A Historical Analysis of Technological Advances to Educational Testing: A Drive for Efficiency and the Interplay with Validity

Sebastian Moncaleano* and Michael Russell
Boston College, Chestnut Hill, MA 02467, USA; moncalea@bc.edu

Abstract
2017 marked a century since the development and administration of the first large-scale group administered standardized test. Since that time, both the importance of testing and the technology of testing have advanced significantly. This paper traces the technological advances that have led to the large-scale administration of educational tests in a digital format. Through this historical review, a drive to develop and apply new technologies to increase efficiency is revealed. In addition, this review reveals a pattern in which each new advance unveils a new drag on efficiency that becomes the focus of future innovation. The interplay between a drive for efficiency and interest in improved validity is also explored. Upon reaching the recent introduction of technology-enhanced items, the paper suggests that it may be advantageous to relax the drive for efficiency with hopes of realizing gains in validity.

Keywords: Automatic Scoring, Educational Measurement, History

1. Introduction
2017 marked a century since the first large-scale standardized test was developed and administered in the United States. Since then, cognitive tests have been used for a variety of purposes and standardized testing has developed into a well-established corporate enterprise.

Like all modern industries, large-scale standardized testing in the United States evolved considerably over the past century (OAT, 1992). This evolution occurred in at least four ways. First, the sheer volume of standardized testing in US K-12 schools rose from none prior to World War I to what is today a multi-billion-dollar industry that impacts the lives of millions of students each year (Gewertz, 2017; Madaus, Russell & Higgins, 2009; US Department of Education, 2016). Second, the instrumentation of testing has become more empirical. This is best evidenced by the use of increasingly sophisticated item response models developed since George Rasch (1960) first introduced his probabilistic model in 1960. Fourth, the theory supporting the validity of test scores and uses has passed through at least three stages of development (Newton & Shaw, 2014). Perhaps most notable was the shift from a trinitarian view of validity, which distinguished among criterion, content, and construct validity, to a unitary view introduced by Messick (1989) during the 1980s, and the subsequent introduction of an argument-based approach to test validation (Kane, 1992).

Since the turn of the century, concerns about validity and the expanding use of standardized tests to inform decisions about students, teachers, and K-12 educational practices have sparked calls to further advance the technology of testing (Bennett, 1999; Pellegrino, Chudowsky, & Glaser, 2001). In particular, the shift to digital delivery of educational tests has ignited interest...
in developing new approaches to collecting evidence of student learning through embedded assessments (Barak & Dori, 2009; Teplovs, Donoahue, Scardamalia, & Phillip, 2007; Wilson & Sloane, 2000), new items types (Bennet, 1999; Scalise & Gifford, 2006; Sireci & Zenisky, 2006), simulations (Bennett, Persky, Weiss, & Jenkins, 2007), and even gaming (McClarty et al., 2012; Mislevy et al., 2012; Shute, Ventura, Bauer, & Zapata-Rivera, 2009). Because efforts to apply technology to improve assessment are not new, we believe this centurion moment provides an important opportunity to recount these prior efforts and to consider the goals we set for our next era of innovation.

The historical review presented below traces the many technological advances made to testing since the early 1900s. The main theme that emerges in this review focuses on a continued effort to increase efficiency through the application of new methods and technologies. This drive for efficiency occurs in a recurring pattern in which each successful advance in efficiency unveils a new drag on efficiency that then becomes the focus for the next advance in efficiency. This drive for efficiency, however, also reveals an interesting relationship with validity. Initially efforts to increase efficiency also implicitly aimed to increase validity. In the middle third of the century, however, efficiency dominated the drive for technological advancement with little concern for improvements to validity. This was largely due to the need to keep pace with the increasing volume of educational testing that occurred during this period. Once standardized testing became nearly universal across K-12 public schools, and its consequences on students and instruction were of concern, validity returned, this time as an explicit goal for technological advancement.

As identified above, the evolution of testing has impacted at least four aspects of testing: volume, technological, technical, and theoretical. Space does not allow us to examine thoroughly all four of these areas of advancement. Instead, the paper focuses narrowly on the many technological advances that have occurred since standardized testing was introduced to K-12 settings. When examining this evolution, changes to the volume of testing that drove several technological advances are noted and the relationship between various technological advances and validity are discussed.

**2. Technology Defined**

To many readers, the word “technology” conjures images of computers and digital devices. In this paper, we use the term more broadly to reference a specialized set of procedures or methods for accomplishing a task (Ellul, 1964; Lowrance, 1986; Winner, 1977). For testing, digital devices and optical scanners clearly fit this definition. But, paper and pencil, the standardized test form, the multiple-choice item, and bubble sheets also qualify as technology. Statistical and psychometric methods for analyzing data, such as the reliability coefficient or an item response theory model, are also considered technologies. In effect, any advance that introduces a new mechanism, product, or process to testing is a new technology (Madaus et al., 2009). In this paper, we limit our focus to technologies that affect the form and processes employed to administer tests and score responses. We exclude statistical and psychometric methods, although we believe a historical review of the development of these technical methods is also warranted given that new innovations in testing may demand new innovations for our psychometric methods.

**3. The Introduction of Standardized Testing in America**

Since the advent of schooling in Colonial America, testing played an important role in assessing student learning. Until the late 1800s, however, tests typically took the form of oral recitations (Madaus et al., 2009). This was due in part to tradition and to the lack of an inexpensive medium on which students could write. By the mid-1800s, however, mass production of paper lowered its cost which made standardized approaches to written tests feasible.

The first large-scale testing project that capitalized on paper administration was undertaken by Horace Mann in 1845, the then school-master of the Boston Public Schools. Concerned about differences in quality and effectiveness of instruction among Boston’s schools, Mann pioneered the use of a standardized test to monitor school quality (Gallagher, 2003; Russell, 2006). Mann hoped that the use of written exams would “provide objective information about the quality of teaching and learning in urban schools, monitor the quality of instruction, and compare schools and teachers within each school” (Gallagher, 2003, p. 84). Mann’s tests did in fact detect noticeable differences in performance among Boston’s schools and his program of standardized testing inspired the development of similar standardized written exams in several school systems across the United States (Russell, 2006).
As Mann's standardized approach to testing achievement gained traction in the United States, interest in measuring intelligence was sparked by Francis Galton's efforts in Europe. While it is questionable whether Galton's tests measured anything related to our current conception of intelligence, his testing centers and fairground booths increased interest in mental measurement (French & Hale, 1990; Gould, 1981). In turn, this work inspired scholars in the budding field of psychology to explore methods for measuring mental abilities (Pearson, 1914; Wolf, 1973; Zenderland, 1998).

The first breakthrough in mental measurement was made in France by Binet and Simon (1905), who developed a measure of mental ability designed to inform decisions about the placement of children in schools. While the Binet-Simon test is often mis-categorized as a test of intelligence, it was in fact designed to assess the extent to which children possessed knowledge and skills that were on-level with their chronological age. Binet and Simon employed this test to identify students whose mental abilities were well below their age and thus were believed to benefit from specialized schooling (French & Hale, 1990; Wolf, 1973). Like Mann's achievement tests, the Binet-Simon test comprised a standardized set of questions and problems that were presented to all students. Unlike Mann's tests, however, the Binet-Simon test of mental ability was administered individually in an oral manner to students (Binet & Simon, 1905).

As the director of research at the Vineyard Training School for Feeble-Minded Girls and Boys in New Jersey, Henry Goddard was also interested in identifying students whose mental development was below age level (Gregory, 1992; Zenderland, 1998). Upon learning of the Binet-Simon scale of mental age, Goddard brought the test to the United States. While his initial use was limited to students in his school, over a ten-year period interest grew and its use beyond Vineyard increased (Zenderland, 1998). Recognizing that some of the items in the Binet-Simon did not translate well to an American setting, American psychologists modified the test (Gallagher, 2003; Gregory, 1992). Louis Terman's refinement, which became known as the Stanford-Binet Intelligence Scales, was the most notable of these efforts (Terman, 1916).

At the same time, the use of written exams continued to expand in school settings. An analysis conducted in 1914 by Fredrick Kelly, a doctoral student working under Edward Thorndike at Teachers College at Columbia University, revealed two shortcomings of this testing format. First, teachers were spending an increasing amount of their time scoring written tests. Second, there was a concerning level of subjectivity in how teachers marked written tests, both among teachers and within teachers (Kelly, 1915). As a remedy to these problems, Kelly's (1915) doctoral dissertation proposed the idea of standardizing the tests with predetermined answers (Davidson, 2011). After earning his degree, Kelly accepted a position at the University of Kansas and developed the Kansas Silent Reading test, which applied his proposed solution to produce a test composed of multiple-choice items that could be scored easily by scanning the page by eye. The Kansas Silent Reading test marked the first timed multiple-choice test (Kamenetz, 2015).

Within a short order, Kelly's innovation influenced ongoing efforts to improve the Binet-Simon test. The most direct application of Kelly's new item type was made by Arthur Otis. As a student of Terman, Otis explored ways to adapt Terman's Stanford-Binet so that it could be administered more efficiently in a group setting (Davidson, 2011; Gallagher, 2003; Kamenetz, 2015). To do so, Otis converted items to a multiple-choice format and then developed a stencil to score responses (Zenderland, 1998).

Shortly after Terman introduced the Stanford-Binet Scale and Otis began working on a group-administered version of the scale, the United States entered World War I in 1917. At the time, the field of psychology was struggling to establish itself as an accepted science. Recognizing the war as an opportunity to firmly demonstrate the scientific nature of psychology and establish its benefit to the nation, Robert Yerkes proposed to the Army that a team of psychologists be convened to develop a psychological test for the Army to use in order to more efficiently classify recruits and inform their assignment to positions within the military (Carson, 1993; Monahan, 1998).

Despite skepticism among some military leaders, Yerkes proposal was approved and a group of psychologists who had been experimenting with various approaches to mental measurement met at Goddard's Vineyard School to launch the development of what became known as the Army Alpha (Zenderland, 1998). Among the team members was Terman, whose promotion of Otis' use of multiple-choice items for the Stanford-Binet Scale had a major influence on the design and development of Alpha. As evident throughout Yerkes' report (1921) documenting the development and use of the Army Alpha, the tasks selected for inclusion on the test battery were efficient.
both in terms of administration time and scoring (Boake, 2002). While several concerns were raised about bias produced by many of the items and the validity of classifications that resulted from the test’s results (Gould, 1981), the Army Alpha exposed millions of recruits to standardized testing and provided ammunition for claims about the efficiency of this new approach to testing (Reed, 1987).

The Army Alpha represents the end of the first era of innovation in testing in the United States. During this period, three major technological advances occurred. Although the concept of validity was not yet established during this era, a desire to improve validity was implicit in all three of these technological advances. The first advance occurred when Horace Mann introduced the written test. His purpose for doing so was to gather empirical evidence of student learning to support claims about differences in the quality of teaching across schools. Undergirding his adoption of written test were two beliefs. First, Mann believed a written test would provide more accurate information about student achievement than was provided by teacher grades. Second, he believed that administering the same test to all students would yield information about student achievement that was directly comparable across teachers and schools (Gallaher, 2003). The second advance was driven by Kelly’s concern about the accuracy of information provided by tests scored by teachers. More specifically, his concern focused on what the field now terms inter- and intra-rater reliability. He believed multiple-choice items would improve consistency among scorers which, implicitly, would increase the validity of information provided by such tests. Finally, although the validity of the actual decisions made based on Army Alpha scores is questionable (Gould, 1981), Yerkes’ intention was to develop a test that could improve the classification of the massive number of recruits entering the Army. While each of these technological advancements increased efficiency of testing, their development was also driven by implicit desires to increase aspects of validity.


Following the end of World War I, enrollment in public schools rose rapidly and schools began placing students into different instructional tracks based on their perceived aptitudes and abilities (Gallagher, 2003). Terman capitalized on these developments by transforming the Army Alpha test into the National Intelligence Tests for School Children. Terman marketed this new test successfully nationwide as a tool for informing decisions about student’s instructional paths (Minton, 1987, 1998). As Gallagher (2003) describes, the use of this test to classify students had lasting effects: “Scores from intelligence tests dramatically changed the ways in which students were classified. Standardized tests were used to stratify students of different abilities into different curricular paths, thereby restricting their academic and social choices” (p. 88).

Shortly after Terman launched the National Intelligence test, several other tests were introduced to American schools. In 1923, Otis published the Stanford Achievement Test (Kamenetz, 2015). Shortly thereafter the College Entrance Examination Board (CEEB) gave birth to the Scholastic Aptitude Test (SAT) (Gallagher, 2003; Lemann, 1999). During this decade, local state tests also emerged. Although New York had first introduced a written testing program in 1865, in 1927 multiple-choice questions were added to the test to increase the number of curricular objectives covered by the test and to increase efficiency (Office of State Assessment, 1987; Russell, 2006).

In 1929, the Iowa Testing Program was launched. This program was an outgrowth of a statewide competition known as the Iowa Academic Meet which sought to “improve educational measurement and stimulate scholarship in Iowa public secondary schools” (Peterson, 1983, p. 2). Everett Lindquist, an assistant professor at the University of Iowa, was responsible for the construction and administration of the examinations employed by Iowa’s meet and in 1930 was appointed director of the state’s testing program. Over the next five years, this program evolved into the Iowa Test of Basic Skills.

All of these testing programs relied heavily on the multiple-choice item format to provide objective test scores in a relatively efficient manner. However, as the volume of tests administered by these programs increased, the effort involved in scoring all of these tests became a drag on the further expansion of these types of testing programs.

An early innovation to increase the efficiency and accuracy of scoring selected-response items was introduced by Otis in 1918. It asked examiners to separate the pages of an answer booklet and place the transparent
Because students recorded their responses directly on each page of the test booklet, the efficiency of Otis's approach was limited by the need to employ a different transparency for each page. To address this, Lindquist created the dedicated answer sheet which was completely separate from the test booklet. Lindquist's innovation eliminated the need for scorers to manually separate pages from the test booklet and enabled the use of a single transparency to score all questions on the test in a single scan. Lindquist also shifted the focus of scoring from Otis's method of identifying incorrect responses to his method of tallying correct responses. Although subtle, this shift simplified the scoring process and set the stage for automated tallying of correct responses.

Ben Wood, a student of Yerkes, was also concerned about the inefficiencies involved in the scoring of large-scale tests. As the director of Columbia University's Bureau of Collegiate Educational Research, Wood worked with the state of New York to develop and score the Regents exam in 1925. His wife directed the brigade of workers in Albany that hand-scored the tests and provided Wood with a first-hand look at the burdens of scoring a large volume of tests. In 1928, Wood served as a technical adviser for The Pennsylvania Study, a longitudinal study that sought to understand the relationship between “secondary schools and colleges [as well as] the best methods for identifying and training preparatory students” (Downey, 1965, p. 19). This study intended to employ a twelve-hour comprehensive examination to measure the accumulated intellectual capital of the class of 1928 across 49 Pennsylvania colleges. This component of the study, however, came under scrutiny due to the high cost of scoring the tests (Downey, 1965).

In response to this concern, Wood reached out to ten corporations that manufactured computing equipment, hoping to entice them to apply their innovations to automate the scoring process. Thomas J. Watson Sr., president of International Business Machines Corporation (IBM), was the only person to respond. During their first meeting, Watson was impressed with Wood's ideas about potential applications of machines and computation to education and hired Wood as a consultant to IBM. In addition, Watson provided Wood with a large amount of computing equipment that became the infrastructure for the Statistical Bureau at Columbia University, of which Wood was appointed director (Downey, 1965).

Wood imagined a scoring method that worked close to the speed of light (Downey, 1965). To obtain such speed, Wood and a team of IBM engineers experimented with electronic scoring. For one approach “the score was recorded on an ammeter in units of electricity when electric circuits in the machine were closed by a graphite mark on the score sheet” (Downey, 1965, p. 52). While this approach greatly increased the speed of scoring, it produced inconsistent results due to variations in the darkness of pencil marks produced by students to indicate their responses which in turn caused variation in the electrical current transferred by the markings.

While Wood continued to experiment with various approaches, Reynold Johnson, a high school science teacher in Michigan, also developed an interest in automated scoring. Johnson's approach also relied on the transfer of electricity through a graphite-marked answer sheet. To address the inconsistencies caused by variations in pencil marks, Johnson used a high resistor unit which raised the total resistance to a point where the variations caused by pencil mark differences no longer impacted the registration of response marks (Downey, 1965).

Believing that his machine would streamline scoring in schools across the nation, in 1934 Johnson began advertising it in newspapers. Shortly thereafter, an IBM traveling salesperson saw the advertisement and notified Wood, who immediately contacted Johnson and arranged a meeting in New York. After reviewing Johnson's prototype, Wood hired Johnson to work with his team to develop IBM's first commercial test scoring machine- the IBM model 805- which was released in 1935 (Downey, 1965). As IBM describes:

…tests to be scored on the 805 were answered by marking spaces on special “mark sense” cards developed by IBM. Inside the 805 was a contact plate with 750 contacts corresponding to the 750 answer positions on the answer cards. When the cards were fed into the 805 for processing, the machine read the pencil marks by sensing the electrical conductivity of graphite pencil lead through the contacts plates. A scoring key separated the contacts into two groups, the “rights” and the “wrongs.” When the operator manipulated the controls, the 805 indicated the scores.
The speed of the IBM 805 was limited only by the operator's ability to insert the answer cards into the machine and record the scores. An experienced operator could record scores at the rate of about 800 cards per hour. (IBM, n.d.)

To function efficiently and accurately, two additional technologies were required. The first was an answer sheet with standardized locations for answer responses. It is this form that became known as the “bubble sheet.” Second, to increase the accuracy with which marked responses were detected by the machine, special “mark-sense” pencils containing electrographic lead were introduced (Koran, 1942; Wood, 1936).

The development of the first automatic scoring machine had an immediate impact on standardized tests of the time. In 1936 the New York Regents exam was administered using the new bubble sheet and mark-sense pencil and became the first large-scale test to be processed by an automatic scoring machine. That same year, Connecticut also relied on the machine for scoring. In this case, however, the machine not only increased the efficiency of scoring, it also saved the program from looming financial insolvency. Connecticut had contracted four tests from Wood's Cooperative Test Service. Aiming to test approximately 50,000 students, the state soon learned that its funds were short of the cost required to print such a large number of test booklets. To reduce costs, the state opted to print an answer sheet for every student but only produced 5,000 test booklets. The booklets were then shared among schools. Soon after, the Educational Testing Service (ETS) adopted this approach to administer the SAT (Downey, 1965).

The introduction of the first automatic scoring machine (the IBM 805) revolutionized the processes for both scoring item responses and calculating total test scores. In turn, the automatically scored bubble sheet rapidly emerged as a standard tool for increasing the efficiency of testing, and reusable test booklets became the norm. However, the high cost of the machine limited its utility to large-scale programs similar to those administered by Wood's testing agency and failed to penetrate schools in the manner Johnson had originally envisioned. In addition, while the machine increased the speed with which item responses were scored, other aspects of score calculation and analyses continued to pose a drag on efficiency.

Between 1920 and 1940, the rapid expansion of standardized testing in the United States was driven by a desire to improve decision-making based on (what was believed to be) objective information about aptitude and/or achievement. The several technological innovations that occurred during these decades, however, focused narrowly on increasing efficiency. The design of scoring templates as introduced by Otis and Lindquist was motivated by a desire to make the labor of human scorers easier. These innovations might have reduced scoring errors and thus increased accuracy of scores, but their main goal was to make scoring more efficient. Similarly, Wood was driven to develop the IBM 805 to address the burden of hand-scoring massive waves of tests. The IBM 805's efficiency and reliability in scoring greatly reduced the amount of time required for scoring. In these ways, the technological innovations that occurred during this period were designed solely to improve efficiency without any implicit or explicit focus on validity.

5. Rise of the Automatic Scoring Industry (1940-1959)

The United States' entry into World War II in late 1941 rekindled large-scale use of tests by the military to support the classification of recruits. To increase the efficiency of this process, the IBM 805 was put to extensive use and further solidified the automated scoring of cognitive and psychological tests. At this same time, IBM focused efforts on reducing the machine's costs in hopes of increasing its attractiveness to universities, colleges and large school districts. And, to increase its utility, other machines were developed as extensions to the IBM 805 (Da Cruz, 2017).

The first of these extensions was the graphic item counter. Rather than tallying correct responses to produce a total test score, this machine kept count of responses for each individual item and greatly increased the speed with which item statistics were calculated (Russell, 2006). Although the graphic item counter slowed the scoring process, it is estimated “item analysis tabulations made by hand [took] more than eight times as long as those made with the Graphic Item Counter” (McNamara & Weitzman, 1946, pg. 90), thus a net gain in efficiency resulted.

The Document-to-Card punch machine was another advance that also increased efficiency. It replaced the laborious process of manually recording scores by reading the output from the scanning machine and then punching
the resulting score into a punch card that was added to a pile of cards that were eventually fed into a punch card reader to perform calculations across test takers. This machine removed the human from the process by automatically recording scanned test or item results on a punch card (Harman & Harper, 1954).

Although not developed specifically for the testing industry, other machines increased the utility and efficiency of International Scoring Machines during the 1940s. For example, the IBM 513 and 514 machines allowed information stored on one punch card to be automatically replicated on one or more cards. Testing programs used these machines to reproduce answer keys when multiple scoring machines were used to further increase the speed of scoring and to transfer test data for analysis by multiple researchers. Similarly, the IBM 604 Electronic Calculating Punch was used to automate the large computations required by testing programs such as the NY Regents exams (Da Cruz, 2017).

During the 1940s, other test developers and administrators recognized the efficiency of automating scoring and analysis of tests and began developing their own scoring machines. In 1940, Elmer J. Hankes, then president on an engineering consulting firm, received a request from a test-publishing company in Minneapolis (the Educational Test Bureau - ETB) to design a test-scoring machine for their tests. While working on this project, Hankes saw opportunities to increase the speed with which several psychological tests were scored (for example the Strong Vocational Interest Blank). Motivated by his success in developing new algorithms and machines, Hankes formed a test-scoring company, TestScor, which became an important player in the rapidly expanding testing industry (Campbell, 1971; Hankes, 1954; Holeman & Docter, 1972).

In 1942, the Iowa Test Programs (ITP) introduced two new programs; the Iowa Tests for Educational Development (ITED) and the Fall Program. Managing three separate testing programs placed strain on scoring capacity and rekindled Lindquist's interests in electronic scoring. Partnering with Phillip Rulon in 1952, Lindquist developed his own scoring machine which became the foundation for the Measurement Research Center (MRC), an independent non-profit corporation that provided scoring services for all Iowa Testing Programs as well as testing programs administered by other organizations (Carroll, 1969; Peterson, 1983). While it is unknown whether interest in cognitive and psychological testing would have evolved regardless of the introduction of automated scoring, there is little doubt that the development of these machines helped firmly establish scoring as an important sector of the testing industry.

5.1 The Industry Reflects on Impacts of Machines and Devices on Developments in Testing

In 1936, the Educational Testing Service (ETS) began hosting an annual Invitational Conference on Testing Problems. In 1953, a main focus of the conference was the Impact of Machines and Devices on Developments in Testing and Related Fields. This conference provides interesting insight into the industry's views at the time about the development of automatic scoring machines, their resulting impacts on the testing industry, early concerns about the widespread use of these machines, and the direction future developments might take.

Arthur Traxler opened this session with an address that described the development of the International Scoring Machine and praised the improvements in efficiency it brought to test scoring procedures (Traxler, 1954). Although Traxler voiced concerns about the growing reliance on automatic scoring and specifically highlighted limitations multiple-choice items have for evaluating some types of skills, he also identified several opportunities for further improving the scoring process.

Harry Harman and Bertha Harper, from the Department of Army's Adjunct General's Office (AGO), presented three machines they used frequently at the AGO. First, the Factor Matrix Rotator was a machine specifically designed to aid the process of factor analysis. Second, they discussed the uses of the Document-to-Card punch machine described earlier. Finally, they mentioned the Card Scoring Punch, which was a machine used to punch the test score onto the answer sheet. This machine was not intended as a replacement to the automatic scoring machine, but rather as an alternative to score responses already registered in punch cards (Harman & Harper, 1954).

Hankes's address began by describing his success developing the automatic scoring machine for the Strong Vocational Interest Inventory. In addition, he presented two other recent innovations. The Testscor Universal Scorer (TUSAC) was designed to score test batteries and was able to score both sides of the answer sheet as well as “gather scores, counting both rights and wrongs, convert
these scores, weight and combine the conversions and print out the resultant index scores” (Hankes, 1954, p. 158). Hankes’ description of the TUSAC resembled a sales pitch that promoted the TUSAC as a machine that “does it all, better and faster”. Hankes closed his presentation by introducing the Digital Universal Scorer (DUS). Still in development, this machine was intended to be a portable and affordable scoring machine that individual schools could use to scan student answer sheets, calculate standardized or percentile scores, and record and analyze test scores for groups of students (e.g., a classroom).

Not to be outdone by preceding speakers, Lindquist introduced an electronic scoring machine that was in a nascent stage of development. In contrast to the single-purpose machines presented by the previous lecturers, he claimed his machine would be capable of doing everything from reading answer sheets to scoring, analyzing and reporting. Lindquist boasted that his machine would reduce the amount of time required of a staff of 40 unskilled and 20 skilled workers to score, tabulate, and generate reports for a 58-page battery of nine tests from five weeks to just 36 hours. Acknowledging that his machine would not be affordable for small colleges and schools, he proposed establishing scoring centers in strategic locations throughout the United States to which schools would send their answer sheets for processing (Lindquist, 1954).

The development of multiple automatic scoring machines and the increasing reliance on these machines created an important market sector and attracted engineers and scientists who continued to improve these machines. However, these machines essentially replicated the mechanical technology of the IBM 805. Although faster and more reliable than manual labor, these machines were still relatively slow and required considerable human interaction to function. These drags on efficiency were the focus of the next generation of innovations which capitalized on optical technology to decrease scoring time and reduce manual operation of the automatic scoring machines.

The machines developed during this period were motivated by a search for more efficient ways of completing steps involved in the assessment process. Many of these machines aimed to reduce human involvement in phases of scoring, analyses, and test score calculation. Granted, reducing the role humans played in these activities reduced human-error and in turn improved score accuracy and reliability. However, improvements in validity and reliability were not motivating factors for these innovations. In fact, many of these machines were originally developed for other industries and adapted for use in testing to increase efficiency of specific steps in the scoring and reporting process. These technological advancements were driven by a desire to keep pace with the increasing expansion of standardized testing. Of note, though, seeds of concern were also expressed at this conference about the impact that standardization of testing through the multiple-choice item might have on test validity. This concern was most pronounced in Traxler’s comment at the 1953 ETS conference on testing. This view, however, did not gain substantial traction until the 1970s and 80s.


Following his bold announcement at the ETS conference, Lindquist invested heavily in creating his do-all machine. The endeavor was more challenging than anticipated, exceeded his initial budget, and took longer to accomplish. Lindquist planned to score the Iowa Fall Program tests in the fall of 1954, but delays in the machine’s development required hand-scoring the tests (Peterson, 1983). Lindquist shifted his targeted roll-out to the next administration of the ITBS in January of 1955. But, when January arrived, the machine was still not functioning properly. Confident that final fixes were near, Lindquist delayed scoring of the ITBS until March, at which time the complete set of tests were scored within a 10-day period. As it turns out, the nearly three-month delay in launching the machine to process student answer sheets resulted in the score reports being issued at the same time they would have had human scorers been used (Peterson, 1983). Nonetheless, the fact that Lindquist’s machine had reduced the time required for scoring and report production from more than three months to just 10 days provided concrete evidence of the machine’s efficiency (Holeman & Docter, 1972).

Rather than relying on electric conductivity to detect answer marks, Lindquist’s machine employed optical electronics. A beam of light was directed to specific locations on an answer sheet. Sensors then measured the amount of light reflected back from the sheet and recorded it in a memory drum. The main idea behind this technique was that pencil marks reflected less light than the white page around them. The locations where
less light was reflected were then compared to the answer key to score each response.

In addition to scoring answer sheets, Lindquist’s patent also laid claim to several other functions:

The present invention [...] counts the correct answers for each test on the answer sheet in a separate counter for each test, and then, before each succeeding answer sheet is scored, automatically converts the raw scores obtained from the previous sheet into desired converted scores, securing various composites of these scores, cumulating the results for groups of answer sheets, and printing the individual and group results. (Lindquist, U.S. Patent nº 3,050,248, 1955, p. 3)

Awarded to Lindquist in 1962, he subsequently transferred the intellectual property rights to his company, MRC, and soon after the machine became known as the MRC-1501.

As predicted, the optical approach was more efficient than the electrical approach. As a result, the MRC-1501 could “read in an hour a volume of data equal to the output of 1440 key punch machine hours” (Holeman & Docter, 1972). The increased efficiency of the machine made it possible to “test every student in a state for about two or three dollars per student and have the results back in two to four weeks” (Clarke, Madaus, Horn, & Ramos, 2000).

A secondary effect of Lindquist’s innovation was eliminating the need for test-takers to record responses with special mark-sense pencils. Instead, optical recognition enabled the use of regular pencils that left marks that were dark enough to alter the reflection of light. At the time, pencils typically contained one of four different types of lead that were graded from 1 to 4. Number 1 lead was soft and tended to smudge easily, while #4 lead was dense and tended to produce light marks. As it turns out, the #2 lead provided marks of considerable darkness that did not smudge. Hence, the use of #2 pencils for selected-response tests became the norm (Veronese, 2012).

Lindquist was the first to employ an optical scoring machine, but he was not the only one exploring this approach. In fact, in the early 1930s, Richard Warren, an IBM engineer, was awarded two patents for optical recognition apparatus (Warren, U.S. Patent nº 2150256, 1939). During the 1950s IBM rekindled Warren’s work and developed several optical scanning machines that were employed outside the field of testing (Harold, U.S. Patent nº 2944734, 1960). These applications of optical recognition were eventually applied to improve the International Scoring Machine, which culminated in a 1960 patent for IBM’s first optical mark-sense automatic scoring machine, the IBM 1230 Optical Mark Scoring Reader (Harold, U.S. Patent nº 2944734, 1960). The machine’s name was soon generalized as Optical Mark Recognition (OMR) (Finger JR., 1966).

Some years after the boom of optical scoring machines, a significant development in automatic scoring emerged when William Sanders founded Scantron Corporation in 1972. Sanders believed the scoring technology used by large-scale testing programs would also benefit testing in local school settings. His vision resembled Hankes’ attempts to bring the Digital Universal Scorer to the school market. Scantron applied OMR technology to produce its own scoring machines which it then sold inexpensively to schools (Poole & Sokolsky, Patent nº 3800439, 1974). Scantron capitalized on the widespread purchase of their machines by then selling the blank answer sheets that were required by the machine for scoring. This business model greatly expanded use of automatic scoring machines such that the term “Scantron” became an eponym for “automatic scoring machine” (Scantron Co., n.d.).

The OMR machine followed the motivations that decades earlier led to the development of the automatic scoring machine, namely, that scoring could be done faster. Additional effects of this innovation were the further standardization of bubble sheets and the No. 2 pencil. Furthermore, this innovation brought the benefits of automatic scoring to schools and teachers. This allowed schools to score reliably and efficiently locally developed tests. While these outcomes likely improved the validity of the inferences teachers made based on test scores, the motivating factor for the development of Optical Mark Recognition technology was further improving the speed of scoring selected response items. Without question, this technology reduced the amount of labor required to score selected-responses. However, two important drags on efficiency remained: namely the distribution and administration of tests and scoring of open-response items. It is these drags that were the focus of the next innovations in testing.

The increasing use of mainframe computers during the 1960s and the introduction of personal computers in the mid-1970s had three major effects on the technology of testing. First, computers provided a platform for test delivery. Second, they enabled recent technical advances in psychometrics to be applied during test delivery to tailor the items presented to each test-taker. And third, computers created opportunities to score written responses more efficiently and objectively. In this section, we treat computer-based and computer-adaptive testing together, and then explore automated essay scoring.

7.1 The Rise of Computer-Based and Computer-Adaptive Testing

The origins of computer-administered tests can be traced to 1964 when A. G. Bayroff published a study on the feasibility “of applying automation and computer techniques to psychometric testing” (p. 1). Bayroff described a machine that combined a 35-mm slide projector with algorithms that could tailor the test questions administered to a test-taker using various adaptive approaches (Weiss & Betz, 1973). Bayroff envisioned that “successful application of machine testing techniques could open up new avenues to testing research and also holds the possibility of facilitating testing operations” (Bayroff, 1964, p. 1).

Although Bayroff’s machine was never built, other innovators did develop such machines. As an example, Edwards (1967) developed the Totally Automated Psychological Assessment Console that administered and scored several different kinds of psychological tests. Similarly, Veldman (1967) developed a computer-based sentence completion method that used a typewriter terminal for stimulus presentation and response recording (Elwood and Griffin, 1972). In addition, Elwood developed a system that specialized in administering the Wechsler Adult Intelligence Scale (Elwood, 1969).

During the 1970s, Fredric Lord’s latent-trait theory—later known as Item Response Theory (IRT)—inspired several studies that explored its application to tailoring tests (Betz & Weiss, 1973; Lord, 1970; Sachar & Fletcher, 1978). While the first empirical studies on adaptive testing were performed with paper-and-pencil tests, the benefits that computer-based administration brought to adaptive testing were obvious to the small but growing field of psychometricians and test developers interested in adaptive testing. The expanding availability of shared computing facilities further increased the feasibility of adaptive testing (Weiss & Betz, 1973). Continuing to rely heavily on tests to inform personnel decisions and assignments, the Army and the Personnel and Training Research Programs of the Office of Naval Research sponsored several of these initial studies that focused on adaptive testing.

In response to growing interest in adaptive testing, in 1975, the Office of Naval Research organized the first Computerized Adaptive Testing Conference. This and the three follow-up conferences (1977, 1979 and 1982) hosted by the University of Minnesota provide rich snapshots of developments in adaptive testing. Topics discussed during these conferences included theoretical approaches to adaptive testing, empirical studies on the use of adaptive tests, discussions on their main advantages and some innovations on test formats that came with these new types of tests (Clark, 1976; Weiss, 1978, 1980, 1985).

During the 1980s, Computer-Based Testing (CBT) and computer adaptive testing (CAT) slowly gained traction in venues beyond Army personnel selection and evaluation. During this period computers also evolved “from being part of the technical hardware that supported data analysis and test-score reporting to being an integral part of the test development and administration process” (Clarke et al., 2000, p. 167).

The penetration of computer-based testing into the educational testing market, however, was curtailed by a backlash against standardized multiple-choice testing that emerged in the late 1980s (Madaus & O’Dwyer, 1999). In place of traditional multiple-choice items, educational

---

1 During the early seventies, terminology wasn’t unified; “tailoring” was the term used by Frederic Lord and his colleagues at ETS, while “adapting” was introduced by David Weiss and used between the University of Minnesota affiliates. Unification of the term to “adapting” may have happened through the titling of the conferences hosted between 1975 and 1982.

2 Computer Assisted Testing and Computer Administered Testing became unified under Computer Based Testing (CBT) to allow CAT to be an exclusive reference to the Adapting feature of the test.
reformers promoted alternative methods of assessment such as portfolios and performance-based tasks (Clarke et al., 2000). States like Vermont, Kentucky, Maryland and Washington pursued these ideas with fervor and implemented large-scale programs that employed alternative approaches to assess students (Stecher, 2010). These approaches were not conducive to computer-based or adaptive test administration, and for a period of time, efficiency was sacrificed for what was believed to be an improvement to validity. Specifically, validity was believed to be improved in at least two ways. First, students were asked to apply knowledge and skills to produce extended and sometimes complex products rather than select from among answer options. Second, the creation of products often required students to combine multiple aspects of knowledge and skills rather than focus on discrete knowledge and skills in isolation. These two factors led some to argue that performance assessments provided a more authentic assessment of student achievement (Darling-Hammond, Ancess, & Falk, 1995). These efforts, however, were short-lived, due in part of studies that revealed inconsistency in scoring and high-costs of administration (Dunbar, Koretz, & Hoover, 1991; Strong & Sexton, 2000).

The launch of the World Wide Web (WWW) in 1989 was a significant breakthrough that impacted all computer-related activity. For the testing industry, the WWW allowed individuals to register on-line for a test administration. In addition, “testing organizations could now electronically exchange questions and examinee responses with test centers and send scores to institutions in a similar fashion” (Clarke et al., 2000, p. 168). In 1992, improvements in the efficiency of communication and transfer of data allowed ETS to transition the delivery of the Graduate Record Exam (GRE) General Test from a Paper-Based Test (PBT) to a computer-based test (CBT) (Briel & Michel, 2014; Clarke et al., 2000). In 1993, CAT technology was added to the computer delivered version of the GRE General Test (Briel & Michel, 2014). And in subsequent years, ETS migrated several of its tests to CBT, including versions of the Graduate Management Admission Test (GMAT) and the SAT I (Bennett, 1999). By the end of the decade computer-based delivery became the norm rather than the exception for many of these tests.

With the passage of the No Child Left Behind Act in 2001, large-scale multiple-choice testing regained its dominance as the primary tool for summative K-12 student assessment. Although adoption of computer-based administration by state testing programs occurred gradually during the first decade of the 21st century, the 2010 Race to the Top Assessment Program provided more than $240 million to develop next-generation assessment programs. In turn, this investment spurred rapid and widespread use of computers for test administration in K-12 schools and led approximately 20 states to fully embrace adaptive testing. Thus, by 2014 computer-based testing was the norm for the efficient assessment of students nationwide.

7.2 Automated Essay Scoring

The wide-scale adoption of optical scanning dramatically increased the efficiency of scoring selected-response items. Yet, the scoring of essays and other open-responses persisted as a drag on efficiency and was prone to human error. Computer-based technologies offered a solution to both of these challenges.

As early as 1966, Ellis Page was inspired by developments in computational linguistics and conceived the idea of using computers to score essays (Ben-Simon & Bennett, 2007; Page, 1966). In 1967, Page developed Project Essay Grader (PEG), the first essay grading computer program. As Rudner and Gagne (2001) later summarized Page’s initial approach, “The underlying theory is that there are intrinsic qualities to a person's writing style called trins that need to be measured, analogous to true scores in measurement theory” (p. 2). Twenty-five years later a revised version of the program was released that capitalized on advancements in natural language processing tools such as grammar checks (Ben-Simon & Bennett, 2007).

The re-release of PEG ignited academic interest in automatic essay scoring (AES), and by 1997 three new AES programs were introduced to the market: ItelliMetric developed by Vantage Technologies, the Intelligent Essay Assessor (IEA) developed by the University of Colorado, and e-rater created by ETS. Although adoption of automated scoring by large-scale testing programs has been slow, the 2010 Race to the Top Assessment program stimulated interest in its use for K-12 summative testing. And in 2012, the Hewlett Foundation sponsored a competition that compared the accuracy and utility of several different essay scoring algorithms (Automated Student Assessment Prize - ASAP).
While use of automated essay scoring remains limited, its existence represents the latest technological advance designed to increase the efficiency of testing. As multiple-choice items have achieved a paragon of efficiency, Automatic Essay Scoring is an attempt to achieve the same for open-ended tasks.

7.3 The Interplay among Computer-based Testing, Efficiency and Validity

The relationship between computer-based testing and both efficiency and validity is a complex one. On the one hand, applications of computers to testing that occurred since 1970 were driven by a desire to increase efficiency of test administration and scoring of open-response items. Adaptive testing aimed to increase efficiency by decreasing test length. Web-delivered tests increased efficiency by eliminating the need to print, ship, sort, and scan tests. And automated essay scoring attempts to increase efficiency by reducing the role of humans in the scoring process. In each case, the initial drive for computer-based administration was driven by desired gains in efficiency.

However, efficiency was not the only driver for adaptive testing and automated essay scoring. For computer adaptive testing, psychometric theory indicated that greater measurement precision could result from selectively tailoring the items administered to each individual. For automated scoring, computer-based scoring held the potential to decrease scorer error that resulted from such factors as fatigue and rater drift that resulted from shifts in the interpretation or application of scoring criteria. Both of these innovations aimed to improve validity by increasing the precision and accuracy of test scores.

Moreover, while the initial driver for adoption of on-line test delivery was efficiency, concerns about validity were promptly raised as testing programs transitioned to computer-based test delivery. These validity concerns fall into two camps. The first camp worried about adverse impacts that may result from the transition to computer-based delivery. Specifically, these concerns about adverse impacts focused on score comparability and potential negative impacts on students with less access to computers (Bennett, 2003; Poggio, Glasnapp, Yang & Poggio, 2005; Russell, Goldberg & O'Connor, 2011). The second camp, however, advocated for the transition to computers and argued that adoption of computer-based testing could positively affect validity. This camp focused on two aspects of computer-based testing. First, they noted that a mismatch had developed between the ways in which students use computers when applying skills and knowledge outside of a testing situation and how they are required to do so on paper-based tests. This argument is most pronounced for writing (Higgins, Russell & Hoffmann, 2005; Horkay, Bennett, Allen, Kaplan, & Yan, 2006; Russell & Haney, 1997; Russell & Plati, 2000) for which negative mode effects have been documented when students who are accustomed to writing on computer must compose test responses on paper. Second, this second camp emphasized the opportunity computer-based administration provides for applying principles of universal design to test delivery systems in order to decrease construct-irrelevant factors that contribute to test performance for students with disabilities and special needs (Russell, Hoffmann, & Higgins, 2009). While the arguments presented by both camps have important bearing on the impact of computer-based testing on validity, they provide clear evidence that the adoption of computer-based testing has not been driven purely by efficiency absent concerns of validity.

In this way, it is fair to conclude that the integration of computer-based technologies with the technology of testing that has occurred since the 1970s was driven both by desires for efficiency envisioned by Wood and Lindquist, and concerns about validity implicit in the early innovations of Otis and Terman.

8. Summary-Validity, Efficiency and Technological Innovation

At first blush, the history of educational testing in the United States can be summed up as a drive for efficiency. Since Mann's introduction of the standardized written test, a long list of technological innovations has helped increase the efficiency of testing. The multiple-choice item, scoring templates, separate answer sheet, scanners, computer-based delivery, computer adaptive delivery, and automated essay scoring have each improved the speed with which response information can be collected, scored, and reported. But a simple focus on increased efficiency captures only one part of the story.

Viewed from a wider perspective, the history of educational testing in the United States reveals an interesting pattern in which each advance sets the table for the future advances. The standardized written test...
allowed the same response information to be collected from a large sample of students in a standardized manner, but it required considerable time to score and scoring was subjective. The multiple-choice item increased the speed with which responses could be scored objectively. Initially, multiple-choice answer options were recorded directly in the items presented on each page of the test booklet. This design made locating answer options and determining correctness cumbersome. The scoring template increased the speed with which scoring in-line responses occurred. The scoring template, however, still required considerable human labor. To reduce this labor, the scoring machine was invented.

The scoring machine greatly increased efficiency by removing humans from the actual scoring process. But, humans were still needed to separate the pages of test booklets and check that student’s responses were properly marked before they were processed by machines. The creation of the single and separate answer sheet addressed part of this inefficiency. Subsequent advances to the scoring machine further increased efficiency and added functionality that further reduced human actions required to replicate punch cards, calculate test scores, and conduct reliability analyses. These advances in response processing greatly increased the speed with which test scores could be calculated once test booklets were received by central processing centers. However, the distribution and return of paper test booklets remained a drag on efficiency. Computer-based testing provided a solution that allowed tests to be distributed and administered via the Internet. Testing time, however, remained a concern and computer-adaptive testing provided a solution. Once test distribution and automated processing of multiple choice items were computerized, open-response items were identified as a remaining drag on efficiency. Although not yet widely adopted, automated essay scoring algorithms and processes were developed to address this inefficiency. In this way, each advance in the efficiency of testing revealed additional drags that in turn became the focus of the next advances in the technology of testing.

8.1 Validity and Efficiency

While there is plenty of evidence that efficiency was an important driver for the many applications of technology to testing over the past century, efficiency was not the only driver. In fact, as conveyed in the sections above, the history of educational testing in the United States reveals an interesting interplay between validity and efficiency. Text books focused on testing, assessment, and cognitive measurement emphasize that the most important characteristic of a test or testing program is validity (Crocker & Algina, 1986; Haladyna, 2012; Russell & Airasian, 2012). For many of the early technological advances that occurred for educational testing in the United States, validity was a concern. Mann introduced the standardized written test in hopes of obtaining a standardized indicator of student achievement to support comparison of achievement across schools. Otis introduced the multiple-choice item and the subsequent scoring template to increase objectivity and accuracy of scoring. Terman and his team of psychologists developed the Army Alpha to improve decision-making about job assignments in the military. In each of these cases, the technological advance was driven by a desire to increase what the field now terms validity. Yet, as noted above, each of these advances also increased efficiency. Thus, this early period of technological advancement can be summarized as having a dual focus on validity and efficiency.

As the uses of selected response tests penetrated the U.S. educational system and the numbers of students performing tests rapidly increased between 1920 and 1950, the aim of additional technological advances focused solely on efficiency, without regard for validity. Machines were developed and then enhanced to provide improvements in the efficiency of scoring responses. To these machines were added tabulators and processors that further increased the speed with which total test scores, reliability estimates, and classifications were made. Over time, use of optical methods further increased the speed with which these scoring machines functioned, and the development of smaller and less expensive machines brought automated scoring from large processing centers directly into schools. In each of these cases, validity was of little concern to the inventors. Rather, this period of continual innovation was spawned by a desire to increase efficiency in order to keep pace with the rapid growth in the volume of educational testing. As a result of these many improvements to efficiency, by the 1970s and 80s, the number 2 pencil, bubble sheet, and Scantron became synonyms with testing.

The rapid adoption of personal computers in the 1980s and early 90s created new opportunities for advancements in testing. Here the story is more complicated. Initial interest in computer-based administration was driven solely by efficiency. Computers and the world-wide web
were viewed as tools for increasing the speed with which tests were distributed to test-takers, item responses were collected and scored, and results returned to end-users. But debates about transitioning educational testing programs to computer during the 2000s were based, in part, on concerns about validity.

For computer adaptive testing and automated essay scoring, efficiency and validity were both motivators. Some advocates of adaptive testing saw it as an approach that could decrease measurement error for test-takers at the extreme ends of a test’s scale (Duncan, 1976). But others have cited decreased testing time as the reason to adopt adaptive testing (Rudner, 1998; Weiss & Betz, 1973). Similarly, efforts to automate scoring of written responses have cited improvements in the consistency of scores awarded. But the primary argument supporting adoption of automated open-response scoring is founded on gains in efficiency (Gierl, Latif, Lai, Boulais, & De Champlain, 2014; Shermis & Burstein, 2003).

The interplay between efficiency gains and validity is not new. During ETS’s 1953 conference on issues in testing, several presentations on the benefits of scoring technologies were made. In one of these, Traxler warned of the negative effects that automated scoring and the resulting reliance on selected response items was having on validity. He also mocked testing programs that aimed to test writing and spelling with multiple-choice items. Similarly, in the 1960s, both Adams (1961) and Hoffmann (1962) heavily criticized multiple-choice test items by raising concerns about the validity of the resulting scores.

Similarly, the 1980s saw a backlash against standardized multiple-choice tests in the United States (Stecher, 2010). The primary argument focused on the weak validity of inferences about student’s higher order thinking skills that resulted from multiple-choice items. This backlash spurred development of performance-based tests (Hamilton & Koretz, 2002; Madaus & O’Dwyer, 1999; Stecher, 2010), exhibitions (Darling-Hammond, Ancess, & Falk, 1995; Sizer, 1992), and portfolio-based approaches for both local and large-scale assessment of student achievement (Stecher, 2010). Performance-based methods of assessment, however, proved to be costly to develop, required considerably more time to administer to students, took human-power to score, and slowed the return of results to schools. In addition, concerns emerged about validity that resulted from both a decrease in the amount of skills and knowledge that could be sampled by performance tasks and decreased reliability of the resulting scores compared to multiple-choice tests (Hamilton & Koretz, 2002; Stecher, 2010). The hit on efficiency that resulted from these alternative methods and their potentially negative impact on some aspects of validity hindered wide-spread adoption of these methods by large-scale testing programs. Instead, selected response items that could be efficiently administered and scored regained dominance as the preferred method of testing.

9. Looking to the Future: A Call to Relax Efficiency to Enhance Validity

The 2010 U.S. Race to the Top Assessment program provided a unique opportunity to spur innovation in testing. Although space does not allow a detailed analysis of the program, at least four outcomes of the Race to the Top Assessment program are evident today. First, the program resulted in two consortia of states, each of which developed their own testing programs, which are commonly referred to as Smarter Balanced and PARCC. Second, the program had a dramatic effect on the adoption of computer-based testing such that the number of states employing computer-based testing nearly tripled between 2010 and 2015. Third, the use of adaptive algorithms by Smarter Balanced led to a large increase in the use of this methodology by state testing programs. Fourth, embrace of computer-based testing spurred development of technology-enhanced items. While there is not consensus on the definition of the term technology-enhanced item, a common characteristic across definitions is that a TEI is something other than a selected-response item. That is, a TEI capitalizes on the digital environment to collect evidence of student achievement by requiring students to manipulate content or produce a product that is something other than a selected response.

The test development proposals for both Smarter Balanced and PARCC specified that 20-25% of items on their tests were to be technology-enhanced (Florida Department of Education, 2010; Washington State, 2010). This requirement led to the development and subsequent use of a significant number of such items.

On the surface, this embrace of technology-enhanced items suggests that this most recent innovation was driven more by validity than efficiency. A recent analysis of publicly released technology-enhanced items, however,
suggests that a substantial proportion of these items are in fact new forms of selected-response items (Russell & Moncaleano, 2017). Moreover, the utility these items have for more authentically assessing the targeted construct is questionable. In a review of just over 300 released technology-enhanced items, 40% were found to have no improvement on utility over a selected response item and 20% had only moderate improvement. While the remaining 40% did yield a meaningful improvement in utility, one must wonder why so many of these new item types had such little impact on validity?

Our hypothesis is that concerns about efficiency created a drag on innovation. While not documented in the technical reports and other documents released by these programs, personal observations of the many discussions that occurred during planning, development, technical, and contract negotiation meetings that occurred over a four-year period, coupled with the resulting products, reveals a tension between innovation and efficiency. While there was a strong desire to produce next-generation assessment programs that capitalized on the flexibility of digital environments, there was also a desire for such items to be developed efficiently and affordably. Furthermore, the responses students produce for these items must be able to be scored in an automated manner. Further, because these items required modifications to existing item authoring and test delivery systems, the organizations responsible for the development and maintenance of these systems rightly raised concerns about the time, effort, and associated costs required to support these new item types. While one cannot ignore costs, and testing programs naturally operate under time constraints, these concerns about efficiency created a drag on innovation. As a result, although the Race to the Top Assessment program spread several innovations throughout the field of testing, the scope and depth of these innovations were limited by concerns about efficiency.

As the field continues its advance, this centurion moment in the history of large-scale testing provides an opportunity to reflect on the many technological innovations that have occurred. As we describe above, technological innovations in testing have been driven consistently by efficiency whereas concerns about validity have been more sporadic. And during two distinct periods when the field actively explored innovations aimed at increasing validity, specifically the performance-based assessment movement of the 1990s and the Race to the Top Assessment program, concerns about efficiency curtailed these efforts. Without question, efficiency is a desirable attribute of a testing program- it affects costs and the speed with which tests are developed, distributed, administered, scored, and reported on. However, as the widespread adoption of digitally-based tests continues to create new opportunities for innovation, we encourage the field to relax its quest for efficiency in order to spur potentially powerful advances that can bring meaningful improvements to validity.

Given the opportunities many in the field believe today’s widespread adoption of digital administration has for advancing testing, the theme and patterns identified in this historical review provide a warning to today’s innovators to be wary of efficiency’s lure as they strive to develop new approaches to testing. Initial solutions may in fact decrease efficiency. As an example outside of the field of testing, initial efforts to modernize shipping through the adoption of steam power simply placed a steam engine on the back of a sail boat, which increased load, decreased cargo space, and had no net effect on the speed of a voyage. But over time improvements to steam power greatly increased the efficiency of the shipping industry (Russell, 2006; Steamship, 2001). Similarly, through persistent ingenuity and refinement, the testing industry will surely develop new methods that enhance the efficiency of innovations that improve validity. Such improvements, however, will likely not come if the last century’s focus on efficiency dominates the initial evaluation of an innovation’s merit. As the history of technology and testing reveals, any drags on efficiency revealed by such innovations will inevitably become the focus of future innovations. If successful, these future innovations will recover the initial loss in efficiency produced by truly next-generation innovations in testing.

10. References


