Validity isn’t just for Psychometricians: Legal Considerations for Assessment and Accountability

Cynthia Roach, Ed.S.
Michelle McKeown, J.D., Ph.D.
I. Introduction

Standardized assessments are used to determine everything from whether or not a student will graduate high school or get into college; will receive licensure in medicine, law, education, or a host of other areas; as well as whether or not a teacher, school, or district is deemed successful in educating its students. What was initially used only to help determine intelligence and placement of candidates for officers’ training has grown into national and state accountability systems that now affect students, teachers, schools, and districts. Unfortunately, there has been strongly divided discourse between viewing assessments as either: 1) sophisticated tools used to identify student abilities and measure teacher and school accountability; versus 2) simple tools that reward rote memorization, unfairly holds teachers and schools accountable without considering external factors, and cause teachers to “teach to the test.” How we move forward and attempt to correct this disconnect will be imperative to the future of standardized assessment use, and ultimately, student success. This paper seeks to offer pragmatic considerations as legal and policy matters for assessment and accountability decisions, in terms of design and use.

Most states and the federal government have passed legislation that involve some level of accountability, whether it be teacher evaluation or school and school corporation accountability. School accountability systems and teacher evaluation systems all use state assessment data as one of the various measures of the system. These systems stem from a policy belief that the government and school leaders have an obligation to collect and track objective data regarding student performance, and that state assessment systems are capable of providing information about whether students are mastering essential standards-based information in school. Indeed,
one of the most divisive and hotly debated areas where the federal government has passed education-related legislation is The No Child Left Behind Act of 2001 (NCLB).

With NCLB, one part of the controversy questions the policy and methodology behind the NCLB’s emphasis on testing\(^1\), and the second part asks whether the federal government ought to be passing such legislation, irrespective of whether the testing-focused policy is agreeable. An example of such criticism is the Education Week article authored by Nel Noddings, in which she stated “The law employs a view of motivation that many of us in education find objectionable. As educators, we would not use threats, punishments, and pernicious comparisons to ‘motivate’ our students. But that is how the No Child Left Behind law treats the school establishment.” (Hess & Petrilli, 2006). Yet in contrast, the same work describes, “Striking a similar note, Ross Wiener, policy director of the liberal Education Trust, argued in summer 2005 that, “NCLB’s accountability provisions are sparking progress. Many states, including Virginia, are narrowing previously stubborn gaps and boosting overall achievement. While the law certainly isn’t perfect, these early results are too encouraging to allow the clock to be turned back on NCLB’s accountability provisions.”

Despite controversy related to methods, NCLB testing exists in all the states, and state-level requirements exist in many. And while arguments have been made under federalism grounds that the grant money violates law, this presentation and paper will focus on the issues related to validity and opportunity to learn that are also legal issues related to NCLB testing. See e.g. Michael D. Barolsky, High Schools are Not Highways: How Dole Frees States from the Unconstitutional coercion of No Child Left Behind, 76 Geo. Wash. L. Rev. 725, 728

\(^1\) See Michael D. Barolsky, High Schools are Not Highways: How Dole Frees States from the Unconstitutional coercion of No Child Left Behind, 76 Geo. Wash. L. Rev. 725, 728
II. History of Standardized Testing

Standardized assessments are defined as such because they are to be administered and scored under the same conditions and criteria and follow the same scoring rubrics. The standardization of these components is imperative to the results being deemed valid, reliable, and in the long run, comparable. While many other factors go into making sure the results of any one test are valid and reliable, the most foundational of those is ensuring that no one student, or group of students, receives an inappropriate advantage, or disadvantage simply because the test was not administered or scored properly. Standardized tests, therefore, were desired for their objectivity. One of the first standardized printed assessments was administered as early as 1845 in Boston and required short answers. Unfortunately, the results on these assessments were so poor, and received such criticism, that within five years they reverted back to “non-standardized exams that were mostly based on oral presentation” (Hoff, 1999). However, by 1915 the multiple-choice test had replaced the short answer format and though these first multiple-choice tests were hand scored (Walsh, 1999), by the late 1930s an electronic scoring machine had been invented that used an optic reader to pick up the graphite content of the pencil used to answer the test (Coles, 1999).

Thus, the first standardized assessments were developed to “measure how well students performed against a prescribed curriculum (Hoff, 1999), however, the push for using standardized testing really occurred during World War I when the American Psychological Association agreed to develop intelligence tests. These tests which were initially only to be used
by the military as a means to “identify candidates for officers’ training school”, but quickly expanded to include other areas “such as the screening of recruits who would have difficulty with the mental demands of military service” (Urban and Wagoner, 2009). The Alpha (used to assess literate recruits) and the Beta (used to assess illiterate) intelligence tests were based on the work by Alfred Binet in 1912. Because of the success of these military tests, once the war was over, testing supporters started looking for other ways in which testing, specifically intelligence testing, could be applied. The obvious next step was their use in education. “School systems soon began developing elaborate bureaus of educational research whose major function was to purchase and administer the standardized tests that were believed to measure the educational potential and achievement of students” (Urban and Wagoner, 2009). At this time, Lewis Terman, who had worked with Edward Thorndike and Robert Yerkes on developing the military’s Alpha-Beta tests, further expanded on Alfred Binet’s work, and created the Intelligence Quotient, or IQ test. However, this time, in what Hoff refers to as “a crucial development”, Terman’s newest version “could be administered to students using pencil and paper, making it much easier and cheaper for schools to administer than earlier versions requiring the services of a specially trained psychologist” (1999).

Nevertheless, Terman did not stop with the IQ test, the next year he went on to create the Stanford Achievement Test which was designed to measure student achievement in different subjects across grades 2-8. During the same time period, in 1929, the University of Iowa had created a series of tests to use as part of a scholarship program. These tests measured content mastery in the areas of “grammar, English, American literature, world history, American History, algebra, geometry, science, physics, typing, and stenography”. These tests contained multiple choice, true-false, fill in the blank, and matching items and were specifically designed
so that they could be scored quickly enough for the student competing to know his or her score before moving on to the next round. However, by 1935 the scholarship program had been replaced with a battery of tests, similar to the Stanford Achievement Tests (Hoff, 1999). While the Stanford Achievement Test and the Iowa Test of Basic Skills are still two of the most widely known standardized assessments available, during the 1930s there were also several competitors emerging, including the California Testing Bureau.

**History of the California Testing Bureau**

At the same time that schools and districts were developing their own bureaus of educational research, Willis W. Clark was working as an assistant director of research for the Los Angeles public schools. During this time he had created the Los Angeles Diagnostic Test in the Fundamentals of Arithmetic, and in 1926, his wife, Ethel M. Clark, who, according to Walsh, was more interested in money making schemes than in testing methodology, asked the district superintendent if she could have the rights to the test. The permission was granted. After working out a royalty agreement with the district, and following some market research, Mrs. Clark proceeded to mail out penny postcards “to 25 large districts, promoting the availability of the test” (1999). It took almost a year before she got her first order, but the Kansas City, MO, district wanted 20,000 copies of the test, and thus the California Testing Bureau, or CTB, was born.

Though his wife had been successful in marketing one of his tests, Willis Clark continued not only writing tests, but also working for the Los Angeles schools. Then, in 1933 he, along with Ernest W. Tiegs of the University of Southern California, developed the Progressive Achievement Tests, a battery of tests similar to both the Stanford Achievement and the Iowa Test of Basic Skills. CTB continued to grow and by the late 1950s had outgrown its Hollywood
Boulevard offices. This was also around the same time that CTB held its annual sales retreat at the Clark summer home, the Del Monte Lodge, on the Monterey peninsula. Because of the beauty of the area, it was decided to move the company to Monterey. Unfortunately the move did not go as planned. Of the 60-70 employees who were expected to move, only about 30 did. This staff shortage took its toll on the business and “CTB struggled through scoring season that fall.” In addition to the difficulty of the move to Monterey, other issues put CTB’s future on even shakier ground. A line of credit needed to update existing tests was withdrawn after the bank who initially accepted the loan changed hands. This made finances tight and loyalties start to wane. In addition, Willis Clark’s health began to deteriorate and in 1964 he died of a cerebral hemorrhage at the age of 69. In 1965 Ethel Clark agreed to sell her 81 percent interest in CTB for $200 a share to McGraw-Hill, “a newcomer to educational publishing in the 1960s” (Walsh, 1999).

The timing of the purchase of CTB was ideal for McGraw-Hill. In 1965, Congress had just enacted the Elementary and Secondary Education Act (ESEA) and according to Urban and Wagoner, “it was by far the most costly and comprehensive federal educational law that had ever been passed” (2009). The purpose of the new ESEA was to provide assistance to schools with high populations of poor children. This act provided funding to schools through Title I programs. Once a school was deemed qualified to receive a Title I grant, “school districts had to show results” (Hoff, 1999) and thus, the government created the Title I Evaluation and Reporting System, or TIERS. In other words, program evaluation was mandated, and a required part of that evaluation translated into an increased demand for norm-referenced tests, which were believed to be, valid, reliable, standardized assessments that would allow student performance to be compared to other students. According to Hoff, this also led to a push for more test
development, and thus the Comprehensive Test of Basic Skills (CTBS) was created. This was the first new test in many years for CTB, and the first of many for the new CTB/McGraw-Hill. Over the years as the standards movement continued to grow, and both federal and state accountability programs grew to require even more testing, CTB has risen to meet these needs. What started as a 25 cent brain-storm investment of the wife of a test developer has now grown into a multi-million dollar industry that has been actively involved with the creation of many of the different state standards aligned custom assessments, and who is also part of the consortiums working towards the new national Common Core assessments.

However, this rise has not been without critics. For example, since McGraw-Hill was, and still is, known as a textbook publisher, some scrutiny regarding the perceived and, potential, inappropriateness of a textbook publisher promoting tests as well as test prep materials along with text books that purportedly align to the tests it also produces has raised questions over the years. As early as 1925, when the first Stanford Achievement Tests were being marketed by the then, World Book text book publisher, “it appears that educators paused to consider whether such an arrangement was appropriate” (Walsh, 1999). However, within a year of the first publication, the brand-new American Educational Research Association, AERA, gave their approval and satisfaction with the product. On the other hand, concerns remain. CTB/McGraw-Hill has had to deal with these same questions regarding a test-practice kit sold by McGraw-Hill in the late 1980s, “which contained the same or similar questions as those that appear on CTB’s California Achievement Test” (Walsh, 1999) and they continue to develop and promote test prep materials along with interim diagnostic and predictive assessments that purportedly align with the state tests they also produce.

III. History of Accountability
A. The Beginning

Webster’s Dictionary (1983) defines accountable as the “subject to giving an account” or “capable of being accounted for” and includes the synonyms “answerable”, “explainable”, and “responsible”. According to Hoff (1999), “In 1965, the US Office of Education contracted with sociologist James S. Coleman to study whether American schools offered equal opportunity to white and black students” (para. 78). The Equality and Educational Opportunity Study was groundbreaking and early results showed not only disparity between white and black students, but that “student’s family backgrounds and the socioeconomic makeup of their schools were more meaningful factors in student achievement than the quality of their schools” (Hoff, 1999, para. 79). These early findings are also what prompted Congress to pass what Parker (2005) considers “one of the most significant and expansive education policy initiatives ever undertaken by the federal government” (p. xii) in order to provide additional financial and educational support to “educationally vulnerable children” (Brady & Thomas, 2005, p. 51). The first Elementary and Secondary Education Act (ESEA) passed in 1965 was part of President Lyndon Johnson’s War on Poverty (Parker, 2005; Brady & Thomas, 2005) and grew out of the Gardner Commission, whose “primary task was to formulate new and innovative thinking on the issue of federal education aid” (Brady & Thomas, 2005, p. 52). The passage of the first ESEA made it possible to provide additional funding to schools that serve high numbers of poor students through Title I grants. According to Standerfer (2006)

By limiting the federal funds to only those schools that had extra needs because of the socioeconomic status of their students, there seemed to be the promise that the federal role in education would lessen the achievement gap between students of different backgrounds without intruding on those schools that were doing well without federal mandate. (p. 1)
Once a school was deemed qualified to receive a Title I grant, “school districts had to show results” (Hoff, 1999, para. 83) and thus, the government created the Title I Evaluation and Reporting System, or TIERS. In other words, program evaluation was mandated, and a required part of that evaluation translated into an increased demand for norm-referenced tests, which were believed to be valid, reliable, standardized assessments that would allow student performance to be compared to other students. Hoff states, this also led to a push for more test development. Where by Urban and Wagoner (2009) further argue, “ESEA was by far the most costly and comprehensive federal educational law that had ever been passed” (p. 373).

**B. Title I becomes Chapter I in the 1980s**

By 1980, ESEA had been reauthorized four times, with each passage focusing on how to further ensure the money was not being abused or misused, as earlier audits had found, but that the federal aid being provided was being allocated in the manner congress had intended, namely, to provide educational supports, according to Peterson, Rabe, and Wong (1991), “to assist educationally disadvantaged students from low-income families” (McDonnell, 2005, p. 24). With the advent of the 1980s, funding and legislative support for ESEA declined and along with the 1981 passage of President Reagan’s Education Consolidation and Improvement Act, which had the goal of “reducing the role of the federal government in domestic policy areas” (Brady & Thomas, 2005, p. 53). The passage of this act also changed the name of Title I to Chapter I; however, the purpose essentially remained the same, to provide funding for educationally disadvantaged students. Unfortunately, the changes not only brought reduced funding, but also relaxed regulations which in turn led to states seeing a dramatic decrease in the number of disadvantaged students eligible to be served.
In 1988 ESEA was yet again amended, only this time states were now mandated to start providing documentation that their disadvantaged students were not only being properly served, but that they were also meeting the goals outlined by the grant. “Consequently, receipt of ESEA funds began to be based on the achievement of educationally deprived students” (Brady & Thomas, 2005, p. 54). The Education Summit of 1989 between President George H. W. Bush and the state governors led to additional requirements such as a need for greater educational accountability of Title I, greater flexibility in how the funds are allocated, and a need for nationally developed educational goals in order to increase student achievement. The failed legislative initiative America 2000 came out of this meeting along with a requirement of national academic standards for all students. Though the bill failed, it still is what spurred states to develop their own state level academic standards.

C. The Standards Movement of the 1990s

The standards based reform initiatives under the Bush administration were continued under President Clinton in the 1990s. With Clinton, however, came yet another reauthorization of ESEA, or otherwise known as the Improving America’s Schools Act (IASA). With the IASA we see the first glimpse of parts of the basic framework that will eventually become No Child Left Behind. IASA contained such components as a requirement to identify whether schools attain Adequate Yearly Progress (AYP) in order to be eligible to receive Title I funds, plus steps for improvement if the school does not; as well as a requirement that all students, regardless of their being eligible for the funds, being held to the same challenging academic standards and curricular goals (McDonnell, 2005).

D. No Child Left Behind: 2001 to Present
The reauthorization of ESEA in 2001, and the subsequent passage of the No Child Left Behind Act (NCLB), meant that ESEA had now been reauthorized six times (Crawford, 2011) and with each passage, the federal government required those receiving the Title I grants to prove accountability through the use of test scores. How the different states have approached the requirements necessary to receive Title I funding has varied as greatly as the varying direction from the federal government about the number of grades that were required to be tested, the type of test required to be used, which students were required to participate in the assessment, or how the results of the students were to be reported (Hoff, 1999). However, this all changed in 2001 and the passage of the NCLB Act. NCLB not only set strict requirements about the grades to be assessed (Grades 3-8 and one grade in high school), the type of test to be used (must be a criterion-referenced assessment and must include applied skills items), and which students were to be included (all students were now required to be assessed and the results of the different populations were required to be reported), but it also required stringent sanctions should schools not meet Adequately Yearly Progress (AYP) in each of the identified categories of students for two years in a row (U.S. Department of Education, 2011).

IV. Legal Parameters

As already mentioned, there are varying opinions about standardized testing and accountability. Despite the ongoing policy conversations, there are two basic legal requirements that must be met for standardized assessments that have any stakes attached, whether teacher evaluation, graduation qualification, traditional school/district accountability, and charter accountability. These include notice and validity.

1. Notice/Opportunity to Learn
The landmark case *Debra P. v. Turlington*, 644 F.2d 397 (5th Cir.1981), established the initial concept of opportunity to learn. As states transition to more rigorous item types under the new Common Core consortia tests, the opportunity to learn concept may no longer simply be about standards and includes access to rigorous assessment item types. Similarly, the underlying concept of allowing students opportunity to learn has transitioned to an “opportunity to teach” argument on behalf of teachers as student performance on assessments has become one of multiple measures in teacher evaluation. There have been a long line of opportunity to learn cases, and the minimum notice period ranges between 2-3 years. *See Rene ex rel v. Reed*, 751 N.E.2d 736 (Ind. Ct. App. 2002) (challenging notice of Indiana assessment requirement as applied to students with disabilities); *GI Forum Image De Tejas v. Texas Educ. Agency*, 87 F. Supp. 2d 667 (W.D. Tex. 2000) (challenging the use of the Texas Assessment of Academic Skills (TAAS) examination as a requirement for high school graduation). See also the “Chapman line”, including *Chapman v. Cal. Dept. of Educ.*, 229 F.Supp.2d 981 (N.D. Cal. 2002)(rev’d in pt by *Smiley v. California Dept. of Educ.*, 53 Fed.Appx. 474 (9th Cir. 2002); *Chapman v. California Dep’t of Educ.*, No. C 01-01780 CRB, 2003 WL 22114264 (N.D. Cal. Sept. 5, 2003).

2. Validity

The bulk of the validity case law has stemmed from teacher evaluation. While many of the cases focus on the statistics of value added models, the heart of the cases involve litigants challenging the validity of the test. *See Wagner v. Haslam*, No. 3:15-CV-115, 2015 WL 3658165, at *7 (M.D. Tenn. June 12, 2015)(discussing State Board policy that it would “continually monitor and revise the list of options under this category based on increasing
availability of high-quality measures of performance,” and that it “will work to develop valid and reliable student growth measures for those areas that do not currently have them.”); See also Sherrod v. Palm Beach Cnty. Sch. Bd, 963 So. 2d 251, 253 (Fla. Dist. Ct. App. 2006)(“For those students utterly dependent on the skills of their teachers, the test may have some validity”); Houston Federation of Teachers v. Houston Ind. Sch. Dist., (S.D. Tx 2014)(specifically challenging validity of the value-added model and arguing that the state assessment system was not aligned to the state standards). Thus, as stakes become associated with standardized tests, validity transitions from being a psychometric concern to a legal one.

Test design has multiple stages, but for any test, the following are design issues that could raise validity and therefore legal concerns:

Have teachers and schools had notice of the accountability system (school grading, evaluation, etc.)?

Is the test aligned to the standards?

Have the test items been properly piloted?

Have the students had opportunity to be exposed to the item types?

Were cut scores developed through a valid and reliable means?

Has paper pencil been compared to online test results? Did you use different scales?

Have schools had an opportunity to review and correct data errors?

If Growth data are used, is the growth statistic valid and reliable?
The Reading cases demonstrate the need to allow for appeal and data corrections, and it further demonstrates the ability of school districts to bring litigation on validity grounds. See Reading Sch. Dist. v. Dep't of Educ., 875 A.2d 1218, 1222 (Pa. Commw. Ct. 2005) (holding that Department's appeal policy that only allowed appeal for statistical error was an unconstitutional restraint on due process); See also Reading Sch. Dist. v. Dep't of Educ., 855 A.2d 166, 172 (Pa. Commw. Ct. 2004) (“District contends that the Department's selection of the “N” number was done arbitrarily and not based on sound statistical methodology”).

V. Conclusion

Courts have upheld and even required school accountability provisions. Claremont Sch. Dist. v. Governor, 147 N.H. 499, 510, 794 A.2d 744, 752 (2002) (holding, “standards of accountability are an essential component of the State's duty to provide a constitutionally adequate education.”) These systems look at performance, and performance metrics are partially fed through assessment data. This paper is designed to offer pragmatic guidance regarding validity issues as they relate to these systems because to the extent that assessment data can be challenged on validity grounds, energy and dollars will be used towards litigation rather than school programming. Moreover, if a test is invalid, then it is unable to serve its intended purpose of accurately identifying student mastery.

References


Indiana Department of Education. (1985). Indiana basic competency skills testing and remediation program manual. Retrieved from http://www.eric.ed.gov/ERICWebPortal/search/recordDetails.jsp?searchtype=basic&page Size=10&ERICExtSearch_SearchValue_0=indiana+basic+competency+skills+testing+and+remediation+program+manual&ERICExtSearch_SearchType=kw&_pageLabel=RecordDetails&objectId=0900019b8008d6e5&accno=ED263130&_nfls=false

Indiana Department of Education. (1988). Indiana statewide testing of educational progress ISTEP+ program manual. Retrieved from http://www.eric.ed.gov/ERICWebPortal/search/recordDetails.jsp?searchtype=advanced&page Size=10&ERICExtSearch_SearchCount=1&ERICExtSearch_SearchValue_0=indiana+statewide+testing+for+educational+progress+ISTEP+program+manual&ERICExtSearch_SearchType=kw&ERICExtSearch_SearchOperator_1=and&ERICExtSearch_SearchValue_1=kw&ERICExtSearch_SearchType_1=kw&_pageLabel=RecordDetails&objectId=0900019b80102397&accno=ED296330&_nfls=false


Legal Considerations for Accountability and Assessment
Michelle McKeown, J.D., Ph.D.
Cynthia Roach, Ed.S.