2016-06-01

Understanding Author Academic Disciplinary Background to Direct A More Effective Use of Standardized Testing Within the School Community

Joseph Jensen
Brigham Young University

Understanding Author Academic Disciplinary Background to Direct A More Effective Use of Standardized Testing Within the School Community

Joseph Jensen

A dissertation submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of

Doctor of Education

Steven J. Hite, Chair
Julie M. Hite
E. Vance Randall
Pamela R. Hallam
David Boren

Department of Educational Leadership & Foundations

Brigham Young University

June 2016

ABSTRACT

Understanding Author Academic Disciplinary Background to Direct
A More Effective Use of Standardized Testing
Within the School Community

Joseph Jensen
Department of Educational Leadership & Foundations, BYU
Doctor of Education

Since the days of Horace Mann, standardized testing has been used as a control mechanism by policy makers to determine who makes decisions about what will happen in public schools. A dynamic struggle for educational control and governance has continued since that time between the local, state, and federal levels. This struggle for control puts school principals in a unique organizational position where they are expected to use standardized tests within the school community with teachers, students, and parents to improve education but at the same time manage external accountability mandates from district, state and federal levels of governance. To further complicate the testing picture, multiple stakeholders from diverse backgrounds write about standardized testing, making the testing literature complex and seemingly contradictory. These competing narratives create distractions and confusion in the standardized testing debate. The purposes of this archival study was to (a) explore the literature about standardized testing to find patterns in the narratives that are being told in the disciplines of education, policy, economics, psychology/psychometry, and history; and, (b) analyze those narratives to determine what major themes emerged from each discipline so that principals can better understand the testing landscape. In each source we tracked first-author characteristics, one of which was author academic disciplinary background—the academic discipline the author primarily trained in during their formal education. With a better understanding of these disciplinary narratives, a principal is in a stronger position to understand and communicate more effectively about standardized testing within their school community, as well as manage the demands from external influences. This study used NVivo software to organize and analyze text from 147 documents from authors representing the five different disciplinary backgrounds. These documents were written by proponents and critics of testing. Patterns emerged that confirm that using standardized testing as a control mechanism is one of the most common themes in the testing literature. Each narrative is influential in unique ways, but the most important finding of this study shows that the two *loudest* narratives are those from education and policy. Both disciplines often focus on the reality that standardized testing is used as a control mechanism. Authors from the discipline of education wrote about this topic from a reactive and defensive position. Educators dominate the professional literature, but don't have nearly as strong of a voice in the mainstream media. On the other hand, the analysis demonstrated that authors in the realm of public policy write about standardized testing in a proactive and assertive tone, and they have a stronger voice in mainstream media. Understanding all five narratives can enable principals to more effectively and proactively take control of the standardized testing narrative in their own school community.

Keywords: standardized testing, accountability, educational leadership

ACKNOWLEDGMENTS

I've been an endurance junkie for decades. In the application essay for this doctoral program I compared my experience to running the Boston Marathon in 2012 to my preparation for this doctoral experience. I wrote, "There are many similarities between my marathon experience and entering BYU's EdD program. The doctoral degree also has a mystique about it, but like the Boston Marathon, I know that my preparation and commitment for the last 18 years are the keys to accomplishing it successfully. It is an endurance event, and endurance is what I do." This EdD has proved to be a worthy endurance challenge. A marathon now looks like a 5K in comparison. It even makes the Leadville 100 bike race look like a cruise on the beach. But it has been worth it. I concluded my application essay with, "Just like my preparation for the Boston Marathon, I am ready to strap on my running shoes for this event. It is three years instead of three hours, but I look forward to the 2016 finish line. Earning an EdD from BYU will train me to run further and faster down this and other paths that lead to greater student achievement." I had no idea how true these words would be. I look forward to where new paths will take me.

Thanks to Dr. Steven Hite for his guidance in both content and logistics as my doctoral chair. He is no micro-manager, and he always helped move me forward in this process. His insights are wise—his vision of the process and product is clear. I also appreciate the insightful contributions of the rest of my committee, and the faculty in BYU's EDLF department. Our cohort was closely bonded and I will cherish my memories with them. Most importantly I appreciate the patience of my wife Julie and my children as I ran this race.

Last, I salute all those who have wrestled with the complexity of standardized testing (on either side of the issue) since the days of Horace Mann.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

DESCRIPTION OF STRUCTURE AND CONTENT

This manuscript is presented in the format of the hybrid dissertation. The hybrid format focuses on producing a journal-ready manuscript. Therefore, this dissertation has fewer chapters than the traditional format, and the manuscript focuses on the presentation of the scholarly article. This hybrid dissertation includes appended materials such as an extended review of literature and a methods section with elaborated detail on the research approach used in this dissertation project.

The targeted journal for this dissertation is the *NASSP Bulletin,* the official journal of the National Association of Secondary School Principals. As clarified in their publishing guidelines online, *NASSP Bulletin* is a peer-refereed journal that publishes scholarly and research-based knowledge that informs practice, supports data-driven decisions, and advances the performance of middle and high school principals. It features a wide range of articles of enduring interest to educators that help promote student learning and achievement, provide insight for strategic planning and decision making in schools, and provide contemporary perspectives on educational reform and policies.

According to their publishing guidelines, manuscripts submitted to *NASSP Bulletin undergo three major stages of review:*

1. Editorial review by staff editors to assess the appropriateness of the manuscript of the journal;
2. Peer review by two or more members of the editorial board of the journal who are recognized scholars and experts in the area of the manuscript; and
3. Final editor review to determine suitability of the manuscript for publication based upon peer review.

All manuscripts are "blind-reviewed." Manuscripts submitted for publication in the *NASSP Bulletin* should not exceed 30 double-spaced pages (excluding references, illustrations, tables, and figures).

NASSP BULLETIN ARTICLE ABSTRACT

Standardized testing is an external control mechanism for K-12 public schools. Principals, nested between internal and external influences, must manage the tension created by testing's roles as both an internal improvement tool and as an external control mechanism. Five competing narratives, each shaped by author academic background, significantly influence this tension. The testing literature is complex, but understanding these five main narratives can enable a principal to more effectively and proactively lead the testing narrative within their own school community.

Background

> "If you ask what problems different advocates [of standards and assessment] are trying to address, it's like the five blind men touching different parts of the elephant. For some it is to motivate students; for others, to achieve world-class standards and compete internationally, to tell Americans which states and local districts are doing well; and for still others, testing is a way of supporting learning. These different goals and definitions of the problem will affect what tests will be designed to do. Will they be for accountability purposes or to improve instruction? If they are designed for one purpose and used for another, then we're back to the old problem" –A Congressional Staffer as quoted in McDonnel, 1994, p.38

This quotation catches the essence of the debate around standardized testing in the United States. Just like the blind men that were each touching a piece of the elephant, each stakeholder in the standardized testing debate sees the elephant differently. In Saxe's famous poem, the blind man touching the leg thought it was a tree trunk, the blind man touching the tail thought it was a rope, etc. Like the elephant, standardized testing is a large and complex enterprise. And, like the metaphor of the blind man's elephant, each stakeholder, from a single student to the federal government, only "sees" a small part of the overall picture. Because each stakeholder only sees a portion of the big picture, the result is what John Godfrey Saxe articulated in the conclusion of his poem:

> "And so these men of Indostan
>
> Disputed loud and long,
>
> Each in his own opinion
>
> Exceeding stiff and strong,
>
> Though each was partly in the right,
>
> And all were in the wrong!" (1873, pp. 77-78).

Thus, the essence of the broader standardized testing debate is that there are multiple, competing narratives about standardized testing, each with "stiff and strong" opinions, but none of the

stakeholders actually get the whole standardized testing picture right. Each stakeholder has a part of the elephant and each therefore is really only "partly right," and each is still at least partly "in the wrong."

What does this mean for principals in traditional, public K-12 schools that must manage standardized testing accountability pressures by external stakeholders, and yet are also expected to use standardized testing effectively within their school community to improve instruction? At the very least principals must get a fuller, albeit not perfect, picture of what all the blind men (stakeholders) are seeing, rather than being guided by one, and only one *blind man's* perspective. If principals can better understand the perspectives of multiple narratives, particularly the powerful and influential academic discipline narratives, that various stakeholders advocate in the standardized testing debate, then they have a much better chance of creating and maintaining the most compelling and productive narrative within their own school community.

Why is it crucial for principals, in particular, to create a compelling and productive standardized testing narrative? First, research has clearly demonstrated that principals are uniquely positioned to influence student learning, since "leadership is second only to classroom instruction as an influence on student learning" (Leithwood & Seashore-Louis, 2012, p. 3). Leithwood and Seashore-Louis further assert that they have "not found a single documented case of a school improving its student achievement record in the absence of talented [school-level] leadership" (2012, p. 3). If a principal does not understand multiple perspectives from these varied narratives, and move forward to create the most compelling narrative possible for their own school community, then that principal will run the risk of being controlled, distracted, or perhaps confused by competing narratives. In this condition, no principal can effectively move their own school community forward in productive ways.

Second, from an organizational theory standpoint, principals are uniquely nested among organizational levels in education. They lead a school community of parents, teachers, and students. At the same time they are held accountable by external influences at the district, state, and federal levels. "To understand the behavior of an organization you must understand the context of that behavior—that is, the ecology of the organization, [and] organizations are inescapably bound up with the conditions of their environment" (Pfeffer & Salancik, 2003, p. 1). Principals are heavily influenced by both internal and external environments. Organizations use control mechanisms to motivate and monitor system effectiveness as they adapt to these changing environments (Hill & Jones, 2013). In the organization of public schools, standardized testing is a significant control mechanism used by the external levels of control (Madaus, Russell, & Higgins, 2009; Mitchell, Crowson, & Shipps, 2011). Yet, if individual school communities are to thrive, a principal—who operates between the school community and external levels—has to create an effective culture to improve teaching and learning in their own school community.

Creating that culture and using standardized testing most productively within the school community has to be the responsibility of the principal. If a principal blindly accepts standardized testing as the primary driver within the school community, it is likely to be counterproductive (Fullan, 2011). As Fullan further articulates:

> To be clear, it is not the presence of standards and assessment that is the problem, but rather the attitude (philosophy or theory of action) that underpins them, and their dominance (as when they become so heavily laden that they crush the system by their sheer weight). If the latter is based on the assumption that massive external pressure will generate intrinsic motivation it is patently false. (2011, p. 8)

The struggle between internal and external stakeholders for control of education is where much of the tension in the standardized testing debate originates. "Because today's data discussion mostly concerns external accountability for schools and educators, it has focused almost exclusively on test scores in reading and math and on graduation rates. Not surprisingly, teachers have viewed this whole enterprise as an intrusion" (Hess & Mehta, 2013, p. 74). Because of this perception, there is a tendency for stakeholders to use standardized testing in ways that can be counterproductive within the school community.

Both internal and external levels of strategic control have a stake in testing. At the same time, both likely have different goals with testing. Phelps, one of the biggest proponents of standardized testing, bluntly asks:

> The key, essential point of debate is who gets to measure school performance—the education 'professionals' or those of us who are footing the bills and giving up our children? The essential point of debate is whether testing, and other methods of quality control, should be done 'internally' or 'externally. (2003, p. 1)

The answer to Phelps's question should be *both*. But it is the principal that has to be the one to see all parts of the elephant so that they can then put that elephant to work in the most productive ways within a school community. In addition, a principal must be well versed in the competing narratives in order to clarify the external narratives that the stakeholders from within the school community constantly read about in the media. In short, a principal must create a culture in which teaching and learning flourishes, where standardized testing is a secondary driver rather than a primary driver (Fullan, 2011), and where that principal can communicate a clear and compelling vision to everyone inside their school community (Leithwood & Seashore-Louis,

2012). A principal has a better chance of doing that effectively if they better understand the different disciplinary backgrounds of authors writing the major testing narratives.

Historical Background

The first standardized tests in America were the invention of Horace Mann in the mid-1840s. These tests were not created to inform teaching and learning. Rather they were created as a control mechanism so that Mann could prove failure of the system so he could spur reform (Office of Technology Assessment, 1992). Mann had visited schools in Europe that he considered to be superior to the school system in Boston, and he wanted to reform Boston's system to be more like the European schools he had seen—thus, "the testing wars had begun" (Reese, 2013, p. 40). For most of the next century, testing was used mainly at the local district level.

Near the beginning of the 20th century, statistical methods were new and in vogue, and a number of psychologists were intrigued with testing and the measurement of intelligence. Testing took on a whole new fervor as the belief that "Whatever exists at all exists in some amount. To know it thoroughly involves knowing its quantity as well as its quality" (Thorndike, 1918, p. 16). Measuring learning (achievement testing) or mental abilities (IQ) became a social and academic fascination. Psychometrics, a field of study dealing with the theory and technique of psychological measurement, took off during this time. In the early testing-enthusiastic environment of World War I, Robert Yerkes created and used the Army Alpha tests to assess over 1.7 million recruits to sort them for their most appropriate role in the army (Lemann, 1995b). Shortly after the war, new technologies were invented to make testing logistically more viable—like the IBM bubble sheet scanner. At the request of Harvard President James Bryan Conant, Henry Chauncey developed the Scholastic Aptitude Test (SAT) as a college entrance

exam shortly after WWI in the hopes of creating a *meritocracy* (Lemann, 2000). Chauncey's SAT test was quickly adopted at many universities, and college entrance in general was changed forever, with a stronger emphasis on academic ability rather than wealth or social position.

Then, Sputnik launched in 1957 creating a new national urgency to improve education, which resulted in the National Defense in Education Act (NDEA) of 1958. Sputnik's influence went far beyond just putting more emphasis on math and science in America's schools. "NDEA was a significant act in U.S. educational history because it was the federal government's first involvement in U.S. education. This involvement, however, was not very strict, but it increased and became stricter over the years" (Turgut, 2013, p. 65). The NDEA was the catalyst for the beginning of national centralization, and testing became one of the core strategies for a more centralized educational system. The question of who would control education started to shift from local districts to the federal government, and testing would begin to play a prominent role in that shift.

Over the course of these last 60 years, testing has taken on ever-increasingly centralized roles. Mitchell, Crowson, and Shipps wrote that "federalization of educational governance over the last 60 years is the most prominent common theme" (2011, p. 36). Daniel Koretz, an educational testing expert wrote:

> The shift from using tests for information to holding students or educators directly accountable for scores is beyond a doubt the single most important change in testing in the past half century. . . .It is not an exaggeration to say that it is now the cornerstone of American education policy. This trend culminated in the enactment of the No Child Left Behind Act in 2001. (2008, pp. 87-88)

At first glance it may seem like Mitchell et al. (2011) and Koretz (2008) are talking about two different things being the most significant change since the early 1960s. But these two educational shifts, centralization and an increase of standardized testing, are inseparably connected. In large part, the latter is the *how* for the former *what*. Standardized testing has become a significant weapon in the war over who controls education.

The centralizing shift in school governance has made the tension between the internal (local school community) and external (district, state, and federal) levels even more intense. Now, the state and the federal governments use testing in ways that would have never been imagined at the school level at the beginning of the 20th century. External uses create interesting and often paradoxical outside pressures on the school organization, and principals must be better prepared to understand and address those uses.

The most extreme critics of testing hope it will go away—it won't. Even while the newest reauthorization of the *Elementary and Secondary Education Act* (ESEA), the *Every Student Succeeds Act* (ESSA), is being hailed as the biggest de-federalization of education in the last 60 years:

> States will still be required to test students annually in math and reading in grades three through eight and once in high school and to publicly report the scores according to race, income, ethnicity, disability and whether students are English-language learners. (Layton, 2015)

The ESSA legislation gave many responsibilities back to the states, but states' plans still "must be approved by the federal Department of Education" (Layton, 2015), and virtually all of the prior testing mandates are still in place.

The Complex Narratives of Standardized Testing

In reviewing the literature about standardized testing, the landscape can quickly become confusing because the testing debate rages with passion, and the debate is complex in both size and scope. In addition to having authors from the academic disciplines of education, policy, economics, psychometry/psychology, and history, there are further divisions within these disciplinary categories because there are both academics and practitioners within each discipline. Additionally critics and proponents exist in every one of these different categories. To add one more level of complexity, many of these different and complex viewpoints have been around for the last 170 years in the United States. Don't forget, testing is an elephant. It is large and complex. Influences, perceptions, uses, and consequences exist that spur these authors to engage in the debate. If a principal can get a grip on the competing narratives, then the testing landscape becomes much more clear and a principal is better equipped to create a vision and strategy to use standardized testing more effectively inside their school community.

However, getting a view of the testing landscape can be challenging because of its breadth and depth. Many elements influence standardized testing. Multiple competing perceptions surround standardized testing and shape the uses of testing. Those uses in turn shape perceptions. Many books, articles, and other literature address the influences, perceptions, and uses of standardized testing.

In the end, standardized tests are only a tool. The tests themselves are neither good nor bad. "Ultimately, the war over testing will be won or lost on the issue of test use. Intelligence and aptitude tests only matter to the extent that they are used" (Chapman, 1988, p. 3). The literature points to many ways the tests are used. For example, tests have been used to sort, select, or classify students (Garrison, 2009; Harris, Smith, & Harris, 2011; Lemann, 2000;

Warren & Grodsky, 2009). Standardized testing results are heavily used to guide policy decisions (Hanushek & Raymond, 2005), and they are used extensively by economists to identify patterns and correlations to study many different facets of education (Cizek, 2005; Goldhaber & Brewer, 1997). Standardized testing can also inform teaching (Chauncey & Dobbin, 1963), diagnose student learning problems (Gandal & McGiffert, 2003; Thorndike, 1918), and certify student competence (Gandal & McGiffert, 2003). These are only a few of the main uses of standardized testing in the United States.

However, the most common use addressed in the literature across the different narratives is that testing is used as a control mechanism (Berliner & Biddle, 1995; Casbarro, 2005; Jacob, 2005; Madaus et al., 2009) or reform lever (Feuer, 2011; Kohn, 2000). This use—addressed by stakeholders with diverse backgrounds—is a main source of tension in the testing debate. Whoever controls standardized testing is able to control many aspects of education, especially if significant consequences are attached to testing results. Harris and Longstreet said of high stakes testing:

> Accountability, which serves as the rationale for the growing use of standardized tests in the United States, has more to do with the locus of power and the control of education than it has to do with the pursuit of excellence. The question of where control and power shall reside is crucial to understanding the pervasive use of these instruments, despite widespread professional dissatisfaction with them. (1990, p. 149)

Three main themes abound regarding how standardized tests are used to control education. First, the tests are used for accountability. By imposing punitive consequences on schools or teachers based on test scores, policy makers exercise considerable control over education (Chiang, 2009; Cizek, 2001; Koretz, 2008). Second, tests are used to control

education by controlling the curriculum. What get tested gets taught (Airasian, 1987; Dorn, 1998; Herman & Linn, 2014; Mehrens, 1998; Resnick & Resnick, 1985). And last, the level at which testing is controlled tends to determine what level of governance controls public education (Berliner & Biddle, 1995; Garrison, 2009; Madaus, 1985; Ravitch, 2013).

Why Understanding the Competing Narratives Matters

The consequences of how testing is used are real and far reaching—for individual children, families, schools, communities, school systems, states, and the nation. Whether testing determines individual opportunity after high school or federal sanctions for underperforming schools, standardized testing carries significant influence. As a result, the body of literature on the issue of standardized testing is highly complex and deeply nuanced. Principals would be better positioned to discern what limitations or abilities tests have, as well as how they might be misunderstood or misinterpreted, by understanding these testing narratives and by distinguishing between critics and proponents from various academic disciplinary backgrounds.

Misinterpreting the results of standardized testing is one of the most prominent themes in the literature regarding the consequences of standardized testing (Betebenner, Wenning, & Briggs, 2011; Bower, 2013; Chapman, 1988; Chappuis, Chappuis, & Stiggins, 2009). Yet, the fact of persistent misinterpretation of testing results should not come as a surprise. Very few stakeholders, including educators, have ever been trained effectively how to understand, use, or interpret the deluge of testing data that creates disparate perspectives among various stakeholders.

If principals can better understand how testing is used as a control mechanism from external levels of governance, they will be more capable to communicate effectively within their school community and to limit the ways testing may contribute to a negative school culture

(Behrent, 2009; Parkinson, 2009; Temin, 2014), such as diminishing important purposes of schooling that the tests do not measure (Bracey, 2003; Cooper, Fusarelli, & Randall, 2004; Friedman & Mandelbaum, 2011), or allowing the tests to narrow the curriculum (Mehrens, 1998; Wiliam, 2010). In short, being fully *standardized testing literate* will enable a principal to use standardized testing at the school level in a way that improves the culture of the school community rather than having standardized testing serve exclusively as an external policy hammer to ostensibly improve scores (Buffum, Mattos, & Weber, 2012; Fullan, 2011; Harvey, Marx, Fowler, & McKay, 2015; Snyder, 2015).

Research Problem

The conflict around test use often leads to confusion and resulting pressures which burdens and, therefore, unavoidably reduces the efficiency and effectiveness of administrators in traditional K-12 schools in their various leadership roles. Ultimately, this conflict surrounding test use can negatively impact the value and power of the educational experience, and the future lives, of millions of American school children each year for whom these administrators have a very real professional and moral responsibility.

This study had two main purposes. First, the study aimed to explore the literature about standardized testing to find patterns in the narratives that are being told in the disciplines of education, policy, economics, psychology/psychometry, and history. Second, this study analyzed those narratives to determine what major themes emerge from each so that a principal can better understand the perspectives and influences in the standardized testing landscape.

Methods

Library databases were searched based on keywords associated with K-12 schooling and standardized testing in the USA. Then, reference sections in those sources were scoured to

search for other qualifying books, book chapters, reports, and articles. Widely-cited authors were sought, as were multiple disciplinary perspectives. Literature from multiple time periods was also sought. The search for representative literature resulted in 171 published works that were gathered on standardized testing, representing the five disciplinary domains, including literature from both critics and opponents in each domain. The 171 sources were narrowed down to 147 to eliminate redundancy and to assure the sampled literature was addressing issues of standardized testing in traditional K-12 public schools in the United States. Table 1 presents the basic distribution of sources ultimately included in this study by academic discipline and type of publication.

Table 1

Type and Number of Documents About Standardized Testing by Disciplinary Background

Academic Discipline	Articles	Books	Book Chapters	Papers/ speeches	Reports	Webpage	Total
Economics	15	0	1	0	4	0	20
Education	28	14	1	2	2	1	48
History	7	7	0	0	0	0	14
Policy	4	5	3	1	5	0	18
Psych	13	15	1	2	4	0	35
Other[a]	7	3	0	1	1	0	12
TOTALS	74	44	6	6	16	1	147

[a] The *other* category includes documents from the disciplinary backgrounds of philosophy, science, math, sociology, and unknown. These were works that were relevant to the issue of standardized testing, but did not fall into the main five disciplinary categories addressed in this study.

Once the sample of works was finalized, representative quotations were extracted from each source and then coded those quotations in QSR International's NVivo 10 Software (Version 10). The coding structure that emerged from the NVivo analysis clustered the data into the following six major categories:

1. Basic standardized testing knowledge
2. Influences on standardized testing
3. Perceptions of standardized testing
4. Uses of standardized testing
5. Consequences resulting from standardized testing
6. Alternatives to traditional standardized testing usage

This study focused specifically on the category of uses of standardized testing *(*#4). Within this category 16 themes emerged about how standardized testing is used. The findings present the data regarding what is written about standardized test use by academic discipline. However, analysis and discussion about the uses of testing also need to consider the consequences and perceptions of standardized testing. Therefore, after the analysis of *test use,* the analyses then focused on the *perceptions* and *consequences* (#3 and #5) of standardized testing by academic discipline.

Using NVivo (QSR, 2014) as the primary analytical platform, we developed a broad profile, a *fingerprint* as it were, of each disciplinary perspective on the varying uses of standardized testing. These fingerprints showed clearly what authors wrote about from varying disciplines addressed in the standardized testing literature. When analyzing these NVivo fingerprints regarding the uses of standardized tests, the need to also consider *how* each of these disciplines treated the prominent themes about test use became apparent. This realization was spurred by the fact that the education and policy fingerprints were similar, yet it became obvious that these two disciplines wrote about the major themes quite differently.

In addition to coding the literature, we also collected and imported author attributes for each first author into NVivo including their primary educational training (disciplinary

background), the domain in which they primarily worked during their career (career function), whether an author was a practitioner or and academic and their author viewpoint (critic, proponent, or neither). The analyses for this study focused specifically on which themes within the three literature categories of use, perceptions, and consequences of standardized testing were most common within and between author's disciplinary backgrounds.

Findings

The findings section was organized into three main areas of focus. The main analysis in this study dealt with the use of standardized testing. Then, as a follow-up, the analyses addressed the consequences of standardized test use, and the resulting perceptions of standardized testing use. The following table provides data on the literature used in the study. Table 2 summarizes the data that was used in the analysis of the uses of standardized testing.

Table 2

Data Summary of the Literature used in the Analysis of the Uses of Standardized Testing

Disciplinary Background	# of Sources	# of Authors	Year Range of Sources
Education	41	28	1919—2015
Policy	16	15	1977—2011
Economics	19	16	1981—2014
Psychology/Psychometry	24	17	1916—2010
History	12	7	1961—2013

The findings on the use of standardized tests in traditional K-12 public education are reported in each of the five author academic disciplines. While other areas of analysis indicated

findings of potential interest for future treatment, the issue of standardized testing as a control mechanism yielded the most prominent and consistent result.

Uses of Standardized Testing by Academic Discipline

The figures in the following sections graphically present the resulting profile of each disciplinary domain. The figures represent the number of sources (documents) that were coded to that theme, not the number of textual references extracted from the sources. Also, there were not the same number of sources for each discipline, so while the profiles are still revealing, the reader will notice that the scales on each graph vary. However, these figures only represent *what* is being written about in each domain. After noting what the authors from different disciplinary backgrounds wrote about, analysis also sought to understand *how* each author, and each disciplinary collection of authors, wrote about specific elements of standardized testing usage. For example, though authors in the disciplines of policy and education both often addressed the idea of using tests for accountability, how they wrote about this issue differed considerably

Education. *Control of education* is clearly the most common theme among authors in the discipline of education (see Figure 1). In analyzing this literature, it seems that authors in this domain rarely write about the ways testing can be beneficial to the educational process. Instead, what ends up dominating their literature is reactive against what is perceived as misuses, or even abuses of standardized testing. For example, one academic (Garrison, 2009) in education wrote, that it should be clear "that neither in the past nor the present is testing mainly about 'improving education.' It is, instead, about control over the purpose and nature of schooling" (p. 2) Hence, the most common themes in the education literature center around ideas such as *control of education*, *manufactured crisis*, and *misuse of defined purpose*.

As a result of the preponderance of the three most common themes, the education disciplinary narrative can be generally described as both *defensive* and *reactive*. Perhaps this is predictable, largely because in this academic domain it is educators themselves that are being *controlled*. For example, one academic critic asserts that, "Standardized tests have become an important tool in the efforts of state governments to improve educational standards and to gain control over the process of education in local school districts" (Airasian, 1987, p. 393). Another author echoed Airasian by saying, "What is clearly at stake here is not only who shall control testing, but also who shall control education" (Harris & Longstreet, 1990, p. 150). One educational practitioner used a more passionate tone when addressing standardized testing: "A troubling reality in today's political climate is that many political leaders actually believe that the best way to change schools is through an 'end of a gun barrel' approach, rather than by building consensus" (Casbarro, 2005, p. 20).

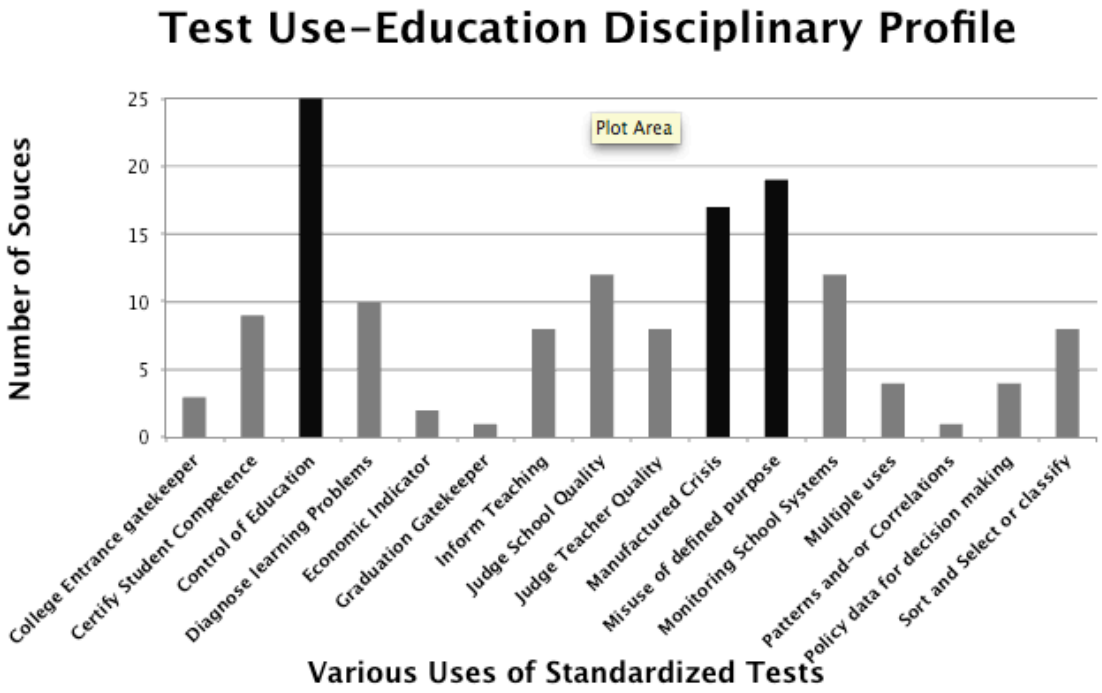

Figure 1. What educational authors write about regarding test use.

One of the reasons this finding is so important is because if educators want to take charge of the testing narrative in their own school community, they need to create a more compelling narrative. Being reactive and defensive is not a narrative that is likely to be compelling or motivating to their students, teachers, or larger school community.

Policy. The policy literature has the same three top themes as education (see Figure 2). It would be simple to misinterpret this similarity by thinking the two disciplines were aligned. However, what is being conveyed in the policy literature is quite different from that of the education literature. For example, one author with an academic background in policy optimistically wrote:

> Finally, we may have reached a point in the United States where standardized tests provide the only pure measure of subject-matter mastery. . . .If standardized tests are, indeed, the only trustworthy measure of academic achievement, can our society afford not to use them? (Phelps, 2003, p. 225)

Another wrote that testing is "a cornerstone of education reform because of its powerful leverage as a policy instrument. . . a growing body of research indicates that school and classroom practices do change in response to these assessments" (McDonnell, 1994, p. 1)

Margaret Spellings, a former U.S. Secretary of Education with an academic background in policy wrote of testing, "Our nation's education report card tells the story. Achievement is up across the board . . .What gets measured does indeed get done. Lesson No. 1: Accountability is a powerful tool and is working to improve learning" (2010, p. 33).

Academics in policy typically say things such as:

> This method of finding out what students know has done more than provide information. Increasingly, it has shaped expectations for what students and teachers do every day in

the classroom. Partly, this has happened by design: recognizing that everyone views tests as important, policy makers have increasingly relied on tests as a lever to improve performance. (Rothman, 1995, p. 36)

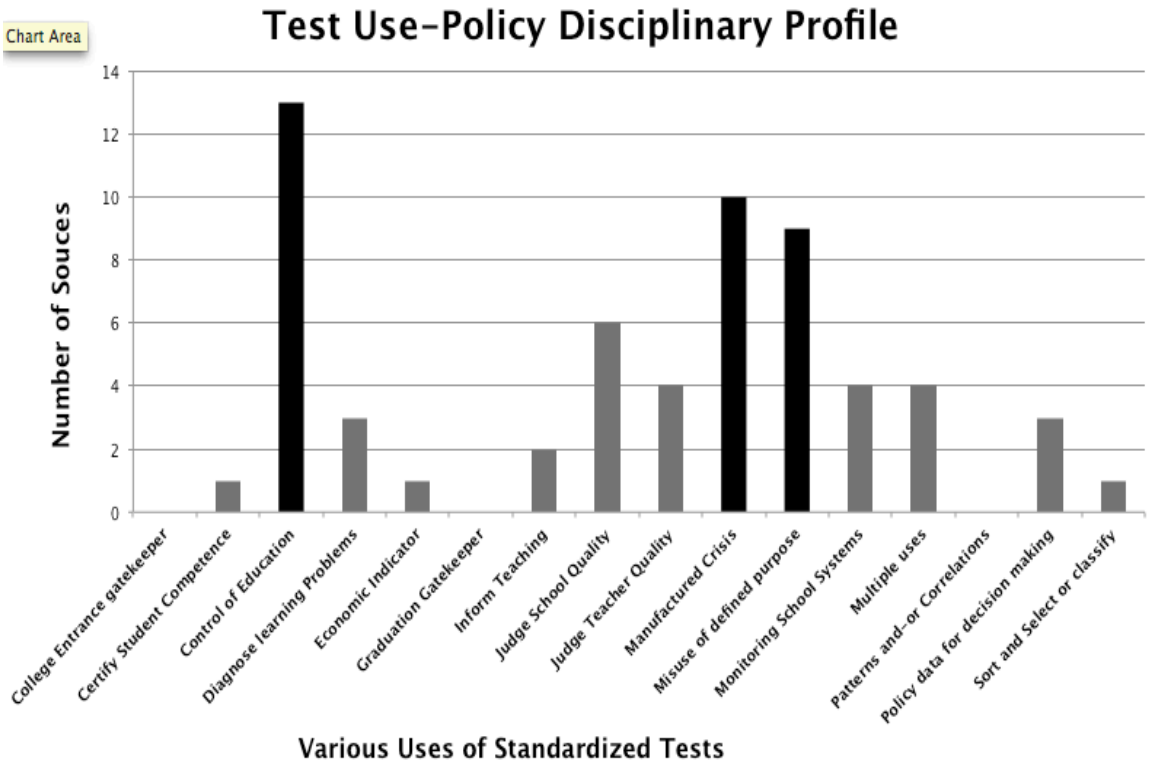

Figure 2. What policy authors write about regarding test use.

Another policy academic simply stated, "education policy today is intensely committed to the use of curriculum standards and high stakes test-based accountability policies to control schools" (Mitchell et al., 2011, p. 295). These statements are typical from authors with an academic background in policy.

As exemplified by these examples, the tone and message from the policy narrative can be described as *proactive* and *progressive*. Even though the discipline of education often views the way policymakers use the tests as political or punitive, the general tone and message from the

viewpoint of those in policy is that testing is going to help us (education as a gestalt) move forward and improve education (Cizek, 2001; Phelps, 2003). In addition, the domain of policy is more likely to have significant influence, to the point of exclusivity, on the narrative presented in most forms of public mass media, especially when the reports are perceived as being negative toward education (Koretz, 2008). Therefore, policy has a more effective way to promote their narrative into the mainstream public than do educators. The policy narrative (especially policymakers) tends to circulate quite effectively in the media (Bracey, 1995).

Economics. The literature from the economic discipline has a very different overall profile than the other four disciplines (see Figure 3), but even economists address this issue of using testing as a control mechanism. For example, Hanushek and Raymond note, "The leading school reform policy in the United States revolves around strong accountability of schools with consequences for performance" (2005, p. 1). Jacob echoes this idea: "Indeed, accountability policies dwarf all other education reforms in scope" (2005, p. 762). Though they occasionally address test use explicitly, economists, for the most part, simply use standardized testing data to find patterns and correlations to use statistically as an economic indicator. The issue of who controls education is also a theme that economists address, but they remain virtually silent on most of the other issues common to the other four disciplines. Economics, then, is the most highly focused of the five disciplines; many of the themes addressed in other disciplines are rarely considered in economics.

The narrative economists tell about standardized testing for the most part is not a narrative at all. Generally, they approach standardized testing from a *utilitarian*, and/or *dispassionate* viewpoint. For the most part, economists use the testing data, but do not very often get into conceptual or practical arguments about how standardized testing should or is used

in schooling. Their job is simply to use the data in predicting economic outcomes or impacts of education. So, what influence do economists have, and why do they matter in the testing conversation? The answer is found by considering who consumes economic literature. Based on citations in the education literature, it seems educators, especially practitioners, rarely consume economic analyses that use testing data. However, policy makers turn to economists constantly to inform their decisions. So, economists become a major influence, albeit an apparently quiet, indirect, and non-intrusive player in the debate. It appears that policy makers rely on economists, while educators (especially practitioners) are often unaware that economists even have a stake or influence in the standardized testing debate.

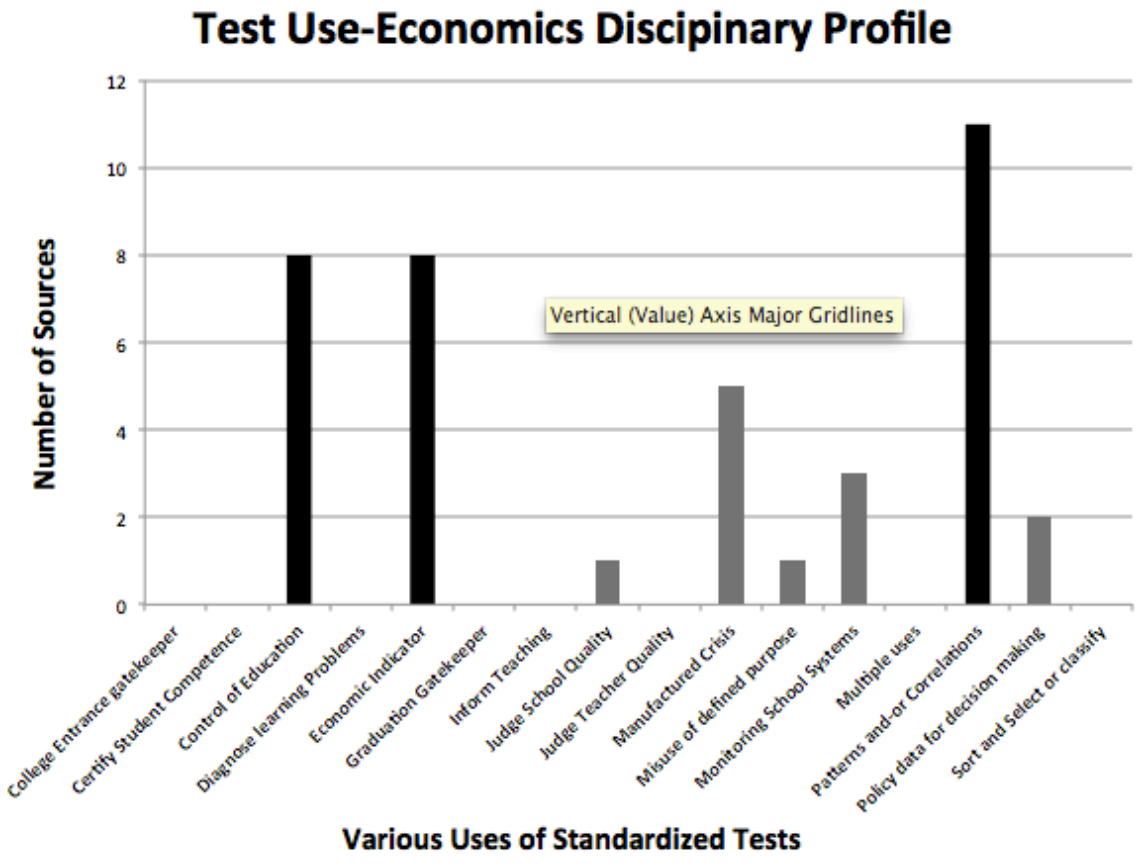

Figure 3. What economists write about regarding test use.

Psychology/Psychometry. The authors in psychology/psychometry wrote most often about how standardized tests are often used in ways that differed from the defined purpose of the test, as articulated by the test makers. This is not surprising since these are the writers of the tests, and tests are often used in ways that the test writers warned against. The second most common theme among this discipline was, again, the idea of using testing as a control mechanism (see Figure 4).

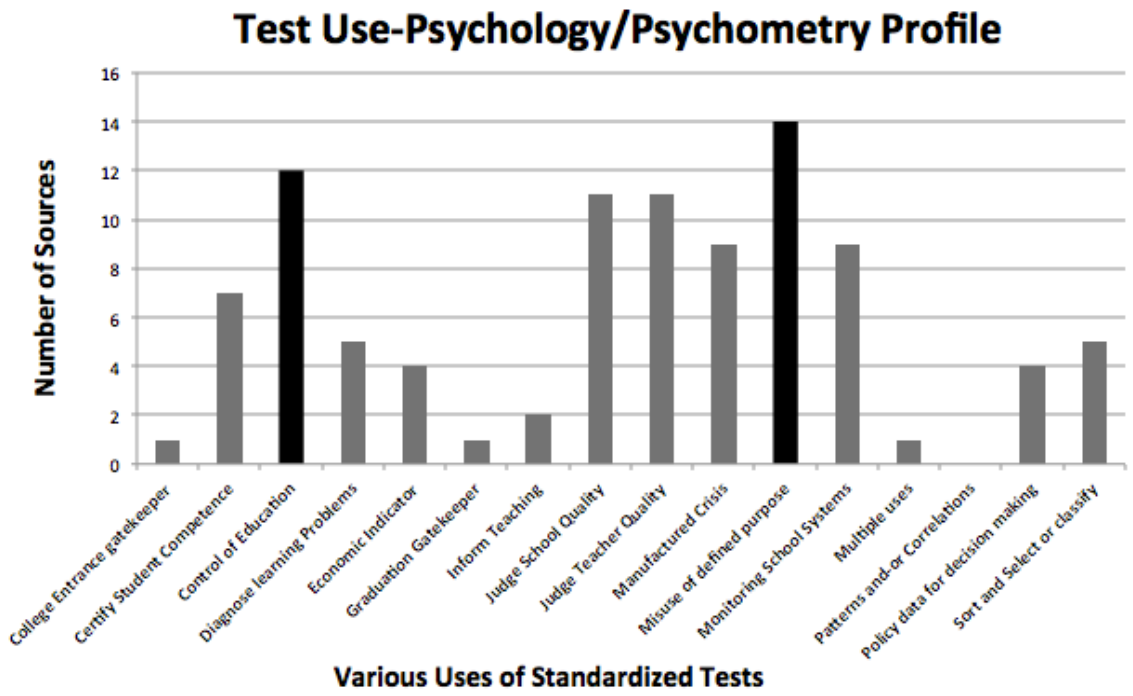

Figure 4. What psychologists/psychometricians write about regarding test use.

It makes sense that the one theme that exceeds the issue of control in this domain was the concern that tests are used in ways that go against the purpose for which they were designed. Test writers often warn against misuse and misinterpretation and are clear about the specific parameters of proper application and interpretation. But, their warnings often go unheeded, and the instruments are commonly used in ways identified as inappropriate by instrument designers.

For example, psychometricians have warned that "one lesson that is repeatedly learned is that data, regardless of their quality, can be used well and can be used poorly" (Betebenner et al., 2011, p. 1).

Yet, once standardized test results are released, they are used however any of the stakeholders choose to use them. Chauncey, the father of Educational Testing Service (ETS) and the Scholastic Aptitude Test (SAT), is one of the most famous psychometricians of the past century. In the 1960's he and Dobbin wrote a book that later went completely unheeded by policy makers. They wrote,

> No test measures accurately enough to support a precise, one-point interpretation of a score or a 'diagnosis' of small differences between scores. Any test that leads the user to believe he has a precise measurement, then, is subtly dishonest and can lead him into trouble. (1963, p. 64)

They went on to say,

> Like a lot of other tools, achievement tests are used in many ways, only some of them the ways in which their makers intended them to be used. Professional test makers, observing the variety of uses to which their instruments are put, occasionally are reminded of the scientist whose delicate micrometer was used by his wife to crack nuts. Whether or not such a comparison is apt, the fact remains that in schools generally there are both appropriate and inappropriate uses of standardized tests. (1963, p. 66)

Chauncey and Dobbin were very clear that the SAT test had limits and should only be used to make judgments about the individual student, yet only a few years later SAT test results were used exactly how he said they should not be used (Salganik, 1985).

Authors in the area of psychology/psychometry continually warn against using tests in ways they were not designed to be used, and then the tests often end up being used in exactly the ways psychologists/psychometricians warned against. Yet, many also address the issue of using the test scores to gain control over education. For example Madaus draws a clear relationship between policy makers and testing as a tool for control:

> The common thread that links these human perspectives on the meaning of test scores is the use of those scores as administrative mechanisms by which to implement one or another policy. In each case, testing as an administrative device has become the linchpin of policy. (1985, p. 612)

He continued,

> The early 1960's brought a slow but inexorable shift in the use of standardized tests. No longer merely tools used by local school district administrators, the tests assumed a central role in establishing and implementing state and federal education policy. (1985, p. 613)

The narrative of psychometry/psychology is very distinct from the other disciplines: test designers are clear about the specific uses and potential misuses of their instruments, but their effective impact on the appropriate use of testing data on policy and practice is apparently negligible. Generally the academics and practitioners in this domain operate in a sphere of *technical purity* that seldom finds adequate voice or responsive application in the worlds of policy or educational practice

History. The discipline of history also addresses the issue of testing as a control mechanism in education. The profile of history (see Figure 5) differs from the others in that one of the most common themes in this discipline is that testing is used to sort and select individuals.

However the idea of using testing as a control mechanism is equally prominent. On the topic of testing as a control mechanism, authors in the discipline of history often reflect Resnick and Resnick's view:

> In our view, two elements have the largest role in shaping what is demanded in schools, and therefore what students can be expected to learn. The first is the curriculum—what is taught. The second is assessment—the way we judge what is taught. (1985, p. 5)

Much of the history literature talks about the importance of controlling testing in order to control curriculum, rather than the larger school setting, per se. While this distinction might appear to be too closely cut, the distinction is significant in practical application and impact.

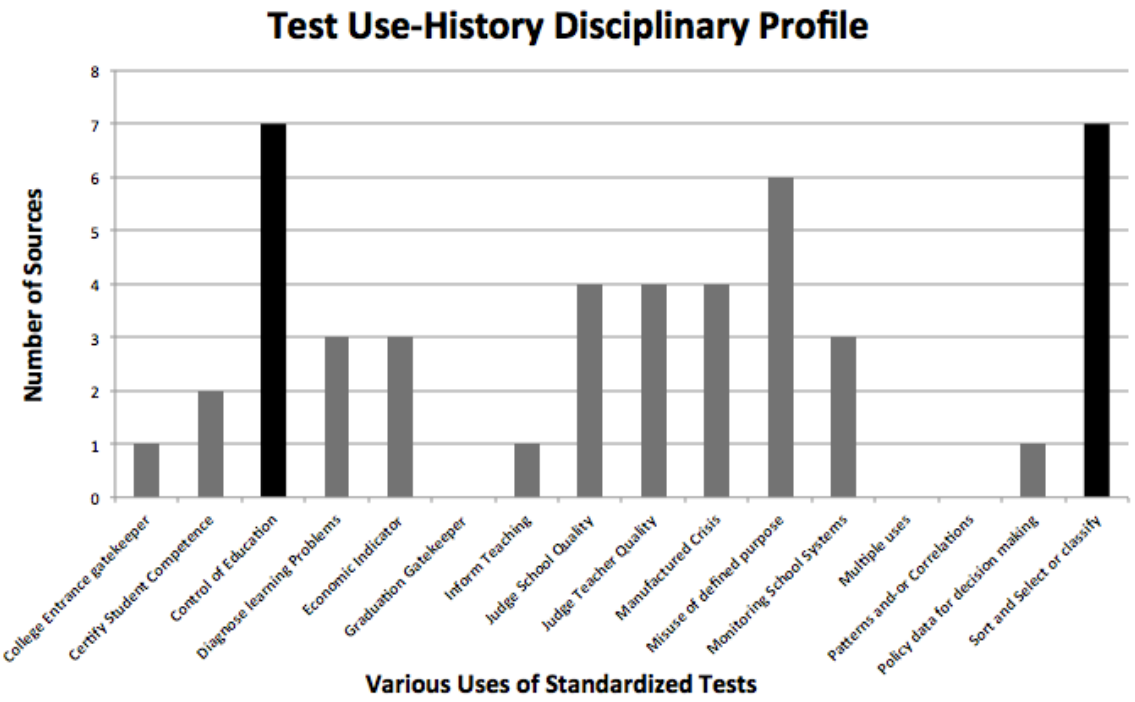

Figure 5 What historians write about regarding test use.

Historians point out that using testing this way is nothing new—it started with Mann in 1845 in Boston:

> The most zealous reformers in Boston believed that written examinations would reveal otherwise hidden truths about the nature of schools. Testing promised to undermine outmoded forms of school organization, identify the best teaching practices, hold teachers and pupils accountable, and provide incontrovertible evidence about what children actually knew: pupil by pupil, school by school, even district by district. (Reese, 2013, p. 129)

Diane Ravitch was trained in history and worked in policy. Once one of testing's strongest proponents Ravitch has become one of its most vocal critics. She voiced a strong opinion on how testing is used as a control mechanism:

> How did testing and accountability become the main levers of school reform? How did our elected officials become convinced that measurement and data would fix the schools? Somehow our nation got off track in its efforts to improve education. What once was the standards movement was replaced by the accountability movement. What once was an effort to improve the quality of education turned into an accounting strategy: Measure, then punish or reward. (2010, p. 16)

The history narrative is varied. Just like the other disciplines, the background of individual historians colors what history has to say about standardized testing. Some historians are passionate critics or proponents, while others are more objective in their view of testing. But, regardless of their viewpoint, historians have a relatively *silent narrative.* The mainstream public is certainly ignorant about the history of testing. Even educators and politicians have virtually no concept of the history of testing; few educators, if any, have a grasp of testing's history, unless they have specifically sought it out themselves. However, the silent narrative of

the history of standardized testing certainly could provide context and perspective as a principal creates and shapes the narrative for their own school community.

Consequences of Standardized Test Use

The analysis of themes in the category of consequences, clearly indicated that across every discipline, the most common theme was the misinterpretation, and therefore the misapplication as a control mechanism, of test scores. The second most common theme was closely related: that test scores end up being used too often as a sole source of high stakes judgment of various elements of K-12 public education.

The misinterpretation of results seemed to be the one theme on the subject of standardized testing where there is near universal agreement across disciplines. Both tone and message are similar—caution must be exercised in the conclusions we draw from testing results. Two prominent education economists said:

> We are most certain of this: To make judgments only on the basis of national average scores, on only one test, at only one point in time, without comparing trends on different tests that purport to measure the same thing, and without disaggregation by social class groups, is the worst possible choice. But, unfortunately, this is how most policymakers and analysts approach the field. (Carnoy & Rothstein, 2013, p. 84)

Other disciplines echo this concern. Ten years after creating his wall charts, even Terrel Bell, famous for using SAT scores as a springboard to the report *A Nation at Risk,* said:

> I included with every chart a cautionary statement on the limitation of these data. But the statement was largely ignored both by the press and by many educational leaders. The national pastime of jumping to conclusions was just as avidly pursued in those days as it is today. (1993, p. 594)

Koretz, from the discipline of education, said, “Achievement testing is a very complex enterprise, and as a result, test scores are widely misunderstood and misused. And precisely because of the importance given to test scores in our society, those mistakes can have serious consequences” (2008, p. 1). Even Thorndike, the early 20th century psychometrician warned, “A pupil's score in a test signifies first, such and such a particular achievement, and second, only whatever has been demonstrated by actual correlations to be implied by it. Nothing should be taken for granted” (1918, p. 22). It is striking that disciplinary views of test use can, in many instances, differ so much, while across all disciplines, concern about the consequences of tests being misinterpreted is uniformly shared.

Disciplinary Perceptions of Using Standardized Testing

So, which did come first, the chicken or the egg? Do the perceptions of testing determine and shape uses and consequences, or do uses and consequences determine perceptions? As with the chicken-and-egg causality dilemma, either can be argued but neither necessarily takes precedence.

The perception that appears most often in the education literature is that standardized testing is a political weapon. It is a clear outlier from all the other perceptions that were evident in the other disciplinary literature bases. For example, one educator wrote, "Standardized tests have become a political tool, one that allows politicians to put on the mantle of educational leadership” (Harris & Longstreet, 1990, p. 149). Another said, “Tests are political weapons instead of tools designed to assess the value and progress of current curricula” (Monroe, 1987, p. 24). Remember from the test use section, educators often feel like victims of testing, so it is no surprise that they perceive tests as being political weapons or tools wielded by politicians or

policy makers. Again, this perception adds to the problem of a narrative that is negatively reactive and defensive.

The area of policy also perceives standardized testing as political. "The nature of education in both its content and access is inherently political and permeated with fundamental values" (Cooper et al., 2004, p. 157). However, instead of a weapon, it is written about as a policy lever. Another perception in the policy literature is that standardized testing has many positive consequences. Though many authors in policy perceive standardized testing as political, the way they talk about it is different in tone and content from educators. For example, note the proactive, assertive tone in this policy narrative:

> Political elites, business leaders, and the general public are looking once again to student assessment as a cornerstone of education reform because of its powerful leverage as a policy instrument . . . a growing body of research indicates that school and classroom practices do change in response to these assessments. (McDonnell, 1994, p. 1)

Economists, for the most part, rarely focus explicitly on standardized testing. This pattern is reflected in the small number of references that are mentioned regarding perceptions in the economic literature. One thing that stands out is that in the few economics references, referencing testing directly, testing is perceived as beneficial. For example, Hanushek and Raymond note, "The most important result is that accountability is important for students in the United States. Despite design flaws in most existing systems we find that they have a positive impact on achievement" (2005, p. 321). Levin sums up the tone of economists' perceptions when he wrote:

> Nothing in this chapter should be interpreted as a rejection of testing when properly used or as a rejection of high standards, which are important for many reasons. Although I

> have some concern about who sets standards and how they are used and assessed by schools, these are issues that are more generally debated among educators and policymakers. (2001, p. 40)

In the discipline of psychometry/psychology, the most common perception is that the tests are *scientific*. The early psychometricians especially believed the tests were truly scientific. Galton argued that "until the phenomena of any branch of knowledge have been subjected to measurement and number, it cannot assume the status and dignity of a science" (1879, p. 149). Many in this domain still address the issue, but they often address it with the understanding that they are not nearly as scientific as the early psychometricians once believed. For example, Madaus, Russell, and Higgins, critical of how tests are often used, still acknowledged this perception: "The numeric scores produced by these programs have an objective, scientific, almost magical persuasiveness about them" (2009, p. 139).

The perception that got the most attention in the discipline of history is that people have viewed the tests as legitimate over the years.

> Despite the clamor of critics, standardized tests enjoy widespread support in U.S. public opinion. Eighty-one percent of a national sample of parents surveyed by the Gallup Organization in 1979 indicated that they thought standardized tests were "useful" or "somewhat useful." Only 17% thought tests were "not too useful." Other polls have shown the same positive attitude toward testing. (Resnick, 1981, p. 625)

In the history of standardized testing in America, the tests have been perceived as legitimate tools to measure learning.

Table 3

Most Common Uses, Consequences, and Perceptions by Academic Discipline in Testing Literature

Disciplinary Background	Most Common Themes Regarding Test Use	Most Common Theme Regarding Consequences of Testing	Most Common Theme Regarding Perceptions of Testing
Education	1. Control of education 2. Manufactured Crisis 3. Misuse of Defined Purpose	Misinterpretation of Results	Testing is Political
Policy	1. Control of education 2. Manufactured Crisis 3. Misuse of Defined Purpose	Misinterpretation of Results	Testing Is Beneficial
Economics	1.Patterns and Correlations T2.Control of Education T2.Economic Indicator	Misinterpretation of Results	Testing is Beneficial
Psychology/ Psychometry	1.Misuse of Defined Purpose 2.Control of Education	Misinterpretation of Results	Testing is Scientific
History	1.Control of Education 2.Sort & Select or Classify	Misinterpretation of Results	Testing is Accepted by the Public as Legitimate

The uses and consequences of standardized testing create perceptions that differ considerably among the five disciplines included in our analysis. Even though similar patterns were found in what different disciplines talked about regarding the use of testing, these disciplines have quite different perceptions that reflect the tension that the different narratives often spark—especially the difference between education and policy. Table 3 provides a summary of the most common themes found in the uses, consequences, and perceptions of the standardized testing literature.

Discussion

A principal cannot control the five competing standardized testing narratives included in this analysis. Nor can a principal control how these narratives are used in the external

environment. However, a principal's standardized testing narrative has the potential to be the most compelling and impactful within their own school community. Often principals parrot other, perhaps negative, narratives. Other times, principals have no specific or structured narrative and therefore default to whatever narrative speaks loudest in their school, district, or state. Or perhaps, principals may just try to discredit the other narratives without providing more compelling narrative themselves. All of these strategies are less effective than a purposefully constructed, positive narrative fitting the challenges and opportunities of the principal's specific school community. In this purposeful, positive scenario, a principal needs to understand the nature of the dominant disciplinary narratives that exist about standardized testing so they can take a more proactive and purposeful stance regarding the competing narratives the stakeholders in their school community are hearing. How principals frame the testing narrative in their own school community has a significant effect on how their teachers, students, and parents view and participate in standardized testing and its consequent applications.

Findings in this study indicate that two disciplinary voices are the loudest in the literature regarding standardized testing—Education and Policy. The other three disciplinary voices have their narratives, and are influential in different ways, but the two narratives of education and policy appear to get the most mainstream attention. Keep in mind that the policy narrative is *proactive* and *assertive* and the education narrative tends to be *reactive* and *defensive*. In addition, policy tends to have a stronger hold on the media to tell their testing narrative. Education, on the other hand, is more pervasive in the professional literature. A classic manifestation of education's reactive nature is found in the opening lines written by Eaker (2015) in the forward of Richard DuFour's newest book titled, *In Praise of American Educators, and How they Can Become Even Better.* These two men are arguably two of the most influential

reformers from the discipline of education in the last 50 years, but even they have a difficult time not being reactive and defensive on this issue. Eaker wrote:

> It should be particularly troubling to realize that a segment of our society—particularly a segment of politicians and the media—has declared war on America's public schools, and more specifically, on America's public school teachers. It is even more troubling that the war on teachers is based on data that are interpreted incorrectly, manipulated, or simply false. (DuFour, 2015, p. xix)

The dynamics of who controls standardized testing, and who most effectively controls the mainstream narrative has significant implications on internal and external stakeholder perceptions. As a result, it appears the discipline of policy has the loudest, clearest, and ultimately the most impactful narrative. It is the narrative that the general public ingests and generally accepts.

The education narrative about standardized testing is more profuse in professional journals or books, but the professional literature is rarely mainstream in the broader American discourse. Also, it seems that practitioner educators tend to only read the literature from the domain of education, which gives them only one perspective from one narrative. One proponent of testing in the discipline of psychometry commented that in a literature search one of his colleagues found 59 articles on standardized tests. Of these articles, 57 confirmed that "high-stakes tests are uniformly bad" (Cizek, 2001, p. 20). This finding is typical when searching academic literature about standardized testing. Unfortunately, much of the education literature ends up being reactive to the policy narrative, and therefore is, and comes across as, defensive.

So, what does all of this mean for a principal that is caught in an external environment of school governance where testing is being *done to them*, and yet that same principal is expected

by the same stakeholders to be an instructional leader and use standardized testing effectively within their own school community to improve student outcomes? The meaning is clear—a principal must take control of the narrative in the school community, and that narrative has to be positive, proactive, and assertive rather than negative, defensive, and reactive. While a principal may not be able to effect systemic change in their external environment, principals are the ones that can, and must, create the testing culture for their teachers and students to thrive daily.

Conclusion

To be an effective instructional leader, a principal must understand the different narratives of standardized testing in order to use such testing in proactive, assertive ways within their own school community. Just like the metaphor of the elephant, the standardized testing debate is large and complex. A principal has to be the one to try and see the whole elephant—not just one part through the perspective of one narrative (typically the reactive and negative education narrative). Unfortunately, most principals in traditional public K-12 schools only really hear two standardized testing narratives. They hear the narrative of the policy discipline through the media, or read about testing from the perspective of the education discipline in its professional literature. Either narrative by itself creates a picture no more complete than any one of the blind men had of the elephant.

Understanding the different standardized testing narratives puts a principal in a position where he or she can parse and digest the narratives being communicated in the external environment and still communicate effectively their own standardized testing narrative within their school community. The purpose of this study is not to create a specific, omnibus framework to outline how every principal should take charge of, and create a narrative for their specific school community—that is a task for each principal. However, it is critical that principals start

thinking about their own narrative and how it can be better informed by understanding the most common disciplinary narratives with influential external voices.

Further, principals would be wise to steal at least one strategy from the policy playbook. A positive, proactive, and assertive tone is much more compelling than a story that is negative, reactive, and defensive. The professional learning communities (PLC) movement has been so compelling nationally largely due to its positive, proactive, and assertive stance. Similarly, a principal must create and promote a similar narrative regarding standardized testing in the school community—instead of being nearly exclusively reactive and defensive about policies imposed by external levels of control. Ironically, such a positive, proactive, and assertive approach would also allow a school community to purposely place standardized testing in its proper role as a "secondary driver" which will, in the long run, get "more accountability all around" (Fullan, 2011, p. 9) than the way most principals currently react to how standardized testing is used as a control mechanism by external stakeholders.

Article References

Airasian, P. (1987). State mandated testing and educational reform: context and consequences. *American Journal of Education, 95*(3), 393-412.

Behrent, M. (2009). Reclaiming our freedom to teach: Education reform in the Obama era. *Harvard Educational Review, 79*(2), 240-247.

Bell, T. (1993). Reflections one decade after "A Nation at Risk". *Phi Delta Kappan, 74*(8), 592-597.

Berliner, D. C., & Biddle, B. J. (1995). *The manufactured crisis: Myths, fraud, and the attack on America's public schools*. Reading, MA: Addison-Wesley.

Betebenner, D. W., Wenning, R. J., & Briggs, D. C. (2011). *Student growth percentiles and shoe leather*. Retrieved from http://nciea.org/publications/BakerResponse_DB11.pdf.

Bower, J. (2013). Telling time with a broken clock: The trouble with standardized testing. *Education Canada, 53*(3), 24-27.

Bracey, G. W. (1995). The fifth Bracey report on the condition of public education. *The Phi Delta Kappan, 77*(2), 149-160.

Bracey, G. W. (2003). *On the death of childhood and the destruction of public schools*. Portsmouth, NH: Heinemann.

Buffum, A, Mattos, M, & Weber, C. (2012). *Simplifying response to intervention: Four essential guiding principles*. Bloomington, IN: Solution Tree Press.

Carnoy, M, & Rothstein, R. (2013). *What do international tests really show about U.S. student performance?* Retrieved from Economic Policy Institute website: http://www.epi.org/publication/us-student-performance-testing/.

Casbarro, J. (2005). The politics of high-stakes testing. *Education Digest*(February 2005), 20-23.

Chapman, P. D. (1988). *Schools as sorters: Lewis M. Terman, applied psychology, and the intelligence testing movement, 1890-1930*. New York, NY: New York University Press.

Chappuis, S, Chappuis, J, & Stiggins, R. (2009). The quest for quality. *Educational Leadership, 67*(3), 14-19.

Chauncey, H, & Dobbin, J. (1963). *Testing: Its place in education today*. New York, NY: Harper & Row.

Chiang, H. (2009). How accountability pressure on failing schools affects student achievement. *Journal of Public Economics, 93*(9), 1045-1057.

Cizek, G. J. (2001). More untintended consequences of high-stakes testing. *Educational Measurement: Issues and Practice, 20*(4), 19-27.

Cizek, G. J. (2005). High stakes testing: Contexts, characteristics, critiques, and consequences. In R. P. Phelps (Ed.), *Defending standardized testing* (pp. 341). Mahwah, NJ: Lawrence Erlbaum Associates.

Cooper, B. S. , Fusarelli, L. D. , & Randall, E. V. (2004). *Better policies, better schools: Theories and applications*. Boston, MA: Pearson Education.

Dorn, S. (1998). The political legacy of school accountability systems. *Education Policy Analysis Archives, 6*(1), n1.

DuFour, R. (2015). *In praise of American educators: And how they can become even better*. Bloomington, IN: Solution Tree Press.

Feuer, M. J. (2011). Politics, economics, and testing-some reflections. *Mid-Western Educational Researcher, 24*(1), 25-29.

Friedman, T. L., & Mandelbaum, M. (2011). *That used to be us: How America fell behind in the world it invented and how we can come back*. New York, NY: Farrar, Struas, and Gireaux.

Fullan, M. (2011). Choosing the wrong drivers for whole system reform, *Paper 204 in Center for Strategic Education Seminar Series*: Center for Strategic Education. Retrieved from http://theeta.org/wp-content/uploads/2011/11/eta-articles-110711.pdf.

Galton, F. (1879). Psychometric experiments. *Brain, 2*(2), 149-162.

Gandal, M, & McGiffert, L. (2003). The power of testing. *Educational Leadership, 60*(5), 39-42.

Garrison, M. J. (2009). *A measure of failure: The political origins of standardized testing*. Albany, NY: State University of New York Press.

Goldhaber, D., & Brewer, D. J. (1997). Why don't schools and teachers seem to matter: Assessing the impact of unobservables on educational productivity. *The Journal of Human Resources, 32*(3), 505-523.

Hanushek, E. A., & Raymond, M. E. (2005). Does school accountability lead to improved student performance? *Journal of Policy Analysis and Management, 24*(2), 297-327.

Harris, K., & Longstreet, W.S. (1990). Alternative testing and the national agenda for control. *Social Studies, 81*(4), 148-152.

Harris, P., Smith, B. M., & Harris, J. (2011). *The myths of standardized tests: Why they don't tell you what you think they do*. New York, NY: Rowman & Littlefield.

Harvey, J, Marx, G, Fowler, C, & McKay, J. (2015). *School performance in context*: The Horace Mann League & the National Superintendents Roundtable. Retrieved from http://www.superintendentsforum.org/wp-content/uploads/2015/01/School-Performance-in-Context.pdf.

Herman, J, & Linn, R. L. (2014). New assessments, new rigor. *Educational Leadership, 71*(6), 34-37.

Hess, F. M., & Mehta, J. (2013). Data: No deus ex machina. *Educational Leadership, 70*(5), 71-75.

Hill, C. W. L. , & Jones, G. R. (2013). *Strategic management theory* (10th ed.). Stamford, CT: South-Western Cengage Learning.

Jacob, B. (2005). Accountability, incentives and behavior: The impact of high-stakes testing in the Chicago public schools. *Journal of Public Economics, 89*(5), 761-796.

Kohn, A. (2000). *The case against standardized testing: Raising test scores, ruining the schools*. Portsmouth, NH: Heinemann.

Koretz, D. (2008). *Measuring up: What educational testing really tells us*. Boston, MA: Harvard University Press.

Layton, L. (2015, December 10, 2015). Obama signs new K-12 education law that ends No Child Left Behind, *The Washington Post*. Retrieved from http://www.washingtonpost.com/local/education/obama-signs-new-k-12-education-law-that-ends-no-child-left-behind/2015/12/10/c9e58d7c-9f51-11e5-a3c5-c77f2cc5a43c_story.html

Leithwood, K. , & Seashore-Louis, K. (2012). *Linking leadership to student learning*. San Fransisco, CA: Jossey-Bass.

Lemann, N. (1995). The structure of success in America. *The Atlantic Monthly, 276*(2), 41-60.

Lemann, N. (2000). *The big test: The secret history of the American meritocracy*. New York, NY: Farrar, Straus, and Giroux.

Levin, H. M. (2001). High-stakes testing and economic productivity. In G. Orfield & M. L. Kornhaber (Eds.), *Raising standards or raising barriers? Inequality and high-stakes testing in public education* (pp. 39-49). New York, NY: The Century Foundation.

Madaus, G. (1985). Test scores as administrative mechanisms in educational policy. *Phi Delta Kappan, 66*(9), 611-617.

Madaus, G., Russell, M., & Higgins, J. (2009). *The paradoxes of high stakes testing: How they affect students, their parents, teachers, principals, schools, and society*. Charlotte, N.C.: Information Age Publishing.

McDonnell, L. M. (1994). Policymakers' views of student assessment. Santa Monica, CA: RAND.

Mehrens, W. A. (1998). Consequences of assessment: What is the evidence? *Education Policy Analysis Archives, 6*(13), 1-30.

Mitchell, D. E. , Crowson, R. L., & Shipps, D. (2011). *Shaping educational policy: Power and process*. NY: Routledge.

Monroe, R. (1987). Testing: A political scalpel. *The English Journal, 76*(8), 24.

Office of Technology Assessment, United States Congress. (1992). *Testing in American schools: Asking the right questions*. Washington, D.C.: United States Government Printing Office.

Parkinson, P. (2009). Political economy and the NCLB regime. *The Educational Forum, 73*, 44-57.

Pfeffer, J., & Salancik, G. R. (2003). *The external control of organizations*. Stanford, CA: Stanford University Press.

Phelps, R. P. (2003). *Kill the messenger: The war on standardized testing*. New Brunswick, NJ: Transaction Publishers.

QSR, International Pty Ltd. (2014): NVivo qualitative data analysis software (Version 10). .

Ravitch, D. (2010). *The death and life of the great American school system*. New York, NY: Basic Books.

Ravitch, D. (2013). *Reign of error: The hoax of the privatization movement and the danger to America's public schools*. New York, NY: Alfred A. Knopf.

Reese, W. J. (2013). *Testing wars in the public schools: A forgotten history*. Cumberland, RI: Harvard University Press.

Resnick, D. P. (1981). Testing in America: A supportive environment. *The Phi Delta Kappan, 62*(9), 625-628.

Resnick, D. P., & Resnick, L. B. (1985). Standards, curriculum, and performance: A historical and comparative perspective. *Educational Researcher, 14*(4), 5-20.

Rothman, R. (1995). *Measuring up: Standards, assessment, and school reform*. San Fransisco, CA: Jossey-Bass Publishers.

Salganik, L. H. (1985). Why testing reforms are so popular and how they are changing education. *The Phi Delta Kappan, 66*(9), 607-610.

Saxe, J. G. (1873). *The Poems of John Godfrey Saxe, Complete Edition*. Boston, MA: James R. Osgood and Co.

Snyder, L. L. (2015). *2016-2018 Strategic Plan: Lakeville Area Public Schools*. Lakeville Area Public Schools. Minnesota. Retrieved from http://isd194.org/about/strategic-plan/

Spellings, M. (2010). Measuring the value of accountability. *U.S. News & World Report, 147,* 33-34.

Temin, P. (2014). Low pay, low quality. *Proquest, 3*(3), 1-8.

Thorndike, E. L. (1918). The nature, purposes and general methods of measurements of educational products. *In The seventeenth yearbook of the national society for the study of education* (Vol. 7). Bloomington, IL: Public School Publishing Company.

Turgut, G. (2013). International tests and the U.S. educational reforms: Can success be replicated? *Clearing House, 86*(2), 64-73. doi: 10.1080/00098655.2012.748640

Warren, J. R., & Grodsky, E. (2009). Exit exams harm students who fail them-and don't benefit students who pass them. *The Phi Delta Kappan, 90*(9), 645-649.

Wiliam, D. (2010). Standardized testing and school accountability. *Educational Psychologist, 45*(2), 107-122. doi: 10.1080/00461521003703060

APPENDIX A: EXTENDED LITERATURE REVIEW

Introduction

The hybrid dissertation format in the Department of Educational Leadership & Foundations (EDLF) requires an "extended, but not comprehensive" review of literature to the journal article attached as Appendix A. In a typical EDLF hybrid dissertation that extended review would expand substantially on the basic elements of the article background or review of literature section to add depth and breadth not typically allowed by the restrictions of a relatively short journal article. Since this dissertation itself was comprised of conducting an extensive in-depth analysis of a very large literature base, which was then focused on the presentation of a single thematic element of that analysis, it was determined that an atypical approach to Appendix A would be more appropriate than the approach taken in a more mainstream EDLF dissertation research project.

In order to provide the broader context from which the article theme emerged, and to portray the breadth and complexity of issues involved in standardized testing in American K-12 school communities, Appendix A presents a narrated overview of the entire spectrum of themes and issues explored in the full analysis created for this dissertation project, rather than focusing only on the one theme presented.

This narrative overview is organized according to the NVivo coding structure created in the larger text analysis. The following table of contents provides a visual overview of that coding structure and orients the reader to the overall content of the literature analyzed.

TABLE OF CONTENTS

Standardized testing in traditional, K-12 public schools in the United States is an incredibly large and complex subject. As a result, the body of literature surrounding the concepts of standardized testing is also extensive. Add the fact that there has been intense debate on nearly every facet of testing for over 150 years, and it is easy to understand why the body of literature that surrounds standardized testing is monolithic. This literature review will not attempt to cover every element of standardized testing. Instead, the purpose is to present the key components surrounding standardized testing.

These six sections represent the categorical coding structure from NVIVO. Each of these categories have multiple themes within them:

1. Basic standardized testing knowledge
2. Influences on standardized testing
3. Perceptions of standardized testing
4. Uses and purposes of standardized testing
5. Consequences resulting from standardized testing
6. Alternatives to traditional standardized testing usage

Basic Standardized Testing Knowledge

There are many technical parts of the standardized testing realm that are important to understand in the literature. First, standardized testing itself is a very broad category, and therefore it is important to distinguish the different types of tests that exist that affect stakeholders in traditional K-12 public schools. Second, understanding some of the key historical events in the standardized testing arena in the United States is valuable to give a context to testing. And, third, there are some key technical terms that appear throughout the literature.

Types of Testing

There is sometimes confusion by what is meant by "standardized testing." As Koretz (2008) clarified:

> People incorrectly use the term standardized test—often with opprobrium—to mean all sorts of things: multiple choice tests, tests designed by commercial firms, and so on. In fact, it means only that the test is uniform. Specifically, it means only that all examinees face the same tasks, administered in the same manner and scored in the same way. (p. 23)

But, even with a definition that simple, confusion abounds. Therefore it is important to have a basic understanding of the type of tests that affect students in traditional K-12 public schools.

> **Achievement and IQ tests.** There are many different tests that all fall under the standardized testing umbrella. The first distinction that should be made is the difference between *achievement tests* and *intelligence quotient (IQ)* tests. Henry Chauncey, the founder of ETS and father of the SAT test quoted E.L. Thorndike, one of the most influential early psychometricians, to explain the purpose of the achievement test. Thorndike said, "The point of the achievement test is to find out whether the student has learned what the teacher has been trying to teach him" (Chauncey & Dobbin, 1963, p. 12). There exists a domain of information or skills that a teacher has tried to teach, and an achievement test is an attempt to find out if the domain was learned by the student.

In contrast, an IQ test attempts to measure something completely different. Invented by Alfred Binet in France in the early 1900s, the IQ test was meant to be only a strategy to classify students as feeble-minded or not. From his writings, it appears as though he did not intend it to

become what Lewis Terman later turned it into—a quantifiable measure of innate ability or intelligence that was fixed by heredity. As Binet (1916) wrote:

> The scale properly speaking does not permit the measure of the intelligence, because intellectual qualities are not superposable, and therefore cannot be measured as linear surfaces are measured, but are on the contrary, a classification, a hierarchy among diverse intelligences; and for the necessities of practice this classification is equivalent to a measure. (p. 41)

However, Binet died prematurely and American Psychometrician, Lewis Terman took the idea of IQ and turned it into a scale with which he believed would measure hereditary intelligence accurately in quantifiable terms. As the great education historian Lawrence Cremin suggested:

> There were numerous adaptations and refinements of the Binet scale, the most important of which was the so called Stanford Revision described by Lewis Terman in *The Measurement of Intelligence* (1916). It was Terman, by the way, who popularized the idea of the Intelligence Quotient, a number expressing the relation of an individual's mental age to his chronological age. (1961, p. 186)

Another term that is often used interchangeably with IQ testing is *aptitude testing*.

In this study, the focus is on achievement testing for the most part. But it is impossible to ignore the idea of IQ testing because they both came to prominence in American education at about the same time, and often, those administering or taking the tests didn't understand the differences. Meier (2002) articulated how hard it is to sometimes distinguish between the two. She wrote:

> Whether tests happen to be called aptitude tests or achievement tests, they are actually much the same thing. The SAT acknowledged this when it decided to change the name

> from Scholastic Aptitude Test to Scholastic Achievement test without making any other changes in the test itself. (p. 107)

Paul Diederich, contemporary of Henry Chauncey's, who was a researcher for decades at Educational Testing Service (ETS), expresses a view common in the field: "IQ tests are reading comprehension and vocabulary doctored up to look like reasoning. To change the SAT to an IQ you'd simply divide the score by an age measure. Basically they're the same thing" (Lemann, 1995a, p. 86). The focus in this study therefore is on achievement testing. However, when sorting through standardized testing literature, the idea of IQ continually inserts itself into the conversation.

Norm and criterion referenced tests. Achievement testing has many facets that need to be distinguished from one another. The first is the difference between norm-referenced testing (NRT) and criterion-referenced testing (CRT). On the surface, these tests do not look a lot different from one another. Both are trying to test whether a student has mastered specific domains. However, there are a couple of key distinguishing features. Popham and Husek were among the first to clarify this distinction:

> At the most elementary level, norm-referenced measures are those which are used to ascertain an individual's performance in relationship to the performance of other individuals on the same measuring device. The meaningfulness of the individual score emerges from the comparison. It is because the individual is compared with some normative group that such measures are described as norm-referenced. Most standardized tests of achievement or intellectual ability can be classified as norm-referenced measures. Criterion-referenced measures are those used to ascertain an individual's status with respect to some criterion, i.e., performance standard. It is because the individual is

> compared with some established criterion, rather than other individuals, that these measures are described as criterion-referenced. The meaningfulness of an individual score is not dependent on comparison with other testees. (1969, p. 2)

Most end-of-level testing required by states or districts are usually CRT's. They are designed to see if a student has learned a specific body of knowledge such as the standards outlined in the state biology core. Tests like the ACT or SAT are norm-referenced tests. They are designed to show the differences between the test takers, so that a comparison is fairly easy for someone reviewing the data such as a college admissions committee. It is easy in such an instance to compare a 35 on the ACT to a 23 and make some assumptions based on those numbers.

College entrance exams. The American College Testing (ACT) and the Scholastic Assessment Test (SAT) are used as college (university) entrance exams. This is a category of tests that plays a significant role in what opportunities a student will have after high school. There are also many other college entrance tests, especially at the graduate level (e.g., the LSAT for law school, GMAT for business school, and the GRE as the general graduate school exam). While these play a role in the standardized testing world, only the ACT and SAT will be considered in this literature, largely because of their central influence in the rise of testing as well the impact they have on traditional K-12 public school students. In addition, different stakeholders have made interesting use of them over the years. For example, the SAT results were used by the Reagan administration to help prove that American public schools were failing (Gardner, 1983).

The NAEP. Another standardized test that receives a lot of attention in the literature is the National Assessment of Educational Progress (NAEP). Known as "The Nation's Report Card,"

the NAEP is designed to be a standard to compare other tests to over time. There is probably no other test that has been used more often to criticize the state of public education in America (Duncan, 2013; Hanushek, 1998; Paige, 2002). At the same time it is also used to do just the opposite—point toward improvements and progress (Bishop, 1998; Wiliam, 2010). The data from this test are widely used by both proponents and critics of testing.

The NAEP is also the target of much criticism in what it is really capable of accomplishing:

> With respect to the validity of NAEP classifications, there has been controversy since their creation in 1990. An independent evaluation found that the National Assessment Governing Board 'failed to produce dependable achievement levels for use in interpreting NAEP results [and] failed to produce evidence that the users of NAEP results can and are likely to make appropriate use of the levels in reaching valid conclusions about the meaning of NAEP results.' The National Assessment Governing Board rejected this evaluation and attempted to cancel the evaluators' contract. However, a subsequent evaluation conducted by the General Accounting Office also found that the standard-setting work was seriously flawed. Since then, several other organizations, including the National Academy of Science, the National Academy of Education, and the Center for Research in Evaluation, Student Standards and Testing have questioned the validity of the NAEP performance levels . . . Remember, the process of establishing cut-scores is arbitrary and depends upon subjective decisions. For this reason, alone, it is imperative that high-stakes criterion-referenced testing programs collect and examine independent external evidence to support their classifications of students. (Madaus et al., 2009, pp. 85-86)

International assessments. The last specific type of test that receives a lot of attention in the United States is international tests. These tests provide a comparison between the achievement of students in the U.S. and students from other countries around the world. The main international tests are the Programme for International Student Assessment (PISA), Trends in International Mathematics and Science Study (TIMMS), and Progress in International Reading Literacy Study (PIRLS). The data from these tests are rarely, if ever, used at the state, district, or school level. They are often used at the federal level in the areas of policy and economics. Critics such as Chester Finn use them to prove failure of the system. "While our outcomes remain flat, theirs rise. Half a dozen nations now surpass our high school and college graduation rates. International tests find young Americans scoring in the middle of the pack" (Finn, 2008, p. A7). Economists also use the data regularly to draw conclusions in many different studies.

Student growth measures. The last element in this section that is important to note is a newer phenomenon of growth measures. For most of the history of standardized testing, the idea of progress or growth was usually just an extension of proficiency. If a student, or a school, or a teacher got better scores from one year to the next, it was considered growth, even though the students were different, or the school had changed. There are now much more technical growth measures, such as the Student Growth Percentile (SGP). SGP is a much different look at growth than we have had in the past. As Betebenner (2009), the main proponent of SGP, described it:

> A student's growth percentile describes how (ab)normal a student's growth is by examining their current achievement relative to their academic peers—those students beginning at the same place. That is, a student growth percentile examines the current achievement of a student relative to other students who have, in the past, "walked the same achievement path." (p. 6)

This idea of measuring growth is important to schools, and school systems as they try to monitor the effectiveness of instruction. Proficiency data do not always tell the whole story and growth measures attempt to at least tell more of the story.

The History of Testing

There were a few key events in the last 170 years that must be understood to clarify the current standardized testing landscape. As Reese (2013) elaborates in this regard:

> Testing has an interesting history in the United States. For the purpose of this study, I am looking only at the history of testing in the United States from Horace Mann to the present that affects students in a traditional K-12 public education setting. We sometimes have a tendency to think that testing has changed considerably over the last 170 years, but the arguments are often surprisingly unchanged. (pg. 233)

Reese concluded his book about Horace Mann's first large scale use of testing in America by saying:

> Anyone who hopes to separate politics from testing, which were intertwined from the start, will have to look to something other than history for guidance. Anyone who imagines that recurrent attacks on high-stakes exams will lead to a diminution in the number and authority of tests is surely mistaken. Anyone who believes that more and better exams will resolve problems endemic to standardized testing, however, can find kinship with numerous Americans who dreamed such dreams before. (2013, p. 233)

Horace Mann and the Boston common schools. The man considered the father of modern standardized testing in America is Horace Mann. In the mid-1840s, the Boston Common Schools were growing rapidly, and for the first time education was becoming much more universal for all students. Mann had visited Europe and when he compared the Boston Schools

he realized there was a lack of "system" and that many of the headmasters were mediocre. He decided to create a standardized test to prove that the schools were doing poorly. "To raise standards across the board and ensure accountability, antebellum reformers were the first to rank urban teachers, students, and schools based on quantitative scores, to shame the worst and honor the best" (Reese, 2013, p. 4).

Once the testing phenomena took off in Boston, it spread to schools throughout the nation. By the 1900s testing was an integral part of the American school culture:

> What reformers wrought in Boston was only the beginning. Less than a generation after Boston's controversial experiment, competitive exams were commonplace. By the 1870s and 1880s, testing was so widespread in America's cities that it generated a backlash. Critics complained that testing narrowed the curriculum, undermined the broad purposes of schooling, ruined children's health, and made teachers automatons, forcing them to teach to the test. But written, timed, in-class exams and statistical measures were here to stay, as traditional means of assessing schools lost their legitimacy. By 1900, their influence seemed unassailable. In subsequent decades, high school and college enrollments boomed, academic credentials grew in importance, and tests helped place pupils in ability groups, vocational tracks, and other pathways to adulthood. (Reese, 2013, p. 5)

Mann set in motion modern testing as we know it. Before Horace Mann's experiment with written tests standardized testing in the United States was virtually non-existent. Not long after his experiment, standardized testing was here to stay.

Early psychometricians. Between the mid 1840s and the early 1900s, testing became strongly rooted in public schooling. The belief in statistics and quantitative measurement also

took off. There were a handful of men with hereditarian viewpoints that wanted to *measure* intelligence. Francis Galton was a key figure of this time. "Quantification was Galton's god, and a strong belief in the inheritance of nearly everything he could measure stood at the right hand" (Gould, 1981, p. 76). As Galton himself wrote, "until the phenomena of any branch of knowledge have been subjected to measurement and number, it cannot assume the status and dignity of a science" (1879, p. 149). Near the turn of the century, many of these psychometricians made it their mission in life to *measure* intelligence and ability and turn it into a science.

One of the chief among them was Edward Thorndike. He summed up the view of the era when he wrote:

> Whatever exists at all exists in some amount. To know it thoroughly involves knowing its quantity as well as its quality. . . . To measure a product well means so to define its amount that competent persons will know how large it is, with some precision, and that this knowledge may be conveniently recorded and used. (Thorndike, 1918, p. 16)

While this is among his most famous quotes, it is often overlooked that he also wrote about the limitations of testing, because he understood and knew that it wasn't nearly as accurate as many believed. However, he did much to move testing forward, and always believed that "[t]he educational measurements now in use are much better than none at all. They do excellent service, provided inferences are made with proper caution" (Thorndike, 1918, p. 23). This advice is certainly relevant in the 21st Century.

World War 1 Alpha Tests. Lewis Terman was a contemporary of Edward Thorndike but was much more interested in measuring innate ability or IQ. He made IQ testing a mainstream activity in schools (Cremin, 1961). Terman took the ideas of Alfred Binet and turned

them into something Binet would have likely rejected had he not died a premature death. He made a test that he claimed could assign a single number representing innate intelligence. Robert Yerkes, another of the early phychometricians convinced the U.S. government to allow him to test all Army recruits to be able to more efficiently classify or sort young recruits into the place they could best serve their country (Lemann, 1995b). This was the first intelligence test given on a massive scale, and the results of this test had ramifications on individual recruits as well as the testing industry for years to come.

While testing progressed because of this experiment in mass testing, there were also many errors. For example, "in one 1921 study, Harvard researcher Robert Yerkes concluded that '37 percent of whites and 89 percent of negroes' could be classified as 'morons'" (Gould, 1981, p. 227). Yerkes had no concerns about the results because the tests were "constructed and administered" to address potential biases and were "definitely known to measure native intellectual ability" (Hess & Mehta, 2013, p. 72).

The meritocracy. The Army Alpha tests led to the creation of the college entrance examination. After the War, the President of Harvard, James Bryant Conant wanted to create a system where the most capable young men across the country could be found to come to Harvard instead of just continuing to bring the sons of the wealthy. Harvard admissions in the early 1900s were in essence still an *aristocracy*. A student very rarely attended Harvard, who did not have parents with vast financial resources and significant indicators of social prestige. Conant encouraged Henry Chauncey to create a test for college entrance with characteristics like the Army Alpha Test that could be used to determine who should be admitted to Harvard. The result was the SAT test and the creation of The Educational Testing Service (ETS), which both still play a prominent role in college admissions testing in America (Lemann, 2000).

Chauncey wanted to take it even a step further. With this scientific instrument he envisioned a test that could help sort and select everyone to their proper spot in society. He wanted to

> mount a vast scientific project that will categorize, sort, and route the entire population. It will be accomplished by administering a series of multiple choice mental tests to everyone, and then by suggesting, on the basis of the scores, what each person's role in society should be . . . the project will be called the Census of Abilities. (Lemann, 2000, p. 5)

His broader vision fell far short, but his influence was massive. Chauncey and Conant changed the structure for success in America. From the adoption of the SAT and later the ACT by our society, Chauncey helped many who would not have gone to college gain access to college opportunities, based on their merit instead of their wealth or position. On the flip side, "A test of one narrow quality, the ability to perform well in school, [stood] firmly athwart the path to success" (Lemann, 2000, p. 6).

The contemporary *rite of pass*age from high school to college did not really exist until Henry Chauncey created the SAT at the behest of Conant. It was quickly adopted in many universities, and within a couple of decades every college that considered itself a serious institution would be using it, or the ACT, which was developed shortly after by E.F. Lindquist. "The machinery that Conant and Chauncey and their allies created is today so familiar and all-encompassing that it seems almost like a natural phenomenon, or at least an organism that evolved spontaneously in response" (Lemann, 2000, p. 6). It is hard to underestimate the influence these tests have had on the American educational system.

Sputnik, the National Defense Education Act (NDEA), and the Great Society of the 1960s. Needless to say, the 1960s were an era of tumult and change. When the Russians launched Sputnik into orbit, there was a collective national anxiety and a renewed will to improve education to keep up with the technological advancements internationally:

> We forget today how Sputnik both electrified and challenged Americans and why it prompted us to update our formula so energetically. Within a year of Sputnik's launch, Congress passed the National Defense Education Act, which supported the study of science, foreign languages, and the history, politics, and economies of foreign countries. (Friedman & Mandelbaum, 2011, p. 40)

Sputnik's influence went far beyond just putting more emphasis on math and science in America's schools. "NDEA was a significant act in U.S. educational history because it was the federal government's first involvement in U.S. education. This involvement, however, was not very strict, but it increased and became stricter over the years" (Turgut, 2013, p. 65). It was really the catalyst for the beginning of centralization, and as we will see later, testing became one of the core strategies for a more centralized educational system.

A Nation at Risk (NAR), 1983. If Sputnik and the NDEA were the catalyst to centralize education at the national level, the government report *A Nation at Risk*, written by Terrel Bell's Commission on Excellence in Education was an accelerator toward centralization. This was another major attempt by the federal government to reform education. "Where did education reform go wrong? Ask the question, and you'll get different answers, depending on whom you ask. But all roads eventually lead back to a major report released in 1983 called *A Nation at Risk*" (Ravitch, 2010, p. 22). While it might appear as hyperbole, the influence of NAR can hardly be overstated. The report was important to testing for two reasons: First, testing results

were used to create a sense of failure in public education, and second, it accelerated the use of testing from that point onward as the barometer for system monitoring.

Education in the United States would never be the same after the release of this report. While it has been oft criticized

> *A Nation at Risk* was notable for what it did not say . . . Far from being a revolutionary document, the report was an impassioned plea to make our schools function better in their core mission as academic institutions and to make our education system live up to our nation's ideals. It warned that the nation would be harmed economically and socially unless education was dramatically improved for all children. (Ravitch, 2010, p. 25)

No more could a school or teacher do their work without significant external scrutiny. "Since the release of NAR, the involvement of business leaders in education and its management as a business establishment has increased steadily" (Turgut, 2013, p. 65). And, that scrutiny would come in large waves of accountability that were mostly composed of standardized testing.

Goals 2000 and No Child Left Behind. The natural outgrowth of the *Nation at Risk* report was more testing to monitor and judge schools and school systems at all levels:

> States began experimenting with school accountability systems during the 1980s, but the decade of 1990s began the age of accountability. States generally worked on developing standards for what should be learned in each grade and subject, and these standards were linked to tests of student performance. Finally, states began to link tests to individual schools and to develop rating systems for performance. (Hanushek & Raymond, 2005, p. 306)

Goals 2000 was a federal law that attempted to unify the governors of the states to raise standards and improve education primarily to assure economic competiveness. It became a

natural precursor to the No Child Left Behind Act (NCLB), which was yet one more major move toward centralization.

The real teeth of NCLB were the testing mandates. And, as Jacob clarified "The passage of No Child Left Behind ensures that test-based accountability will be a pervasive force in elementary and secondary education for years to come" (Jacob, 2005, p. 791). That has certainly been the reality of the last 15 years of education. "Relationships between and among academic standards, standardized tests, test scores, and accountability measures provide an indication of the audit culture that frames the NCLB regime" (Parkinson, 2009, p. 45). Though it has been heavily criticized for some elements that were not realistic, like assuring that EVERY child be proficient by the year 2014, "No Child Left Behind—or NCLB—changed the nature of public schooling across the nation by making standardized test scores the primary measure of school quality" (Ravitch, 2010, p. 15).

We are now beginning a new chapter in standardized testing. The *Every Student Succeeds Act* (ESSA) is the newest iteration of the *Elementary and Secondary Education Act* (ESEA) and replaces the 15 year-old NCLB. It is being hailed widely as a major decentralization. But, read carefully and it becomes apparent that most of the testing mandates from the federal government are still in place. Testing will continue to play a very significant role in education and educational reform for the foreseeable future.

Keywords in Standardized Testing

The last section of testing elements that are important to understand revolves around some of the technical elements of testing. While this will only clarify some of the main concepts in testing, they are elements that have to be understood to have any intelligent understanding of the debate about standardized testing:

> The core principles and concepts are truly essential. Without an understanding of validity, reliability, bias, scaling, and standard setting, for example, one cannot fully make sense of the information yielded by tests or find sensible resolutions to the currently bitter controversies about testing in American education. Many people simply dismiss these complexities, treating them as unimportant precisely because they seem technical and esoteric. (Koretz, 2008, p. 14)

The discussion of many of these concepts comes from Daniel Koretz, a testing expert and professor at Harvard University.

Reliability. Reliability simply means that a person is likely to get the same score (or very close to the same score) each time he or she takes the test. **"**Reliable scores show little inconsistency from one measurement to the next—that is, they contain relatively little measurement error. Reliability is often incorrectly used to mean 'accurate' or 'valid' but it properly refers only to the consistency of measurement" (Koretz, 2008, p. 30).

For example, the ACT would not be reliable if a student took it 3 times and got a 17, a 32, and then a 24. When students take the ACT, they rarely fluctuate more than a couple of points. It is more likely that a student might get a 24, a 25, and a 24. That type of between-testing scoring consistency for an individual tester is what is referred to as reliability.

Validity. If we had to pick one single element that makes or breaks standardized testing, validity is it. Validity simply refers to whether the test is actually measuring what the test maker claims it is measuring. As Rothman (1995) notes:

> Perhaps the most important criterion in evaluating tests is validity. Generally speaking, a test is valid if it measures what it is supposed to measure. A calendar is a valid measure of time but not a valid measure of temperature, even though it is usually colder in winter

> than in summer. Similarly, a test might be a valid measure of a student's mathematics achievement but not a valid measure of a school's quality, even though many schools with large numbers of high achievers are good. (p. 152)

Validity is an interesting construct because most of the time we talk about validity in terms of whether the instrument itself is valid. With any credible test, it usually does address what its creators say it measures.

However, the real threat to validity has more to do with the uses made of test scores. The SAT is a statistically valid predictor of how well college freshman are likely to perform, but in the early 1980s, it was used to show that declining trend scores in the U.S. were proof that we were "A Nation at Risk." That use was not a valid use of the SAT data. Koretz (2008) emphasizes this overlooked element of validity. He says:

> Validity is the single most important criterion for evaluating achievement testing. In public debate, and sometimes in statutes and regulations as well, we find reference to "valid tests," but tests themselves are not valid or invalid. Rather, it is the inference based on test scores that is valid or not. A given test might provide good support for one inference but weak support for another. (p. 31)

He also points out that we should be aware of the elements that undermine validity:

> Before considering the evidence used to evaluate validity, we should start by asking what factors could undermine validity, making our conclusions unjustified. There are many of these, of course, but they fall into three broad categories: failing to measure adequately what ought to be measured, measuring something that shouldn't be measured, and using a test in a manner that undermines validity. (pp. 219-220)

The largest overall area of focus in this study deals with the use of standardized testing. The use of standardized testing data is what drives the majority of tension in the testing debate. Therefore, Koretz's third factor—using a test in a manner that undermines validity—becomes a crucial element in the study of standardized testing.

Bias. Bias refers to the idea that test questions or tests themselves favor some students or group of students over others. For example, if there is a question on a test that refers to farming, it could benefit a student familiar with farming while hindering students that grew up in large cities and have never been on a farm. Other examples have more significant implications when it affects groups based on socio-economic status, gender, race or other groups. When writing tests for millions of students, it difficult to avoid bias. No matter what the question is, it is likely favor some students over others.

While test makers work diligently to limit bias as much as possible in individual test questions, Meier (2002) points out that tests are designed to expose the differences between students, and therefore are biased in and of themselves:

> The bias is in the nature of the tool . . .it is necessarily steeped in prior cultural assumptions—norms—that favor some kids over others. This is not a question of test makers having anything against any particular group of test takers; the nature of such tests requires that they must discriminate and rank order on some basis. If all testees responded the same way, the question would be a bad item; if the "wrong" people got it right, that would also make it a bad item. (p. 111)

In addition to that form of bias, there are other forms that can affect whole groups of students. "For example, a mathematics test that requires reading complex text and writing long answers

may be biased against immigrant students who are competent in mathematics but have not yet achieved fluency in English" (Koretz, 2008, p. 13).

Error. "Standardized Tests are not perfect and neither are the human beings who develop, administer, or score them" (Phelps, 2003, p. 265). There are many different points that can produce mistakes in the results of standardized testing:

> Human error can be, and often is, present in all phases of the testing process. Error can creep into the development of items. It can be made in the setting of a passing score. It can occur in the establishment of norming groups, and it is sometimes found in the scoring of questions. (Rhoades & Madaus, 2003, p. 28)

Measurement. We have come to accept standardized testing as measurement to the extent that any reader would likely wonder why this would be considered an important testing term in this study. Already in this appendix *measure* or *measurement* has appeared multiple times. Why is this an important term? Interestingly enough, the concept of measurement is crucial to the standardized testing debate. It is a term that has become nearly universal, but it is also a term that can be misleading.

The earliest psychometricians laid claim to measurement. "Psychometry, it is hardly necessary to say, means the art of imposing measurement and number upon operations of the mind" (Galton, 1879, p. 149). This enthusiasm for testing was echoed by practitioners in education in the early 1900s. One school practitioner stated that "accurate measurements of the abilities of students may be made by using standardized tests" (Monroe, 1918, p. 19).

But not all people adopted these beliefs early on. The arguments 100 years ago are eerily similar to the arguments that we hear today. Walter Lippmann, an American writer, political commentator, and reporter criticized early on this idea that testing was a measurement:

> Because the results are expressed in numbers, it is easy to make the mistake of thinking that the intelligence test is a measure like a foot rule or a pair of scales. It is, of course, a quite different sort of measure. For length and weight are qualities which men have learned how to isolate no matter whether they are found in an army of soldiers, a heap of bricks, or a collection of chlorine molecules. Provided the footrule and the scales agree with the arbitrarily accepted standard foot and standard pound in the Bureau of Standards at Washington they can be used with confidence. But 'intelligence' is not an abstraction like length and weight; it is an exceedingly complicated notion which nobody has as yet succeeded in defining. (1922d, p. 247)

Yet, the arguments have remained largely the same. The tension around this idea of measurement is nearly identical to the debate of 100 years ago. The Secretary of Education behind the report *A Nation at Risk* wrote, "All these goals should be quantified, so that the nation's progress toward achieving them can be measured annually" (Bell, 1988, p. 402). President George H.W. Bush said, "As a first step in this strategy, we must challenge not only the methods and the means that we've used in the past, but also the yardsticks that we've used to measure our progress. Let's stop trying to measure progress in terms of money spent" (Bush, 1991, p. 7). The general public accepts testing as measurement because our testing paradigm was shaped that way.

The father of ETS captured the essence of this measurement debate as well as anyone. We would do well to heed his wisdom from 50 years ago:

> It is important here to point out a fact that has escaped most laymen. No test or technique measures mental ability directly. What Binet did, what all other "intelligence test" builders after him have done, was to set up some tasks for the young intellect to attack

> and then to observe what happened when that intellect was put to work on them. (Chauncey & Dobbin, 1963, p. 3)

As Chauncey and Dobbin point out, even Alfred Binet never considered testing as measurement. Instead he looked to intelligence testing as a classification. The difference between measurement and classification is significant, and framing it by one or the other makes a significant difference in how we view and use standardized testing data.

Testing for classification is certainly not new. Binet himself wrote about his own tests that the "scale properly speaking does not permit the measure of the intelligence, because intellectual qualities are not superposable, and therefore cannot be measured as linear surfaces are measured, but are on the contrary, a classification, a hierarchy among diverse intelligences" (Binet & Simon, 1916, p. 41). But, Binet was not completely free of the idea of measurement either. Because he followed up the last sentence with "and for the necessities of practice this classification is equivalent to a measure."

Testing as classification instead of measurement is articulated well by Garrison (2004). He points out that there are many things that can be classified effectively, but not necessarily measured accurately. He wrote:

> A key problem, one pointed to above with the definition of psychometry as mind measurement, is the assumption that the mind or a purported function of mind is a property capable of gradation. There are many properties that do not permit gradation—such as Pilsner, feline, wooden, and human. In other words, the psychometric dictum of E. L. Thorndike that if something exists, it must exist in some amount is false. (2004, pp. 64-65)

Regardless of how someone views standardized testing, this understanding of the difference between "measurement" and "classification" is an important distinction. It helps shape our paradigm about testing to consider what standardized tests are really doing. Popham (1999) gives us an example of why measurement is not always an effective framework to work from—it skews what we view as the purpose of testing. "Employing standardized achievement tests to ascertain educational quality is like measuring temperature with a tablespoon . . . Standardized achievement tests have a different measurement mission than indicating how good or bad a school is" (p. 10).

Achievement gap. The achievement gap simply refers to the differences in performance on standardized tests between groups. For example, how do minority groups perform compared to the majority group? How does the special education group or English as a second language group perform compared to the rest of the population? How does the low socio-economic group perform compared to the rest? One of the most significant mandates behind NCLB was that these achievement gaps must be monitored and closed. The whole point of NCLB was to raise the achievement of all groups so that achievement gaps don't exist.

However, many authors point out that testing is not going to solve the problem of achievement gaps. Ravitch (2013) criticized the NCLB testing mandates by saying that

> Even more curious is the unwarranted belief that more testing and accountability will close the achievement gaps between rich and poor, blacks and whites, and Hispanics and whites. Since the source of the gaps is socioeconomic inequality, it is sheer fantasy to believe that the test scores of these groups will converge if only there are higher standards plus more testing and accountability. The assumption is that those who teach

the low-performing groups are not really trying, and a carrot or a stick will motivate them to try harder." (p. 266).

Simpson's paradox. This is a phenomena related to achievement gaps. It is a situation in standardized testing statistics where every subgroup could be improving on tests, yet the entire group proficiency totals could actually be going down. It happens when there is a significant shift in demographics in a population. So, for example, let us create a simplistic situation to demonstrate. If we have a school with exactly 100 students, we can easily see how this works. Let us assume that the school only has 2 subgroups—regular education students and ESL students. The first year they test, there were only 10% ESL students and the other 90% of students were drawn from the so-called regular education group. The groups scored like this the first year:

Table 1

Simpson's Paradox Example 1

Group	% proficient	% of Population	# proficient
ESL	50%	10%	5
Regular Ed	80%	90%	72

So, if we do the math, 5 ESL students passed the test, and 72 regular education students passed the exam. So, we have a total of 77 students that passed the exam. Or, since we have exactly 100 students in the school, 77% of the students passed.

Then, the next year, there was a major shift in the economy and there was a lot of mobility in the community. Many regular education students moved out, and many ESL students

moved in. (But we still ended up with exactly 100 students) The second year, the test results looked like this:

Table 2

Simpson's Paradox Example 2

Group	% proficient	% of population	# proficient
ESL	56%	50%	28
Regular Ed	84%	50%	42

The regular education students increased their proficiency by 4% and the ESL improved by 6% (the achievement gap was even slightly closed) yet we went from 77% of our students being proficient down to 70%. There was a 7% drop in overall proficiency even though both subgroups improved.

This is obviously an extremely simplified example to illustrate the point, but this is something that happens, and is often overlooked. No school is likely to have those kinds of demographic shifts over only one year, but over longer periods those kinds of shifts in subgroups are quite common.

Campbell's law. This is another interesting phenomenon that can affect the world of standardized testing. Basically it is the idea that

> the more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it was intended to monitor. In the case of standardized testing, corruption and distortion can come in a variety of ways. (Bower, 2013, p. 26)

For example, when a school is graded or judged solely on the percentage of students that are proficient, it can be very easy for a school to focus solely on the *bubble kids* (the students that are right on the verge of passing or not passing) and do test prep with them. More learning might not actually happen, but the score will go up because the school did just enough to get that percentage of children to pass the test.

There are many examples of this phenomenon outside of schools as well. One of the most obvious examples is the "on-time rates" for airlines. If you have a flight cancelled for weather, rather than putting you and your fellow passengers on the next plane out, you will find yourself sitting instead, watching many flights leave to your destination filled with other passengers, and eventually you will get a plane later. Why? Because if the airline gives you and your fellow passengers the next plane, then that means both your plane and the next plane will be counted as delayed, rather than just yours. And, since "on time ratings" are one of the most important data points for the airline industry, they make decisions that don't make sense in order to influence the ratings. There are many other examples that are quite interesting such as Vietnam body counts during the war, speeding ticket quotas, TV sweeps months, among many others" (Rothstein, 2008, pp. 15-17).

NCLB-type attempts to raise test scores puts Campbell's law into motion in most educational settings where there is a lot of pressure to improve a particular data point.

Influences on Standardized Testing and Test Results Data

The meaning of the data collected from standardized testing means can't be fully understood without understanding some of the context in which standardized tests are created, administered and interpreted. This section will address the most significant influences on standardized tests. Much of the literature addresses these elements, because without considering

the influences on testing, anyone using the data is likely to misinterpret the data. For example, Alfie Kohn, an outspoken critic of testing, emphasizes that

> Results should always be evaluated in light of the special challenges faced by a given school or district: A large number of students with special needs, or a very low-income community, provides a necessary context in which to understand a set of results. (2000, p. 47)

Whether a person is a critic or proponent of testing, this is wise advice.

Demographics

There is a statistical correlation between demographics and performance on standardized tests. There are many reasons for this, but it is a well-documented reality. "Research dating back to the 1966 release of Equality of Educational Opportunity (the 'Coleman Report') shows that student performance is only weakly related to school quality. The report concluded that students' socioeconomic background was a far more influential factor" (Goldhaber, 2002, p. 51). A student's background is one of the biggest influences on how they do in school and therefore has a significant impact on how that student will likely perform on standardized tests:

> Demographic differences should not be an excuse for low performance, but critics who ignore the impact of social factors on test scores miss the point: the reason to acknowledge their influence is not to let anyone off the hook but to get the right answer. Certainly, low scores are a sign that something is amiss; after all, finding out where performance is strong or weak is one of the primary reasons for administering tests. But the low scores by themselves don't tell *why* achievement is low and are usually insufficient to tell us where instruction is good or bad, just as a fever by itself is insufficient to reveal what illness a child has. Disappointing scores can mask good

> instruction, and high scores can hide problems that need to be addressed. (Koretz, 2008, p. 120)

Addressing and acknowledging the demographics of any school community should help educators ask important questions in both high and low socioeconomic communities. Often, demographics are only a consideration in low SES schools. It is just as important to ask demographics questions in high SES schools, though it rarely happens. It is also important to recognize demographic shifts:

> The American family structure is changing, and teachers are encountering more children from single parent homes and homes where both parents work. These demographic changes are real, persistent, and accelerating. They will drive change in education, and other social institutions as well, especially since we continue to accept the challenge to educate all of our youth. (Carson, Huelskamp, & Woodall, 1992, p. 304)

Poverty

Obviously a subset of demographics, but seemingly one of the most influential elements of demographics, poverty deserves some attention of its own:

> If school opportunities are equal, then the results of schooling should be affected primarily by such characteristics as ability and willingness to learn. But if poverty, let us say, affects the student's ability or willingness to learn, then equal schooling will result in unequal achievement. Poor students will do less well than their affluent peers because they are less capable of profiting from the opportunity. The result of equal opportunity is that those who enter school behind will leave it behind. (Cooper et al., 2004, pp. 209-210)

Poverty is rarely included in the common reports like newspapers or other news stories about standardized testing—when was the last time you saw a grading schools report that also included poverty data in the report? But, as one economist points out, socioeconomic factors matter. "As with their earlier analysis of the wage gap, the explanatory power of test scores is overstated because they do not include individuals' socioeconomic status and education in the equations" (Levin, 2001, p. 44).

From a teacher's perspective, it is pretty obvious that this element of poverty matters. One teacher commented, "I don't want to sound like a whiner looking for excuses, but parents don't send me standardized kids. At least one-fourth of the kids in the U.S. live in poverty and sit in school buildings that are crumbling around them" (Ohanian, 1997, p. 33). Ravitch (2013) sums up this idea pretty well, and emphasizes that schools can't overcome this influence alone:

> The schools did not cause the achievement gaps, and the schools alone are not powerful enough to close them. So long as our society is indifferent to poverty, so long as we are willing to look the other way rather than act vigorously to improve the conditions of families and communities, there will always be achievement gaps. (p. 62)

Parents' Education Level and Home Culture

A student's ability to succeed in school and on standardized tests is influenced by the culture in the home and by parents' education level:

> Student test score gains are also strongly influenced by school attendance and a variety of out-of-school learning experiences at home, with peers, at museums and libraries, in summer programs, on-line, and in the community. Well-educated and supportive parents can help their children with homework and secure a wide variety of other advantages for them. Other children have parents who, for a variety of reasons, are unable to support

> their learning academically. Student test score gains are also influenced by family resources, student health, family mobility, and the influence of neighborhood peers and of classmates who may be relatively more advantaged or disadvantaged. (Baker et al., 2010, p. 3)

Friedman talked about a study in which

> children growing up in homes with many books get 3 years more schooling than children from bookless homes, independent of their parents' education, occupation, and class. This is as great an advantage as having university-educated rather than unschooled parents, and twice the advantage of having a professional rather than an unskilled father. It holds equally in rich nations and in poor; in the past and the present; under communism, capitalism, and Apartheid; and most strongly in China. The study went on to say that Chinese children who had five hundred or more books at home got 6.6 years more schooling than Chinese children without books. As few as twenty books in a home made an appreciable difference. (Friedman & Mandelbaum, 2011, p. 114)

Heredity

Heredity plays a role in how well students can do on standardized tests. It is clear that learning disabilities exist, and that some individuals clearly have strengths and talents that exceed others. However, the role of heredity was believed to be much stronger in the early years of testing. Heredity was believed to be *the* factor that mattered most. One famous early psychometrician wrote:

> Within the past ten years we have also worked out and perfected another new and very important means whereby it is now possible to measure and classify children on the basis of their intellectual capacities. From the use so far made of this new measuring stick in

> retesting children once measured, at a later age, it is now confidently asserted that the degree of intelligence which a child has at six or eight or ten years of age is the degree of intelligence he will retain through life. . . .This is a matter of his racial and family inheritance, and nothing within the gift of the schools or our democratic form of government. (Cubberley, 1919, pp. 450-451)

This paradigm about heredity being a fixed trait led to many erroneous interpretations of standardized testing. Lewis Terman, the father of American IQ testing, concluded, "The children of successful and cultured parents test higher than children from wretched homes for the simple reason that their heredity is better" (Gould, 1981, p. 183). Another typical misinterpretation of the time was that "failure of blacks to attend school, he argued, must reflect a disinclination based on low innate intelligence. Not a word about segregation, poor conditions in black schools, or economic necessities for working among the impoverished" (Gould, 1981, p. 219). It is obvious how dangerous these kinds of assumptions and conclusions can be when testing is approached with the idea that heredity is one of the biggest contributing factors.

But, not everyone bought into the heredity argument. In the early 1900s, Walter Lippmann wrote a scathing critique of the hereditarian claims:

> The claim that we have learned how to measure hereditary intelligence has no scientific foundation. We cannot measure intelligence when we have never defined it, and we cannot speak of its hereditary basis after it has been indistinguishably fused with a thousand educational and environmental influences from the time of conception to the school age. The claim that Mr. Terman or anyone else is measuring hereditary intelligence has no more scientific foundation than a hundred other fads, vitamins and glands and amateur psychoanalysis and correspondence courses in will power, and it will

> pass them into that limbo where phrenology and palmistry and characterology and the other Babu sciences are to be found. In all of these there was some admixture of primitive truth which the conscientious scientist retains long after the wave of popular credulity has spent itself. (Lippmann, 1922b, p. 11)

Policy

Public policy is one of the biggest influences on standardized testing. Policymakers are heavily influenced by economic literature. Interestingly enough, the primary tool of most educational economists is standardized testing data. Therefore it creates an interesting cycle of using the data from the tools that policymakers use to control education in the first place. "The tests offer administrators and politicians a set of numbers to justify whatever 'policy du jour' they are pursuing" (Harris et al., 2011, p. 68). These ideas will be explored further in other sections.

However, policy is crucial to education, and as a corollary, to testing:

> Whatever their age and stage, better policies hold the key to improving education for all children. These policies seek to change the vision and mission of education; enhance the way teachers work and what they teach; increase the human, capital, and financial resources to schools; and ensure that standards are high and results forthcoming. In fact, all these policy areas need to work in concert—purpose, teaching, funding, standards, and assessment—if schools are to improve. (Cooper et al., 2004, p. 2)

Up until about the 1960s standardized testing policy usually took the form of what inputs were needed to improve education. After that, the focus shifted to outputs. That shift was accelerated by NCLB:

> The landmark NCLB codified a developing policy view that standards, testing, and accountability were the path to improved performance. Much of earlier educational policy, both at the federal and state level, concentrated on providing greater resources—especially for the education of disadvantaged students. But student out-comes proved noticeably impervious to these policy initiatives. As a result, federal policy made a distinct shift in focus to emphasizing performance objectives and outcomes rather than school inputs. (Hanushek & Raymond, 2005, pp. 297-298)

Just before testing really got a boost from NCLB, Lattimore (2001) wrote:

> Testing has become an attractive option for policymakers both because it has the potential to affect the behavior of educators in the educational system and because it is often viewed by the public as a way to guarantee a basic level of quality education. Whatever the reasons, formal testing tied to grade promotion and graduation continues to spread throughout the United States. (p. 57)

Lattimore's predictions of the spread of testing have certainly been realized in the last 15 years with NCLB. Many others came to the same realization—testing is one of the great tools to drive policy. "The common thread that links these human perspectives on the meaning of test scores as administrative mechanisms by which to implement one or another policy. In each case, testing as an administrative device has become the linchpin of policy" (Madaus, 1985, p. 612). Another author wrote, "Evaluation and testing have become the engine for implementing educational policy" (Mehrens, 1998, p. 22).

Stakeholder Influence

The people and organizations affected by standardized testing work to influence testing. That would include those with high-stakes consequences from the tests such as students, parents,

principals, and teachers. It also includes those who write policy that affects testing, and business leaders interested in worker productivity, as well as organizations that are critics or proponents of testing such as The American Educational Research Association (AERA), American Psychological Association (APA), and the National Council on Measurement in Education (NCME) (AERA, 2000).

As mentioned multiple times in the literature, most tension revolves around the issue of test data use:

> Emphasizing data use begs the question: Data use by whom? Which stakeholders (e.g., policy makers, administrators, teachers, and/or parents) should use data, which data should they use, and how do we envision them using it? Answering this question involves a thorough explication of the theory of action associated with increases in education efficacy. Breiter and Light (2006) reported on the difficulties of data use and pointed out several impediments to effective data use, particularly in situations involving practitioners. Much greater care must be taken to get the right data, to the right people, at the right time, and in the right format. Turning data into information and ultimately into knowledge requires concerted effort that involves striking an ideal balance between data quality, data availability, and data use. (Betebenner & Linn, 2010, p. 20)

One of the biggest problems surrounding this use by stakeholders is the misinterpretation of test data. An entire section will be committed to this idea, but it is important to note here that

> it is commonly assumed that typical users of assessment results, such as policymakers, educators, and members of the general public, understand the information that is typically included in test results reports. However, a body of research has been compiled in the

> past decade that indicates this assumption is, unfortunately, untrue. (Goodman & Hambleton, 2005, p. 104)

Because very few stakeholders have the knowledge to be able to make accurate or appropriate interpretations of the data, most stakeholders resort to making the data say what they want it to say:

> People search for explanations of specific test scores that are, for whatever reason, of particular interest to them. Parents want to identify effective schools for their children; politicians want to claim credit for successful initiatives or to use low scores to justify reforms; newspapers want to highlight supposed differences in school quality; critics want to identify failures; educators want to claim success and so on. (Koretz, 2008, pp. 132-133)

Madaus et al. (2009) give valuable insight to stakeholders that are seeking out the data for their individual purposes. Members of the testing community "value testing, but recognize its limitations. It is crucial that the public also understand that high-stakes testing is a paradoxical policy strategy that affects—for both good and ill—individuals students, teachers, schools, and communities (p. 3).

Testing as Big Business

An often-overlooked influence on standardized testing is the fact that testing is big business. Money, especially in large amounts, will always be a significant influence. Even back in the early 1900s "testing soon became a multi-million dollar industry" (Gould, 1981, p. 177). Reese's (2013) history of testing concluded that

> Testing is now a multi-billion dollar enterprise. Test prep companies abound, and reformers try to link teachers' salaries to student scores. Charter school advocates

> promise better measurable results than regular schools. States compete for lucrative federal grants in a frantic "race to the top," an educational Mount Olympus ruled by school innovators, high priced consultants, and the testing gods. (p. 3)

Testing as a big business is not likely to change anytime in the near future. One of the unique characteristics of the standardized testing industry, however, is its lack of oversight. Critics are vocal on this point. To have an industry that affects nearly every person in the United States with almost no government regulation or oversight is problematic.

The U.S. government regulates virtually everything from financial transactions to how much insulation is required in new home construction, yet,

> Today those who take and use many tests have less consumer protection than those who buy a toy, a toaster, or a plane ticket. Rarely is an important test or its use subject to formal, systematic, independent professional scrutiny or audit. Civil servants who contract to have a test built, or who purchase commercial tests in education, have only the testing companies' assurances that their product is technically sound and appropriate for its stated purpose. Further, those who have no choice but to take a particular test–often having to pay to take it–have inadequate protection against either a faulty instrument or the misuse of a well-constructed one. Although the American Psychological Association, the American Educational Research Association, and the National Council for Measurement in Education have formulated professional standards for test development and use in education and employment, they lack any effective enforcement mechanism. Despite widespread use of testing in education and employment, there is no US agency (analogous to the Federal Trade Commission or the Federal Aviation Administration) that independently audits the processes and products of testing agencies. The lack of oversight

> makes errors difficult to detect. Individuals harmed by a flawed test may not even be aware of the harm. Although consumers who become aware of a problem with a test can contact the educational agency that commissioned it, or the testing company; it is likely that many problems go unnoticed. (Rhoades & Madaus, 2003, p. 7)

It is notable that something that has high-stakes consequences on nearly every American child has no government oversight. Madaus et al. (2009) ask a pertinent question. "What other entity in society could subject 30 million children to a treatment without an independent mechanism to monitor the quality and effects of that treatment?" (pp. 197-198).

Perceptions of Testing

Another major category in the literature revolves around the perceptions of testing. After all, testing is merely a tool. The tool in and of itself is neither good nor bad, effective or ineffective, meaningful or meaningless. It is the way the tool is used and perceived that creates complications. The next major category will address what the literature says about the uses and purposes of standardized tests, which is where most of the real tension comes from in the standardized testing debate. First, however, we will explore the perceptions surrounding standardized testing, because both the uses of and the criticism of standardized testing stem from these perceptions.

Tests Are Accepted by the Public as Legitimate

Standardized testing has become such an integral part of our educational process that there are few people who haven't been affected in some substantial way by their use. For the most part, "high-stakes tests are viewed by the public as accurate measures of a student's ability and skills. The higher the score, the smarter the student. The higher the aggregate scores of a

given school, the 'better' the school" (Casbarro, 2005, pp. 22-23). One proponent of testing goes so far as to say,

> Public support for widespread and consequential use of standardized testing is overwhelming, and has been since pollsters first posed questions about tests. Over several decades, the scale of the large magnitude of public support has barely budged. Indeed, the public would like to see standardized tests administered: more often (more than once a year), and for all the purposes for which they are now administered, as well as some for some others. (Phelps, 2005, p. 21)

Even the critics of testing acknowledge that

> Standardized tests enjoy widespread support in U.S. public opinion. Eighty-one percent of a national sample of parents surveyed by the Gallup Organization in 1979 indicated that they thought standardized tests were "useful" or "somewhat useful." Only 17% thought tests were "not too useful." Other polls have shown the same positive attitude toward testing. (Resnick, 1981, p. 625)

Since the beginning of standardized testing in the United States with Horace Mann, tests have been accepted generally by the public as a legitimate way to check on students and the schools. "Standardized testing programs possess a number of seductive aspects that make policymakers and the public at large amenable to implementing them in the name of educational improvement. Tests are trusted and desired by a majority of American adults" (Airasian, 1987, p. 394). Madaus et al. (2009) sum up the way the public's acceptance of standardized testing and the place it holds in our discourse quite effectively:

> Today, testing is woven into the fabric of our nation's culture and psyche. Chatter about test results can be heard on the playground, at book groups, offices, dinner parties, and

> the supper table. Testing is a focus of business executives, politicians, policymakers, and think tanks from the right, left, and center. It is an issue of concern for diverse organizations from the Business Roundtable to the National Council of Churches, from the National Governors Association to the Children's Defense Fund. It is a topic for talk show hosts, the media and entertainment industries. Even Disney has tackled the issue of testing by creating a web-site called "What Every Parent Should Know About Standardized Testing." (p. 4)

Historically, there have been waves of backlash against standardized testing, but even though they seem to gain limited traction at times, they rarely endure. The current public school environment with the common core seems to be in the middle of one of these backlashes against testing, but if history is any indication, testing will continue, and just as importantly, the public in general will continue to view standardized testing as legitimate.

Standardized tests are fair and objective. This perception is related to the fact that standardized tests are accepted by the public as legitimate measures. If the tests were not perceived as fair and objective they likely would not be accepted by the general public. Tests are perceived by most people to be "'scientific' because they produce a numerical score, 'fair' because all students are required to take and pass the identical test, and 'objective' because decisions made from their scores are not greatly influenced by teachers', principals', or parents' personal biases" (Airasian, 1987, p. 394).

Because testing is viewed as fair and objective, we often accept testing without a lot of scrutiny. Testing has the potential to heavily influence an individual's education and future opportunities:

> Educational testing results can open or close doors of opportunity from kindergarten through school and beyond, into one's job as a firefighter, sales clerk, or lawyer. Decisions based on state testing programs in elementary school can influence the type of secondary education one receives, and decisions at the high school level can affect one's path after graduation. All of these decisions are based on the quantification of performance—on numbers—and this bestows on them the appearance of fairness, impartiality, authority and precision. (Rhoades & Madaus, 2003, p. 1)

The view that standardized testing is objective and fair is one of the most desirable elements of testing—it has almost become a core American value:

> There is a deeply American quality to this reliance on tests; they were a remarkable invention of social engineering in large part because they did not appear to require a tradeoff between efficiency and fairness—they rather spectacularly seemed to achieve both goals at once. I would argue that standardized testing became a symbol of the aspiration for fairness and universal access that distinguished American schools from European and Asian schools. (Feuer, 2011, p. 26)

One facet that creates the perception of fairness is the quantitative nature of standardized tests. The general public tends to accept standardized testing because numbers don't lie:

> A number connotes objectivity or, at the very least, legitimacy. Because we perceive numbers and statistics as having a certain force on its face (just by being quantitative), we allow statistics to shape our perception of the world and the issues we perceive as important. (Dorn, 1998, p. 2)

Madaus et al. (2009) echo this sentiment when they say that "Our society trusts numbers. We like tests because they provide numeric scores. Our society believes test scores are fair, impartial, and precise" (p. 22).

Standardized testing has taken on many new uses:

> In recent years, several states have adopted accountability systems that provide financial rewards for principals and teachers that are tied directly to school-level performance on standardized tests. These programs are politically appealing because test scores are commonly accepted as objective measures of school performance . . . However, the early returns from research on accountability systems are mixed. (Neal, 2002, p. 35)

Sometimes these uses are easy to implement because they seem objective and fair even though they may not be in reality.

It seems that nearly all types of standardized tests take on this aura of being fair and objective, even though the people reading about the results, or even *using* the results, often know nothing about the tests. A good example is the international tests. Very few Americans have any idea what is on these tests, or who made them, or who is tested, or how reliable or valid they are. Yet we often just accept the fact that America is behind the rest of the world on these tests:

> The major international tests are the PIRLS, TIMSS, and PISA. Data for these tests are meticulously collected and analyzed. Therefore, they are perceived as objective and accurate tools with which to compare different countries, evaluate educational standards, and discuss potential needs for reforms. At least this is how U.S. bureaucrats seem to make use of these standardized tests. (Turgut, 2013, p. 66)

Critics of testing are often blunt in asserting how unfair and un-objective tests really are. "What was meant as a tool for diagnosing learning needs becomes the 'objective' basis for denying learning opportunity" (Robinson, 1990, p. 90).

Standardized testing is scientific. This perception is closely related to the perception that testing is fair and objective. Because of the view that a test is scientific in nature, people are more likely to view it as impartial. The early psychometricians were determined to create a science of psychology. Monroe (1918) simply stated, "Standardized tests have been scientifically devised" (p. 19).

As Galton (1879) articulated, "until the phenomena of any branch of knowledge have been subjected to measurement and number, it cannot assume the status and dignity of a science" (p. 149). Cattell echoed his view. "Psychology cannot attain the certainty and exactness of the physical sciences, unless it rests on a foundation of experiment and measurement . . .The results would be of considerable scientific value in discovering the constancy of mental processes . . ." (p. 373). Yet another early psychometrician compared testing to the new scientific methods of farming:

> Teaching without a measuring stick of standardized length, and without definite standards for the work of the different grades, is much like the old time luck-and-chance farming, and there is no reason to think that the introduction of well-tested standards for accomplishment in school work will not do for education what has been done for agriculture as a result of the application of scientific knowledge and methods. (Cubberley, 1919, p. 450)

Another early psychometrician, Harold Rugg, originally viewed testing with enthusiasm but later had a significant paradigm shift. He explains the viewpoint of the psychometric view as

he clarifies that "the most respected concepts and methods were those of the physical sciences and those were taken over. And with them were taken the theory, the outlook on life and education, and the assumptions of the physical scientists" (Rugg, 1934, p. 116). He goes on to criticize the testing frenzy because, as he called it, it led to an "orgy of tabulation" (p. 115).

The view that testing is scientific is more scrutinized nowadays. Not all stakeholders will admit that testing has inherent weaknesses. Nor will some stakeholders admit that how the results are used is always appropriate. However, many will admit that testing is fallible:

> Our misfounded faith that everything can be reduced to the precision of some of the hard sciences and math leads sensible and otherwise compassionate psychometricians and politicians to foolishness. They are left to conclude that they must rely on test scores to make decisions, even when they themselves acknowledge that real-life hard data suggest it is wrong to do so . . . Tests become the definition of success, not merely the predictor of it. (Meier, 2002, p. 117)

Tests certainly can be useful, but they are often not nearly as accurate or precise and we sometimes want to believe them to be. Ravitch (2010) sums up the problem of accepting tests as scientific:

> The problem with using tests to make important decisions about people's lives is that standardized tests are not precise instruments. Unfortunately, most elected officials do not realize this, nor does the general public. The public thinks the tests have scientific validity, like that of a thermometer or a barometer, and that they are objective, not tainted by fallible human judgment. But test scores are not comparable to standard weights and measures; they do not have the precision of a doctor's scale or a yardstick. (p. 152)

Tests Are Accurate

A major perception that allows standardized testing to maintain a prominent place in schools is that they are accurate. If testing is viewed as scientific and fair or objective, then it will, as a result, be viewed as accurate. Interestingly enough, in the literature, the claims of accuracy were much more common in the era of the early psychometricians in the early 1900s. In the environment of test-mania that existed then, many claimed flatly that "accurate measurements of the abilities of students may be made by using standardized tests" (Monroe, 1918, p. 19).

The psychometricians of this era were not the only ones that bought into the standardized testing frenzy. Many school leaders were also believers in the accuracy of the tests. One superintendent from the period described it thus:

> Lord Kelvin, the great British scientist, is quoted as saying: "When you express what you are talking about in numbers, you know something about it; but when you cannot measure it, when you cannot express it in numbers, you knowledge is of a meager and unsatisfactory kind . . ." Until we can really measure educational progress—express it in numbers, as Lord Kelvin says—our knowledge is likely to be a "meager and unsatisfactory kind." (Madsen, 1930, p. 16)

The enthusiasm for testing in that early era set a path for standardized testing that endures today. Even today, seniors in high school will look at the ACT results of different friends as if someone had taken out a brain-o-meter and accurately summed up the learning of the previous 12 years for all their friends, whether they understand what the 1-36 scale means or not. And most really have no idea what the scale means.

Cubberley's words in 1923 are often perceived to be as accurate now as they were almost 100 years ago:

> The use of these tests enables the principal to substitute measurable and standardized results for personal opinion, and provided him with a series of clear and incontestable records of the achievement of the pupils and teachers in his school . . . So important is this new method in education that a principal can no longer be considered prepared for his work unless he is familiar with the use of the standard tests and measures and with simple statistical procedure. (pp. 485-486)

However, there is plenty of evidence and literature to the contrary—standardized tests are not nearly as accurate as we sometimes are led to believe. While there are still many proponents of testing, most proponents will acknowledge now that the tests themselves are not nearly as accurate as the claims from the early days.

There Are Positive Consequences of Standardized Testing

When tests are perceived as accurate, by extension, they are usually also perceived to have positive consequences. Literature critical of standardized testing is much easier to find than literature that condones it, but there is some literature that outlines the benefits of testing. One of the biggest proponents of testing said, "On average, however, the use of testing tends to improve academic achievement. The evidence for this proposition is overwhelming and voluminous" (Phelps, 2004, p. 84).

Henry Chauncey, one of the most significant psychometricians of the twentieth century had a real belief in the positive consequences of standardized testing:

> Thus in a period of sixty years educational testing has developed from a part time chore of psychologists to a set of techniques that affects every student in school and college.

> The history of testing has been one of dedicated effort by a great many people to shape and sharpen and wield a tool that can be of great help to educators everywhere. (Chauncey & Dobbin, 1963, p. 20)

Though strong proponents of testing, Chauncey and Dobbin were also very aware of its limitations.

One author points out that though there are many benefits, he says "the benefits of high-stakes tests have been assumed, unrecognized, or unarticulated" (Cizek, 2001, p. 23). He went on to articulate a list of 10 major benefits of testing, including that tests provide school leaders the ability to offer more focused professional development, and that tests boost efficiency in accommodating students with special needs, and ultimately he argues that testing improves student learning. Feuer (2011) provides a similar list of benefits. Hanushek (2005) echoes the issue of using tests to improve student learning when he states that "despite design flaws in most existing systems, we find that they have a positive impact on achievement" (p. 321).

The Tests Are Not Accurate

Although some see standardized tests as accurate, most current literature recognizes them as having clear limitations. Not even the early psychometricians were blind to the weaknesses of standardized testing. Thorndike (1918) was very aware of some of the inaccuracies when he wrote that the "zeroes of the scales for the educational measures and the equivalence of their units are only imperfectly known. As a consequence, we can add, subtract, multiply, and divide educational quantities with much less surety and precision than is desirable" (p. 17).

The literature is extensive on how standardized testing is not nearly as accurate as we often infer. That does not mean testing cannot still be useful, but one of the biggest sections of the literature that will be addressed later discusses misinterpreting the results of standardized

testing. One of the biggest sources of misinterpretation comes from the perception that tests are accurate. The literature clarfies testing's limitations. Chauncey and Dobbin said it clearly:

> No test measures accurately enough to support a precise, one-point interpretation of a score or a 'diagnosis' of small differences between scores. Any test that leads the user to believe he has a precise measurement, then, is subtly dishonest and can lead him into trouble. (Chauncey & Dobbin, 1963, p. 64)

Many others from multiple disciplinary backgrounds have echoed his sentiment. One writer with a policy background wrote, "Tests are imprecise tools of estimation that provide only a partial view of selected aspects of what students know. Using tests as a basis for more comprehensive judgments is usually inappropriate" (Feuer, 2011, p. 27).

The most prominent educational economist of the last 30 years, who has used test data extensively for a variety of purposes, asserted:

> Measurements of school quality, including those that are incorporated into research, tend to be much narrower, frequently coming down to such things as performance on a specific standardized test. This narrow assessment is unfortunate. Many researchers and decision-makers rightfully question whether individual standardized tests adequately measure the relevant skills, and many object to relying on existing tests as the only measure of quality. Even if quality differences are very important for subsequent success, they argue, specific, narrow measures are unlikely to capture the full picture of how quality differs across schools and over time. (Hanushek, 1994, p. 19)

Another prominent economist added that

> Using tests to ascertain how well schools are doing in preparing a productive workforce and to determine who is likely to be a productive employee requires not only that such

> tests be valid predictors of workplace productivity, but also that the link between test criteria and worker productivity is large enough to make productivity predictions with great precision. I will assert that this link is missing. (Levin, 2001, p. 39)

Most teachers that have had highly capable students perform poorly on a CRT, college entrance exam, or AP test sense intuitively that tests do not always accurately reflect what students know or can do. One writer with an education background summed it up by writing that the kinds of measurements made in education "vary a great deal in their precision and accuracy. While the goal should be, of course, to measure as precisely as possible, an ordering or ranking of individuals is sometimes the best that can be done" (Lindeman, 1967, p. 2).

Negative Effects of Testing Use

One of the results of the perception that tests are not accurate is that there are many who perceive significant negative side effects from standardized testing. It is easier to find literature that criticizes standardized testing. One proponent of testing described what it is like to review literature on the subject:

> If nothing else, published commentary concerning high-stakes testing has been remarkable for its uniformity: the conclusion—high-stakes tests are uniformly bad. A colleague of mine recently performed a literature search to locate information about the effects of high-stakes tests. She found 59 entries over the last 10 years. A review of the results revealed that only 2 of the 59 could be categorized as favorably inclined toward testing . . .The other 57 entries reflected the accepted articles of faith concerning high-stakes tests. (Cizek, 2001, p. 20)

It is simply easier to find literature that criticizes standardized testing than literature that is in favor of it.

But once you get into the literature critical of testing, it becomes quickly apparent that "Most current criticisms of tests are clearly identifiable as criticisms of test use (or misuse), rather than criticisms of the tests themselves" (Anastasi, 1990, p. 15). In reading a wide variety of testing literature, it becomes apparent that "scholars seem to agree that it is unwise, illogical, and unscholarly to just assume that assessments will have positive consequences. There is the potential for both positive consequences and negative consequences" (Mehrens, 1998, p. 22). The next major section of this literature review will address consequences. Even though the literature, if weighed by sheer content, would lean heavily to criticisms of testing, there clearly are both positive and negative consequences of testing.

Critics often make lists of criticisms (Feuer, 2011). Proponents will fight the criticisms and rebut lists of criticisms (Cizek, 2001; Goodman & Hambleton, 2005). For example, one proponent of testing wrote that many of the criticisms we hear about educational assessments appear to be based on misconceptions: "Critics sometimes misrepresent the available information, if they have any at all, and the public hears it and often believes the criticisms to be true. After all educational assessments are easy to dislike" (Goodman & Hambleton, 2005, p. 107).

But there are many legitimate criticisms of standardized testing. Many of the main criticisms will be addressed in the section on consequences of testing, but the following is a typical criticism of testing as articulated by Gerald Bracey (2003):

> One of the undesirable by-products of testing practice has been the emphasis on academic talent with its accompanying indifference to other kind of talent. Tests have fostered a narrow conception of ability and restricted the diversity of talent, which might be brought to the attention of young people considering various professions. It is small wonder that

> some people have mistakenly interpreted test scores as a measure of personal worth and have mistakenly assumed that academic talent, as evidenced in school, is related in a major way to adult accomplishment." Let's repeat that message: Test scores are not a measure of your worth, nor are they related to what you will accomplish as an adult. (p. 33)

The critiques of testing range from mild annoyances to heated tirades. In a fairly moderate critique, William (2010) gives us another sampling of testing criticism:

> The introduction of high-stakes testing regimes was associated, in some cases, with increased student drop-out rates, inappropriate test preparation practices (up to and including cheating), and decreased teacher morale, leading to increased teacher defection from the profession. A subsequent analysis, involving the 27 states with the highest stakes associated with test score outcomes in Grades 1 through 8 confirmed these findings and indicated that the introduction of high school graduation examinations was associated with a lowering of average academic achievement. (p. 117)

Standardized Testing Is Arbitrary

One of the main elements contributing to the tests not being accurate is the reality that testing is in some ways arbitrary. For example,

> The process of setting standards—deciding just how much students have to do to pass muster—is technically complex and has a scientific aura, but in fact the standards are quite arbitrary. The simplicity of the form of reporting is therefore more apparent than real, and most people do not really have clear idea of what the standards actually mean. (Koretz, 2008, p. 87)

Even though the literature is replete with assertions that testing is quite arbitrary, especially in areas such as standard setting and cut scores, many stakeholders still treat the results of a test as if they were accurate. Recently, the states of Ohio and Arkansas showed just how arbitrary standardized tests can be when they unilaterally decided to make a 3 proficient in their states on the Common Core PARCC test when the test writers and all the other states had agreed that only 4s and 5s were proficient. A reporter on the story pointed out that "Arkansas claims that 60 percent of its Algebra I students are proficient, while fewer than half that many — just 28 percent — would be considered on track had Arkansas stuck with PARCC's more stringent definition of 'proficient'" (Brown, 2015).

Dorn (1988) described this arbitrary nature of testing. "The mundane details of statistical accountability systems encourage fads. Without a concrete sense of what children and teachers should be or are doing, the public compares statistics against a set of arbitrary benchmarks" (p. 15).

All tests can display this arbitrary dimension. College entrance tests are not exempt from this possibility. For example:

> In 1996, concerned by the misuse to which SAT scores have been put in debates about public education, the College Board "recentered" the scale on which SAT scores are reported. Instead of calculating scores with reference to the average performance of 1941 test takers, the new scale defines average 1990 achievement as being 500. On the new scale, a 600 score now means the test taker did better than about five-sixths of test takers in 1990, not in 1941. (Rothstein, 1998, p. 61)

This same phenomenon on a much more drastic scale has happened with the new common core tests. Where many states were reporting proficiency rates in the 80s or 90s, now few states are ever above 50% proficient—unless of course they arbitrarily decide to lower the bar.

There is not necessarily any magic or science regarding how cut scores are determined. Koretz (2008) provides sage advice on keeping this arbitrary nature in mind when considering standardized testing results:

> There is no reason to expect that if you and your friends lined up 100 students in order, ranging from the lowest-performing to the highest, and examined their work, you would end up placing a "proficient" cut anywhere near where your state education department placed it by using the bookmark method, the modified Angoff method, or any other. I think you are more likely to be misled by taking the descriptions of standards at face value than by treating the standards as arbitrary classifications. (p. 325)

Standardized Testing Is Simple

Standardized testing is a very complex technology and appropriate uses and applications of testing are also complex. One of the quickest ways to get in trouble with testing data is to assume it is simple and straightforward:

> The first piece of advice I would offer those making decisions about testing is to avoid unrealistic expectations. This might be called the Rolling Stones principle: "You can't always get what you want . . . and if you try sometime, you find you get what you need." Unrealistic expectations about testing are everywhere. They seem to rest on an inconsistent, even paradoxical view of the complexities of measurement and of the advice offered by people like me. On the one hand, the complexities of testing are widely discounted, and the complications raised by experts are often derided as being too arcane

> to matter. But on the other hand, there seems to be a widespread faith in the wizardry of psychometrics, a tacit belief that no matter what policymakers and educators want a test to do, we can somehow figure out how to make it work. One widespread unreasonable expectation is that a test created for one purpose will do just fine for many others. But a single test cannot serve all masters. (Koretz, 2008, p. 327)

Other writers echo Koretz's words:

> It is surprising that so many education policy makers have been seduced into thinking that simple quantitative measures like test scores can be used to hold schools accountable for achieving complex educational outcomes. After all similar accountability systems have been attempted, and have been found lacking, in other sectors, both private and public, many times before. (Harris et al., 2011, p. 144)

Often stakeholders will use test results as though they are simple and straightforward—just watch the U.S. Secretary of Education the next time the results of an international test are released. But, in the professional literature about testing, no one will try to convince you that anything around the world of standardized testing is simple.

Banesh Hoffman, a British mathematician and physicist, weighed in on standardized testing in 1962. He emphasized the complexity of testing:

> There is no satisfactory method of testing--nor is there likely to be. Human abilities are too intricately interactive to be measured satisfactorily by present techniques. There is reason to doubt even that they can be meaningfully measured at all in numerical terms."

But he went on to say that even though it is complex and imperfect, we must keep doing it, because it does have value. "Yet measurement, assessment, estimation, guesswork—call it what you will—can not cease" (p. 30).

Tests Are a Political Tool or Weapon

Tests as tools or weapons are two common metaphors used throughout the testing literature. For example, some claim that standardized tests "have become a political tool, one that allows politicians to put on the mantle of educational leadership. By berating low-scoring schools or by identifying instances of improvement, politicians give the impression of being in the forefront of academic excellence" (Harris & Longstreet, 1990, p. 149). Another author offered this view of testing as politics:

> Tests are political weapons instead of tools designed to assess the value and progress of current curricula. Because commercial tests are relatively inexpensive to administer, and because they provide simplified data on student "learning"—through percentage points and bar graphs—legislators and administrators, communities and the media embrace them, holding test results up as milestones of competence or deficiency. (Monroe, 1987, p. 24)

No matter the metaphor, it is clear that

> High-stakes testing is a politically charged issue that has had a tremendous impact on the way our schools operate. But educators must not be afraid to keep their perspective. They must encourage a healthy, honest dialogue about the role of testing and, most importantly, engage in the political debate. Their students deserve nothing less. (Casbarro, 2005, p. 23)

Airasian (1987) probably captured the political nature of testing most succinctly. "In essence, test scores become a medium of exchange to be bartered for educational, social, and economic benefits or rewards" (p. 405).

Why would testing be a political tool or weapon? People care about education, and therefore they pay attention to what schools are doing. Often standardized test scores are the only indicator that is easily accessible and therefore it is almost always what is used to judge schools. And, if a politician can convince their constituency that education is broken, and they have the answer for fixing it, then it gives that politician significant political capital:

> We must recognize that good news about public schools serves no one's political education reform agenda even if it does make teachers, kids, parents, and administrators feel a little better. Conservatives want vouchers and tuition tax credits; liberals want more resources for schools; free marketers want to privatize the schools and make money; fundamentalists want to teach religion and not worry about the First Amendment; Catholic schools want to stanch their student hemorrhage (and create more Catholics); home schooling advocates want just that; and various other groups no doubt just want to be with their "own kind." All groups believe they improve their chances of getting what they variously want if they pummel the public schools. (Bracey, 2003, p. 59)

And the way they pummel the public schools is through standardized testing data.

Many critics of testing have viewed testing as that reform tool, used to prove the schools are broken to further their own agenda. Garrison (2004) took it even a step further by saying that testing is in itself a political act:

> Thus it may be more useful in analyzing psychometry to view it as political theory. . .It is no wonder that results obtained by these methods closely parallel the inequalities upon which the entire economic and political order is based. (p. 72)

One other major part of standardized testing as a political tool has to do with the particularistic system of governance of public schools in America. There is always a tension

between levels of governmental control. The issue of levels of control will be explored in greater detail in a later section, but Airasian (1987) illustrates well the role testing plays as a political tool to move control to the central level:

> Since the mid-1960s, when educational reform became national in scope and social and economic in intent, standardized tests have assumed primacy as instruments both to monitor and to implement public policies. The linkage of standardized tests, particularly of the state certification variety, to educational standards, numerical scores, centralized control, and traditional educational values gives the tests substantial popular and political appeal. (p. 405)

Uses and Purposes of Testing

In reality, testing is simply a tool or a technology. The large majority of the tension exists in the standardized testing debate around the use or misuse (perceived or real) of testing *results*:

> Ultimately, the war over testing will be won or lost on the issue of test use. Intelligence and aptitude tests only matter to the extent that they are used, and therefore the most important question one can ask of these tests is: "What good are they?" Are they the efficient decision-making tools they are purported to be, or are they biased, invalid instruments and therefore undesirable selection tools? (Chapman, 1988, p. 3)

Understanding how tests are used, rather than merely how they are perceived, helps us enter the realm of understanding where the tension around testing originates.

When Tests Are Used Beyond Their Defined Purpose

The first major theme in the literature regarding use of standardized testing results is the idea that tests are often used in ways that go against their defined purposes:

> Like a lot of other tools, achievement tests are used in many ways, only some of them the ways in which their makers intended them to be used. Professional test makers, observing the variety of uses to which their instruments are put, occasionally are reminded of the scientist whose delicate micrometer was used by his wife to crack nuts. Whether or not such a comparison is apt, the fact remains that in schools generally there are both appropriate and inappropriate uses of standardized tests. (Chauncey & Dobbin, 1963, p. 66)

Wiliam (2010) echoes this concern. "This distinction is particularly important in view of the fact that test scores are often interpreted in ways that differ significantly from those intended by the designers of the test" (p. 107).

If you Google "hammer use," you will find hundreds of uses for a hammer. Just because a hammer can be used for hundreds of different tasks does not mean it should be used that way. For example, a hammer could certainly be used as a murder weapon, but we can all agree that is not what a hammer was designed to do. What does test misuse look like in education? Chappuis, Chappuis, and Stiggins (2009) suggest several ways that assessment-literate teachers would *not* use standardized testing:

> Use a reading score from a state accountability test as a diagnostic instrument for reading group placement. Use SAT scores to determine instructional effectiveness. Rely solely on performance assessments to test factual knowledge and recall. Assess learning targets requiring the "doing" of science with a multiple-choice test. (p. 19)

Yet, all four of these misuses are fairly common uses by different stakeholders: Schools are guilty of using state accountability tests for placement. The federal government was guilty of using SAT scores in the early 80s to show that American education had lost effectiveness. Most

State CRTs assess the "doing" learning targets with multiple-choice tests. Why? Every stakeholder has an agenda, and test data is an easy way to provide evidence to support each of those agendas.

Therefore, test use becomes central to the issues around testing. "Standardized tests are, in themselves, neither good nor bad. I will contend that everything depends on the paradigm we adopt in determining their purposes and goals" (Finn, 2008; Robinson, 1990, p. 88).

One of the interesting ironies in testing is that nowadays we are practically drowning in testing data, but there is little training or instruction about how we should actually use the data—for practically any stakeholder. Therefore testing data get used, but they may or may not be used for the purpose for which the test was designed:

> In such an intentionally designed and comprehensive system, a wealth of data emerges. Inherent in its design is the need for all assessors and users of assessment results to be assessment literate—to know what constitutes appropriate and inappropriate uses of assessment results—thereby reducing the risk of applying data to decisions for which they aren't suited. (Chappuis et al., 2009, p. 19)

Hess and Mehta (2013) discuss the lack of professional training for data use:

> To date, there's been more interest in data systems, unit records, and the machinery of data than in how educators are supposed to use these to improve teaching and learning. Educators may be awash in data, but failures in teacher preparation, professional development, and district practices mean that few are equipped to take full advantage of new tools. Few schools have provided more opportunities for teachers to develop expertise. Instead, we just ask teachers to use data "more often" and "better" on top of everything else they already do. (p. 72)

This is a problem throughout the educational system. Stakeholders have large amounts of data at their fingertips, but are rarely, if ever trained how to actually interpret or use it. They further explain,

> We're often imprecise about what kind of data to use for what purpose. There are two related issues here. The first is that crude data that were designed for public accountability are now being used to manage performance in ways that were never intended. The second is that we don't collect the kinds of data that would be more broadly useful for organizational improvement. (Hess & Mehta, 2013, p. 73)

The criticism surrounding using standardized test data in ways not intended seems to often come from test writers. They are the experts regarding what a test can and can't d, or how we should or shouldn't use a test:

> Testing experts frequently remind school officials that standardized test scores should be used not in isolation to make consequential decisions about students, but only in conjunction with other measures of student performance, such as grades, class participation, homework, and teachers' recommendations. Testing experts also warn that test scores should be used only for the purpose for which the test was designed: For example, a fifth-grade reading test measures fifth-grade reading skills and cannot reliably serve as a measure of the teacher's skill. (Ravitch, 2010, pp. 152-153)

Richards (2004) explains that historically tests have been misused when compared to what they were designed for, but he distinguishes that the motive behind the misuse has also shifted. Testing gained popularity in the era of measurement frenzy at the beginning of the 20th century. Now, the motive appears to be more political as the struggle for control between levels of government has increased:

> The beginning of the 20th century witnessed the gathering of some very talented, dedicated, optimistic and somewhat naive educational scientists seeking more effective ways to measure everything possible. Like the little boy who discovers a hammer, everything looked like a nail. They became enraptured by the task of measuring and sought to measure everything. Eventually, educators recognized that the empirical tools that had been developed had an important use, but needed to be seen as one of many potential research tools in a very complex enterprise.
>
> In the beginning of the 21st century, on the other hand, politicians have finally discovered the same hammer—but more like the adolescent they too are pounding everything as if it were a nail—not because they genuinely believe they are nails but because they love to see themselves swinging the hammer. Then it was used as an instrument of curiosity, interest and hope—now it is being used as an instrument of power and mechanism of control. (p. 5)

Multiple Uses for the Same Tool

This is a corollary of the first issue of use. One of the practices that create problems of use is when we try to use tests for multiple concurrent purposes. AERA, in their guidelines for test use, make it clear that

> Tests valid for one use may be invalid for another. Each separate use of a high-stakes test, for individual certification, for school evaluation, for curricular improvement, for increasing student motivation, or for other uses requires a separate evaluation of the strengths and limitations of both the testing program and the test itself. (AERA, 2000)

Chauncey and Dobbin (1963) express this a little bit differently. "Standardized tests of student achievement are such useful teaching tools that it is often a mistake to try to make them

do double duty as measures of the teacher as well" (p. 106). Koretz (2008) makes it clear that "one widespread unreasonable expectation is that a test created for one purpose will do just fine for many others. But a single test cannot serve all masters" (p. 327).

In essence, this issue of multiple use comes back to the key element of validity. As Rothman (1995) wrote:

> A calendar is a valid measure of time but not a valid measure of temperature, even though it is usually colder in winter than in summer. Similarly, a test might be a valid measure of a student's mathematics achievement but not a valid measure of a school's quality, even though many schools with large numbers of high achievers are good. (p. 152)

Standardized testing stakeholders should constantly be asking if each use they are applying for the test has been validated for that use. Most tests have not been validated for all the different uses:

> Multiple and diverse expectations for what assessment can accomplish translate into multiple policy targets, disparate notions of the process by which change occurs, and competing uses for the results of student assessments . . . .The question then becomes whether a single assessment system can serve these diverse purposes. Testing experts have warned about the difficulties inherent in relying on the same assessment system to serve multiple purposes. (McDonnell, 1994, p. 12)

Testing as a Gatekeeper

Testing is often employed as a gatekeeper. This gatekeeper role is often high stakes in nature, meaning that it has significant consequences that either allow or deny future opportunities. The two most common gatekeeper functions in the United States testing culture

are graduation gatekeepers and college entrance gatekeepers. A graduation gatekeeper would be a test like the former Utah Basic Skills Competency Test (UBSCT) that a student has to pass in order to receive a diploma. A college entrance gatekeeper would be the SAT or ACT test. Student performance on such tests is one of the most influential data points that colleges use to determine enrollment acceptance or rejection

This function is heavily criticized in the literature because of the sorting nature of the tests. Critics of graduation gatekeeping wrote,

> State exit exams harm students who fail them and provide no discernable benefits to students who pass them. Obviously, states didn't intend to implement ineffective and punitive education policies. Exit exam policies are broken, and states should either fix them or get rid of them, but either option requires a political will that is in scarce supply among policy makers and politicians. (Warren & Grodsky, 2009, p. 649)

Another critic wrote:

> We are compelled to admit that current policy regarding test use is a Gordian knot of conflicting motives and a classic mismatch of purpose and technology. Just how tangled this knot is becomes obvious if we examine the purposes for which standardized tests have been used. Tests have been used as gatekeepers, as quality indicators, as indicators of educational progress, and as vehicles to ensure accountability. (Robinson, 1990, p. 90)

Harris et al. (2011) wrote that because of the way these tests are used as gatekeepers, "they potentially damage society in ways that may far outweigh the lesser benefits they confer—whether the tests are used for "measuring achievement" in the K-12 schools or for helping determine admission to college" (p. 1).

Yet, many stakeholders believe there should be some minimum competency for a student to graduate, and the college admissions tests do give a clear, albeit highly focused indication of whether a student is capable of succeeding in college. They also make it much easier for colleges to determine who to admit to their institution by relying heavily on at least this one purportedly objective measure.

Testing as a Control Mechanism

One of the most significant overall uses of standardized testing is that it becomes a significant player in who controls education:

> In the field of education, the key governance question is who controls education and the formal institutions called schools that are organized to carry out the critical process of social reproduction and creation of individuals. The answer to this question has profound implications for the future of society and the various social organizations and individuals that comprise a society. Education is the process by which a society with its own particular culture reproduces itself. Through education, norms for proper conduct are established, social life and political institutions are legitimated and worldviews are created and justified. "Education, both in its content and pedagogy, is the cultural furnace where a particular image of mankind and the world is forged and a way of living is pass on" (Randall, 1997, p. 71). Simply an enormous amount is at stake when deciding who is to be educated, what will be taught, and who will control the process of educating the children. (Cooper et al., 2004, p. 136)

Because who controls education is so important, it is one of the reasons standardized testing becomes so important. There are four ways testing can be used for control that emerge from the literature.

Control through accountability systems. Accountability is a term that was rarely used before the 1960s in education. The focus in education before that was always on providing more inputs. The idea of accountability changed the focus from inputs to outputs. Thus, standardized testing became the key ingredient to accountability because it is the one output that is relatively cheap and easy to gather. Testing provides readily available data:

> But, A TROUBLING reality in today's political climate is that many political leaders actually believe that the best way to change schools is through an 'end of a gun barrel' approach, rather than by building consensus. Accountability, as prescribed by No Child Left Behind (NCLB) and its corollary regulations at state levels, clearly supports this approach. [emphasis in original] (Casbarro, 2005, p. 20)

That end-of-a-gun-barrel approach is exemplified often by policy makers or politicians. For example, former U.S. Secretary of Education Margaret Spellings said, "What gets measured does indeed get done. Lesson No. 1: Accountability is a powerful tool and is working to improve learning. Lesson No. 2: Accountability also makes people uncomfortable" (2010). One economist clarified how central the idea of accountability has become in American politics: "Indeed, accountability policies dwarf all other education reforms in scope" (Jacob, 2005, p. 762).

Jacob (2005) continues, "Despite its increasing popularity within education, there is little empirical evidence on test-based accountability (also referred to as high-stakes testing)" (p. 762) Even 15 years before NCLB and high stakes tests as we know them today, one author warned about the testing ramifications surrounding the accountability movement:

> Of late, the evidence has been disheartening, provoking a call for higher standards, more discipline, and greater school and teacher accountability, which has been translated into

> more intrusive forms of standardized testing. The breadth of the audiences for testing means that discussion and debate over testing are no longer technical and arcane but encompass political, legal, social, and economic as well as educational issues. (Airasian, 1987, p. 406)

Testing can definitely have productive uses, but Ravitch (2013) exposed one weakness of relying solely on an accountability system:

> Even more curious is the unwarranted belief that more testing and accountability will close the achievement gaps between rich and poor, blacks and whites, and Hispanics and whites. Since the source of the gaps is socioeconomic inequality, it is sheer fantasy to believe that the test scores of these groups will converge if only there are higher standards plus more testing and accountability. The assumption is that those who teach the low-performing groups are not really trying, and a carrot or a stick will motivate them to try harder. (p. 266)

Koretz (2008) offers a final, and politically unpalatable, piece of advice:

> We need to be more realistic about using tests as a part of educational accountability systems. Systems that simply pressure teachers to raise scores on one test (or one set of tests in a few subjects) are not likely to work as advertised, particularly if the increases demanded are large and inexorable. (2008, p. 330)

Accountability has merit, and accountability is important. But accountability needs to make sense to all stakeholders to some degree of equity. A comparison to another profession is apt. If we were to judge dentists on how many cavities their patients have, we might make poor judgments on the quality of the dentist. The same is true of standardized testing data. Care should be taken how accountability systems and their consequences are designed and applied.

Control of curriculum. A common theme throughout the literature is that the easiest way to control curriculum is to control testing. "Neither in the past nor the present is testing mainly about 'improving education.' It is, instead, about control over the purpose and nature of schooling" (Garrison, 2009, p. 2). While Garrison is a clear critic, most of the literature from other disciplines, and both critics and proponents alike, seems to support his premise on this aspect of testing.

For example, "Standardized tests have become an important tool in the efforts of state governments to improve educational standards and to gain control over the process of education in local school districts" (Airasian, 1987, p. 393). Another author wrote, "In a broad sense, standardized testing supports the determination or control of curriculum content at the state and national levels" (Dorn, 1998, p. 11). Others echo the idea as well: "What is clearly at stake here is not only who shall control testing, but also who shall control education" (Harris & Longstreet, 1990, p. 150). There seems to be near-universal agreement on this point: "The lesson of history, as well as that of contemporary experience, is clear: to change curriculum and instruction, change the mandated universal examination" (Madaus, 1985, p. 616).

As a corollary to this idea, the literature also supports the notion that controlling testing changes the behavior of the teachers in the classroom. "Extensive research demonstrates the principle, what you test is what you get. Study after study shows that teachers tend to focus on tested content and formats and to ignore what's not tested" (Herman & Linn, 2014, p. 34). McDonnell (1994) further states:

> Political elites, business leaders, and the general public are looking once again to student assessment as a cornerstone of education reform because of its powerful leverage as a

> policy instrument . . . a growing body of research indicates that school and classroom practices do change in response to these assessments. (p. 1)

This is reflected often in the literature. "Minimum competency tests, aimed at the academically weakest of our students, also tend to function as examinations that control the content of teaching" (Resnick & Resnick, 1985, p. 14). The proposition that whoever controls the test controls education probably explains the testing opt-out movement of the past few years. Parents have been unhappy about the common core, and whether they realize it or not, their boycott of the tests is their only way of protesting the control from the federal level.

Controlling shifts in educational governance. The centralization of education in the United States is accomplished, at least in part, by taking control of testing at a more central level. Centralization would have been unlikely without standardized testing, because it is an effective tool for shifting school governance. If a group or level of governance wants to control the content of education, then the easiest way to gain that control is by controlling testing:

> To be clear, we are not suggesting that a shift in control of the curriculum and educational policies from the district to the state or from the state to the federal government is either desirable or undesirable—that is an ideological, political judgment. It is clear, however, that mandating a high-stakes testing program shifts power to those who establish and control the testing program. (Madaus et al., 2009, p. 157)

This shift has been occurring for a long time. It wasn't something that happened overnight. Thirty years ago Airasian (1987) talked about the shift that had begun nearly 20 years earlier (that totals nearly 50 years ago):

> In the past 20 or so years, a number of trends in the wider society have led to educational growth, to shifts in the locus of school control, and to politicization of educational

> decision making. As new educational roles and expectations emerged in response to these changes, new roles and expectations for standardized testing also emerged to complement altered educational priorities. In particular, new forms and uses of standardized testing have arisen. Traditional, instructionally oriented testing that is controlled by local school districts and used primarily in the service of classroom teaching and learning now coexists with policy-oriented testing that is controlled by agencies external to the local district and used to implement or to assess the effect of an educational policy or practice. (p. 409)

It appears that this shift will continue. Passage of the Every Student Succeeds Act (ESSA) in December of 2015 has been hailed by most stakeholders as the biggest de-federalization in the many years. Yet, it seems evident that the federal government has maintained nearly all the testing mandates.

This use of testing to shift the locus of governmental control is nothing new. Going back to the early days in testing in the United States, testing was used to shift the locus of control. "The drive to build a public school system that shifted responsibility for education from the family, church, and community to the state is more likely the decisive factor influencing the rise of standardized testing" (Garrison, 2009, p. 63). He went on to write:

> The main role of the academic achievement test adopted by Horace Mann was to increase supervisory authority of the state. While the new exams functioned to make public education more accountable to a central state authority, it also served as an instantiation of reformers' educational philosophy . . . .As was the case with Mann, Binet's standard served as a means to further legitimate state control over education. The emergence of

> this standard is linked to secular governing principles derived from the new social sciences. (Garrison, 2009, p. 73)

Building on the earlier section of this review that lightly touched on the history of standardized testing in America, it becomes more apparent why some of the events of the past 60 years are so significant. "NDEA [of 1958] was a significant act in U.S. educational history because it was the federal government's first involvement in U.S. education. This involvement, however, was not very strict, but it increased and became stricter over the years" (Turgut, 2013, p. 65). The catalyst was Sputnik, but it truly was just the beginning. The early 1960s "brought a slow but inexorable shift in the use of standardized tests. No longer merely tools used by local school district administrators, the tests assumed a central role in establishing and implementing state and federal education policy" (Madaus, 1985, p. 613).

A Nation at Risk was another major event that furthered the shift in control toward the federal government. Subsequently, the most comprehensive shift toward centralization in U.S. history was NCLB. Even former U.S. Secretary of Education, Rod Paige, acknowledged NCLB's role. "The No Child Left Behind law dramatically reshapes the federal role in education. It authorizes the federal government to demand results from our schools" (2002, p. 713). Educational Historian, Diane Ravitch (2013), confirms what Paige said:

> NCLB put the federal government, with its relatively minor financial contribution, in the driver's seat . . . After NCLB the federal government assumed a command-and-control role that was never envisioned in 1965 or 1979. For the first time in history, school districts and states had to ask permission from the U.S. department of Education to change their plans to meet federal goals. (p. 281)

A common theme among the literature is that standardized testing has been crucial to centralization at the federal level. Some authors feel more or less strongly about the implications of this shift in power, but it seems to be well agreed that testing has been used this way. Garrison (2009) is direct about what he sees as the implications of this shift in power and the role testing played. He wrote:

> The standards used to judge the success of schools have changed, and that this change in standards is about shifts in power and purpose, not "school improvement." What schools are expected to do by those officials who now wield power is different from the past, and this difference is reflected in their adoption of different standards than those put into place by their predecessors. That business leaders were, for example, directed by the Business Roundtable to sit on "cut score committees" to ensure high levels of failure is a remarkable example of the role of standards in establishing power (Business Roundtable 1998) and more generally of standardized testing as a measure of failure. (p. 4)

Despite Garrison's view of this facet of testing, it is interesting to note that the new 2015 ESSA maintains practically all the testing mandates, and states still have to answer to the federal government on the issue of testing and accountability.

Control by privatization. This is a corollary of using testing to shift power. Among the literature, many authors criticize the use of test scores to push an agenda of privatizing public education. "NCLB has opened the door to the privatization of public schools through some of the more insidious penalties imposed on 'failing' schools" (Behrent, 2009, p. 242). Another author said,

> Other people, meanwhile, are determined to cast public schools in the worst possible light as a way of paving the way for the privatization of education. After all, if your goal was

> to serve up our schools to the marketplace, where the point of reference is what maximizes profit rather than what benefits children, it would be perfectly logical for you to administer a test that many students would fail in order to create the impression that public schools were worthless. (Kohn, 2000, p. 2)

Ravitch (2010) warns against this use of the testing data:

> I started to see the danger of the culture of testing that was spreading through every school in every community, town, city, and state. I began to question ideas that I once embraced, such as choice and accountability, that were central to NCLB. As time went by, my doubts multiplied. I came to realize that the sanctions embedded in NCLB were, in fact, not only ineffective but certain to contribute to the privatization of large chunks of public education. (p. 102)

Manufacturing a Crisis

One major element of controlling education relates to what Berliner (1995) called the "manufactured crisis." He said, "American education has recently been subjected to an unwarranted, vigorous, and damaging attack—a Manufactured Crisis" (p. 343) This is, in essence, an educational boy-who-cried-wolf phenomenon. He goes on to explain:

> Early in the 1980s, prominent figures in our federal government unleashed an unprecedented onslaught on America's schools, claiming that those schools had recently deteriorated, that they now compared badly with schools from other advanced countries, and that as a result our economy and the future of our nation were seriously threatened. These claims were said to be supported by evidence, although somehow that evidence was rarely cited or appeared only as simple, misleading analyses of limited data. Nevertheless, this attack was waged with great vigor, was eagerly supported by

> prominent figures in industry, and was widely reported and endlessly elaborated on by a compliant press. And as a result, many of the claims of this attack came to be accepted by good-hearted Americans, including a lot of powerful people and leaders in the educational community; and great mischief resulted because of the misunderstandings and poor policies this attack created. (Berliner & Biddle, 1995, p. 343)

Testing becomes a significant tool in creating a crisis mindset, and throughout the history of testing, it has been often been used for that purpose.

This manufactured crisis is meant to be a reform lever. There are a couple of major elements to create the crisis mindset. The first is to create a false narrative of failure, and the second is to make the schools the source of that failure.

A false narrative of failure. To effectively create a manufactured crisis, the first thing that has to be done is to create a false narrative of failure. Horace Mann accomplished this effectively in the mid-1840s with his tests, but the most nationally comprehensive example was *A Nation at Risk* in 1983. It is obvious in reading the report that it is based on military themed language. It was written at the height of the Cold War, and the threat of communism to other governance structures was real and nearly palpable. For example, the report starts with the following assertion:

> Our nation is at risk. Our once unchallenged preeminence in commerce, industry, science, and technological innovation is being overtaken by competitors throughout the world . . .The educational foundations of our society are presently being eroded by a rising tide of mediocrity that threatens our very future as a nation and a people. . . If an unfriendly power had attempted to impose on America the mediocre educational performance that exists today, we might well have viewed it as an act of war. As it

> stands, we have allowed this to happen to ourselves . . .We have in effect, been committing an act of unthinking, unilateral educational disarmament. (Gardner, 1983, p. 9)

He then used test data from tests that were not designed to do what he was trying to prove in order to accomplish this crisis mindset, namely declining SAT scores.

This idea that test data could be used to create a militarized crisis mindset is nothing new, and was well documented in a report commissioned by the government 10 years after *A Nation at Risk.* The Office of Technology Assessment (OTA) did a thoroughly comprehensive report that considered the issues around standardized testing. In this report, which certainly did not get the same public or media attention as *A Nation at Risk,* the OTA articulated well the environment in which testing has often been used to manufacture a crisis. They referred specifically to early testing in the time of Horace Mann:

> The idea underlying the implementation of written examinations (in the nineteenth century), that they could provide information about student learning, was born in the minds of individuals already convinced that education was substandard in quality. The sequence--perception of failure followed by the collection of data designed to document failure . . .offers early evidence of what has become a tradition of school reform and a truism of student testing: tests are often administered not just to discover how well schools or kids are doing, but rather to obtain external *confirmation*--validation--of the hypothesis that they are not doing well at all. (Office of Technology Assessment, 1992, p. 108)

Contemporary critics often see the use of tests to prove failure as a new phenomenon. It is not. This use has been around since the first standardized tests in the Boston Common

Schools. In a written communication from Horace Mann to the man in charge of administering the first exams in Boston, Mann coached him on how to create the crisis mindset and make sure the blame got placed where it should—on the educators:

> Once parents saw the printed report, with the actual statistics showing the low test scores, they would demand an explanation. Then came a warning to Howe: "If the odium of such a disclosure is to fall upon the children, the parents will be disposed to punish you for it. If on the other hand, it can be fastened where it belongs, they will condemn the teachers. . . . Emphasize the great difference existing between different schools, on the same subject, showing that the children could learn, if the teachers had taught" effectively. Otherwise the examiners, not the masters, would be vilified, since parents in particular would not blame the children. Mann told Howe to take heart: "You will be hereafter hailed as the regenerator of the Boston schools." (Reese, 2013, pp. 119-120)

This creation of a false narrative of failure is a common theme throughout the literature, and represents a major use of standardized tests. If a stakeholder can create a mindset of failure, it becomes easier for that stakeholder to use the data as a lever for reform. Garrison (2009) echoed the Office of Technology Assessment report: "Standardized exams were developed as markers of failure, and stood as justifications for and symbols of the changes reformers sought" (p. 3).

Many stakeholders use tests for this purpose. Another example is the proficiency levels of the NAEP test, as Rothstein (1998) explains:

> If the NAEP achievement levels established by the NAGB are to be believed, only 30 percent of U.S. nine-year-olds are proficient in reading. This is simply not plausible, and it raises questions about how the proficiency levels are determined by the National

> Assessment Governing Board before they are broadcast to the American people in support of a "failing schools" story. The procedure for defining these achievement levels, in reality, is both ideologically and technically suspect. The standards seem to have been established primarily for the purpose of confirming preconceptions about the poor performance of American schools. The specification of such levels is an extraordinarily complex undertaking; it would challenge even the most expert psychometricians. (p. 72)

Blaming schools as the source of failure. One of the strategies to creating the manufactured crisis is to blame the schools for society's ills. For example, the former U.S. Secretary of Education Arne Duncan said the following in a speech in 2010:

> The chief reason that U.S. students lag behind their peers in high-performing countries is not their diversity, or the fact that a significant number of public school students come from disadvantaged backgrounds. The problem, OECD concludes, is that "socioeconomic disadvantage leads more directly to poor educational performance in the United States than is the case in many other countries." Disadvantaged Canadians are much less at risk of poor educational performance than their counterparts here. Our schools, in other words, are not doing nearly as much as they could to close achievement gaps. As schoolchildren age in America, they "make less progress each year than children in the best-performing countries," according to the OECD. (2010)

If schools are to blame and the *scientific* backing of standardized testing proves that failure, then an effective educational reform lever has been shaped.

Testing as a reform lever. Why create a false narrative of failure? Why blame the schools and make them the source of failure? Because by doing so, tests become a reform lever to change the education landscape. This is a common theme in the literature. Kohn (2000)

claims that, "Tests have lately become a mechanism by which public officials can impose their will on schools, and they are doing so with a vengeance" (p. 2). This idea has been around as long as testing. Feuer (2011) notes:

> Standardized educational tests have been a staple of public accountability in education for almost two centuries, and that from their inception they have been popular devices used for both good and mischief. Horace Mann and his partners in the great reform movement were not only brilliant social reformers intent on expanding the educational franchise, but they were shrewd politicians too, who understood long before the ascendance of professional communications experts and policy wonks that by including certain questions on the tests they could expose the failures of school masters they were battling with, and, as one of our preeminent education historians noted, use testing as a "bludgeon of reform . . ." (Tyack, 1974). In a word, if you think some teachers and principals are feeling pressured by NCLB testing, you are right; but based on the historical evidence one cannot help think that today's test-based accountability pales in its ferocity when compared with the earliest episodes of the "bludgeoning. (p. 26)

The use of tests for “bludgeoning” schools has only accelerated in the last 60 years:

> In the 1950s, policies regarding testing began to change. Over the next fifty years, a number of events and federal legislative acts solidified the importance of high-stakes testing in American society. Today high-stakes testing is the primary strategy employed by federal and state governments to monitor and reform the educational system. (Madaus et al., 2009, p. 16)

It is easier to target a tangible, easily scored assessment as a reform strategy than to look deeply at the causes behind test scores. One congressional staffer expressed this sentiment as to why tests are so readily used:

> People settle on assessment as a cheap way to fix problems. One of our most prominent governors sees assessment as an important lever to change American education . . . It's a lever for change without having to spend a lot of money. (McDonnell, 1994, p. 23)

"Indeed, accountability policies dwarf all other education reforms in scope" (Jacob, 2005, p. 762). It seems to be *the* tool for reform.

Test Data as an Economic Indicator

The literature is replete with assertions and claims that the connection between education and economics is crucial. Eric Hanushek (1994) asserts that "education is ultimately an investment and thus is best evaluated alongside other, perhaps more mundane, forms of capital spending such as roads and machinery" (p. 13). Hanushek is considered by many to be the most renowned American education economist, and he often makes these types of claims. "Without doubt, the achievement of our students has direct ramifications for the future well-being of our society" (Hanushek & Raymond, 2005, p. 323). Most understand intuitively that there is a connection between the two elements.

Where the literature becomes oddly interesting is when claims are made that standardized testing results affect the economy. For example, Arne Duncan, former U.S. Secretary of Education, states that "President Obama has repeatedly warned that the nation that 'out-educates us today will out-compete us tomorrow.' And the PISA results, to be brutally honest, show that a host of developed nations are out-educating us" (2010).

While there is little dispute about how important education is to the future economy, citing test scores as the cause of economic change is a doubtful connection. The great economist, Henry Levin disputes these claims. For example, he writes:

> If the focus shifts from curriculum to measures of schooling like test scores, there still seems to be little connection between these measures of education and the earnings of high school graduates. There is a long history of researchers failing to find an economically significant relationship between scores on achievement tests and wages. Researchers have also looked at possible effects of secondary school grades and class rank on wages" (Kang & Bishop, 1986; Meyer, 1982). In some regressions, small results are found some of the time. But the main message from studies on course work, test scores, and grades is that learning in high school does not seem to be a significant factor in explaining the correlation between secondary schooling and wages" (Weiss, 1995, p. 141). In summary, the general notion that the competitive economic position of the U.S. can only be sustained if we can out-compete students from other countries in scores on achievement tests is naive and hardly supported by the overall empirical data. (Levin & Kelley, 1994, p. 100)

Data Patterns and Correlations

Looking for patterns and correlations in the data is where the real value of the data lies. Data is readily available these days. "State and federal mandates for the collection and reporting of this information have resulted in unparalleled access to the data" (Cizek, 2005, pp. 38-39). Cizek further points out:

> Increasingly, from the classroom to the school board room, educators are making use of student performance data to help them refine programs, channel funding, and identify

> roots of success. If the data—in particular achievement test data—weren't so important, it is unlikely that this would be the case. (p. 39)

The group that appears to look for these patterns and correlations the most is economists. There may not be a legitimate causal connection between economic prosperity and test scores themselves, as Levin and Kelley (1994) point out, but economists certainly use standardized testing to assert a wide range of other statistical connections.

For example, Ferguson (1991) used educational data to show that teacher quality has a distinguishable relationship to student learning, that teachers are attracted to higher paying districts, and that teacher experience matters. Goldhaber (2002) found that "the effectiveness of teachers has more of an influence on student achievement than any other schooling factor" (p. 52). He earlier found that "teachers who are certified in mathematics, and those with bachelor's or master's degrees in math, are identified with higher test scores" (Goldhaber & Brewer, 1997, p. 520). Hanushek and Raymond (2005) also found that "the large differences in spending per pupil never influence scores. Consistent with past evidence on the impacts of resources, the pattern of NAEP scores across states is not explained by spending" (p. 310). What all of these economists have in common is that they use test scores to arrive at their conclusions.

Patterns and Correlations in Decision Making

Why would it matter that economists look for patterns and correlations using the data from standardized testing? Because policymakers seek out the conclusions that economists make. Policymakers often turn to economists to get information that shapes policy. So much of the testing data gets to politicians through the filter of the economists. Informed decision-making is important because "One assumption underlying high-stakes testing has received particularly scant attention—the need to make decisions. There is simply no way to escape

making decisions about students" (Cizek, 2001, p. 21). And policymakers do use the data as explained by Salganik (1995), when she states that "governors and chief state school officers seem certain to view public comparisons of average test scores across the states as a justification for reasserting their own decision-making authority in education" (pp. 609-610).

Educational stakeholders are awash in data. However, very few stakeholders are trained in how to use the data. Because test scores can often be misinterpreted, it is important to be careful consumers of the data:

> Our misfounded faith that everything can be reduced to the precision of some of the hard sciences and math leads sensible and otherwise compassionate psychometricians and politicians to foolishness. They are left to conclude that they must rely on test scores to make decisions, even when they themselves acknowledge that real-life hard data suggest it is wrong to do so . . . Tests become the definition of success, not merely the predictor of it. (Meier, 2002, p. 117)

Teaching

From the early 1900s to the 1960s, one of the main uses and purposes of testing was to inform teaching and help teachers know where students were scoring. That use became less prominent when testing data started to be used to monitor teachers, schools, and whole systems. But test data still can be used for these purposes. Achievement tests

> were originally designed primarily for diagnostic purposes, to help teachers and administrators identify relative strengths and weaknesses in their students' achievement. They were also intended to identify areas of strength and weakness within schools and school districts, in order to facilitate improved instruction. However, they were not

> intended to provide summary evaluations of the performance of schools, districts, states, or nations, or to hold educators accountable. (Koretz, 2008, p. 48)

Chauncey and Dobbin (1963), right before the major shift in test use in the mid-1960s gave a typical psychometricians' view on test use: "Testing in schools is intended to improve the instruction and guidance of students. Any testing that does not contribute substantially to the quality of instruction or guidance is too much testing" (p. 81). One of testing's biggest proponents, Richard Phelps, validates that we still need to be using the test data this way:

> What we need is more testing, not less—and more willingness to act on the results of those tests so that the poor of whatever race are no longer given such a rotten deal by America's schools. Perhaps the simplest, and least disputed, beneficial information use of standardized tests is in diagnosis. Test results can pinpoint a student's academic strengths and weaknesses, areas that need work, and areas of particular promise. (2003, p. 225)

Certifying Student Competence

Part of using the data to inform teaching is being able to certify student competence. This construct is expressed in different ways, such as the following:

> A few critics will always condemn the use of testing in schools. However, with students' futures at stake, we must not abandon the very tools that have the power to transform teaching and learning. We must make our education assessment stronger and take advantage of the information they provide to ensure that all of our graduates are academically healthy. (Gandal & McGiffert, 2003, p. 42)

However, even the strongest critics can see that this use is meaningful and can help in the teaching and learning process:

> The information derived from tests can be extremely valuable, if the tests are valid and reliable. The results can show students what they have learned, what they have not learned, and where they need to improve. They can tell parents how their children are doing as compared to others of the age and grade. They can inform teachers whether their students understood what they were taught. They can enable teachers and school administrators to determine which students need additional help or different methods of instruction. They can identify student who need help in learning English or special education services. They can inform educational leaders and policymakers about the progress of the education system as a whole. They can show which programs are making a difference and which are not, which should be expanded and which should be terminated. They can help to direct additional support, training, and resources to teachers and schools that need them . . . .used judiciously, this is valuable information. (Ravitch, 2010, pp. 150-151)

In the literature there seems to be little argument that using standardized tests for these purposes can help inform teaching.

Diagnosing Learning Problems

Closely related to certifying student competence is the idea of diagnosing student learning problems. In the past this has sometimes been a problem because the significant time lag between students taking the tests and getting data back from the tests to teachers and leaders was often not conducive to using the data well. But, in recent years this has improved due to technological advances. From the beginning, psychometricians recognized the diagnostic value of the tests. Early psychometricians viewed it thusly:

> The value of a test depends upon the use made of the measures . . . Without this step standardized tests become mere "playthings" and their use cannot be justified. The omission of this step creates a situation similar to that which would exist if a physician examined a patient carefully and determined the nature of his ailment, but did not prescribe any remedial treatment. (Monroe, De Voss, & Kelly, 1917, pp. 432-433).

Thorndike (1918) echoed his statements one year later: "Another important group of uses centers around the problem of giving the individual pupil the information about his own achievement and improvement which he needs as a motive and a guide" (p. 19).

Even policymakers in the 21st century can agree with these early purposes of testing. Former U.S. Secretary of Education, Rod Paige (2002) said:

> Teachers will be able to use individual student data to tailor their teaching to the specific needs of each student. Principals will be able to use the data to make informed decisions about what their schools need in order to improve students' performance. Parents will no longer wonder whether or not their children's schools are teaching them. (p. 711)

Meaningful data that identifies learning problems "turn a diagnosis into action, thereby enabling educators to respond to individual student needs. They also make assessments a helpful tool for educators rather than simply an accountability hammer" (Gandal & McGiffert, 2003, p. 41). Testing data uses that inform instruction are at the heart of the numerous potential benefits of standardized testing.

Monitoring Whole School Systems

As mentioned previously the use of monitoring whole systems as well as judging teacher quality or school quality with standardized testing data has been used to some degree since

Horace Mann, but this approach accelerated when the federal government got more involved in American education in the early 1960s:

> Prior to the 1970s, the functions of the achievement testing involved the monitoring, diagnosing, and placement of individual students; the impact on classroom behavior was minimal. . .tests are now used to "monitor, not individual children, but rather educational systems--programs, schools, entire districts, even states." (Harris & Longstreet, 1990, p. 150)

This role for testing became even more important as education became more centralized. Airasian (1987) noted that "in the 1960s and 1970s, when education became a central feature of state and national policy agendas, tests assumed an important role in monitoring the status of the educational system" (Airasian, 1987, p. 402). Proponents of this use point out several of the benefits for using testing data to judge whole school systems:

> Information benefits can also be associated with accountability. Information can be used by higher-level system administrators to make judgments about the performance at the school or school district level and to make changes to increase efficiency. In an environment of school choice (e.g., school districts with open enrollment), information about school performance can help parent-student school shoppers to make a more informed selection. (Phelps, 2003, p. 225)

Judging Individual School Quality

Related to monitoring whole systems is using standardized testing data to judge school quality. While test data is used for this purpose often, especially by policy makers and the media, this use of test data has been criticized in the literature in almost every era of testing. An example from the early 1900s is instructive:

> From the results obtained from intelligence testing we now know that it is not safe to assume that poor teaching is the cause of the failure to make proper progress in their school work, either of pupils or of classes. Other factors enter into the problem. (Cubberley, 1923, p. 497)

The same concern was voiced in the mid-1960s:

> Too many people, particularly those who are critical of a school or system, or who for some reason with to denounce public education, seek to prove that one school has a lower average score—or a higher one—than some other school on a "standardized school achievement test." Standardized tests cannot be used alone to judge schools, for even a whole battery of them will not measure all kinds of learning students achieve. (Chauncey & Dobbin, 1963, p. 67)

Later in the late 1900s, the same concern continued to be voiced: "Despite evidence that high-stakes testing possibly corrupts teaching and does not provide stable information about school performance, test results continue to dominate the way government officials, politicians, newspaper editors, and others describe the performance of schools" (Lattimore, 2001, p. 57). A similar sentiment echoes from the post-NCLB era regarding how we use testing: "To be clear about our own opinions on the subject: The results of large-scale assessments should never be used as the sole determinant of education/educator quality" (Betebenner et al., 2011).

Using tests this way is one of the more contentious uses of standardized testing data. Horace Mann used this way in 1845, and it continues today. Initiatives to grade schools across the country are a typical example of this use in the contemporary era.

Judging Teacher Quality

Similar to the use of judging school quality is the practice of judging teachers by the test scores of their students. Judging teacher quality by the scores of their students is a contentious use, because teachers feel victimized when they are judged on the results of students when many student characteristics are out of their control. "Educators and researchers have amassed evidence of significant problems with the growing reliance on tests, confirming suspicions that teachers since Horace Mann's time have held: namely, that using tests as gauges of teacher or school performance is unfair" (Rothman, 1995, p. 49). A current example is the requirement, that, as part of federal NCLB waivers, states must tie part of their teacher evaluations to their students' test scores. The large majority of the literature outside of policy makers criticizes this type of usage for the many reasons.

Student test scores provide only an incomplete, limited picture of what a teacher's impact and quality really is. A typical example for the literature puts it this way:

> Administrators, either on their own, or at the insistence of parents and school board members, all too often judge the quality of a teacher's instruction by the average scores earned by that teacher's students on a standardized test. This can be far more dangerous than even the most knowledgeable advocates of educational measurement are likely to know. The danger lies in the fact that it is so easy to accept test results as the only evidence of teaching quality—when, at their best, tests can yield only a small part of the evidence necessary to make a sound judgment . . . Using tests as a basis for more comprehensive judgments is usually inappropriate. Tests alone offer preliminary clues, at best, as to how students learned whatever it is they demonstrate on the test. Inferences

> about teachers, schools, principals, class size, and other possible causes require substantially more data than score reports. (Chauncey & Dobbin, 1963, pp. 102-103)

Other authors caution that:

> Another warning about the dangerous side effects of high-stakes testing surfaced when a plan to pay teachers on the basis of their students' scores was offered, making student test scores very high stakes for teachers. A schoolmaster noted that under these conditions, "a teacher knows that his whole professional status depends on the results he produces and he is really turned into a machine for producing these results; that is, I think, unaccompanied by any substantial gain to the whole cause of education." This concern about testing students to judge a teacher's worth first surfaced in the year 1887, but it is as fresh as recent headlines about pay-for-performance in Denver, Colorado; Houston, Texas; Florida; Minnesota; and Iowa. (Nichols & Berliner, 2007, p. 6)

Sorting and Classifying Students

One of the main early purposes/uses of standardized testing was to sort and select students. For example, an educator in the early 1900s explained Terman's classification system:

> L.M. Terman has advocated a five-track plan based in part upon intelligence test results. According to this plan pupils would be tested and classified on entrance to school under five types: (a) the gifted, (b) the bright, (c) the average, (d) the slow, and (e) the special or very slow. The pupil would then be assigned to the course having subject matter suited to his group. (Madsen, 1930, p. 225)

The different ways standardized tests have been used to sort and classify have been numerous, whether to promote or retain students at grade level, or to accept or deny them to college, or to determine what role they should play in the military. Tests have always been used

for these purposes and continue to be used these ways today. There are strong critics and proponents on each side of this debate. Proponents tend to look at this use this way:

> Very superior minds are usually (but not always!) discovered sooner or later; the advantage of the intelligence test is that they are sifted out at once for the special attention which they deserve and which the higher interests of society demand that they should get. (Wood, 1923, p. 275)

Lippmann, a critic, wrote one year earlier that:

> They believe that they are measuring the capacity of a human being for all time and that his capacity is fatally fixed by the child's heredity. Intelligence testing in the hands of men who hold this dogma could not but lead to an intellectual caste system in which the task of education had given way to the doctrine of predestination and infant damnation. If the intelligence test really measured the unchangeable hereditary capacity of human beings, as so many assert, it would inevitably evolve from an administrative convenience into a basis for hereditary caste. (Lippmann, 1922a, p. 298)

In trying to create a meritocracy, Chauncey created the SAT to base college admissions on academic merit rather than money or societal position. But he actually had a larger target that was never realized. He wanted to be able to sort all people in the society so that society would be more efficient, and people could be best utilized according to their abilities. It was a dream of sorting and classifying that has never been fully implemented. He wanted to

> mount a vast scientific project that will categorize, sort, and route the entire population. It will be accomplished by administering a series of multiple choice mental tests to everyone, and then by suggesting, on the basis of the scores, what each person's role in

> society should be . . . the project will be called the Census of Abilities. (Lemann, 2000, p. 5)

The criticisms of using testing to sort or classify are much louder than the proponents: "Using test scores to pick winners and losers—whether states, districts, schools, or individuals—is misguided at best and truly harmful at worst" (Cremin, 1961, p. 190). Feuer (2011) pointed out that even John Dewey got into the fray:

> The IQ, Dewey argued, "is an indication of risks and probabilities. Its practical value lies in the stimulus it gives to more intimate and intensive inquiry into individualized abilities and disabilities." Barring complete imbecility, he continued, even the most limited member of the citizenry had potentialities that could be enhanced by a genuine education for individuality. "Democracy will not be democracy until education makes it its chief concern to release distinctive aptitudes in art, thought, and companionship." Insofar as tests assisted this goal, they could serve the cause of progress; insofar as they tended in the name of science to sink individuals into numerical classes, they were essentially antithetical to democratic social policy. (pp. 26-27)

Garrison (2009) echoes the criticism of sorting and classifying when he says that "Hence, we have the psychometric practice of equal treatment as a basis for social differentiation" (p. 103).

Consequences or Results of Testing

Uses lead to consequences. With each use of standardized testing data there are many possible consequences. The perceptions of testing are both a result of and a cause for the following major consequences of test use.

Testing as Tool for Centralization

The influence of testing on centralization in the U.S is one of the more intriguing ideas that manifests in the literature:

> Responsibility for setting standards and enforcing consequences has moved from more than 18,000 districts in mid-century to the 50 states in the 1980s. And, by 2002, the federal Department of Education (DOE), and agency that did not even enjoy cabinet status until the late 1970s, had taken over responsibility for providing core direction to the nation's 90,000 schools. Mid-twentieth century state governments had delegated virtually all of their constitutional responsibility for education policy to local school districts. By the first decade of the 21st century, however, states find themselves pressured by federal mandates, competing with one another for federal incentive grants, and looking to Washington for fiscal bailouts. (Mitchell et al., 2011, p. 286)

Mitchell, et al. (2011) clearly articulate that the "federalization of educational governance over the last 60 years is the most prominent common theme" (p. 36). Koretz (2008), an educational testing expert from Harvard, wrote:

> The shift from using tests for information to holding students or educators directly accountable for scores is beyond a doubt the single most important change in testing in the past half century . . . It is not an exaggeration to say that it is now the cornerstone of American education policy. This trend culminated in the enactment of the No Child Left Behind Act in 2001. (pp. 87-88)

At first glance it may seem like Mitchell et al. and Koretz are talking about two different things being the most significant change in the last fifty years. But, these two educational shifts,

centralization and an increase of standardized testing, are inseparably connected. In large part, the latter is the "how" for the former "what."

Airasian (1987) saw this trend toward using testing as a tool for centralization of public education 30 years ago:

> The increasingly tighter link between test results and decision making poses the serious threat of further erosion of local school control. As tests developed and/or administered by agencies external to the local school district take on heightened importance, control of the curriculum and other school prerogatives shifts from the local school or district level to that of the test developer or administrator, usually the state . . . the agency that controls the important tests or certifying devices in education exerts substantial influence over the curriculum of schools (Madaus and Airasian 1977). State-mandated pupil and teacher certification testing programs represent a move by legislatures and state departments of education to exert greater power and control over local schools. (p. 406)

Quite a few authors have made the connection between testing and centralization, as articulated in this statement from Turgut (2013):

> Over the years, educational reforms in the United States have become more centralized, standardized, measurable, and strict with the increasing involvement of the federal government in education. Currently, the unsatisfactory ranking of the United States in international tests is used as the major reason for establishing nationwide standards to "race to the top" of international test rankings. (p. 72)

Here is a similar message with slightly different details:

> But, the ramifications of using output measures, in the form of tests, as an organizational control in public schooling are far from politically neutral. Like other reforms that rely

> on a technical model of the schooling process, the recent reforms that rely on testing have helped to change the tools and the language of political debate, to weaken the authority of professional judgment, and to centralize school governance. (Salganik, 1985, p. 609)

Tension Between Levels of Control

The historical trend toward centralization has created significant tension between local control at the school district and state offices of education levels, and external control at the federal level. In the midst of this tension, one of the main uses of tests has been to help facilitate the move toward federal control. Airasian (1987) noted that in particular that

> new forms and uses of standardized testing have arisen. . .Traditional, instructionally oriented testing that is controlled by local school districts and used primarily in the service of classroom teaching and learning now coexists with policy-oriented testing that is controlled by agencies external to the local district and used to implement or to assess the effect of an educational policy or practice. (p. 409)

Koretz (2008) concurs on this function of standardized testing:

> There has been a fundamental change in the primary functions of large-scale achievement testing, with accountability gradually superseding diagnosis of the strengths and weaknesses of individual students' learning. This shift in how tests are used has been accompanied by changes in the types of conclusions test scores are used to support. Inferences about individual students remain important—indeed, in many states and localities these conclusions have much more serious consequences than they did three for four decades ago—but in many cases, conclusions about the performance of groups, in particular the performance of schools and districts, are far more consequential. (p. 47)

Many authors address this tension between external and internal levels and how they are using standardized testing. For example, Hess and Mehta (2013) assert "Because today's data discussion mostly concerns external accountability for schools and educators, it has focused almost exclusively on test scores in reading and math and on graduation rates. Not surprisingly, teachers have viewed this whole enterprise as an intrusion" (p. 74).

Others see this as a necessary intrusion to spur the needed reforms because they don't think the internal stakeholders are capable of reform without this external pressure. Phelps (2003), an influential proponent of standardized testing, bluntly asks:

> The key, essential point of debate is who gets to measure school performance—the education "professionals" or those of us who are footing the bills and giving up our children. The essential point of debate is whether testing, and other methods of quality control, should be done "internally" or "externally." (p. 1)

Movement toward centralization has been facilitated partly by using standardized testing as a control mechanism. But centralization has also created more intense tension between the different levels of school governance. Rothman (1995) agrees that centralization is a consequence of using tests to control education, but he questions whether or not these policies get us to where we want to be:

> But it is external tests that have increasingly driven what is taught in schools. And this is no accident. As part of their attempt to hold schools accountable for student performance, states and school districts have not only implemented testing programs but have also made sure that there are consequences—real or perceived—attached to the results. That way, students and schools have an incentive to keep their 'eyes on the prize' and to improve performance. Thus in recent years a growing number of states have made

> sure that good things happen to schools where test scores go up and (in some cases) bad things happen when they go down. As we will see, these policies have had the desired effect of making teachers and schools pay attention to the tests and strive to boost scores, but these efforts have not always ended up the way public officials intended. (p. 43)

Watching the current backlash against common core testing is certainly an indication that some stakeholders are not pleased with the struggle for local and/or federal control.

Media Interest in Testing

Ever since Horace Mann gave the first tests in Boston, media interest has been high. Sometimes it seems that media coverage of test scores is a relatively new activity, but it was around from the beginning:

> More than a glimpse of the future had appeared in 1845 . . . For the first, but hardly the last time, citizens read about the shocking test results in newspapers and magazines and debated whether they signified a school system in decline. The major political parties and their intensely partisan newspapers—the leading "media" of the day—were forced to take a stand. Their flawed statistics notwithstanding, reformers then and later compared cities unfavorably with suburbs or smaller communities, whether Roxbury or Quincy, based on test scores. In response to their numerous critics, examiners also claimed on the front page of the Boston Daily Atlas in 1846 that the local schools were inferior to "the public schools of Scotland and Holland, to say nothing of Prussia," anticipating the modern movement to rank schools across national boundaries. (Reese, 2013, pp. 226-227)

Media involvement has accelerated in the past few years. Now releases of international scores are usually front-page news, as well as the annual "grading schools" reports that are common around the country.

No single event got as much media attention as the famous report commissioned by Terrel Bell of the U.S. Department of Education in 1983 entitled *A Nation at Risk*. At the time of Bell's report, the U.S. Department of Education was actually on the chopping block, but with the release of *A Nation at Risk,* the department survived and took on a more central role:

> The report was an immediate sensation. Its conclusions were alarming, and its language was blunt to the point of being incendiary . . . The national news media featured stories about the crisis in education. The report got what it wanted: the public's attention. (Ravitch, 2010, pp. 24-25)

Interestingly enough, 10 years later in 1993, Bell commented on his use of SAT scores that were very influential in the report. He had used SAT scores to show that the overall performance in American schools was declining. This was a function the SAT was never designed to do. In this regard, he said,

> When I published a ranking of the states with these indicators, I included with every chart a cautionary statement on the limitation of these data. But the statement was largely ignored both by the press and by many educational leaders. The national pastime of jumping to conclusions was just as avidly pursued in those days as it is today. (1993, p. 594)

He claims to have warned the public against the misuse of standardized tests, yet much of the foundation for *A Nation at Risk* was built on faulty application of test scores. *A Nation at Risk* continued to influence policy and school governance for many years after its release:

> This attack was waged with great vigor, was eagerly supported by prominent figures in industry, and was widely reported and endlessly elaborated on by a compliant press. And as a result, many of the claims of this attack came to be accepted by good-hearted Americans, including a lot of powerful people and leaders in the educational community; and great mischief resulted because of the misunderstandings and poor policies this attack created. (Berliner & Biddle, 1995, p. 343)

Whether it is an appropriate use of the test or not, media interest and coverage of standardized testing results have certainly become mainstream. Just as with Bell in 1993, there are significant limitations to test data, but most of those limitations are ignored because like a competition on the sports page, the public wants to see the statistics and see who is winning the testing game:

> It didn't take the press long to figure out that something potentially newsworthy was taking place. Newspapers could easily write stories that compared schools within a district on the basis of competency test failure rates (and subsequent diploma denial). Consequently, a public perception began to emerge that schools in which few students failed were good schools, and schools in which many students failed were bad schools. The quality of schooling was being linked to the quality of students' test scores. And, as we shall see, once this approach to judging schools took root, it flourished. (Popham, 2001, p. 8)

One of the results of the media playing a major role in communicating information about standardized testing is that often that the information reported can be misleading, inaccurate, or just plain wrong:

> But anyone who tries to follow this information by reading newspaper accounts, press releases, or public statements of education reformers or district and state administrators

> can be excused for being somewhat confused. Accounts are often inconsistent, even when the same data are referenced. Claims about scores are often exaggerated or simply wrong. Scores are routinely reported in forms that make it hard to know whether a change in scores or a difference between groups is relatively good news or unusually bad. Changes in context that should shape the interpretation of scores—such as trends in the mix of students tested—are typically ignored entirely. Completely unsubstantiated claims about he causes of changes in scores are ubiquitous. It is important to get the story straight. (Koretz, 2008, pp. 74-75)

Misinterpretation of Testing Data

Bell's reference to the media ignoring his warnings about testing data limitations is just the tip of the iceberg in what the literature says about misinterpretation of results. This is one of the most frequently addressed issues throughout the literature. Both critics and proponents address this issue with regularity. Misinterpretation is probably the biggest downside to testing because as we have seen, the perceptions of testing (as previously presented) lend significant credibility to test results. Yet, if a skewed vision of what tests are meant to convey results in misinterpretation, then using tests becomes counterproductive. Bower (2013) used an analogy of a broken clock to make his point:

> Bestselling author and blogger Seth Godin reminds us that the worst kind of clock is a clock that randomly runs fast or slow. "If there's no clock," Godin writes, "we go seeking the right time. But a wrong clock? We're going to be tempted to accept what it tells us." Godin's message is that tracking the wrong data or misreading good data can get us into trouble. What if standardized test scores aren't telling us what we think they are telling us? What if the scores are illusions that are giving us false confidence? What if our

> reliance on standardized testing to judge our schools is like relying on a broken clock for time? (p. 24)

How are test scores so easily misinterpreted? Possibly because so very few people know the intricacies or nature of testing data, including many of those who use them. It becomes a short jump to immediately draw conclusions. Another reason may be that

> Often—people search for explanations of specific test scores that are, for whatever reason, of particular interest to them. Parents want to identify effective schools for their children; politicians want to claim credit for successful initiatives or to use low scores to justify reforms; news papers want to highlight supposed differences in school quality; critics want to identify failures; educators want to claim success and so on. (Koretz, 2008, pp. 132-133)

Therefore, caution must be used in what accountability means in practice, because how accountability is enforced could have unintended consequences. And, if teachers are expected to be the main ingredient in the recipe to fix what is purportedly wrong with our schools, failure may be unavoidable:

> Ultimately, great teachers make great schools, but great teachers can't do it alone—they require the support of an equitable society. If we are not careful, we risk misinterpreting the scores, and instead of waging war on poverty and inequity, we end up waging war on teachers and schools. (Bower, 2013, p. 26)

One of the causes of test misinterpretation is that the tests are often used in ways that the test designers said were not an appropriate use of the test. As Wiliam (2010) points out: "This distinction is particularly important in view of the fact that test scores are often interpreted in

ways that differ significantly from those intended by the designers of the test" (p. 107). For example:

> Though appropriate for making judgments about the achievement level of students at a school for a given year, they are inappropriate for judgments about educational effectiveness. In this regard, status measures are blind to the possibility of low achieving students attending effective schools. It is this possibility that has led some critics of No Child Left Behind (NCLB) to label its accountability provisions as unfair and misguided and to demand the use of growth analyses as a better means of auditing school quality. (Betebenner, 2009, p. 3)

Throughout the literature, one of the most common examples of test misinterpretation is judging teacher quality by standardized testing data. The literature includes persistent and consistent warnings that judging teacher quality using test scores is inappropriate. Yet that very use of the test data continues to be one of the most popular uses among policymakers. Many, if not every state policy ties at least a piece of teacher evaluations to their students' test scores. Advice from over 50 years ago should be heeded today:

> Administrators, either on their own, or at the insistence of parents and school board members, all too often judge the quality of a teacher's instruction by the average scores earned by that teacher's students on a standardized test. This can be far more dangerous than even the most knowledgeable advocates of educational measurement are likely to know. The danger lies in the fact that it is so easy to accept test results as the only evidence of teaching quality—when, at their best, tests can yield only a small part of the evidence necessary to make a sound judgment. (Chauncey & Dobbin, 1963, pp. 102-103)

Unfortunately, there are some stakeholders who willfully misuse the standardized testing data that go against the purpose for which the tests were designed.

One of the elements in the literature that is important to consider is context. The context of the data must be considered in order to have a better chance of interpreting scores appropriately and learning what we should from them. Results should always be evaluated "in light of the special challenges faced by a given school or district: A large number of student with special needs, or a very low-income community, provides a necessary context in which to understand a set of results" (Kohn, 2000, p. 47). By considering the context, data is more likely to be used appropriately and in more constructive ways:

> In such an intentionally designed and comprehensive system, a wealth of data emerges. Inherent in its design is the need for all assessors and users of assessment results to be assessment literate—to know what constitutes appropriate and inappropriate uses of assessment results—thereby reducing the risk of applying data to decisions for which they aren't suited. (Chappuis et al., 2009, p. 19)

Tests as a Sole Source of High-Stakes Judgment

One of the leading causes of misinterpretation of testing data is a results from putting a tremendous amount of weight on one source of information. Using tests as the sole source of many high-stakes judgments is common in many contexts. This is done by policy makers, economists, and sometimes teachers and educational leaders (at all levels).

As mentioned several times previously, the assessment frenzy that surrounded the early 1900s led to a reliance on the tests that gave inordinate importance to them. One early psychometrician claimed, "Time and again, it has been shown that the scores on a single intelligence examination enable us to predict college success as accurately as we can predict it

from four years of high school marks" (Wood, 1923, p. 5). This type of reliance continues in some ways today. Ask any adult what their ACT score was, and they often remember it either with pride or shame.

At the same time, the AERA guidelines on standardized testing clarify that "decisions that affect individual students' life chances or educational opportunities should not be made on the basis of test scores alone" (AERA, 2000). Contemporary literature echoes AERA's guidelines. For example:

> A review of the technical evidence leads us to conclude that, although standardized test scores of students are one piece of information for school leaders to use to make table of contents judgments about teacher effectiveness, such scores should be only a part of an overall comprehensive evaluation. (Baker et al., 2010, pp. 1-2)

Both proponents of testing and critics agree on this application of testing. Betebenner et al (2011) noted that "to be clear about our own opinions on the subject: The results of large-scale assessments should never be used as the sole determinant of education/educator quality" (p. 2). Another author wrote:

> We've said before that we don't oppose the judicious use of standardized tests. But, the inherent unfairness of allowing the scores on standardized tests to be our primary--in some cases, our only—way of judging school quality is one of the cruelest ironies in the way public education in America has evolved. (Harris et al., 2011, p. 45)

Ironically, though the literature is clear on this point, tests continue to be used as the sole source of various high-stakes judgments. NCLB accelerated this kind of use in the United States. Rothstein (2008) voiced the consequence of this use: "Under No Child Left Behind,

reliance solely on numerical measures, principally math and reading scores, to evaluate performance has corrupted schooling" (p. 14).

Narrowing the Curriculum

One of the negative consequences that has been articulated from very early on is that standardized testing narrows the curriculum. Because of the pressure the tests exert, curriculum is consequently narrowed. What gets tested gets taught. Ravitch (2013) addresses one manner in which testing has this negative affect:

> Certainly teachers should be evaluated, but evaluating them by the rise or fall of their students' test scores is fraught with perverse consequences. It encourages teaching to multiple-choice tests; narrowing the curriculum only to the tested subjects; gaming the system by states and districts to inflate their scores; and cheating by desperate educators who don't want to lose their jobs or who hope to earn a bonus. When the tests become more important than instruction, something fundamental is amiss in our thinking. (p. 111)

There are some proponents of testing that minimize the effect that standardized tests have:

> There are studies suggesting that multiple-choice tests result in a narrowing of the curriculum and more drill work in teaching. But, in fact, the studies are few in number and critics of traditional basic skills testing accept the studies somewhat uncritically. In my opinion, the evidence is not as strong as the rhetoric of those reporting the research would suggest and there is some research evidence that teachers do not choose topics based on the test content. (Mehrens, 1998, p. 8)

But, though it may be difficult to ever quantify how much testing narrows the curriculum, Koretz (2008) shares some wisdom on the subject:

> A final, and politically unpalatable, piece of advice: we need to be more realistic about using tests as a part of educational accountability systems. Systems that simply pressure teachers to raise scores on one test (or one set of tests in a few subjects) are not likely to work as advertised, particularly if the increases demanded are large and inexorable. They are likely instead to produce substantial inflation of scores and a variety of undesirable changes in instruction, such as an excessive focus on old tests, an inappropriate narrowing of instruction, and a reliance on teaching test-taking tricks. (p. 330)

Teaching to and Gaming Tests

The tendency to teach to and game tests is closely related to narrowing the curriculum. One of the consequences of testing is that efforts will be made to make test scores go up, whether actual learning improves or not. This section addresses the idea that there are many instances where that happens, especially in the face of tests with high stakes attached:

> Insofar as repeated failure is meaningfully penalized, struggling schools face a powerful incentive to raise their performance ratings. However, schools may have at their disposal a range of mechanisms for improving their ratings. The mechanisms consistent with policymakers' intent are those that reform the inputs and processes of educational production within failing schools, but schools may also choose to "game" or manipulate the accountability system in ways that raise test scores without contributing to students' knowledge and skills. (Chiang, 2009, p. 1045)

Chiang concludes by explaining, "The threat of sanctions from school accountability systems provides powerful incentives for low-performing schools to raise test scores, but there is the potential for observed test score gains to stem from non-educational manipulation of testing conditions" (p. 1056).

Again, like so many other issues surrounding testing, this is not a new problem. It has been around as long as the tests have. Talking about testing in the late 1800s, Reese (2013) explained:

> Once competitive testing entered the schools, however, the issue became not whether there would be written exams, but whether, as the authors of quiz books believed, educators could teach pupils subject matter more effectively and enhance their test taking skills. As one commentator wryly commented in 1897, "passing the hot-house training keeps a youth at this 'trick' every week or two for years, until he is as skillful with the question as is a baseball expert with the twirled sphere." The sports metaphor was well chosen, and critics ever since have complained about how some pupils game the system. (p. 224)

What Tests Do Not Measure

Another major consequence of standardized testing is that because so much emphasis is put on the test, many other important elements of education are not given much attention. Critics argue that we are hurting society by not putting more emphasis on the things that matter more than reading and math. The tests

> tell us next to nothing about where anything in particular has been learned, about the relation between what is learned in one institution and what is learned in another, about how different individuals synthesize what they have learned in various institutions, and about what might be the best possible combinations of institutions for teaching particular kinds of knowledge or skills. (Cremin, 1976, p. 89)

This is a result of testing being an easy way to gather data. It is much easier to do a standardized test that generates quickly and easily to classify students than other methods of evaluation:

> In education, it is much easier to quantify student learning in terms of test scores rather than portfolios or students' performance. Similarly, it is easier to emphasize student achievement as the primary goal or purpose of education because this goal is presumably quantifiable and measurable, whereas other goals such as preparing students to live in a democracy are more difficult to capture. (Cooper et al., 2004, p. 47)

But, as Friedman and Mandelbaum (2011) point out, our future economy demands workers that can do much more than just be good at the skills on a standardized test. They say we need our "education system not only to strengthen everyone's basics—reading, writing, and arithmetic—but to teach and inspire all Americans to start something new, to add something extra, or to adapt something old in whatever job they are doing" (p. 102).

Several authors provide lists of things the tests do not measure, including things such as creativity, motivation, curiosity, etc. (Bracey, 2003, p. 31; Kohn, 2000, p. 17). These types of lists give us pause and help us remember that educating children is about a lot more than scores from standardized testing, yet we continue to measure the educational progress of students through these narrow lenses. But what can be quantified isn't always what matters most:

> Our schools will not improve if we value only what tests measure. The tests we have now provide useful information about students' progress in reading and mathematics, but they cannot measure what matters most in education. Not everything that matters can be quantified. What is tested may ultimately be less important that what is untested, such as a student's ability to seek alternative explanations, to raise questions, to pursue knowledge on his own, and to think differently. If we do not treasure our individualists, we will lose the spirit of innovation, inquiry, imagination, and dissent that has contributed

> powerfully to the success of our society in many different fields of endeavor. (Ravitch, 2010, p. 226)

In a study that analyzed thousands of questions from standardized tests in California, they found some interesting results regarding what the tests don't measure: Researchers at the Midcontinent Regional Educational Laboratory analyzed 6,942 items for the Stanford achievement batteries and the California Test of Basic Skills to identify the general cognitive abilities tested and study their relationship to student performance.

Two major findings emerged from the analysis of the 6,942 items. First the test items included only 9 of the 22 general cognitive operations, and second, the general cognitive operations required to answer the questions had very little to do with student achievement on those tests. (Marzano & Costa, 1988, p. 67)

So, the tests as they have been are not enough. Fullan (2011) warns:

> What sets out as progressive for the 21st century ends up going backwards. Make no mistake about it, the higher-order skills—critical thinking and reasoning, problem solving, communication (including listening), collaboration, digitally-based learning, citizenship—will become the new average for the rest of this century. The four wrong drivers block any possibility of heading down this critical path. (p. 9)

Robinson (1990) also clarified poignantly the paradigm shifts we need to make if we are to look to more complete views of what matters in education. We must "insist on measuring what is educationally significant, not just what is technically convenient. We must insist that there is much of value that cannot be quantified. We must say that there is much that counts that cannot be counted" (p. 89).

Negative School Culture

Standardized testing, when used poorly, can influence a school culture negatively. Anyone who has ever worked in a school knows that the culture of the school is very important not only to achievement but also for many other reasons that can't be measured through typical tests:

> Using tests as the sole basis for measuring adequate yearly progress can lead to huge numbers of schools being misclassified, even though the source of their failure can be quite random. The effects of such misclassification on resource allocation, student mobility, parental support for schools, and morale can be costly in ways we don't really know how to measure. (Feuer, 2011, pp. 27-28)

Berhent (2009) comments on this concept of culture and why we need to be careful how we use testing data:

> The teachers with whom I work are incredibly dedicated, creative, and compassionate individuals . . . Too often, however, their efforts are stymied by a lack of resources and support, overcrowded buildings and classrooms, outrageous amounts of paperwork, and pressure to become mindless drones of the testing industry. The high teacher turnover rate in public schools is a testament to the pressure placed on teachers. If we are to stem this tide, we need to abandon the blame-the-teacher rhetoric that is so fashionable today. (p. 244)

Another author warned of how using standardized tests could contribute to a negative culture for teachers:

> Adopting an invalid teacher evaluation system and tying it to rewards and sanctions is likely to lead to inaccurate personnel decisions and to demoralize teachers, causing talented teachers to avoid high-needs students and schools, or to leave the profession

> entirely, and discouraging potentially effective teachers from entering it. Legislatures should not mandate a test-based approach to teacher evaluation that is unproven and likely to harm not only teachers, but also the children they instruct. (Baker et al., 2010, p. 4)

The issue of teacher retention and recruitment presents significant challenges. In nearly all studies, teacher quality is the number one factor in student achievement. The challenging culture that is being created with standardized testing contributes to difficulties in recruiting and retaining quality teachers:

> There is evidence that high-stakes accountability testing makes it harder to keep teachers (Clotfelter, Ladd, Vigdor, & Diaz, 2003), that teachers of disadvantaged students are likely to experience greater pressure to improve their test scores and to focus on test content than teachers of more advantaged students (Herman, Abedi, & Golan, 1994), as well as a host of other unintended outcomes. (Wiliam, 2010, p. 118)

Selling Real Estate

As discussed, demographics can have an effect on testing data: additionally, and perhaps surprisingly, student testing data can also have an affect on demographics. Test results may affect people's decisions regarding where they live. Home values can actually be affected by test scores: "Schools are publicly rated and ranked based on test scores. In turn, the valuation of homes in a community can increase or decrease based on these rankings" (Madaus et al., 2009, p. 5). This connection is one example of how these quantitative indicators are given a huge weight in our society. Madaus et al. (2009) point out that "policy makers and the general public accept these scores as a symbol of educational quality and, in turn, use scores to make decisions about

school choice, hiring and firing of school administrators, restructuring of schools, and even home buying" (p. 139).

Daniel Koretz had a friend who asked him what school he should send his child to based on test scores. Koretz (2008) replied:

> If all you want is high average test scores, tell your realtor that you want to buy into the highest-income neighborhood you can manage. That will buy you the highest average score you can afford . . . The homebuyer's phone call reflected two misunderstandings of achievement testing: that scores on a single test tell us all we need to know about student achievement, and that this information tells us all we need to know about school quality . . . . a common misconception is that testing is simple and straightforward. (p. 6)

This reality is common throughout the literature on standardized testing:

> One can easily find critics arguing that the best predictor of a student's or school's score on any given standardized test is their associated social class—what one commentator called the Volvo Effect: family wealth as indicated by brand of car predicts student test performance. (Garrison, 2004, p. 62)

This issue of test scores influencing real estate gets enough traction in the literature that it can be considered as a separate consequence of standardized testing.

Using Test Scores Productively

The literature points to a disconnect between standardized testing and using the data in ways that actually improve teaching and learning. This is likely a result of the shift from using testing as a diagnostic tool that informs teaching to an accountability mechanism over the past 60 years as education has become more centralized in the United States. It appears "we've overinvested in data that are useful for public accountability, and we're underinvested in data

that improve management or instruction. If we want increased performance, we need to reverse these priorities" (Hess & Mehta, 2013, p. 74). Yet, because of the nature of school governance, it is easier said than done. A principal has almost no power in changing the testing mandates at the federal, state, or even the district levels of organization. There are some things they can do, and principals should be active in spreading their influence at those levels—but, for most principals, their influence is seriously limited. Despite these external pressures and demands, principals can still determine how they will discuss and use standardized testing data within their own schools.

Principals "must start by asking the right questions. The right questions shape organizational thinking and lead to answers and actions that improve learning for all students. Unfortunately, we find that far too many schools are asking the wrong questions" (Buffum et al., 2012, p. 3). Buffum et al. (2012) articulated further:

> What is the wrong question? How do we raise our test scores? This is the most pervasive, misguided (and misguiding) question. While high-stakes testing is an undeniable reality in public education, this fatally flawed initial question leads to the wrong answers for achieving deep levels of student learning. We are constantly confronted with situations in which district leadership gives lip service to RTI while reinforcing the need for schools to "get scores up now, or else." As a result, schools often seek out the quick fix that will result in a sudden bump in test scores rather than investing in the long-term work. (p. 3)

Yet, the accountability pressure that principals feel from external levels of organization makes it hard not to focus on raising test scores.

After all is said and done, the literature points to data use as the core issue regarding standardized testing. If we are being counterproductive with testing data, we better look to how it is being used at different organizational levels. If data use is being productive for teaching and learning, then we should also take heed and learn from it:

> Accountability system results can have value without making causal inferences about school quality, solely from the results of student achievement measures and demographic characteristics. Treating the results as descriptive information and for identification of schools that require more intensive investigation of organizational and instructional process characteristics are potentially of considerable value. Rather than using the results of the accountability system as the sole determiner of sanctions for schools, they could be used to flag schools that need more intensive investigation to reach sound conclusions about needed improvements or judgments about quality. (Betebenner & Linn, 2010, p. 18)

There are three ways data can be used by principals more effectively to improve teaching and learning in their school community: control the narrative, use data in context, and build a positive school data culture.

Control the Narrative

Principals must be clear about what they, the teachers in their school, and the school patrons care about and value. If principals were to survey their parents and ask the question, "What is the number one result you want for your child from a public education?" they would likely not get very many answers along the lines of "I hope they can pass the standardized test." Instead, parents are more likely to talk about intrinsic, holistic elements. Parents want their students to develop capabilities and characteristics like confidence, responsibility, healthy social

interactions, etc. Therefore, school leaders need to be more deliberate in defining, with their communities, what they mean by student learning.

What is student learning? Few schools or districts are very good at defining this construct. All schools and districts have mission statements, but they are usually vague and don't really get clear or specific about what a school is really striving for. That may be because other organizational levels have defined it for so long in terms of standardized testing, that it is difficult for many schools to think beyond raising scores. One school district that has done a great job of defining their purpose is the Lakeview Area Public Schools in Minnesota. In their 2016-18 strategic plan, they have clearly identified what their purpose is and articulated it in a way that is much more compelling than a typical school improvement plan that might say they are going to improve 2% in proficiency on standardized math, science, and language arts tests.

They have articulated that what matters to their school community is that they prepare students to be *future ready*. Instead of just academic targets, they have three main categories that they strive for. As shown in Figure 1, these categories are (a) Foundational Literacies, (b) Competencies, and (c) Character Qualities. Then, each of those categories is further defined in more specific terms (Snyder, 2015).

If a parent, a teacher, or a student were presented with the option to focus on these outcomes, or to try to raise test scores, it would likely be a rare stakeholder that would choose raising test scores. The irony of course is that if these other elements are being systematically addressed and taught, test scores will go up. The majority of the elements that testing addresses are in the column under foundational literacies.

Figure 1. *Lakeview Area Public Schools Definition of Student Learning*

Using the Data in Context

Another crucial element in the literature regarding how to use the testing data most effectively is that principals must always use data in the context of their school community. To compare scores between schools without considering context is misleading at best and destructive at worst. The Horace Mann League and the National Superintendent's Roundtable presented a much more comprehensive view of international large-scale assessment (ILSAs) results. The authors wrote:

> While results of ILSAs are potentially valuable, they are simply one of many potential indicators. Others should also be considered . . . viewed holistically, a portfolio of indicators can provide a more comprehensive view of the context in which any nation's public schools operate—and a more accurate guide for action. (Harvey et al., 2015, p. 3)

They continued:

> Educators understand the importance of assessments and accountability. However, most express concern that any assessment should help them improve education for the students

> in their classrooms. Simply developing a scoreboard without identifying the societal factors that influence results does not help the education system become more legitimately accountable to those it serves. (Harvey et al., 2015, pp. 4-5).

Harvey et al. then compare nine nations on six different domains, only one of which—the domain of *student outcomes*—takes data from ILSAs like PIRLS, PISA, TIMMSS. In addition to that domain, they also consider five other domains that give a more complete context to overall outcomes. The other five domains are economic equity, social stress, support for families, support for schools, and system outcomes. Each of these domains is then broken down into components with attached indicators, so that it is quite clear what is being considered. This approach gives a much more holistic view of what a nation is doing with their educational system, and it provides a more honest context for the standardized testing data:

> Too often, as the president of the Horace Mann League pointed out recently, policymakers, educators, and the public are inclined to narrow their focus to a few things that are easily tested. They become captives of the results and their goal becomes raising test scores rather than developing fully educated people. To avoid that mentality, the Horace Mann League and the Roundtable want to emphasize the power of a consistent and comprehensive framework that looks at all the measures involved in shaping our future citizens. (Harvey et al., 2015, p. 5)

In their conclusion, they reaffirm that context matters, and when context is not considered, "In too many cases, reports on international assessments encourage national leaders to consider education to be a 'horse race' in which nations compete with each other around educational outcomes, whatever the initial goals and purposes of individual national systems" (2015, p. 39).

School Data Culture

Schools need to provide better, more professional working environments in which students can learn. Berliner and Biddle (1995) state that "school improvement must also be concerned with creating environments in which teachers can succeed too" (p. 340). Just trying to raise test scores does not inspire meaningful, lasting change in schools. The literature is clear on that point. What does create meaningful and lasting change is improving school culture, and in the current testing environment, specifically a school data culture. The people required to influence culture are the people inside the schools. As Berliner and Biddle point out, "The crucial players needed to transform, reform, or improve schools are teachers and other educators. Literally nothing good will happen in our schools unless the professionals who run those schools make it happen" (1995, p. 336). The educational change expert, Michael Fullan has outlined a framework that comprehensively addresses this issue. He wrote. "The key to system-wide success is to situate the energy of educators and students as the central driving force. This means aligning the goals of reform and the intrinsic motivation of participants" (2011, p. 5). He then outlines four *wrong* drivers with their accompanying *right* drivers. His first *wrong* driver is "accountability: using test results, and teacher appraisal, to reward or punish teachers and schools" (p. 5)

The driver that Fullan calls the right driver is *capacity building*. He says, "the right drivers are effective because they work directly on changing the culture" (p. 4). Then he explains why putting drivers in their proper role is so important:

> In the rush to move forward, leaders, especially from countries that have not been progressing, tend to choose the wrong drivers. Such ineffective drivers fundamentally miss the target . . . Although the four "wrong" components have a place in the reform

> constellation, they can never be successful drivers. It is, in other words, a mistake to lead with them. Countries that do lead with them (efforts such as are currently underway in the US and Australia, for example) will fail to achieve whole system reform. Even worse, chances are that such strategies will cause backward movement relative to other countries that are using the right drivers. (p. 5)

One of the keys to his argument is about how much emphasis we put on each driver. Fullan is not anti-testing; in fact, he sees it as an important part of the educational system. But how much emphasis testing is given and where testing is placed as a driver matters:

> I need to be clear here. The four "wrong drivers" are not forever wrong. They are just badly placed as lead drivers. The four "right drivers"—capacity building, group work, pedagogy, and "systemness"—are the anchors of whole system reform. You don't have to give up your affinity to accountability, individual quality, technology, and favored quality components of the reform package. Stated another way, I am not talking about presence or absence or even sequence, but rather dominance. Dominance is another word for saying what system leaders state and acknowledge as the anointed, explicitly articulated lead drivers. The encouraging news is that the judicious use of the four right drivers ends up accomplishing better the goals that those espousing the wrong drivers are seeking. And it does so in a fundamentally more powerful and sustainable manner. The right drivers—capacity building, group work, instruction, and systemic solutions—are effective because they work directly on changing the culture of school systems (values, norms, skills, practices, relationships); by contrast the wrong drivers alter structure, procedures and other formal attributes of the system without reaching the internal substance of reform—and that is why they fail. (Fullan, 2011, p. 5)

Fullan clarifies that testing plays an important role, but only as a secondary driver. It can actually become counterproductive when placed in the role of primary driver:

> It is okay to use the full constellation of eight drivers along the way, as long as you make sure the less effective four play a decidedly second fiddle role to the right four. This distinction is critical because the evidence is clear: the wrong four as drivers de-motivate the masses whose energy is required for success; the right four drivers do the opposite. Countries that are successful (increasingly on a sustained basis) have figured this out and will only get stronger. (Fullan, 2011, pp. 5-6)

He sums up his framework by arguing that

> Strange as it sounds, leading with accountability is not the best way to get accountability, let alone whole system reform. The four right drivers actually produce deeper, more built-in accountability of action and results . . .To be clear, it is not the presence of standards and assessment that is the problem, but rather the attitude (philosophy or theory of action) that underpins them, and their dominance (as when they become so heavily laden that they crush the system by their sheer weight). If the latter is based on the assumption that massive external pressure will generate intrinsic motivation it is patently false. Instead (and this will require combining the right elements of all four driver sets) what is required is to build the new skills, and generate deeper motivation. Change the underlying attitude toward respecting and building the profession and you get a totally different dynamic around the same standards and assessment tools. (Fullan, 2011, p. 8)

Conclusion

The question of how to most effectively use standardized testing data has been elusive and often contentious throughout the history of the American education system. The reasons for

this struggle are complex, as is demonstrated by the number of influences and competing interests at play in the United States regarding the use of standardized tests. But using testing data effectively in a school community, regardless of how they are used at any other level, is the crucial issue. As Ravitch (2010) clarifies while quoting from other experts:

> School reform will continue to fail, Cohn warned, until we recognize that "there are not quick fixes or perfect educational theories. School reform is a slow, steady labor-intensive process" that depends on "harnessing the talent of individuals instead of punishing them for noncompliance with bureaucratic mandates and destroying their initiative." He predicted that "ground level solutions, such as high quality leadership, staff collaboration, committed teachers, and clean and safe environments, have the best chance of success. These solutions are not easily quantified. They cannot be experimented on by researchers or mandated by the federal government." . . . in my conversation with him, Cohn cited the work of sociologists Anthony Bryk and Barbara Schneider, who maintain in their study *Trust in Schools* that successful school reform depends on an atmosphere of trust. Trust "foments a moral imperative to take on the hard work of school improvement." Trust, not coercion, is a necessary precondition for school reform. (p. 66)

Ravitch goes on to write, "A good accountability system, whether for schools, teachers, or students, must include a variety of measures, not only test scores. To use a phrase I first heard from educator Deborah Meier, our schools should be 'data-informed,' not 'data-driven'" (Ravitch, 2010, p. 228).

In addition, all stakeholders should work more collaboratively to create this kind of a culture to improve teaching and learning. Tests have been used for multiple tasks, often in ways they were never designed to be used. We should be more transparent about how we use tests:

> While educational legislators might be literate in policy making, they should collaborate with educational researchers that are literate in interpreting test results before using them for reforms. Accurate interpretation of results by experts is as important as the validity and reliability of tests because tests are informative only if they are interpreted accurately and meaningfully. Furthermore, inclusion of educational researchers', teachers', and school leaders' perspectives can provide the additional information missing from the results and can help the nation accurately make meaning of the results by considering contextual factors. For educational reforms to be successful, reformers should feel required to embrace not only top-down but also bottom-up reforms. In the collaborative world in which we live, policy makers should learn how to collaborate and take input from the people who will be effected by the reforms, such as school administrators, researchers, teachers, and even students and parents. (Turgut, 2013, p. 70)

Ultimately, the comprehensive and effective use of testing data rests in the hands of a confident, capable principal who takes responsibility to guide their own school community in building a positive data culture. Despite, or perhaps because of, external control mechanisms which at times seem to work against this end, a visionary principal has the professional and moral responsibility to do whatever is needed on behalf of the children and school community served. The task is daunting, but the path forward must be created and walked—together if possible, but alone if necessary.

APPENDIX B: DETAILED METHODS

Research Problem

The conflict around test use leads to confusion and resulting pressures that burdens and, therefore, unavoidably reduces the efficiency and effectiveness of administrators in traditional K-12 schools in their various leadership roles. Ultimately, this conflict surrounding test use potentially impacts the value and power of the educational experience, and the future lives, of millions of American school children each year for whom these administrators have a very real professional and moral responsibility.

This study had two main purposes. First, the study aimed to explore the literature about standardized testing to find patterns in the narratives that are being told in the disciplines of education, policy, economics, psychology/psychometry, and history. Second, this study analyzed those narratives to determine what major themes emerge from each so that a principal can better understand the perspectives and influences in the standardized testing landscape.

Methods

Library databases were searched based on keywords associated with K-12 schooling and standardized testing in the USA. Then, reference sections in those sources were scoured to search for other qualifying books, book chapters, reports, and articles. Widely cited authors were sought, as were multiple disciplinary perspectives. Literature from multiple time periods was also sought. The search for representative literature resulted in 171 published works that were gathered on standardized testing, representing the five disciplinary domains, including literature from both critics and opponents in each domain. The 171 sources were narrowed down to 147 to eliminate redundancy and to ensure the sampled literature was addressing issues of standardized testing in traditional K-12 public schools in the United States. Table 1 presents the basic

distribution of sources ultimately included in this study by academic discipline and type of publication.

Table 1

Type and Number of Documents by Disciplinary Background

Academic Discipline	Articles	Books	Book Chapters	Papers/ speeches	Reports	Webpage	Total
Economics	15	0	1	0	4	0	20
Education	28	14	1	2	2	1	48
History	7	7	0	0	0	0	14
Policy	4	5	3	1	5	0	18
Psych	13	15	1	2	4	0	35
Other[a]	7	3	0	1	1	0	12
TOTALS	74	44	6	6	16	1	147

[a] The *other* category includes documents from the disciplinary backgrounds of philosophy, science, math, sociology, and unknown. These were works that were relevant to the issue of standardized testing, but didn't fall into the main five disciplinary categories addressed in this study.

Books, articles, reports, conference papers, and testing materials/instructions on the use or misuse of standardized testing were sought from multiple fields/disciplines—education, policy, psychology, psychometry, economics, and history. A wide variety of perspectives was sought until the body of literature had multiple items representing each of the disciplines in question as well as differing viewpoints from those disciplines. For a complete list of literature used in the study, see Appendix B.1.

The research started with library database searches. Diligently searching reference sections of significant works led to other valuable resources that either were apparent in the titles, were written by the same authors, or appeared consistently between items. Becoming a wise consumer of the literature was important because the body of literature is far larger than is possible to read.

Literature that is favorable toward standardized testing was specifically sought out because it is less readily available and viewpoint balance was crucial to this study. Frankly, it is a

lot easier to find literature that is critical of standardized testing. Also sought were authors exerting the largest influence (numbers of citations in other articles/books, references within the literature as to the magnitude of influence, etc.) in each discipline to get an appropriate representative sample of the differing viewpoints. Table 2 summarizes the literature used in the study.

Data Management

Once the sample of works was finalized, representative quotations were extracted from each source and then coded in QSR International's NVivo 10 Software (2014). After exporting the entire database from Endnote X6 (2012), there were 596 double-spaced pages of testing literature extracts that were coded. The coding structure that emerged (in a more emic approach) from the NVivo analysis clustered the data into the following six major categories:

1. Basic standardized testing knowledge
2. Influences on standardized testing
3. Perceptions of standardized testing
4. Uses of standardized testing
5. Consequences resulting from standardized testing
6. Alternatives to traditional standardized testing usage

A reader may question why the review of literature in this study seems unorthodox in length and style. The review of literature is also much more broad in scope than the article that emerged from this process. The nature of this archival study required a broad survey of literature about standardized testing. The narrow focus of the article did not fully emerge until after the NVivo analysis. Therefore, the review of literature still reflects all of the literature that was gathered for the study while the article focuses in on a more specific part of the literature. It

would have been easier to seek out only literature on the article topic, but the topic would have likely never emerged the way it did without the NVivo analysis of the entire body of literature.

This study eventually focused specifically on the category of *uses of standardized testing (*#4). Within this category 16 themes emerged about how standardized testing is used. The findings present the data regarding what is written about standardized test use by academic discipline. However, analysis and discussion about the uses of testing also need to consider the consequences and perceptions of standardized testing. Therefore, after the analysis of *test use,* the analyses then focused on the *perceptions* and *consequences* (#3 and #5) of standardized testing and were also considered by academic discipline.

Using NVivo as the primary analytical platform, we developed a broad profile, a *fingerprint* if you will, of each disciplinary perspective on the varying uses of standardized testing. These fingerprints showed clearly what authors from varying disciplines addressed in the standardized testing literature. When analyzing these NVivo fingerprints regarding the uses of standardized tests, the need to also consider *how* each of these disciplines treated the prominent themes about test use became apparent. This realization was spurred by the fact that the education and policy fingerprints were similar, yet it became obvious that these two disciplines wrote about the major themes quite differently.

In addition to coding the literature, we also collected and imported author attributes for each first author into NVivo including their primary educational training (disciplinary background), the domain in which they primarily worked during their career (career function), whether an author was a practitioner or an academic and their author viewpoint (critic, proponent, or neither), among others. The analyses for this study focused specifically on which themes within the three literature categories of use, perceptions, and consequences of

Table 2

Author Attributes and Definitions

Attribute Title	Definition
Title	Title of the source
Primary Author	In cases where there was more than one author I tracked the attributes of the primary author.
Author Academic Background	The primary disciplinary background the author received in their formal education.
Career Function	The disciplinary area in which they worked as a career.
Career Role	Was this author an academic (professor) or a practitioner?
Viewpoint	Was the author a critic, a proponent, or neither?
Flipped (Yes or No)	Did the author have a significant change in their viewpoint from one point in their career to another? (e.g.—Diane Ravitch)
Time Period	Four major eras of standardized testing in the United states are considered: • 1845-1900—The Beginnings • 1900-1965—The Frenzy • 1965-2000—Early Accountability • 2001-Present—Accelerated Accountability These different time periods have unique characteristics and are defined by certain turning points.
Year	The year the actual source was written.
Source	The journal or publisher that printed the source
Type of Literature	What type of literature was it? (Journal article, book, book chapter, speech, report, conference paper, webpage, etc.)
Cited By	The number of times the source has been quoted in other legitimate sources. This data point was obtained via Google Scholar.

standardized testing were most common within and between author's disciplinary backgrounds. This focus of this study had to be narrowed considerably because of space limitations for the *NASSP Bulletin*. Table 3 provides a summary of the author attributes that were tracked in this study.

The nature of the data set in this study will allow for many future articles. This first article establishes a foundation for many potential follow-up articles. This first article establishes *why* it is important for a principal to understand the different standardized testing narratives. One obvious article to follow would be to address *how* a principal can go about creating a compelling testing narrative. There are many other possibilities for analysis and publication from the data set.

APPENDIX C: DISSERTATION LITERATURE

Title	Author 1	Author Academic Discipline	Career Function	Viewpoint
State Mandated Testing and Educational Reform: Context and Consequences	Airasian, Peter	Psychometry	Education	Critic
AERA Position Statement on High-Stakes Testing in Pre-K – 12 Education	American Educational Research Association	Policy	Policy	Critic
What is Test Misuse? Perspectives of a Measurement Expert	Anastasi, Anne	Psychology	Psychometry	Proponent
Problems with the Use of Student Test Scores to Evaluate Teachers	Baker, Eva L	Psychology	Policy	Critic
Reclaiming our Freedom to Teach: Education Reform in the Obama Era	Behrent, Megan	Education	Education	Critic
Parting Words of the 13th Man	Bell, Terrel	Education	Policy	Proponent
Reflections One Decade after A Nation at Risk	Bell, Terrel	Education	Policy	Proponent
The Manufactured Crisis: Myths, Fraud, and the Attack on America's Public Schools	Berliner, David C.	Psychology	Psychology	Critic
A Primer on Student Growth Percentiles	Betebenner, D.A.	Psychology	Psychometry	Neither
Growth in student achievement: Issues of measurement, longitudinal data analysis and accountability	Betebenner, D.A.	Psychology	Psychometry	Neither
Student Growth Percentiles and Shoe Leather	Betebenner, D.A.	Psychology	Psychometry	Neither
The Development of Intelligence in Children	Binet, Alfred	Psychology	Psychometry	Proponent
Is the Test Score Decline Responsible for the Productivity Growth Decline?	Bishop, John H.	Economics	Economics	Proponent
Do curriculum-based external exit exam systems enhance student Achievement?	Bishop, John H.	Economics	Economics	Proponent
Telling Time with a Broken Clock: The Trouble with Standardized Testing	Bower, Joe	Education	Education	Critic
Why Can't They Be Like We Were?	Bracey, Gerald W.	Psychology	Education	Critic
The Fifth Bracey Report on the Condition of Public Education	Bracey, Gerald W.	Psychology	Education	Critic
On the Difficulty of Knowing Much of Anything about How Schools Reform over Time	Bracey, Gerald W.	Psychology	Education	Critic
On the Death of Childhood and the Destruction of Public Schools	Bracey, Gerald W.	Psychology	Education	Critic
Improving Schools by Standardized Tests	Brooks, Samuel	Education	Education	Proponent
Simplifying Response to Intervention: Four Essential Guiding Principles	Buffum, Austin	Education	Education	Critic
America 2000: An Education Strategy	Bush, George H.W.	Economics	Policy	Proponent

Title	Author 1	Author Academic Discipline	Career Function	Viewpoint
School Resources and Student Outcomes: An Overview of the Literature and New Outcomes	Card, David	Economics	Economics	Critic
What do International Tests Really Show about U.S. Student Performance?	Carnoy, Martin	Economics	Economics	Critic
Perspectives on Education in America: An Annotated Briefing, April 1992	Carson, C.C.	Unknown	Policy	Critic
The Politics of High Stakes Testing	Casbarro, Joseph	Education	Education	Critic
Mental Tests and Measurements	Cattell, J. McK	Psychology	Psychometry	Proponent
Schools as Sorters: Lewis M. Terman, Applied Psychology, and the Intelligence Testing Movement, 1890-1930	Chapman, Paul Davis	Education	Education	Critic
The Quest for Quality	Chappuis, Stephen	Education	Education	Proponent
Testing: Its Place in Education Today	Chauncey, Henry	Psychology	Psychometry	Proponent
How accountability pressure on failing schools affects student achievement	Chiang, Hanley	Economics	Policy	Neither
More Unintended Consequences of High-Stakes Testing	Cizek, Gregory J.	Psychometry	Psychometry	Proponent
High Stakes Testing: Contexts, Characteristics, Critiques, and Consequences	Cizek, Gregory J.	Psychometry	Psychometry	Proponent
Better Policies, Better Schools: Theories and Applications	Cooper, Bruce S.	Policy	Policy	Neither
The Transformation of the School: Progressivism in American Education, 1876-1957	Cremin, Lawrence A.	History	History	Neither
Public Education	Cremin, Lawrence A.	History	History	Proponent
Public Education in the United States: A Study and Interpretation of American Educational History	Cubberley, Ellwood P.	Education	Psychometry	Proponent
The Principal and his School	Cubberley, Ellwood P.	Education	Psychometry	Proponent
Does Teacher Preparation Matter? Evidence about Teacher Certification, Teach for America, and Teacher Effectiveness	Darling-Hammond, Linda	Education	Policy	Neither
Criteria for High-Quality Assessment	Darling-Hammond, Linda	Education	Policy	Proponent
Evaluating NCLB	Dee, Thomas S.	Economics	Economics	Neither
The Political Legacy of School Accountability Systems	Dorn, Sherman	History	Education	Critic
Secretary Arne Duncan's Remarks at OECD's Release of the Program for International Student Assessment (PISA) 2009 Results	Duncan, Arne	Education	Policy	Proponent
The Threat of Educational Stagnation and Complacency	Duncan, Arne	Education	Policy	Proponent

Title	Author 1	Author Academic Discipline	Career Function	Viewpoint
Essentials of Educational Measurement	Ebel, R.L.	Psychology	Psychometry	Proponent
PAYING FOR PUBLIC EDUCATION: NEW EVIDENCE ON HOW AND WHY MONEY MATTERS	Ferguson, Ronald F.	Economics	Policy	Neither
Politics, Economics, and Testing: Some Reflections	Feuer, Michael J.	Policy	Policy	Neither
A Nation Still at Risk	Finn, Chester	Education	Policy	Proponent
Twenty-Five Years Later, A Nation Still at Risk	Finn, Chester	Education	Policy	Proponent
That Used to Be Us: How America Fell Behind in the World it invented and How We Can Come Back	Friedman, Thomas	History	History	Neither
Choosing the Wrong Drivers for Whole System Reform	Fullan, Michael	Education	Education	Critic
Psychometric experiments	Galton, Francis	Psychology	Psychometry	Proponent
The Power of Testing	Gandal, Matthew	Unknown	Policy	Proponent
A Nation at Risk	Gardner, David P.	Policy	Education	Proponent
Measure, Mismeasure, or not measurement at All: Psychometrics as political Theory	Garrison, Mark J.	Education	Education	Critic
A Measure of Failure: The Political Origins of Standardized Testing	Garrison, Mark J.	Education	Education	Critic
Instructional Technology and the Measurement of Learning Outcomes	Glaser, Robert	Psychology	Psychometry	Neither
The Mystery of Good Teaching	Goldhaber, Dan D.	Economics	Economics	Neither
Why Don't Schools and Teachers Seem to Matter: Assessing the Impact of Unobservables on Educational Productivity	Goldhaber, Dan D.	Economics	Economics	Neither
Some Misconceptions about Large-Scale Educational Assessments	Goodman, Dean	Policy	Policy	Proponent
The Mismeasure of Man	Gould, Stephen Jay	Other-Science	History	Critic
Throwing Money at Schools	Hanushek, Eric A.	Economics	Economics	Neither
Making Schools Work	Hanushek, Eric A.	Economics	Economics	Proponent
The Seeds of Growth	Hanushek, Eric A.	Economics	Economics	Proponent
Does School Accountability Lead to Improved Student Performance?	Hanushek, Eric A.	Economics	Economics	Proponent
Alternative Testing and the National Agenda for Control	Harris, Karen	Education	Education	Critic
The Myths of Standardized Tests: Why They Don't Tell You What You Think They Do	Harris, Phillip	Psychology	Education	Critic
School Performance in Context	Harvey, James	Education	Education	Critic

Title	Author 1	Author Academic Discipline	Career Function	Viewpoint
New Assessments, New Rigor	Herman, Joan	Education	Psychometry	Proponent
NCLB's Critical Design Flaw and the Lesson to Take	Hess, Frederick M.	Education	Policy	Critic
Data: No Deus Ex Machina	Hess, Frederick M.	Education	Policy	Critic
The Tyranny of Testing	Hoffmann, Banesh	Other-math	Other	Critic
Accountability, incentives and behavior: the impact of high-stakes testing in the Chicago Public Schools	Jacob, Brian	Economics	Economics	Critic
The Case Against Standardized Testing: Raising Test Scores, Ruining the Schools	Kohn, Alfie	Other-sociology	Education	Critic
Measuring Up	Koretz, Daniel	Psychology	Education	Critic
The Plural Worlds of Educational Research	Lagemann, Ellen	History	History	Neither
The Wrath of High -Stakes Tests	Lattimore, Randy	Education	Education	Critic
The Structure of Success in American	Lemann, Nicholas	History	History	Critic
The Great Sorting	Lemann, Nicholas	History	History	Critic
The Big Test: The secret History of the American Meritocracy	Lemann, Nicholas	History	History	Critic
Can Education Do It Alone?	Levin, Henry M.	Economics	Economics	Critic
High-Stakes Testing and Economic Productivity	Levin, Henry M.	Economics	Economics	Critic
Educational Measurement	Lindeman, Richard H.	Education	Education	Proponent
Measurement and Assessment in Teaching	Linn, Robert L.	Psychology	Psychometry	Neither
A New Era of Test Based Educational Accountability	Linn, Robert L.	Psychology	Psychometry	Critic
The Mental Age of Americans	Lippmann, Walter	Other-philosophy	Other	Critic
The Mystery of the "A" Men	Lippmann, Walter	Other-philosophy	Other	Critic
The Reliability of Intelligence Tests	Lippmann, Walter	Other-philosophy	Other	Critic
The Abuse of the Tests	Lippmann, Walter	Other-philosophy	Other	Critic
A Future for the Tests	Lippmann, Walter	Other-philosophy	Other	Critic
Test Scores as Administrative Mechanisms in Educational Policy	Madaus, George	Psychology	Psychometry	Critic
The Paradoxes of High Stakes Testing: How they affect students, their parents, teachers, principals, schools, and society	Madaus, George	Psychology	Psychometry	Critic
Educational Measurement in the Elementary Grades	Madsen, I.N.	Education	Education	Proponent
Question: Do Standardized Tests Measure General Cognitive Skills? Answer: No.	Marzano, R.J.	Education	Education	Critic
Policymakers' Views of Student	McDonnell, Lorraine M.	Policy	Policy	Neither

Title	Author 1	Author Academic Discipline	Career Function	Viewpoint
Assessment				
Consequences of Assessment: What is the Evidence?	Mehrens, W. A.	Education	Policy	Neither
In Schools We Trust: Creating Communities of Learning in an Era of Testing and Standardization	Meier, Deborah	Education	Education	Critic
What Have We learned about Shaping Educational Policy?	Mitchell, Douglas E.	Policy	Policy	Neither
Testing: A Political Scalpel	Monroe, Rick	Education	Education	Critic
Educational Tests and Measurements	Monroe, Walter Scott	Education	Education	Proponent
Measuring the Results of Teaching	Monroe, Walter Scott	Education	Education	Proponent
The War Against Testing	Murray, David W.	Policy	Policy	Proponent
Standardized Achievement Testing	National School Boards Association	Policy	Policy	Neither
How Vouchers Could Change the Market for Education	Neal, Derek	Economics	Economics	Neither
Collateral Damage: How High-Stakes Testing Corrupts America's Schools	Nichols, Sharon	Psychology	Psychology	Critic
Why has High-Stakes Testing so Easily Slipped into Contemporary American Life?	Nichols, Sharon	Psychology	Psychology	Critic
Insults to the Soul	Ohanian, Susan	Education	Education	Critic
Testing in American Schools: Asking the Right Questions	OTA	Policy	Policy	Critic
An overview of America's education agenda	Paige, Rod	Education	Policy	Proponent
No Child Left Behind: The Ongoing Movement for Public Education Reform	Paige, Rod	Education	Policy	Proponent
Political Economy and the NCLB Regime	Parkinson, Paul	Education	Education	Critic
Kill the Messenger: The War on Standardized Testing	Phelps, Richard P.	Policy	Policy	Proponent
Persistently Positive: Forty Years of Public Opinion on Standardized Testing	Phelps, Richard P.	Policy	Policy	Proponent
The Rich, Robust Research Literature on Testing's Achievement Benefits	Phelps, Richard P.	Policy	Policy	Proponent
Implications of Criterion Referenced Measurement	Popham, James W.	Education	Psychometry	Neither
Why Standardized Tests Don't Measure Educational Quality	Popham, James W.	Education	Psychometry	Critic
The Truth about Testing: An Educator's Call to Action	Popham, James W.	Education	Psychometry	Critic
Accountability Tests' Instructional Insensitivity: The Time Bomb Ticketh	Popham, James W.	Education	Psychometry	Critic
Anchoring Down the Data	Popham, James W.	Education	Psychometry	Critic

Title	Author 1	Author Academic Discipline	Career Function	Viewpoint
Schooling, Statistics, and Poverty	Raudenbush, Stephen W.	Policy	Policy	Proponent
Why We're Behind: What Top Nations Teach Their Students But We Don't	Ravitch, Diane	History	Policy	Critic
The Death and Life of the Great American School System: How Testing and Choice are Undermining Education	Ravitch, Diane	History	Policy	Critic
Reign of Error: The Hoax of the Privatization Movement and the Danger to America's Public schools	Ravitch, Diane	History	Policy	Critic
Testing Wars in the Public Schools: A Forgotten History	Reese, William	History	History	Neither
Testing in America: A Supportive Environment	Resnick, Daniel P.	History	History	Proponent
Standards, Curriculum, and Performance: A Historical and Comparative Perspective	Resnick, Daniel P.	History	History	Proponent
Errors in Standardized Tests: A Systemic Problem	Rhoades, Kathleen	Psychology	Policy	Critic
Orgy of Tabulation	Richards, LeGrand	Other-philosophy	Education	Critic
The Agenda for Reform in the Use of Standardized Tests: Achieving the Ideal of Inclusiveness	Robinson, Sharon P.	Education	Policy	Neither
Measurement in Today's Schools	Ross, C.C.	Psychology	Education	Proponent
Measuring Up: Standards, Assessment, and School Reform	Rothman, Robert	Policy	Policy	Critic
The Way We Were? The Myths and Realities of America's Achievement	Rothstein, Richard	Policy	Policy	Critic
The Corruption of School Accountability	Rothstein, Richard	Policy	Policy	Critic
After Three Decades of Scientific Method in Education	Rugg, Harold	Psychology	Education	Critic
Why Testing Reforms Are so Popular and How they are Changing Education	Salganik, Laura Hersh	Psychology	Psychometry	Critic
Measuring the Value of Accountability	Spellings, Margaret	Policy	Policy	Proponent
Low Pay, Low Quality	Temin, Peter	Economics	Economics	Critic
The Measurement of Intelligence	Terman, Lewis M.	Psychology	Psychometry	Proponent
An Introduction to the Theory of Mental and Social Measurements	Thorndike, Edward L.	Psychology	Psychometry	Proponent
The Nature, Purposes and General Methods of Measurements of Educational Products	Thorndike, Edward L.	Psychology	Psychometry	Proponent
The Measurement of Intelligence	Thorndike, Edward L.	Psychology	Psychometry	Proponent
International Tests and the U.S. Educational Reforms: Can Success Be	Turgut, Guliz	Education	Education	Critic

Title	Author 1	Author Academic Discipline	Career Function	Viewpoint
Replicated?				
Exit Exams Harm Students Who Fail Them--And Don't Benefit Students Who Pass Them	Warren, John Robert	Other-sociology	Other	Critic
Testing in the Elementary School	Webb, L.W.	Education	Education	Proponent
Human Capital vs. Signaling Explanations of Wages	Weiss, Andrew	Economics	Economics	Neither
Standardized Testing and School Accountability	Wiliam, Dylan	Education	Education	Proponent
Measurement in Higher Education	Wood, Ben D.	Education	Education	Proponent

APPENDIX D: REFERENCES

AERA. (2000). AERA Position Statement on High-Stakes Testing in Pre-K – 12 Education. Retrieved from http://www.aera.net/AboutAERA/AERARulesPolicies/AssociationPolicies/PositionStatementonHigh-StakesTesting/tabid/11083/Default.aspx.

Airasian, P. (1987). State mandated testing and educational reform: context and consequences. *American Journal of Education, 95*(3), 393-412.

Anastasi, A. (1990). What is test misuse? Perspectives of a measurement expert. In ETS (Ed.), *The uses of standardized tests in American education*. Princeton, NJ: Educational Testing Service.

Baker, E. L., Barton, P. E., Darling-Hammond, L, Haertel, E, Ladd, H. F., Linn, R. L., . . . Shepard, L. A. (2010). Problems with the use of student test scores to evaluate teachers (EPI Briefing Paper #278 ed.). Washington, D.C.: Economic Policy Institute.

Behrent, M. (2009). Reclaiming our freedom to teach: Education reform in the Obama era. *Harvard Educational Review, 79*(2), 240-247.

Bell, T. (1988). Parting words of the 13th man. *The Phi Delta Kappan, 69*(6), 400-407.

Bell, T. (1993). Reflections one decade after "A Nation at Risk". *Phi Delta Kappan, 74*(8), 592-597.

Berliner, D. C., & Biddle, B. J. (1995). *The manufactured crisis: Myths, fraud, and the attack on America's public schools*. Reading, MA: Addison-Wesley.

Betebenner, D. W. (2009). Primer on student growth percentiles - 2009 (pp. 22): The Center for Assessment.

Betebenner, D. W. , & Linn, R. L. (2010). *Growth in student achievement: Issues of measurement, longitudinal data analysis and accountability*. Retrieved from http://www.k12center.org/publications.html

Betebenner, D. W., Wenning, R. J., & Briggs, D. C. (2011). Student growth percentiles and shoe leather. Retrieved from http://nciea.org/publications/BakerResponse_DB11.pdf.

Binet, A., & Simon, T. (1916). *The development of intelligence in children* (E. S. Kite, Trans.). Baltimore, MD: Williams and Wilkens Company.

Bishop, J. (1998). Do curriculum-based external exit exam systems enhance student achievement? *CPRE Research Report Series* (Vol. 40, pp. 1-42). Washington, D.C.: Cosortium for Policy Research and Improvement.

Bower, J. (2013). Telling time with a broken clock: The trouble with standardized testing. *Education Canada, 53*(3), 24-27.

Bracey, G. W. (1991). Why can't they be like we were? *The Phi Delta Kappan, 73*(2), 104-117.

Bracey, G. W. (1995). The fifth Bracey report on the condition of public education. *The Phi Delta Kappan, 77*(2), 149-160.

Bracey, G. W. (1997). On the difficulty of knowing much of anything about how schools reform over time. *The Phi Delta Kappan, 79*(1), 86-88.

Bracey, G. W. (2003). *On the death of childhood and the destruction of public schools*. Portsmouth, NH: Heinemann.

Brooks, S. (1922). *Improving schools by standardized tests*. Cambridge, MA: The Riverside Press.

Brown, E. (2015, October 12, 2015). Another state redefines 'proficiency' on Common Core tests, inflating performance, *The Washington Post*. Retrieved from

https://http://www.washingtonpost.com/news/education/wp/2015/10/12/another-state-redefines-proficiency-on-common-core-tests-inflating-performance/

Buffum, A, Mattos, M, & Weber, C. (2012). *Simplifying response to intervention: Four essential guiding principles*. Bloomington, IN: Solution Tree Press.

Bush, G. H.W. (1991). America 2000: An education strategy. Washginton, D.C.: Department of Education.

Carnoy, M, & Rothstein, R. (2013). What do international tests really show about U.S. student performance? Retrieved from Economic Policy Institute website: http://www.epi.org/publication/us-student-performance-testing/.

Carson, C.C., Huelskamp, R.M., & Woodall, T.D. (1992). Perspectives on education in America: An annotated briefing. *The Journal of Educational Research, 86*(5), 259-310.

Casbarro, J. (2005). The politics of high-stakes testing. *Education Digest*(February 2005), 20-23.

Cattell, J. M., & Galton, F. (1890). Mental tests and measurements. *Mind, 15*(59), 373-381.

Chapman, P. D. (1988). *Schools as sorters: Lewis M. Terman, applied psychology, and the intelligence testing movement, 1890-1930*. New York, NY: New York University Press.

Chappuis, S, Chappuis, J, & Stiggins, R. (2009). The quest for quality. *Educational Leadership, 67*(3), 14-19.

Chauncey, H, & Dobbin, J. (1963). *Testing: Its place in education today*. New York, NY: Harper & Row.

Chiang, H. (2009). How accountability pressure on failing schools affects student achievement. *Journal of Public Economics, 93*(9), 1045-1057.

Cizek, G. J. (2001). More untintended consequences of high-stakes testing. *Educational Measurement: Issues and Practice, 20*(4), 19-27.

Cizek, G. J. (2005). High stakes testing: Contexts, characteristics, critiques, and consequences. In R. P. Phelps (Ed.), *Defending standardized testing* (pp. 341). Mahwah, NJ: Lawrence Erlbaum Associates.

Cooper, B. S. , Fusarelli, L. D. , & Randall, E. V. (2004). *Better policies, better schools: Theories and applications*. Boston, MA: Pearson Education.

Cremin, L. A. (1961). *The transformation of the school: Progressivism in American education, 1876-1957*. New York, NY: Alfred A Knopf.

Cremin, L. A. (1976). *Public education*. New York, NY: Basic Books.

Cubberley, E. P. (1919). *Public educaiton in the United States: A study and interpretation of American educational history*. Boston, MA: Houghton Mifflin Company.

Cubberley, E. P. (1923). *The principal and his school*. Boston, MA: Houghton Mifflin Company.

Darling-Hammond, L. (2005). Does teacher preparation matter? Evidence about teacher certification, teach for America, and teacher effectiveness. *Education Policy Analysis Archives, 13*(42), 1-48.

Darling-Hammond, L. , Herman, Joan, Pellegrino, James, Abedi, Jamal, Aber, J. Lawrence, Baker, Eva L., . . . Steele, Claude M. (2013). Criteria for high-quality assessment Stanford, CA: Stanford Center for Opportunity Policy in Education.

Dorn, S. (1998). The political legacy of school accountability systems. *Education Policy Analysis Archives, 6*(1), n1.

DuFour, R. (2015). *In praise of American educators: And how they can become even better*. Bloomington, IN: Solution Tree Press.

Duncan, A. (2010). Secretary Arne Duncan's remarks at OECD's release of the program for international student assessment (PISA) 2009 results: U.S. Department of Education.

Duncan, A. (2013). The threat of educational stagnation and complacency. Washington, D.C.: U.S. Department of Education.

Ebel, R.L., & Frisbie, D.A. (1991). *Essentials of educational measurement*. Englewood Cliffs, NJ: Prentice Hall.

Ferguson, R. F. (1991). Paying for public education: New evidence on how and why money matters. *Harvard Journal on Legislation, 28*, 465-498.

Feuer, M. J. (2011). Politics, economics, and testing-some reflections. *Mid-Western Educational Researcher, 24*(1), 25-29.

Finn, C. (1989). A nation still at risk. *Commentary, 87*(5), 17-25.

Finn, C. (2008). Twenty-five years later, a nation still at risk, *The Wall Street Journal,* p. A.7. Retrieved from http://www.wsj.com/articles/SB120916804732546311

Friedman, T. L., & Mandelbaum, M. (2011). *That used to be us: How America fell behind in the world it invented and how we can come back*. New York, NY: Farrar, Struas, and Gireaux.

Fullan, M. (2011). Choosing the wrong drivers for whole system reform, *Paper 204 in Center for Strategic Education Seminar Series*: Center for Strategic Education. Retrieved from http://theeta.org/wp-content/uploads/2011/11/eta-articles-110711.pdf.

Galton, F. (1879). Psychometric experiments. *Brain, 2*(2), 149-162.

Gandal, M, & McGiffert, L. (2003). The power of testing. *Educational Leadership, 60*(5), 39-42.

Gardner, D. P. (1983). *A Nation at Risk*. Washington, DC: The National Commission on Excellence in Education.

Garrison, M. J. (2004). Measure, mismeasure or not measurement at all: Psychometrics as political theory. *Scholar Practitioner Quarterly, 2*(4), 61-76.

Garrison, M. J. (2009). *A measure of failure: The political origins of standardized testing*. Albany, NY: State University of New York Press.

Glaser, R. (1994). Instructional technology and the measurement of learning outcomes: Some questions. *American Psychologist, 18*(8), 519-521.

Goldhaber, D. (2002). The mystery of good teaching. *Education Next, 2*(1), 50-55.

Goldhaber, D., & Brewer, D. J. (1997). Why don't schools and teachers seem to matter: Assessing the impact of unobservables on educational productivity. *The journal of Human Resources, 32*(3), 505-523.

Goodman, D, & Hambleton, R. (2005). Some misconceptions about large-scale educational assessments. In R. P. Phelps (Ed.), *Defending standardized testing* (pp. 341). New Jersey: Lawrence Erlbaum Associates.

Gould, S. J. (1981). *The mismeasure of man*. New York, NY: W.W. Norton & Company.

Hanushek, E. A. (1981). Throwing money at schools. *Journal of Policy Analysis and Management, 1*(1), 19-41.

Hanushek, E. A. (1994). *Making schools work*. Washington DC: Brookings Institution Press.

Hanushek, E. A. (1998). Conclusions and controversies about the effectiveness of school resources. *Economic Policy Review*(March 1998), 11-27.

Hanushek, E. A. (2002). The seeds of growth. *Education Next, 2*(3), 10-17.

Hanushek, E. A., & Raymond, M. E. (2005). Does school accountability lead to improved student performance? *Journal of Policy Analysis and Management, 24*(2), 297-327.

Harris, K., & Longstreet, W.S. (1990). Alternative testing and the national agenda for control. *Social Studies, 81*(4), 148-152.

Harris, P., Smith, B. M., & Harris, J. (2011). *The myths of standardized tests: Why they don't tell you what you think they do*. New York, NY: Rowman & Littlefield.

Harvey, J, Marx, G, Fowler, C, & McKay, J. (2015). School performance in context: The Horace Mann League & the National Superintendents Roundtable. Retrieved from http://www.superintendentsforum.org/wp-content/uploads/2015/01/School-Performance-in-Context.pdf.

Herman, J, & Linn, R. L. (2014). New assessments, new rigor. *Educational Leadership, 71*(6), 34-37.

Hess, F. M. (2012). Edu-leaders get over your policy allergies. *Education Week*(August 6, 2012).

Hess, F. M., & Mehta, J. (2013). Data: No deus ex machina. *Educational Leadership, 70*(5), 71-75.

Hill, C. W. L. , & Jones, G. R. (2013). *Strategic management theory* (10th ed.). Stamford, CT: South-Western Cengage Learning.

Hoffman, B. (1962). *The tyranny of testing*. New York, NY: The Crowell-Collier Publishing Company.

Jacob, B. (2005). Accountability, incentives and behavior: The impact of high-stakes testing in the Chicago public schools. *Journal of Public Economics, 89*(5), 761-796.

Kohn, A. (2000). *The case against standardized testing: Raising test scores, ruining the schools*. Portsmouth, NH: Heinemann.

Koretz, D. (2008). *Measuring up: What educational testing really tells us*. Boston, MA: Harvard University Press.

Lagemann, E. (1989). The plural worlds of educational research. *History of Education Quarterly, 29*(2), 185-214.

Lattimore, R. (2001). The wrath of high-stakes tests. *The Urban Review, 33*(1), 57-67.

Layton, L. (2015, December 10, 2015). Obama signs new K-12 education law that ends No Child Left Behind, *The Washington Post*. Retrieved from c77f2cc5a43c_story.htmlhttp://www.washingtonpost.com/local/education/obama-signs-new-k-12-education-law-that-ends-no-child-left-behind/2015/12/10/c9e58d7c-9f51-11e5-a3c5-c77f2cc5a43c_story.html

Leithwood, K. , & Seashore-Louis, K. (2012). *Linking leadership to student learning*. San Fransisco, CA: Jossey-Bass.

Lemann, N. (1995a). The great sorting. *The Atlantic Monthly, 276*(3), 84-100.

Lemann, N. (1995b). The structure of success in America. *The Atlantic Monthly, 276*(2), 41-60.

Lemann, N. (2000). *The big test: The secret history of the American meritocracy*. New York, NY: Farrar, Straus, and Giroux.

Levin, H. M. (2001). High-stakes testing and economic productivity. In G. Orfield & M. L. Kornhaber (Eds.), *Raising standards or raising barriers? Inequality and high-stakes testing in public education* (pp. 39-49). New York, NY: The Century Foundation.

Levin, H.M. , & Kelley, C. (1994). Can education do it alone? *Economics of Education Review, 13*(2), 97-108.

Lindeman, R. H. (1967). *Educational measurement*. Glenview, IL: Scott, Foresman and Co.

Linn, R. L. (2010). A new era of test-based educational accountability. *Measurement: Interdisciplinary Research & Perspective, 8*(2-3), 145-149. doi: 10.1080/15366367.2010.508692

Linn, R. L., & Gronlund, N. E. (2000). *Measurement and assessment in teaching* (K. M. Davis Ed. 8th ed.). Upper Saddle River, NJ: Prentice Hall.

Lippmann, W. (1922a). The abuse of the tests. *The New Republic, 32*(415), 297-298.

Lippmann, W. (1922b). A future for the tests. *The New Republic, 32*(417), 9-11.

Lippmann, W. (1922c). The mental age of Americans. *The New Republic, 32*(412), 213-215.

Lippmann, W. (1922d). The mystery of the "A" men. *The New Republic, 32*(413), 246-248.

Lippmann, W. (1922e). The reliability of intelligence tests. *The New Republic, 32*(414), 275-277.

Ltd., QSR International Pty. (2014). NVivo qualitative data analysis Software (Vol. Version 10).

Madaus, G. (1985). Test scores as administrative mechanisms in educational policy. *Phi Delta Kappan, 66*(9), 611-617.

Madaus, G., Russell, M., & Higgins, J. (2009). *The paradoxes of high stakes testing: How they affect students, their parents, teachers, principals, schools, and society*. Charlotte, N.C.: Information Age Publishing.

Madsen, I.N. (1930). *Educational measurement in the elementary grades*. Yonkers on Hudson, NY: World Book Company.

Marzano, R.J., & Costa, A.L. (1988). Question: Do standardized tests measure general cognitive skills? Answer: No. *Educational Leadership, 45*(8), 66-71.

McDonnell, L. M. (1994). Policymakers' views of student assessment. Santa Monica, CA: RAND.

Mehrens, W. A. (1998). Consequences of assessment: What is the evidence? *Education Policy Analysis Archives, 6*(13), 1-30.

Meier, D. (2002). *In schools we trust: Creating communities of learning in an era of testing and standardization*. Boston, MA: Beacon Press.

Mitchell, D. E., Crowson, R. L., & Shipps, D. (2011). What have we learned about shaping educational policy? In D. E. Mitchell, R. L. Crowson & D. Shipps (Eds.), *Shaping education policy: Power and process* (pp. 286-296). New York, NY: Routledge.

Monroe, R. (1987). Testing: A political scalpel. *The English Journal, 76*(8), 24.

Monroe, W. S. (1918). *Measuring the results of teaching*. Boston, MA: The Riverside Press Cambridge.

Monroe, W. S., De Voss, J. C. , & Kelly, F. J. (1917). *Educational tests and measurements*. New York, NY: Houghton Mifflin Company.

Murray, D. W. (1998). The war against testing. *Commentary*(September 1998), 34-37.

Neal, D. (2002). How vouchers could change the market for education. *The Journal of Economic Perspectives, 16*(4), 25-44.

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

Nichols, S. L., & Berliner, D. C. (2008). Why has high stakes testing so easily slipped into contemporary American life? *The Education Digest*(December 2008), 41-47.

NSBA. (1977). Standardized achievement testing (pp. 54): National School Boards Association.

Office of Technology Assessment, United States Congress. (1992). *Testing in American schools: Asking the right questions*. Washington, D.C.: United States Government Printing Office.

Ohanian, S. (1997). Insults to the soul. *English Journal, 86*(5), 32-35.

Paige, R. (2002). An overview of America's education agenda. *Phi Delta Kappan, 83*(9), 708-713.

Paige, R. (2006). No Child Left Behind: The ongoing movement for public education reform. *Harvard Educational Review, 76*(4), 461-473.

Parkinson, P. (2009). Political economy and the NCLB regime. *The Educational Forum, 73*, 44-57.

Pfeffer, J., & Salancik, G. R. (2003). *The external control of organizations*. Stanford, CA: Stanford University Press.

Phelps, R. P. (2003). *Kill the messenger: The war on standardized testing*. New Brunswick, NJ: Transaction Publishers.

Phelps, R. P. (2004). The rich, robust research literature on testing's achivement benefits. In R. P. Phelps (Ed.), *Defending standardized testing* (pp. 341). Mahwah, NJ: Lawrence Erlbaum Associates.

Phelps, R. P. (2005). Persistently positive: Forty years of public opinion on standardized testing. In R. P. Phelps (Ed.), *Defending standardized testing* (pp. 341). Mahwah, NJ: Lawrence Erlbaum Associates.

Popham, W. J. (1999). Why standardized tests don't measure educational quality. *Educational Leadership, 56*, 8-16.

Popham, W. J. (2001). *The truth about testing: An educator's call to action*. Alexandria, VA: ASCD.

Popham, W. J. (2007). Accountability tests' instructional insensitivity: The time bomb ticketh. *Education Week, 27*(12), 30-31.

Popham, W. J. (2009). Anchoring down the data. *Educational Leadership, 66*(4), 85-86.

Popham, W. J., & Husek, T.R. (1969). Implications of criterion referenced measurement. *Journal of Educational Measurement, 6*(1), 1-9.

QSR, International Pty Ltd. (2014): NVivo qualitative data analysis software (Version 10). .

Raudenbush, S. W. (2004). *Schooling, statistics, and poverty*. Paper presented at the William H. Angoff Memorial Lecture Series, ETS, Princeton, New Jersey.

Ravitch, D. (2010). *The death and life of the great American school system*. New York, NY: Basic Books.

Ravitch, D. (2013). *Reign of error: The hoax of the privatization movement and the danger to America's public schools*. New York, NY: Alfred A. Knopf.

Ravitch, D, & Cortese, A. (2009). Why we're behind: What top nations teach their students but we don't. *The Education Digest, 75*(1), 35-38.

Reese, W. J. (2013). *Testing wars in the public schools: A forgotten history*. Cumberland, RI: Harvard University Press.

Resnick, D. P. (1981). Testing in America: A supportive environment. *The Phi Delta Kappan, 62*(9), 625-628.

Resnick, D. P., & Resnick, L. B. (1985). Standards, curriculum, and performance: A historical and comparative perspective. *Educational Researcher, 14*(4), 5-20.

Rhoades, K, & Madaus, G. (2003). Errors in standardized tests: A systemic problem. Boston, MA: National Board on Educational Testing and Public Policy.

Richards, L. (2004). Orgy of Tabulation. *Far Western Philosophy of Education Society*.

Robinson, S. P. (1990). The agenda for reform in the use of standardized tests: Achieving the ideal of inclusiveness. In ETS (Ed.), *The uses of standardized tests in American education* (pp. 98). Princeton, NJ: Educational Testing Services.

Ross, C.C. (1941). *Measurement in today's schools*. New York, NY: Prentice-Hall.

Rothman, R. (1995). *Measuring up: Standards, assessment, and school reform*. San Fransisco, CA: Jossey-Bass Publishers.

Rothstein, R. (1998). *The way we were? The myths and realities of America's achievement*. New York, NY: The Century Foundation Press.

Rothstein, R. (2008). The corruption of school accountability. *The School Administrator, 65*(6), 14-18.

Rugg, H. (1934). After three decades of scientific method in education. *Teachers College Record, 36*(1), 111-122.

Salganik, L. H. (1985). Why testing reforms are so popular and how they are changing education. *The Phi Delta Kappan, 66*(9), 607-610.

Saxe, J. G. (1873). *The Poems of John Godfrey Saxe, Complete Edition*. Boston, MA: James R. Osgood and Co.

Snyder, L. L. (2015). *2016-2018 Strategic Plan: Lakeville Area Public Schools*. Lakeville Area Public Schools. Minnesota. Retrieved from http://isd194.org/about/strategic-plan/

Spellings, M. (2010). Measuring the value of accountability. *U.S. News & World Report, 147,* 33-34.

Temin, P. (2014). Low pay, low quality. *Proquest, 3*(3), 1-8.

Terman, L. M. (1916). *The measurement of intelligence*. Boston, MA: Houghton Mifflin Company.

Thorndike, E. L. (1913). *An introduction to the theory of mental and social measurements*. Teachers College, Columbia University.

Thorndike, E. L. (1918). The nature, purposes and general methods of measurements of educational products. *In The seventeenth yearbook of the National Society for the Study of Education* (Vol. 7). Bloomington, IL: Public School Publishing Company.

Thorndike, E. L. (1926). *The measurement of intelligence*. New York, NY: Columbia University.

Turgut, G. (2013). International tests and the U.S. educational reforms: Can success be replicated? *Clearing House, 86*(2), 64-73. doi: 10.1080/00098655.2012.748640

Warren, J. R., & Grodsky, E. (2009). Exit exams harm students who fail them-and don't benefit students who pass them. *The Phi Delta Kappan, 90*(9), 645-649.

Webb, L.W., & Shotwell, M. A. R. M. (1939). *Testing in the elementary school*. New York, NY: Farrar & Rinehart.

Weiss, A. (1995). Human capital vs. signalling explanations of wages. *The Journal of Economic Perspectives, 9*(4), 133-154.

Wiliam, D. (2010). Standardized testing and school accountability. *Educational Psychologist, 45*(2), 107-122. doi: 10.1080/00461521003703060

Wood, B. D. (1923). *Measurement in higher education*. New York, NY: World Book Company.