cultural one as well. This cultural component can be thought of as consisting of two distinct but interrelated parts. The first is the gradual acceptance of standardized tests as an important part of the work of the operation of schools and test results as a commonsense way to understand what goes on in schools. The second part is that, as testing became an accepted part of schooling, it increasingly signaled the social importance of particular subjects—some were worth testing in, and some not. The more commonplace testing has become, the more the baseline of public conversations about public schools seem to accept testing as a default practice, restricting the debate to questions of how much and how often.

In the early twentieth century, as standardized tests began to be used more commonly, a culture of modernity shaped the way educators and the public understood the process of testing. Testing advocates of the period focused on the ostensible upside of testing—its usefulness in creating universal standards and presumably objective ratings. But they also tended to emphasize the degree to which standardized tests were an expression of education as a technical and efficient affair that drew on the latest scientific techniques. As one author wrote in 1921, "measurement work is now generally regarded in leading educational circles as one of the earmarks of an up-to-date school system," and brought with it a certain "prestige value." Teachers and communities, he noted, may be somewhat skeptical of the work. But they could certainly be overcome, he argued, by the suggestion "that those opposing measurement work are thereby in danger of being considered unprogressive" (Alexander, 1921, 347-48).

The speed at which standardized tests became available for easy integration into school systems is impressive. By 1917, tests were "available in hundreds of school systems in the United States" and had been "printed and reprinted and placed before teachers very generally" (Gray, 1917, 767). And by 1922 there were over 250 tests commercially available (Buckingham, 1919, 46). Throughout the Progressive Era, tests benefited from a culture that valued science and the scientific. "Measurements of achievement, through the use of educational tests, have come to be a common feature of the public schools," wrote one advocate in 1922. "During 1921 over two

million of American school pupils were tested by educational tests," he added—tests that had been "scientifically selected and scientifically constructed" (Hines, 1922, 37).

One result of the proliferation of testing was that it became an increasingly normal and accepted part of schooling. While in the late nineteenth century, periodicals were full of articles decrying the harmful mental and physical effects of "over-study" brought on by the introduction of competitive standardized testing (Reese, 2013, 195), those concerns had almost disappeared by the 1920s. By then, testing was no longer a brand new technology, but rather a widespread practice. In that cultural moment, medicalizing anxieties about standardized testing no longer held sway.

By the 1950s, children entered school in a different world—a world in which their parents and grandparents had taken standardized tests not only in schools, but also upon entry into an ever growing number of professional settings including the military (Kett, 2014). Insofar as that was the case, Americans became increasingly comfortable with the idea that standardized tests, and their results, could govern elements of life from the educational to vocational and the professional. Thus, as one observer noted, a "naively trusting spirit" about testing—one "that prevailed in the Nineteen Twenties"—began "seeing a revival" in the early 1960s. "What greater triumph can a parent report," he asked, "than 'My child is in the third grade, but he scored grade 6, second month, in reading?'" (Brodkin, 1968, SM12).

Acceptance of testing was also shaped by the fact that teachers—key stakeholders in the educational system—were not only increasingly likely to have taken standardized tests, but also to have done relatively well on them. Thus, despite some discomfort around testing, teachers at various points have been found to support testing, at least in its broad contours. As a 1965 report put it: "A person's test taking experience and his attitudes toward tests are not isolated parts of his total experience" (Brim, 6). And as a 1980 study found, "data do not support the popularly held notion promoted by test critics that 'most' teachers feel too much standardized testing takes place in schools" (Beck, 7).

The advent of systems analysis in the early 1960s elevated standardized test scores to a key output in efforts to determine the production functions of school systems or features of school systems (e.g. Kershaw & McKean, 1959). Test scores as key outputs were further institutionalized with the passage of the Elementary and Secondary Education Act (ESEA) in 1965. Title I of ESEA included a provision—the first of its kind—requiring local school districts to conduct annual evaluations of program effectiveness involving, among other things, "objective measurement of educational achievement" (ESEA, Title I, sec. 205 (5)).

Though the exact testing and evaluation requirements evolved over time, the general influence of these required evaluations is hard to understate. A federal government study of testing in American schools conducted in 1992 concluded that the testing requirements of Title I had "helped create an enormous system of local testing" (U.S. Office of Technology Assessment, 1992, 85). In 1987, over 1.6 million students were tested as a result of Title I evaluation requirements—testing that had to be conducted twice a year in order to provide the necessary information about learning growth (U.S. Office of Technology Assessment, 1992).

Many districts had so many Title I students that it became prudent to test all students rather than try to distinguish Title I and non-Title I beneficiaries. Of course, this distinction ultimately became moot with the reauthorization of ESEA through NCLB, which mandated testing for all students in tested grades and subjects. Though the information provided by Title I-initiated standardized testing rarely provided the kind of cut and dry effectiveness data that lawmakers hoped for (e.g. McLaughlin 1975), lawmakers and advocates quickly realized that the test score information could serve important symbolic and political purposes. As an advisory committee on Title I evaluation observed, "In the aggregate, standardized test scores can also be a readily grasped symbol of success or failure. 'A steep trend line on a graph can be strong ammunition in political struggles over the quality off schools.'" (Advisory Committee on Testing, 1993, 8; US Technology Assessment, 1992, 54).

The increasing familiarity and comfort with viewing the school system through test scores can likewise be seen in the introduction of the National Assessment of Educational Progress (NAEP) in 1969. A contemporaneous development with ESEA and its Title I evaluation requirements, NAEP was designed to provide a general achievement measure of American students from select age and demographic groups via a single test score (Vinovskis, 1999). In the same way that economists could track the general health of the economy through the newly created Gross Domestic Product (GDP) statistic, policymakers could track the NAEP metric as a way of keeping tabs on the general health and effectiveness of the American school system as a whole. These new uses of test scores for analytical and diagnostic purposes, well removed from individual students or classrooms, reflected not only the increasing sophistication of test design and policy analysis, but also an increasing familiarity with testing and comfort with a testing culture.

The further penetration of testing culture did not mean that anyone began to enjoy taking tests or that they believed tests to be perfect. But the more familiar standardized tests became, the more they became part of the accepted experience of schooling and, in turn, a way of signifying the importance of a skill, subject, or transition. For instance, after World War II the GED was created primarily as a way to honor the military service of veterans and symbolically validate the academic worth of their military experience rather than as a way of assessing their scholastic abilities—a fact underscored by the elaborate organizational structure created to develop and administer the test so as to maintain civilian oversight of the test, as well as by a passing score set close to the level of random guessing (Hutt & Stevens, forthcoming). Similarly, in the 1970s, lawmakers sought to emphasize the importance of getting back to basic skills by implementing minimum competency testing and high school exit examinations (Baker, Myers, & Vasquez, 2014). While a useful political signal for politicians trying to show that they are holding the line on high standards, the reality of exit examinations in practice is that they rarely improve academic achievement or standards (Holme et al., 2010). Rather than hold the line, most states lower standards or eliminate the exit examination altogether in order to avoid the political fallout from having a large number of students fail to receive their diplomas. When California eliminated its exit examination, the governor signed a bill granting a diploma to all those who completed their course work but had failed the test that year (Mason, 2015).

No Child Left Behind, likewise, reflected the desire to express the importance of standards and achievement—at least in math and reading—by requiring that schools test schools annually in

grades 3-8, and once in high school. The fact that so many commentators considered the goals set by NCLB—100 percent proficiency by 2014—unattainable is less relevant for our purposes than the fact that NCLB sought to enforce the commitment to educational improvement, and the demonstration of that improvement, by way of standardized tests. Complaints about the "narrowing" of the curriculum that resulted from the push to meet the NCLB's proficiency targets often resulted in calls to expand the number of subjects, even among general critics of standardized testing (e.g. Ravitch & Chubb, 2009). Likewise, the identification of non-cognitive skills that are associated with improved school and life outcomes has led to calls that they too be subject to standardized tests (e.g. Zerkine, 2016; cf. Duckworth, 2016) The centrality of standardized testing to the operation of a major piece of legislation like NCLB, in short, is a sign of how much testing has become part of the basic culture of American schooling.

Even the most recent outspoken opposition to standardized testing serves to highlight how deeply the culture of standardized testing is embedded in American schools. For instance, when parents sought to express frustration with what they considered to be the excessive amount of testing done in public schools by having their children "opt out" of testing, they met stiff resistance from a wide range of organizations including prominent civil rights groups like the NAACP and mainstream school organizations like the National Parent Teacher Association and the National Association of Secondary School Principals. These groups countered efforts to oppose testing and thereby reduce the value of the information gleaned from test scores (by making samples non-representative) by arguing that these efforts "rob us of the right to know how our students are faring" (Leadership Conference on Civil and Human Rights, 2015). Whether one agrees with this characterization, it is clear from this statement that test scores have become so much a part of the culture of schooling that, for some, they are inextricably linked to claims for equity, justice, and reform. Similarly, the debate over the testing requirements included in the Every Student Succeeds Act (ESSA) was framed by reports and discussions about what constituted *too much* testing, not whether the federal government should mandate testing at all (e.g. Council of Great City Schools, 2015).

3. Testing Infrastructure

Standardized tests were strengthened by the work of experts, who consistently made the case that flawed tests had been replaced by better ones. And Americans increasingly came of age having been exposed to standardized tests, establishing cultural norms around the enterprise. Whether or not these factors had been in place, however, local and state achievement tests would likely have persisted because of how deeply integrated they became in policy and structures. Not long after their introduction, the tests had been baked into the system; it could not function without them.

What do we mean by this? Consider what a "system" is—a set of interrelated components that form a collective whole. In education, the system is both vertical, with students progressing from kindergartens to college, and horizontal, in that students from California to Maine ostensibly learn comparable material despite much that separates them. Governance in the educational system is also vertical and horizontal. Actors at the federal and state level—at a great distance from the school—are presumed to possess sufficient knowledge and control to reasonably act upon the schools. And the many stakeholders in public education—parents, teachers,

policymakers, colleges and universities, community members—are imagined to possess similar understandings of individual schools, even if their particular values and priorities are different.

In order for this system to function, vast amounts of information are required—information about what is being taught and how much is being learned. Though it is possible to collect it, a serious challenge is presented by the volume and density of that information. Teaching and learning are not easily distilled; they are complex and multifaceted enterprises. Additionally, several million teachers work with tens of millions of students—representing a staggering multiplier effect.

Test scores have served as a solution to this problem.

There are other solutions, as well. One notable example is the age-graded school. Without age-grading—the process of sorting students by birthdates—the work of placing students at different levels in the system would require massive coordination of educator judgments about student ability. Instead, stakeholders have accepted a "good enough" approach that replaces teacher judgment—a compromise that keeps the system functioning.

Test scores were not entirely different. For, imperfect though they were, they were viewed as a useful tool for addressing a core dilemma. As one 1920 journal article made clear, the primary aim of adopting standardized tests was to create a simple and portable currency for evaluation that could be applied widely across contexts. Test scores, the author wrote, had the advantage of being "concrete and graphic enough to be clearly understood by teachers, pupils, and parents." Not lost on him was the fact that they could also be used to begin structuring more of a system in education. They could, he argued, "be used as the beginning of a continuous record to measure progress of pupils and ability of teachers'" (Brooks, 1920, 730).

Others saw test scores as a solution to different challenges in the system-building quest. A 1930 *New York Times* story, for instance, reported that the superintendent of the high schools in New York City had long struggled to coordinate the efforts of secondary schools with elementary schools. In seeking to determine "what is to be done about the children who do not survive in the high schools and what are the causes for their failure," he turned to test scores, which would presumably pinpoint the area to be addressed ("Pupils Here Lead," 1930, 21).

Not surprisingly, then, tests were incorporated into the emerging structures at district, state, and national levels. Each time a system-building challenge emerged, standardized tests were an available technology deemed suitable for the problem. And each time they were adopted, they were sealed even more tightly into the underlying framework of the system.

By mid-century, standardized tests had become an integral part of school structures. A 1943 recommendation to University of Michigan pre-service teachers, for instance, suggested that aspiring educators "become familiar with the best standardized tests in your major and minors." And they were told explicitly that their training would include "administer[ing] a standardized test" and "interpret[ing] the pupil responses" (Schroling, 1943, 98). Similarly, a 1954 report by the New York State Education Department observed that the Regents test was serving "as a basis for admission to college and as a supervisory device for maintaining and improving the quality of instruction in the major secondary school subjects" (Buder, 1954, 31).

Certainly there were critics who lamented the degree to which tests were being baked into the structures of the educational system. Yet even those critics often recognized the usefulness of tests for the purpose of system-building. One observer, for instance, wrote in 1963 that "It is regrettable that many schools appear to appraise teaching competence by making comparisons of groups of pupils on the results of standardized achievement tests. When teachers realize that their teaching effectiveness is being evaluated by this method, many of them find ways of teaching for the tests, thus reducing possible contributions the tests might make toward genuine program improvement" (Snider, 1963).

Structures, of course, are not static. They change over time. Yet rarely are they disassembled and rebuilt from scratch, at least not in enterprises as vast as public education. Thus, while they do change, their transformation is evolutionary—altering, rather than replacing, the original structure.

This is true in the case of standardized testing. New tests were designed, new ways of testing were developed, and new ways of analyzing and disseminating test results were devised. Yet the general practice of standardized testing remained at the core of policy and governance structures. In 1970, for instance, a report from the New York State Education Department (1970) noted that computers were being used to tabulate standardized test scores, explaining that "the distributions of scores are processed by computer, and reports which summarize the results in conveniently interpretable form are returned to each school and central office" (10). It was a significant shift from the way tests were managed just a decade earlier. Yet it was hardly a departure in the core practice—of administering a standard test to students and then treating the results as valuable information.

Automation was a particularly notable aspect of the change that testing structures underwent—notable because it was a major departure from previous practices, and also because it enabled a veritable revolution in the way test scores were used. As scoring machines, and eventually complex computer data systems, were increasingly relied upon, they made it possible not only to expand the scope of testing, but also to interpret the resulting data in new ways.

These new technologies, however, did not replace standardized testing. Which is not to say that new technologies *could not have* replaced standardized testing. Rather, it is to say that standardized testing was so much a part of the existing system that it was not even questioned. The incorporation of computers into educational data collection and interpretation—beginning in earnest in the 1960s and increasing ever since—created an inertia around the use of test scores to provide to communicate the general health of the system. Hence the introduction of a NAEP "thermometer" for the education system and the national outcry to the reported decline of average SAT scores in the 1970s, which was interpreted as evidence of the declining quality of American schools (e.g. National Academy of Education, 1978). More recently, the continuation of yearly testing has been a major point of contention in NCLB reauthorization with supporters arguing that at least some of the value of past data is tied to the continued collection of data in the future. How else, the argument runs, will we know if the schools of the coming decades are as good as those of the previous ones, if we do not maintain test scores as data streams?

As David Tyack and Larry Cuban have observed, the general trend in the organization of American schooling has been toward the steady accretion of features of the system with new elements being added in addition to, not in place of, existing ones (1995). This is no less true with tests and test scores. Over time, a very large infrastructure has developed to support the creation, implementation, and interpretation of test scores. This infrastructure is held in place by an increasing number of state and federal laws and is populated by a variety of professionals (e.g. psychometricians, policy analysts, state officials, etc.) in a range of organizational contexts (e.g. districts, universities, state departments of education). The more elaborate this infrastructure has become, the more test scores have become part of the "grammar of schooling"— elements, whether hated or loved, that constitute the core of the school system (Tyack & Cuban, 1995).

A Paradox Shaped by Equal and Opposing Forces

Standardized tests have faced valid criticisms over the past century. Among others, the consistent themes of distortion, waste, and misclassification represent substantive critiques of the use and value of standardized tests. And there has seemingly been sufficient time for those critiques to penetrate the consciousness of various stakeholder groups to create a massive anti-testing movement.

Yet tests were never dislodged from the school system. Such criticisms never led to policymakers reversing their positions on testing, never led teachers to strike, and never spurred widespread outrage. If anything, testing become more central to the operation of schools over time. This, in many ways, is the story of a revolution that never came.

Why?

Our answer is that, though there have been pockets of discontent and consistent objection to testing practice, there has never been sufficient fuel for a movement. Though engagement with quantified information is sometimes framed as a binary between trust and distrust of numbers (e.g. Porter, 1996), we think the history of standardized testing in American schools requires us to incorporate a much more dynamic picture. Following Power's (2004) general argument about counting and control, we argue this story is not about a hard line between trust and distrust, but rather, an evolving boundary demarcating areas of trust, distrust, and ongoing contention. The more standardized testing was considered a normal school practice, the more its byproducts—determinations of competency, test scores, etc.—could be incorporated into school governance and incorporated into second-order calculations to evaluate other aspects like teacher quality or resource efficiency.

This evolution was facilitated in important ways by the technical evolution of the tests themselves. Testing experts were not immune from criticism, but instead responded to critiques by aligning tests in response to policy changes, articulating new rationales for testing, and developing new ways of communicating score results to the public. Despite considerable evolution, the stability of standardized testing as a practice fostered a popular cultural viewpoint that testing is a foundational and inescapable part schooling. And, because policymakers

understood that the tests helped the system run, the side-effects of testing were often dismissed as simply the cost of doing business.

Consequently, potential energy never became kinetic energy. Each time the criticisms grew, they were matched.

These oppositional forces have been in constant interaction with each other across several generations. But as should be clear from this kind of dynamic model, test makers have not won every battle over the use of tests. In many cases they have had to give ground or cede whole areas completely. IQ tests, once the primary tool for student placement, are now scarcely used and only for very specialized purposes. Exit examinations, once introduced with lofty rhetoric about raising academic standards and bolstering the value of the diploma, have been found decades later to have no effect on achievement (e.g. Reardon, Arshan, Atteberry, & Kurlaender, 2010); in some cases, they have even been mothballed (Tucker, 2015). Attempts to increase the rigor of tests by increasing the difficulty or raising cut scores—and, inevitably lowering the pass rate—have often been scuttled in response to public outcry and political pressure to maintain the status quo (e.g. Edelman, 2014; Herzenhorn, 2005; Medina, 2010).

Even when these kinds of adjustments are made, however, the tests have usually just evolved and endured. Consider, for example, how in the last decade the preference and emphasis on proficiency scores has given way to measures of student "growth"—a shift that has attempted to inure standardized tests from critique of unfairness by changing the operative mode of interpretation. It is a move that also has the effect of realigning the constituencies of support. Schools that previously scored at the bottom on Adequate Yearly Progress (AYP), can now claim pride of place at the top of student growth metrics (e.g. Dyslin, 2012; Turque 2011).

Writing in 2017, we would be remiss not to mention the growing "opt-out" movement, which may alter the trajectory of our conclusion. Recently, for instance, roughly 200,000 students boycotted testing in New York—a significant show of protest. Still, roughly one million students took the tests as expected. Indeed, the largely critical response, including the charge that parents are selfish or irrational to withhold their children from being tested (Hess, 2015), only underscores how much test scores have become part of the "common sense" of schooling—to raise questions is put oneself well out of the mainstream. And despite some policy adjustments, or calls for testing moratoriums, the tests roll on.

Standardized testing is a manifestation of the inherent contradictions within, and necessary compromises required by, the American educational system. To those hoping for an end or reduction in the use of standardized tests in American education, this might seem like a rather depressing message to hear from a pair of historians: testing has long been a central practice in American schools, and will likely remain so. The structural and cultural factors that keep testing in place have been well matched to the available critiques. Insofar as that is the case, victories against standardized tests have come largely at the margins, and have done little to halt the general proliferation of testing.

So what is the aspiring reformer or policymaker to make of this history? To this we have several replies. The first, perhaps a touch unsatisfying, is that we still need more systematic examination

of the interaction between standardized testing and its critics. Understanding is the first step toward effective reform. We have tried here to lay out a dynamic framework that we hope others will take up, apply, and deepen as they investigate specific uses of standardized testing.

Second, as we have tried to make clear, the history of testing is not a story about evil testing experts imposing their will on teachers and students. Rather, it is a story about cross-cutting motivations and shifting alliances. This suggests that any meaningful change will come not from demonization or prohibitions, but from reforms that address the multiple functions served by standardized testing. Thus, for example, it is possible to imagine the sophisticated sampling techniques used for NAEP being taken up in place of the testing of every student—something that would serve the demand for accountability, and for standardized, comparable information, but doing so in a way that responds to concerns about over-testing.

Third, given the many constituencies for test production and test score information consumption, it seems worth considering whether one way to diminish the influence of tests is to develop constituencies around other forms of school information. Historically speaking, it is only relatively recently that facts have come to mean quantitative information and school accountability to mean test scores. There are many other ways one can measure schools and there are likewise many other ways we can provide an account of how teachers and students spend their time in class. It seems unlikely that this will provide a total solution to the problem, as people will likely gravitate toward the evidence—in whatever form— that tends to reinforce their existing view. Still, more constituencies engaged with more diverse measures at least creates the opportunity for more nuanced assessments and conversations about school.

As we observed at the outset, as long as there have been standardized tests, there have been critiques of tests. In 1922, Lewis Terman responded to critics of his efforts to quantify intelligence with the sarcastic quip that "in the interest of freedom of opinion there ought to be a law passed forbidding the encroachment of quantitative methods upon those fields which from time immemorial have been reserved for the play of sentiment and opinion" (116). While, to our knowledge, no such law has ever passed, neither has it been necessary. Despite the aspirations of Terman and generations of test-makers since, the advance of testing has never been total, nor has it driven "sentiment or opinion" from schooling. Instead, standardized tests—though embedded in legislation, well supported via political and organizational infrastructures, and part of the cultural currency of schooling—remain a source of constant and evolving struggle between their proponents and their detractors.

Bibliography

Abbott, A. (1988). *The system of professions: An essay on the division of expert labor*. Chicago: University of Chicago Press.

Advisory Committee on Testing in Chapter 1. (1993). *Reinforcing the Promise, Reforming the Paradigm. Report of the Advisory Committee on Testing in Chapter 1*. Washington, DC: Government Printing Office.

Alexander, C. (1919). *School statistics and publicity*. Boston, MA: Silver, Burdett.

Alexander, C. (1921). Presenting educational measurements so as to influence the public favorably. *The Journal of Educational Research, 3*(5), 345-358.

Association for Supervision and Curriculum Development (2015, March). Policy points. Retrieved from http://www.ascd.org/ASCD/pdf/siteASCD/publications/policypoints/Testing-Time-Mar-15.pdf

Beadie, N. (1999). From student markets to credential markets: The creation of the Regents examination system in New York State, 1864-1890. *History of Education Quarterly*, 1-30.

Beck, M.D. (1980). Student and teacher attitudes toward standardized tests: A summary of two surveys. Paper presented at the National Institute of Education Invitational Conference on Test Use (Washington, DC).

Bianchini, J. C., & Loret, P. G. (1974). *Anchor Test Study Supplement. Final Report. Volumes 1-31, Project Report*. Washington, DC: Office of Education.

Bowles, S. (1970). Toward an education production function. In W.L. Hanson (Ed.), *Education, income, and human capital* (pp. 11-61). New York: Columbia University Press.

Blame regents test for faulty teaching (1927, November, 14) *New York Times*, 14.

Briggs, D. C., & Weeks, J. P. (2009). The impact of vertical scaling decisions on growth interpretations. *Educational Measurement: Issues and Practice, 28*(4), 3-14.

Brim, Jr., O.G., Neulinger, J., & Glass, D.C. (1965). Experiences and attitudes of American adults concerning standardized intelligence tests. Technical Report No. 1 on the Social Consequences of Testing. New York: Russell Sage Foundation.

Brodkin, A.M. (1960, July 3). The question of 'standardized test.' *New York Times*, SM18.

Brooks, S. S. (1920). Using standardized tests in rural schools for grading purposes. *The Journal of Educational Research, 2*(4), 729-735.

Brown, E. (2015, January 25). What happens when the Common Core becomes less…common. *Washington Post*. Retrieved from: http://www.washingtonpost.com/local/education/what-happens-when-the-common-core-becomes-less--common/2015/01/25/33b8eb58-a2bf-11e4-b146-577832eafcb4_story.html

Buckingham, B.R. (1919). *Bureau of Educational Research announcement, 1918-1919*. Urbana, IL: University of Illinois Press.

Buder, L. (1954, September 28). Educators split on regents test. *New York Times*, 31.

Buros Institute of Mental Measurements. (1938). *Mental measurements yearbook.* Ipswich, Mass.: EBSCO Publishing. http://search.ebscohost.com/login.asp?profile=web&defaultdb=mmt

Bushaw, W. J., & Lopez, S. J. (2012). A nation divided: The 44th annual Phi Delta Kappa/Gallup Poll of the public's attitudes toward the public schools. *Phi Delta Kappan*, *94*(1), 8-25.

Canning, J. B. (1916). The meaning of student marks. *The School Review*, 196-202.

Cardozo, F. L. (1924). Test and measurements in public schools. *School and Society*, *20*, 797-798.

Carson, J. (2007). *The measure of merit: Talents, intelligence, and inequality in the French and American republics, 1750-1940*. Princeton University Press.

Chingos, M. (2012). Strength in Numbers: State Spending on K-12 Assessment Systems. Washington, D.C.: Brookings Institute. Retrieved from: http://www.brookings.edu/research/reports/2012/11/29-cost-of-ed-assessment-chingos

Cook, J. (1964, September 24). Unrelenting pressure on students brings varied assessment of tests for intelligence and ability. *New York Times*, 51.

Council of Great City Schools. (2015). *Student testing in America's great city schools*: *An inventory and preliminary analysis.* Retrieved from http://www.cgcs.org/cms/lib/DC00001581/Centricity/Domain/87/Testing%20Report.pdf.

Dorn, S. (2014). Testing like William the Conqueror: Cultural and instrumental uses of examinations. *Education Policy Analysis Archives*, 22(119): 1-15.

Duckworth, A. (2016, March 26). Don't grade schools on grit. *New York Times*. Retrieved from https://www.nytimes.com/2016/03/27/opinion/sunday/dont-grade-schools-on-grit.html.

Duncan, A. (2009, May 29). Excerpts from Secretary Arne Ducan's remarks at the National Press Club. Retrieved from http://www.ed.gov/blog/2009/06/excepts-from-secretary-arne-duncan%E2%80%99s-remarks-at-the-national-press-club/

Dyslin, A. (2012, May 22). Local schools shine in new evaluations. *Mankato Free Press*. Retrieved from http://www.mankatofreepress.com/news/local_news/local-schools-shine-in-new-evaluations/article_4ff1ef73-bc21-56ae-8339-dba1d1177ebc.html

Edelman, S. (2014, August 17). NY 'fixed' Common Core tests—and scores surged. *New York Post*. Retrieved from http://nypost.com/2014/08/17/dept-of-ed-officials-adjust-proficiency-thresholds-for-common-core/

Elementary and Secondary Education Act. (1965). Pub. L. 89-10. 20 U.S.C. ch. 70.

Geyer, D.L. (1922). Introduction to the use of standardized tests. Chicago: Plymouth Press.

Gray, W.S. (1917). Educational writings. *The Elementary School Journal* 17(10), 767.

Grodsky, E., Warren, J. R., & Kalogrides, D. (2009). State high school exit examinations and NAEP long-term trends in reading and mathematics, 1971-2004. *Educational Policy, 23*(4), 589-614.

Harris, D.N. (2011). *Value-added measures in education: What every educator needs to know*. Cambridge, Mass.: Harvard Education Press.

Haertel, E.H. (2013). Reliability and validity of inferences about teachers based on student test scores. Princeton, NJ: Educational Testing Service. Retrieved from https://www.ets.org/Media/Research/pdf/PICANG14.pdf

Herzenhorn, D.M. (2005, June 25). State lowers passing score for a regents math test. *New York Times*. Retrieved from http://www.nytimes.com/2005/06/25/nyregion/state-lowers-passing-score-for-a-regents-math-exam.html

Hess, F.M. (2015, May 5). Opt-out parents have a point. *US News and World Report*. Retrieved from http://www.usnews.com/opinion/knowledge-bank/2015/05/05/parents-opting-out-of-common-core-tests-have-a-point

Hines, H. C. (1922). Measuring the achievement of school pupils. *American School Board Journal, 65*, 37-38.

Hutt, E. L. (2014). The GED and the rise of contextless accountability. *Teachers College Record, 116*(9), 1–20.

Hutt, E.L., & Stevens, M.S. (Forthcoming). From soldiers to students: The Tests of General Educational Development (GED) as diplomatic measurement. *Social Science History*.

Kaestle, C. F., & Lodewick, A. E. (Eds.). (2007). *To educate a nation: Federal and national strategies of school reform*. Lawrence, KS: University Press of Kansas.

Karabel, J. (2005). *The chosen: The hidden history of admission and exclusion at Harvard, Yale, and Princeton.* Boston: Houghton Mifflin.

Kett, J. F. (2012). *Merit: The history of a founding ideal from the American Revolution to the twenty-first century*. Ithaca, NY: Cornell University Press.

Koretz, D. M. (2002). Limitations in the use of achievement tests as measures of educators' productivity. *Journal of Human Resources,* 752-777.

Lagemann, E. C. (2002). *An elusive science: The troubling history of education research*. Chicago: University of Chicago Press.

Labaree, D. F. (1997). *How to succeed in school without really learning*. New Haven: Yale University Press.

Labaree, D. F. (2004). *The trouble with ed schools*. New Haven: Yale University Press.

Labaree, D.F. (2012). *Someone has to fail: The zero-sum game of public schooling.* Cambridge, MA: Harvard University Press.

Leadership Conference on Civil and Human Rights (2015, May 5). We oppose anti-testing efforts. Retrieved from: http://www.civilrights.org/press/2015/anti-testing-efforts.html

Lederman v. King. (2016). Supreme Court, New York, Index No. 5443-14.

Lemann, N. (2000). *The big test: The secret history of the American meritocracy*. New York: Macmillan.

Lindblom, C. E., & Cohen, D. K. (1979). *Usable knowledge: Social science and social problem solving* (Vol. 21). Yale University Press.

Linn, R. L. (2005, December). Adjusting for differences in tests. In *A Symposium on the Use of School-Level Data for Evaluating Federal Education Programs, National Academies, Board on Testing and Assessment*.

Linville Assails Tests by Regents, (1930, February 26). *New York Times*, p. 19.

Lockwood, J. R., McCaffrey, D. F., Hamilton, L. S., Stecher, B., Le, V.-N., & Martinez, J. F. (2007). The sensitivity of value-added teacher effect estimates to different mathematics achievement measures. *Journal of Educational Measurement, 44*(1), 47–67.

Marsh, J. A., Sloan McCombs, J., & Martorell, F. (2010). How instructional coaches support data-driven decision making: Policy implementation and effects in Florida middle schools. *Educational Policy, 24*(6), 872-907.

McDonnell M., L. (2015). Stability and Change in Title I Testing Policy. *The Russell Sage Foundation Journal of the Social Sciences, 1*(3), 170–186.

McGuinn, P.J. (2006). *No Child Left Behind and the transformation of federal education policy, 1965–2005*. Lawrence, KS: University Press of Kansas.

McLaughlin, M. W. (1975). *Evaluation and reform : the Elementary and secondary education act of 1965, Title I*. Cambridge, Mass. : Ballinger Pub. Co.,.

Medina, J. (2010, October 10). On New York school tests, warning signs ignored. *New York Times*. Retrieved from http://www.nytimes.com/2010/10/11/education/11scores.html?pagewanted=all

Mehta, J. (2015). *The allure of order: High hopes, dashed expectations, and the troubled quest to remake American schooling*. Oxford University Press.

Monroe, W. S. (1924). Written examinations versus standardized tests. *The School Review, 32*(4), 253-265.

National Academy of Education. Committee on Skills; United States. Dept. of Health, Education Welfare. Assistant Secretary of Education, T. B. (1978). *Improving educational achievement: Report of the National Academy of Education, Committee on Testing and Basic Skills to the Assistant Secretary for Education*. Washington, D.C.: The Academy.

National Association of Secondary School Principals. Opt-out policies for student participation in standardized assessments. Retrieved from https://www.nassp.org/who-we-are/board-of-directors/position-statements/opt-out-policies-for-student-participation-in-standardized-assessments?SSO=true

National Parent Teacher Association (2016, Jan. 21). National PTA Board of Directors adopts position statement on student assessment and opt-out policies. Retrieved from http://www.pta.org/about/newsdetail.cfm?ItemNumber=4719

New York State Education Department (1970). *New York state pupil evaluation program: School administrator's manual*. Albany, NY: New York State Education Department.

Nichols, S. L., & Berliner, D. C. (2007). *Collateral damage: How high-stakes testing corrupts America's schools*. Cambridge, MA: Harvard Education Press.

Oakes, J., & Guiton, G. (1995). Matchmaking: The dynamics of high school tracking decisions. *American Educational Research Journal, 32*(1), 3-33.

Porter, T. M. (1996). *Trust in numbers: The pursuit of objectivity in science and public life*. Princeton, NJ: Princeton University Press.

Power, M. (2004). Counting, control, and calculation: Reflections on measuring Management. *Human Relations, 57*(6), 765–783.

Pupils here lead in regents test (1930, February 24). *New York Times*, p. 21.

Pressey, S. L., & Cole, L. (1922). *Introduction to the use of standard tests: a brief manual in the use of tests of both ability and achievement in the school subjects*. World Book Company.

Ravitch, D., &Chubb, J. (2009). The future of No Child Left Behind. *Education Next*. Retrieved from http://educationnext.org/the-future-of-no-child-left-behind/.

Reardon, S. F., Arshan, N., Atteberry, A., & Kurlaender, M. (2010). Effects of failing a high school exit exam on course taking, achievement, persistence, and graduation. *Educational Evaluation and Policy Analysis, 32*(4), 498-520.

Reese, W. J. (2013). *Testing wars: A forgotten history*. Cambridge, MA: Harvard University Press.

Reilly, P. (2016, January 15). PARCC is a test worth taking. *The State Journal Register*. Retrieved from http://www.sj-r.com/opinion/20160115/pam-reilly-parcc-is-test-worth-taking-.

Schorling, R. (1943). Experiences in directed teaching. *The University of Michigan School of Education Bulletin, 14*(6), 98.

Schudson, M. (1972). Organizing the 'meritocracy': A history of the College Entrance Examination Board. *Harvard Educational Review, 42*(1), 34-69.

Schneider, J. (2009). Privilege, equity, and the Advanced Placement program: Tug of war. *Journal of Curriculum Studies, 41*(6), 813–831.

Schneider, J. (2011). *Excellence for all: How a new breed of reformers is transforming America's public schools*. Nashville: Vanderbilt University Press.

Schneider, J., & Hutt, E. (2014). Making the grade: a history of the A–F marking scheme. *Journal of Curriculum Studies, 46*(2), 201–224.

Shanker, A. (1988, April 24). Exams fail the test. *New York Times*.

Shavelson, R.J., Linn, R.L., Baker, E.L., Ladd, H.F., Darling-Hammond, L., Shepard, L.A., Barton, P.E., Haertel, E., Ravitch, D., & Rothstein, R. (2010). Problems with the use of student test scores to evaluate teachers. Washington, DC: Economic Policy Institute.

Snider, G. (1963). The secondary school and testing programs. *The Teachers College Record, 65*(1), 57-67.

Steffes, T. L. (2012). *School, society, and state: A new education to govern modern America, 1890-1940*. Chicago: University of Chicago Press.

Stevens, W.K., (1971, June 18). Once-feared regents test face hazy future. *New York Times*, 41.

Tallmadge, G. K., & Wood, C. T. (1976). *ESEA Title I Evaluation and Reporting System: User's Guide*. Mountain View, CA: RMC Research Corporation.

Terman, L.M., (1922, December 27). The great conspiracy or the impulse imperious of intelligence tests, psychoanalyzed and exposed by Mr. Lipmann. *The New Republic*, 116-120.

Torrance, E. (1972). Predictive validity of the Torrance tests of creative thinking. *The Journal of Creative Behavior*, *6*(4), 236-262.

Tong, Y. & Kolen, M.J. (2007). Comparisons of methodologies and results in vertical scaling for educational achievement tests. *Applied Measurement in Education*, 20(2), 227-253.

Tröhler, D., Meyer, H.-D., Labaree, D. F., & Hutt, E. L. (2014). Accountability: Antecedents, power, and processes. *Teachers College Record, 116*(9), 1-12.

Trouble with testing mania (2013, July 14). *New York Times*. Retrieved from http://www.nytimes.com/2013/07/14/opinion/sunday/the-trouble-with-testing-mania.html

Tucker, J. (2015, July 10). Bill would give diplomas to students who fail state exit exam. *San Francisco Chronicle*. Retrieved from http://www.sfchronicle.com/education/article/Bill-would-give-diplomas-to-students-who-fail-6378867.php

Turque, B. (2011, December 6) District unveils first ranking of charter schools. *Washington Post*. Retrieved from http://www.washingtonpost.com/local/education/district-unveils-first-ranking-of-public-charter-schools/2011/12/06/gIQAJkYraO_story.html

U.S. Office of Technology Assessment. (1992). *Testing in American Schools: Asking the Right Questions.*. Retrieved from https://eric.ed.gov/?q=Testing+in+American+schools%3a+Asking+the+right+questions&id=ED340770

Vinovskis, M. (1999). *Overseeing the nation's report card: The creation and evolution of the National Assessment Governing Board (NAGB)*. Washington, D.C.: National Assessment Governing Board.

Waugh, K.T. (1916, January 2). A new mental diagnosis of the college student. *New York Times*, SM12.

Wechsler, H. S. (1977). *The qualified student: A history of selective college admission in America*. New York: Wiley.

Willing, M. H. (1920). The encouragement of individual instruction by means of standardized tests. *The Journal of Educational Research*, *1*(3), 193-198.

Woody, C. (1924). The meaning, use and development of educational measurements. *The Teachers College Record*, *26*(2), 95-106.

Zernike, K. (2016, February 29). Testing for joy and grit? Schools nationwide push to measure students' emotional skills. *New York Times*. Retrieved from https://www.nytimes.com/2016/03/01/us/testing-for-joy-and-grit-schools-nationwide-push-to-measure-students-emotional-skills.html.