

**Validity of high stakes standardized test requirements for  
homeschoolers: a psychometric analysis**

March 30 2008  
Nicky Hardenbergh

**Introduction**

The issue of high stakes testing occupies a large portion of today's educational policy discussions. Public and private debates about testing policies are animated and multi-faceted. One very small facet of the topic, however, rarely emerges into public view. I refer to those policies that require the use of standardized tests as a high stakes assessment of home education.

For students in home education, as well as students in public school classrooms, the term "high stakes" signifies that something of high value will be gained or lost depending on test results alone, without reference to any other indicator of student performance. Students in public school may be retained in grade or denied a diploma if they fail to score above a predetermined *cut point* (the numerical score that separates those who "measure up" from those who fail to do so). Homeschooled students who do not score above that cut point may be required to enroll in school.

Given the relatively small number of students educated at home, the lack of attention to this issue is not surprising. Only about two percent of the nation's students are homeschooled, and only a small fraction of those students live in the diminishing number of jurisdictions that have high stakes testing policies. Nonetheless, each legislative session seems to bring yet another call for increased testing for homeschoolers.

As a homeschool advocate, I see no need for such policies. My purpose in this paper, however, is not to make the case that these policies serve no useful purpose. Rather, my purpose is to make the case that these policies, useful or not, are inherently flawed. As we shall see, the test instruments specified in the policies cannot properly be used to measure what policymakers hope to measure, and the cut points specified in the policies are nonsensical.

The continued existence of these flawed policies reveals a lack of knowledge, among policymakers and the general public, about *psychometrics*, literally "measurement of the mind," the field of expertise that deals with the design, construction, and interpretation of psychological and educational tests. The tests

referred to in this paper are properly termed standardized *achievement* tests, in order to distinguish them from other types of standardized tests that are used for psychological or cognitive assessment. These tests are properly viewed as tools for measuring mental capacities.

A tool used to measure a mental capacity, such as reading ability, is fundamentally different from a tool used to measure a physical attribute, such as height. If we want to measure someone's height, we stand that person up to a measuring tape and simply read off the results. We can immediately perceive what measurement task the tool is performing.

With mental measurement tools, however, we cannot immediately perceive what task the measuring tool is performing. A test might be called a test of reading ability, but until we understand a great deal about how that test is designed, constructed, and interpreted, we cannot hope to evaluate how well that particular measurement tool works. And until we understand test instruments themselves, we cannot hope to evaluate whether or not any test instrument is *valid* for the prescribed purpose.

In the field of mental measurements, *validity* has a particular meaning. The professional publication *Standards for Educational and Psychological Testing* characterizes "validity" as "the most fundamental consideration in developing and evaluating tests" (American Educational Research Association 1999, p. 9). "Validity refers to the weight of accumulated evidence supporting a particular use of test scores" (Phillips 2000, p. 348). In other words, given a certain student score, can we make sound assumptions about that student's abilities based on that score? Or, more simply stated, does the instrument actually work as intended?

In this paper, I demonstrate, through reference to the extensive psychometric literature, that the psychometric tool prescribed in current high stakes homeschool policies, a *norm-referenced* standardized test, is invalid for use in a high stakes testing policy. Norm-referenced test scores may not validly be used to determine if a student meets a given standard of performance.

I go on to examine another testing tool proposed by some policymakers: the state-specific high stakes *criterion-referenced* tests administered to public school students in every state. While theoretically valid for determining a standard of performance, such tests would be problematic for use in the homeschooling context. I end by reviewing the setting of cut points on high stakes tests, showing that, to a very large extent, the entire controversy of high stakes testing can be reduced to the question of the validity of the cut point.

After considering the psychometric evidence, I conclude that

current and proposed high stakes standardized requirements for homeschoolers are baseless. Policies based on such requirements are a waste of taxpayer dollars and a needless imposition on homeschooling families.

### Who makes the tests?

We are all familiar with standardized tests. We've done our share of filling in the little ovals on the answer sheet. But we might never have considered that these tests are produced by commercial testing companies. Three companies, companies that also produce textbooks, account for the publishing of most of the norm-referenced tests on the market. (We will learn about norm-referenced tests in the next section.) These companies are: McGraw Hill (Terra Nova and California Achievement Test), Riverside/Houghton Mifflin (Iowa Test of Basic Skills), and Harcourt-Brace (Stanford Achievement Test).

Traditionally, almost all standardized tests administered in schools were norm-referenced tests. Starting in the 1990s, however, testing companies began to produce criterion-referenced tests in order to meet the demands of large clients, namely departments of education in various states that were involved in standards-based education reform. (We will learn about the characteristics of criterion-referenced tests in the following sections.) The market for criterion-referenced testing mushroomed following the passage of the federal No Child Left Behind Act of 2001 (NCLB). Other testing companies are now growing in size (Olson 2004).

Massachusetts hired the Measured Progress company to develop the Massachusetts Comprehensive Assessment Test (MCAS). California hired the Educational Testing Service to develop its STAR standards-based tests. Texas hired Harcourt Educational Measurement (which had the Massachusetts contract for four years from 2000-2004) to develop its Texas Assessment of Academic Skills (TAAS).

The passage of the NCLB is also responsible for creating a new hybrid type of test: the *augmented norm-referenced* test. Two years after the passage of the Act, the United States Department of Education approved this new type of test (2003, F-5), effectively allowing companies to adapt their traditional norm-referenced test to fit the requirements of the Act by adding some criterion-referenced items.

I point out the existence of these commercial test publishing companies so that we do not operate under the illusion that standardized tests magically appear on testing day. These measurement tools are created by professional test makers. We need to understand more about the design, construction, and

scoring of the tests themselves in order to analyze the validity of their use for homeschoolers.

### **Validity of norm-referenced test requirements**

When a state or local homeschool policy calls for testing, that policy will almost certainly specify the use of a nationally normed standardized test. The following New York policy is one example of such a regulation:

[The homeschool] program will be placed on probation only if the composite score of the student is below the thirty-third percentile on national norms or the score fails to reflect one academic year of growth when compared to a prior test. (New York State Department of Education, 2005)

The use of the terms "percentile" and "norm" indicates that the type of test specified is *norm-referenced*.

A norm-referenced test is used to compare one student's performance to that of other students. The "reference" is the student's peer group or "norm" group (more about this later). The term "norm" is a technical, statistical term and should not be confused with the everyday use of the word "normal," as in speaking about someone's "normal" (or typical) behavior or condition.

### **Characteristics of norm-referenced tests**

The most important point to keep in mind when thinking about norm-referenced tests is that they are *rankings*. Ranking is exactly what is sought in selective situations where the goal is to distinguish the most promising candidates for some position or reward, such as in admission to a particular college. The most selective colleges use a norm-referenced test, usually the College Board Scholastic Achievement Test (SAT), to sort out, or rank, the top prospects to win one of the limited number of places available.

A norm-referenced score is one that compares the test-taker with other students who took the test. Which students are test-takers being compared with? On the norm-referenced tests commonly used in public schools, the "norm group" is generally a sample of several thousand students from public and private schools who are chosen to be representative of the entire student population. From this norm group, psychometricians first determine the raw scores, or the number of test items a student answered correctly. The raw score is then "scaled," that is, transformed mathematically, using a function derived by psychometricians for

this purpose. For a norm referenced test, the raw scores will be transformed into a *scaled score*. It is this scaled score that is reported to the student and used for policy purposes. Norm-referenced tests and (jumping ahead a bit) criterion-referenced tests both yield scores that are computed first as raw scores and then as scaled scores, though the scaling mechanisms are different. For policy purposes, it is important to note that sometimes very small differences in raw scores can result in disproportionately large differences in scaled scores, and vice versa.<sup>1</sup>

The scale typically employed for norm-referenced tests is a *percentile scale*, though sometimes *stanines*<sup>2</sup> are used. The percentile scale ranges from one to 99. A student's percentile score should not be confused with the percent of correctly answered questions. A "percentile" is linked to a "percent" only in connection to rank: a student's *percentile score* indicates what *percentage of students* in the group scored lower than that individual student. For example, student A, who has a percentile rank in the 45<sup>th</sup> percentile, scored above 45 percent of students in the norm group.

By definition, only one percent of all the students in the norm group will score in the 99<sup>th</sup> percentile and, likewise, only one percent of all students in the norm group will score in the first percentile. By definition, half of all the students in the norm group will fall at or below the 50<sup>th</sup> percentile. It is theoretically impossible, then, for more than half of all students in the norm group to score above the 50<sup>th</sup> percentile.

In practice, however, more than half of students often do score in the upper 50 percentiles. This seemingly impossible situation occurs because of the upward creep, over time, of raw test scores due to improved academic performance, familiarity with the test, or other reasons. While the raw scores, as a whole, may be going up, the original norm group will continue to serve as the reference scale for all subsequent test-takers until the testing company renorms the test. So when the Lake Wobegon

---

<sup>1</sup> Let's look at an actual example from one norm-referenced test. Student A, with a percentile score of 48<sup>th</sup> percentile, answered only one more question correctly than student B, who received a scaled score of the 41<sup>st</sup> percentile. For all practical purposes, there is no difference between the two students, yet they are separated by 7 percentile points. (Rudner 1993, p. 3). Similarly, on criterion-based tests, relatively large differences in scaled scores can mask relatively small differences in raw scores.

<sup>2</sup> Stanines are generated from percentile scores, using a conversion chart. For example, percentiles 96-99 comprise stanine 9, while percentiles 41-59 comprise stanine 5. Typically, stanine scores are interpreted as above average (9, 8, 7), average (6, 5, 4), and below average (3, 2, 1) (Pearson Educational Measurement 2007c).

district claims that 75 percent of its students are "above average," those students are only "above average" as compared to students in past years, not in comparison to their true peer group, which is the group of fellow test-takers in the same year.

***Norm-based policies for homeschoolers psychometrically invalid***

When homeschoolers are required to take standardized tests, the test instrument, as I indicated previously, is generally specified as a nationally normed standardized test. Some lawmakers seem to think, erroneously, that norm-based tests are an appropriate vehicle for setting a standard of performance. Although these norm-referenced testing policies cannot withstand psychometric scrutiny, they continue to exist.

Prominent measurement expert Gregory J. Cizek comments on this poor practice:

The practice of using norms as standards - despite being universally condemned by testing specialists - continues in many testing contexts ... One example is the use in many states of percentile standards for accountability, in which parents must submit evidence that a student attained a certain score (i.e., performance at the 30<sup>th</sup> percentile) on a nationally-normed standardized test in order to remain eligible for home schooling. (Cizek, p. 43)

Some jurisdictions, in an effort to make fair provision for low-scoring homeschool students, have policies that call for use of test scores to provide evidence of "one year of academic growth" (New York State Department of Education, 2005). While the provision may sound reasonable, it does not make psychometric sense. Any student's norm-based score will be determined not only by that student's performance but also by the performance of the numerous other students in the norm group. Thus it is perfectly possible for a student to make progress but score lower than the previous year. Measuring "one year of academic growth" using a norm-referenced test is a psychometric task that sounds straightforward but is, in fact, highly complex and not at all straightforward. (See footnote 4 for more on the difficulties of measuring yearly progress.)

These flawed psychometric policies receive little scrutiny. Even when, in two cases, courts have reviewed these policies in the context of low-scoring homeschooled students, judges have not been asked to examine issues of psychometric validity.

### **Review of court cases concerned with norm-referenced testing for homeschoolers**

To the limited extent that homeschool testing issues have been litigated, the courts have not made their decisions based on psychometric considerations, because the cases turned on whether or not a family homeschooling for religious reasons must submit to any state regulation. (Answer: yes.) In their decisions, the judges did not comment on the suitability of the test instrument itself.

A West Virginia court decision (Null v. Bd. of Educ., 1993) illustrates this lack of psychometric analysis. A homeschooled child in West Virginia failed to score above the 40<sup>th</sup> percentile cut score, even after two tries. The relevant statute at that time, as quoted in the Court's decision, provided that if a child's scores on a nationally normed test "are not above the fortieth percentile level, home instruction shall no longer satisfy the compulsory school attendance requirement exemption" (\*938). The parents sought a preliminary injunction to keep the child from being ordered back to school. At the hearing on the motion for injunctive relief, a measurement expert testified about the lack of precision of percentile scores. Nonetheless, the Court, citing the reasonableness of the state statute, denied the request for an injunction, even though the child's score was in the 38<sup>th</sup> percentile, only two percentile points below the prescribed cut score. (See footnote 1 for a specific example of the imprecision of percentile scores.)

The Court, in its ruling, upheld the 40<sup>th</sup> percentile cut point, asserting, with no analysis, that "the state statute's 40 percent [sic] cutoff reasonably may be intended to promote above average scores" (\*940). This passage from the Court's decision is remarkable. First, the West Virginia court made the common error of referring to percentile scores as percentage scores! Second, the Court seems to think that the 40<sup>th</sup> percentile represents a cut score "above average." Both comments demonstrate a lack of psychometric literacy.

A decision from an Arkansas court (Murphy v. Arkansas, 1988) also reveals a lack of knowledge about norm-referenced tests. In this case, the Court upheld a requirement that homeschooled students be placed in public, private, or parochial school if they do not score "within eight months of grade level"<sup>3</sup> (\*1041).

---

<sup>3</sup> "Grade level" is another way of reporting norm-referenced scores. As stated in a glossary on the Pearson Educational Measurement website:

A grade equivalent (GE) is a score reported on norm-referenced tests that allows educators and parents to compare students based on the performance of other students relative to the school year. Based on a 9-month school year (typically September through May), the score represents a period

The Court holds that, since "the state must have a mechanism by which it can confidently and objectively be assured that its citizens are being adequately educated," the testing requirement is appropriate.

The wording of the Court's decision is suggestive of the faith the Court has in the test. The validity of the measurement instrument is simply assumed. The Court accepted completely the notion that a certain score on the norm-referenced test was adequate proof of an inadequate home education. As we have seen, that notion is false. Norm-referenced tests are not designed to measure the adequacy of one individual's education, let alone to do so "objectively and confidently."

To my knowledge, the holdings of the West Virginia and Arkansas courts have not been revisited, nor has any court been presented with expert evidence about the misuse of norm-referenced standardized tests to set performance standards for homeschoolers.

#### **Validity of criterion-referenced test requirements**

The informed reader will already be asking whether the use of criterion-referenced testing would solve the psychometric problems presented by the use of norm-referenced tests. After all, a criterion-referenced test is designed to "measure a level of mastery according to a specific set of performance standards" (Zucker 2003, p. 6). True, the use of criterion-referenced tests for high stakes homeschool evaluation would eliminate the psychometric problem of attempting to use norm-referenced tests to set standards. Other, new psychometric problems would, however, be raised. Criterion-referenced tests contain their own set of problematic issues, as we shall see.

Criterion-referenced tests, also termed "standards-based academic performance tests" have achieved public prominence because of their use in today's public school high stakes testing programs. While such tests are designed for use in the public school, legislators in some states have proposed requiring homeschool students to take the same state-specific standardized tests that are required of students in the public schools. This bill introduced in the New Mexico legislature is one example:

---

during the school year, displayed as a number to show a grade and a month. The score is an estimate of the performance that an average student at a grade level is assumed to demonstrate on the test at a particular time in the school year. For example, a score of 5.8 represents a performance level typical of fifth-grade students in the eighth month (April) of the school year (Pearson Educational Measurement 2007a).

All home school students shall participate in the state's academic assessment program and shall take the standards-based academic performance tests in their respective school districts that are required of public school students ... The department may require home school students who do not demonstrate adequate yearly progress<sup>4</sup> to attend a public or private school. (New Mexico House Bill 158, 2005)

While this particular bill did not become a law, such proposals continue to surface. Lawmakers continue to suggest that homeschoolers should be assessed by means of the same "standards-based" (criterion-referenced) assessments that are in use for public school students in that state. To understand why such proposals are problematic, we first need to know more about the nature of criterion-referenced testing.

### ***What is tested must be aligned with what is taught***

High stakes criterion-referenced tests are only considered valid if students have been given an opportunity to learn the subject matter and skills on which they will be tested. The legal system and the measurement community agree that high stakes tests are only acceptable if what is tested is aligned with what is taught. Courts have determined that high stakes testing programs in schools are legally permissible only if the tests measure content that students have had an opportunity to learn. (Debra P. v. Turlington, 1984, and G.I. Forum v. Tex. Ed. Agency, 2000).

Homeschoolers, like private school students, are not required to follow the state's detailed curricular frameworks. As the Massachusetts Supreme Judicial Court determined some time ago: "[t]he great object of these provisions of the [compulsory attendance] statutes has been that all the children shall be educated, not that they shall be educated in any particular way" (Commonwealth v. Roberts, 1893). Lawmakers in Massachusetts and elsewhere have long recognized the fact that a variety of successful educational methods exist, and that children who are not educated at public expense need not follow a curriculum

---

<sup>4</sup>"Adequate yearly progress" is a term used in high stakes school testing programs. States require evidence, via test scores, of progress from year to year. The problem with such a regulation is that it presumes "adequate yearly progress" can be accurately measured. In fact, the measurement of yearly progress is fraught with psychometric difficulties. While policymakers seem unaware of these difficulties, measurement experts themselves discuss them at length in their professional publications. The authors of one such article remind us that "the difficulty of accurate measurement, especially with regard to annual progress, should not be underestimated" (Camilli 2006, p. 12). The authors, making use of the item parameters published by the Massachusetts Department of Education, found that purported large gains on MCAS scores were, in fact, modest gains, with some plausible evidence that changes in scoring methods may also have influenced the score gains.

identical to that of the public school system.

The curricular sequence followed in home education need not match the sequence followed in the public school classroom. Students who have been educated via a non-conventional curriculum may not have been taught the content assumed by test-makers and thus will not be properly assessed by a test designed for the majority of students. The students in the non-conventional curriculum may be either ahead or behind those in the regular public school classrooms.

Charter schools provide small laboratories for studying the effects of non-standard curriculum. Test scores from one small charter school in California reflect the effects of late reading instruction.<sup>5</sup> A Waldorf-inspired school in California, Yuba River Charter School, does not introduce formal reading lessons until the third grade. On the 2004 state tests, only one of the 21 Yuba River second grade students scored above the "basic" cut-off level, while in the district as a whole 38% of second graders scored above that level. Not surprisingly, students not yet exposed to reading instruction were decidedly subnormal on the state tests. Older Yuba River students, who *had* received formal reading instruction, however, scored ahead of district averages.

A second charter school, this one in Wisconsin shows the reverse effect: an initial edge in reading scores disappeared in later years. In this case, third grade students in a Montessori charter school scored higher on reading comprehension than their conventionally educated peers. By the eighth grade, however, this lead had disappeared. In both the California and Wisconsin examples, students' future reading ability could not be accurately extrapolated from the second or third grade scores.

The Wisconsin researchers termed their results "surprising, because early reading skills normally predict later reading" (Lillard 2006, p. 1894). The conventional wisdom is that if a student is not reading by third grade, then the student will never catch up. My experience with homeschooling, however, tells me that the phenomenon of "never catching up" is an artifact of the school curriculum. Anecdotal reports of many, many homeschoolers who were "late" readers demonstrate that older non-readers can blossom into fluent readers in a matter of months, provided they have not internalized the idea that "I can't read." Classroom instruction, of course, cannot accommodate such wide variations in learning timetables. In most schools, there is no place other than special education for a child who does not read on the conventional timetable.

In home education, by contrast, the curriculum can be

---

<sup>5</sup> This data can be extracted from yearly test score data provided by the Educational Testing Service and published on the California Department of Education website (California Department of Education, 2004).

tailored to meet the needs and interests of the student. Such an individualized curriculum will not be able to be fairly assessed by a standardized criterion-referenced test. Because homeschoolers are not required to follow exactly the public school curriculum timetable and frameworks, a requirement that they be tested on those frameworks would be psychometrically invalid.

### ***Mistaken assumptions about today's high stakes criterion-referenced tests***

While the mismatch of curriculum to test is a problem limited to non-conventional educational settings, there are other criterion-referenced issues that are relevant to the public school setting as well as the homeschool setting. The measurement task that these criterion-referenced tests are supposed to accomplish is incredibly complex. Policymakers and the general public are mostly unaware of these complexities; we tend to view these high stakes tests as simply larger, longer classroom tests. In this view, however, we are very much mistaken. To gain an accurate understanding of criterion-referenced standardized tests, we need to examine some mistaken assumptions.

#### *a. Assumption about the meaning of "criterion"*

It is hard to define a criterion-referenced test precisely, because the term "criterion" has developed two conceptually distinct meanings. Both meanings are, unfortunately, used interchangeably when referring to high stakes criterion-referenced testing. It is critical, however, to distinguish the two meanings.

The first meaning of "criterion" in a testing context denotes the set of specific behaviors or tasks that a student needs to be capable of performing. The term "criterion-referenced measurement" was originally used in a seminal article<sup>6</sup> by educational psychologist Robert Glaser, who discussed issues relevant to the "measurement of subject matter proficiency, as it may be defined by subject matter scholars" (Glaser 1963, p. 519). For Glaser, a score on a criterion-referenced test is meant to provide "explicit information as to what the individual can and cannot do" (p. 520). In this sense, the term "criterion" can be equated with "skills to be learned."

Not long after the first usage was coined, the word took on a second usage. Rather than referring to the specific behaviors

---

<sup>6</sup> I am indebted to Gene V. Glass for pointing me to Glaser's article. It was Glass' article, "Standards and Criteria Redux" (Glass 2003), that illuminated for me the conceptual confusions caused by the double meaning of the term "criterion."

that a student must learn, "criterion" took on a more abstract significance. Now the term denoted not specific skills or behaviors, but a certain number that corresponded to the *level of performance* required to meet a certain standard (Glass 2003). In this sense, the term "criterion" is used interchangeably with the phrase "cut point."

In certain situations, we might use both senses of the word "criterion" interchangeably with relative impunity. For example, when talking about a simple teacher-made test, such as a spelling test of ten words, we might deem that the *skills to be learned* (definition #1 of "criterion") would be mastering the spelling of ten specific words. We might further deem that the *cut score* (definition #2 of "criterion") for passing would be seven words spelled correctly. We could easily equate a student's passing score of seven out of ten correct with explicit information about how much of the criterion (skills) the student had mastered.

By contrast, a student's score on a complex, standardized test does not directly correspond to any explicit set of skills (criterion). Unlike a simple spelling test, a high stakes criterion-referenced test is designed to test a student's knowledge of an unmanageably large sets of facts, skills, and concepts. In actual practice, as we will see in the next section, students do not have a chance to learn the prescribed set of skills (criterion). Rather, the standard (or criterion) becomes the cut score itself. In other words, to use both meanings of the term, the cut score (criterion) does not equate to mastery of a specific set of skills or knowledge (criterion).

This subtle, but essential, difference in concept between the two meanings of the term "criterion" generates much confusion in discussions of standards-based reform. The term "criterion-referenced" leads policymakers and the general public to assume, mistakenly, that students' scores on high stakes criterion-referenced tests are linked directly to a level of mastery of particular skills and subject matter (criterion).

#### *b. Assumption that all frameworks can be covered*

High stakes criterion-referenced tests are constructed to assess the breadth of material in the state frameworks. Theoretically, students have been taught the skills and subject matter prescribed in the frameworks for their grade level. However, the curricular frameworks in most states represent an unmanageably large set of facts, skills, and concepts, more than can be covered during the class hours available.

In Massachusetts, the curricular frameworks were developed over time, subject area by subject area. As far as I could tell, no official reviewed the frameworks as a whole to determine

whether or not all the subjects could be covered in the time available. Researchers who have investigated the topic of mismatch between frameworks and time in the school day have found that the hours needed to cover the prescribed subjects far exceed the number of hours of instructional time available (Marzano 1999). My own analysis, as a former social studies teacher, tells me that there is not nearly enough time in the school year to cover adequately the topics listed in the Massachusetts frameworks for junior high social studies. I suspect that to the extent that students gain exceptionally high scores, they likely learned a fair amount of their information outside of the classroom.

The goal of standards-based high stakes testing is to hold schools and students accountable for learning the content in the state frameworks. Yet, if the content in the state frameworks is too vast to be covered in the school hours available, the goal of accountability makes little sense. For a high stakes test to be fair to students, students need to have an opportunity to learn all the material that might appear on the test.

*c. Assumption that items are selected based on educational importance*

When constructing a standardized test, test makers must consider the statistical properties of each item or question on the test. In order to understand just how different that process is from the process used by a teacher in making a classroom test, we need to know more about the topic of item selection. Item selection is accomplished through a set of technical procedures that are so complex that they would be impossible without today's high speed computers. These procedures are beyond the scope of this paper. It is enough that we appreciate the fact that these procedures were developed in order to construct a *reliable* measurement instrument. In psychometric terms, a reliable test is one that will return similar results for similar groups of students. To develop a psychometrically reliable test, test makers need to pilot test each item in order to predict how students will respond to that question on the final test.

During the construction of both norm-referenced and criterion-referenced tests, test makers use a pilot test to determine which items (test questions) will appear on the finished test administered to students on testing day. The survivor items are ones that possess the proper statistical profile; survivor items are those that "discriminate" between high-scoring and low-scoring students. In other words, test makers only want items that are answered correctly by high scoring students and answered incorrectly by low scoring

students. Any items that are answered correctly by almost all students will be eliminated from the final test. Items that almost everyone answers correctly are considered "too easy for the target population" (Massachusetts Department of Education, 2005, p. 102). Such "easy" items provide little helpful psychometric information; in selecting test items, test-makers want to choose the items that will be most helpful in distinguishing among students of differing abilities.

We might not be surprised when such item selection methods are used to construct norm-referenced tests. After all, the stated purpose of the test is to rank all test takers from highest to lowest. We might, however, be surprised to discover those same selection methods being used for criterion-reference tests. After all, the stated goal of the criterion-referenced test is not to rank test takers but, as we saw, to "measure a level of mastery according to a specific set of performance standards." Presumably, there are certain performance standards that are so important that we would expect teachers to emphasize them heavily. As a result of such attention to those standards, we would expect most students to know the correct answers to those essential items. The finished test, however, will most likely not contain questions about such essential items; if the pilot test reveals that most students know the answer to a particular item, that item will not generate the proper statistical profile. Items without the proper statistical profile are eliminated from the finished test.<sup>7</sup>

Clearly, this method of test construction is very different from that used by classroom teachers. Let's see what might happen if a similar method were applied to our earlier example of a ten word spelling test. In that example, students were assigned to learn ten words and then were tested on those ten words. Students' scores on such a test clearly showed how well students had mastered the spelling of those ten words. Now let's imagine that ten word spelling test is used as a pilot test from which to create a final test. Using the results of the original test, test makers eliminate any spelling word that *all* students spelled correctly. Test makers then construct a second and final test using only the surviving words. This second test is no longer a simple criterion-referenced test. The second test is, in an important sense, norm-referenced, because test items were chosen by direct reference to how other students performed. Depending on how others performed, a student who correctly spelled seven out

---

<sup>7</sup> I am indebted to Walt Haney of Boston College for his work on MCAS item selection. He examined the MCAS technical reports (supplied to the public on the Massachusetts Department of Education website, <http://www.doe.mass.edu/>) and determined that, indeed, "items have been selected for inclusion on MCAS tests by using norm-referenced test construction procedures" (Haney 2002, following section heading "Why are MCAS score averages poor indicators of school quality?").

of ten words on the first test could receive a score of two out of five on the second test. Unless we knew how the second test had been constructed, we would most probably draw very inaccurate conclusions based on that student's score of 2 out of 5 correct. Similarly, without an understanding of how standardized tests are constructed, policymakers may well draw inaccurate conclusions from students' scores.

At the very least, policymakers need to recognize that scores on the standardized tests required for today's high stakes assessments are not easily or directly correlated with a student's level of achievement or performance. Yet, by definition, a high stakes test does just that: pegs a student score to a particular level of performance. And there is one peg, or point, that is of overwhelming significance for the student. That is the point that separates passing from failing students.

*d. Assumption that cut points are objectively determined*

Even if the psychometric problems discussed above could somehow disappear, we would still be left with the most meaningful issue of high stakes testing: where do you draw the line? There must always be a line when using test scores to sort students into categories of passing and failing. That line is marked by the cut point.<sup>8</sup> Does the cut point accurately separate adequate from inadequate students? The legitimacy of any high stakes testing program depends on the validity of the cut point. Without a psychometrically valid cut point, there is no legitimate way to implement a high stakes testing program. Thus, to a very large extent, the entire controversy of high stakes testing can be reduced to the question of the validity of the cut point. For that reason, I examine the question in greater detail in the next section.

**A closer look at the validity of high stakes testing and cut points**

The all-important question of cut point validity has been the subject of some litigation, and I discuss two of these cases in this section. Both cases raised civil rights issues. I emphasize this point because, without a claim of some constitutional right being violated, as in the case of disparate impact on minorities, federal courts really do not have

---

<sup>8</sup> While today's high stakes testing programs may designate two or three different cut points, only the bottom cut point is actually high stakes for students. That's the one that determines which students fail, and that's the one that most needs to be set in a manner that can withstand scrutiny.

jurisdiction to evaluate the validity of standardized tests or their cut points. Since I am not making a legal argument, I need not try to make the issue of homeschool testing into a constitutional matter. Rather, I will make use of the psychometric evidence presented in both cases, because in both cases the judges placed great weight on the testimony of psychometric experts when determining cut score validity.

### **Validity of cut scores in public school high stakes tests**

A landmark school testing case, GI Forum v. Texas Education Agency, was decided in 2000. The decision covered several aspects of high stakes standardized testing, some of particular relevance to the use of cut scores. In this case, the United States District Court was asked to consider whether the cut score on the Texas Assessment of Academic Skills (TAAS) was properly set. The Court, in its decision, stated, "Whether the use of a given cut score, or any cut score, is proper depends on whether the use of the score is justified" (\*680). The Court declared that "the relevant criterion here is whether the cut score is related to the quality the test purports to measure" (\*680). The Court determined that Texas education officials had acted properly in setting the cut score. Two points raised in the decision deserve a closer look for our purposes.

#### *a. Cut score setting is an exercise in professional judgment*

The Court determined that deciding where to set the cut score was properly left in the hands of the legislature and, by delegation of the legislature, the state's education officials. School officials, the Court observed, are the ones who properly "decide how much a student should be required to learn - the cut score issue," stating: "This Court has no authority to tell the State of Texas what a well-educated high school graduate should demonstrably know at the end of twelve years of education" (\*670). The Court thus left the issue of test validity and cut score validity squarely in the hands of the state's education officials as long as those officials used standard professional procedures in reaching their decisions.

Standard professional procedures for setting cut scores vary somewhat, but in concept they are all, of necessity, subjective. As explained by University of Massachusetts measurement specialist Ronald Hambrelton in his discussion of cut scores on Massachusetts high stakes MCAS test:

There are no "true performance standards" waiting to be discovered from careful educational research. Setting

performance standards on educational tests like the MCAS involves careful professional judgment being made by persons who are viewed as suitable for providing the judgments. (Hambleton 2003, p.14)

In other words, at the end of the mathematical calculations involved in test construction, the final determination of what constitutes passing is reached using the professional judgment of a committee of educators. As Hambleton explains, different committees, or "panels," examine each individual section of the MCAS tests. Each panel consists of 12 to 15 people, about half teachers and half administrators. Sometimes members of the general public are also included on panels.

The panels gather for a two or three day period, during which the panel members view actual student test booklets that have been corrected and assigned a raw score. The panelists, first individually and later as a group, determine what minimum raw score should correspond to a scaled score of 220, a score which is officially termed "needs improvement" but more accurately could be termed "just passing." The panels perform the same deliberations to set the two cut points for "proficient" and for "advanced." The panels then present their recommendations to the Department of Education, which virtually always accepts the panels' recommendations.

Some observers in Massachusetts have commented that the MCAS cut score may be set too low. According to a *Boston Globe* article (Sachetti 2006), Massachusetts state college administrators report that they have seen no reduction in the need for remedial courses for students graduating from Massachusetts high schools, even though these students have all scored above the 220 points needed to pass the MCAS.

In response to this criticism, the Massachusetts Commissioner of Education reportedly replied that the situation would be remedied by raising the passing score twenty points, from 220 to 240 (Sachetti 2006). However, there was no indication in the article that the remediation classes were composed only of students with scores between 220 and 240. Without this kind of hard data, how could the Commissioner be so certain that students who scored 240 actually did have a solid command of the material and did not need remediation?

Here we have an example of the problem generated by the two meanings of the term "criterion." We know that there is a set of skills and competencies that the MCAS is designed to assess. Those are the "criteria" that the MCAS tests. But the "criterion" for being designated as "proficient" is an abstract numerical score, the cut score. Whether or not students who are "proficient" will be able to do college level work is actually an

empirical question. If we wanted to determine whether or not MCAS scores possessed *predictive validity* for college performance, we would need to do a rigorous study of student outcomes in the years following their MCAS testing in order to determine if there were, in fact, a correspondence between their MCAS scores and their college success.

One standardized test that has been extensively studied for its predictive validity is the College Board SAT test. The College Board has determined the SAT test to be valid for predicting a student's performance during the first year of college (Camara 2000). Those validity studies, of course, have their critics who point to other statistics than those used by the test makers. Given the complexities and the controversies of determining the predictive validity of the SAT, we can assume that determining the predictive validity of the MCAS would be a mammoth undertaking. In any case, state departments of education are under no obligation to subject their high stakes tests to any examination of predictive validity, at least if we extrapolate from the ruling in the Texas case.

*b. High stakes tests need only show content validity*

The Texas Court determined that state education officials had sufficiently demonstrated that the TAAS possessed *content validity*. The test is intended to determine if students have sufficiently mastered the content prescribed by the state. In a review of the findings of this case, a prominent measurement expert comments:

The most important evidence of validity in this situation is a measure of the degree to which the items on each subject-matter test measure the knowledge and skills prescribed by the state-mandated curriculum ... [I]n achievement testing applications [this type of validity evidence] is usually referred to as content validity evidence (Phillips 348).

*Content validity* evidence is different from *predictive validity* evidence. Plaintiffs attempted to have the TAAS reviewed by more stringent standards such as those used to validate cut scores used on employment screening tests, but the Court decided otherwise, commenting that the TAAS is a "conceptually different exercise from that of predicting ... success in employment ..." (\*680). Thus, the Court supported the notion that the content validity standard was the proper standard to follow: if questions on the TAAS properly corresponded to the content on state curricular frameworks, that was all the validity evidence needed.

By contrast, in the employment screening tests, as we shall see, the test content and the cut score are subjected to a more thorough examination in terms of their predictive validity.

### **Validity of cut scores in employment screening tests**

United States v. Delaware (2004) dealt with a standardized test used to select Delaware state troopers. On this test, prospective candidates who did not score above a certain cut score were not considered further for hiring. The fairness of this cut score was the subject of the Court's deliberations. The Delaware judge determined, following precedent in dealing with these kinds of cases, that a decision in the case must involve two steps. First, the Court needed to ascertain whether the skills being tested matched the skills necessary to perform as a state trooper. Second, it needed to determine whether the cut score was placed correctly. A correctly placed cut score would screen out unqualified candidates but would not eliminate qualified candidates.

The screening exam was a test of reading comprehension and writing skills. To determine if the skills tested matched the skills needed on the job, researchers canvassed supervisors to ask what skills troopers needed in their work. Evidence from this research indicated that the skills needed were, indeed, the skills being tested. The disputed issue was where to set the cut score.

Experts on both sides examined data gathered from three groups who had taken the exam: (1) those who scored above the cut point and went on to become state troopers, (2) those who scored below the cut point initially but subsequently scored higher and became state troopers, and (3) those who scored below the cut point but went on to secure employment elsewhere in law enforcement. The evidence revealed that a significant number of rejected applicants had gone on to become successful law enforcement officers elsewhere.

Lawyers for both sides introduced psychometric experts to explain the proper way to set cut points, but the experts did not agree with each other. The judge, as demonstrated in his extensive written decision, carefully analyzed the psychometric evidence presented by the experts. After my own investigations, reading any number of conflicting experts, I particularly appreciated reading the judge's comments about how statistical evidence is not as objective as we might imagine. He stated, "the purported objectiveness of the statistical evidence in this case seemed to melt away as well-respected, highly qualified statistical experts drew widely varying conclusions from the data" (\*95).

Nonetheless, the judge found that if the cut score had been several points lower, most of the otherwise qualified candidates would have passed the exam; he ruled accordingly. The judge determined that lowering the cut score a small amount would still protect the public safety interest by setting the cut score at a level that indicated "a high likelihood of being able to do the job" (\*90).

The Delaware decision indicates that for a predictive test, two hurdles have to be overcome: first, the skills tested must match the skills needed on the job, and, second, the cut point must be set at the point that indicates a high likelihood of being able to do that job, but not so high that it eliminates qualified candidates. The predictive validity of a cut point, following these guidelines, can only be determined empirically, after an examination of the real world consequences of using that particular cut point. The cut point must correspond to the minimal skills necessary for the job.

### ***Today's high stakes tests are not tests of minimal skills***

We saw earlier how the term "criterion" has two different significances, which should not be used interchangeably. We revisit that same issue when we look at the term "minimal skills." "Minimal" obviously means the least amount necessary. But necessary for what? In the Delaware employment screening test, the particular skills needed to perform the job were first outlined, and then the screening test was developed in order to measure those particular skills. A valid cut score indicated the point that separated those applicants who possessed the necessary skills from those applicants who did not. The *minimal score* needed to pass the test thus directly corresponded to the *minimal skills* needed to perform as a state trooper.

On school high stakes tests, by contrast, there is no "job" to analyze in terms of skills needed. Test takers are not seeking employment in a particular field; rather, they are seeking to obtain a high school diploma. The skills necessary to obtain that diploma are outlined by the state education officials in the state curricular frameworks. There are no performance standards outside of the curricular frameworks themselves. Thus, the minimal skills needed to obtain a diploma may or may not correspond to the minimal skills needed to function successfully in the adult world.

Yet policymakers seem to assume such a correspondence, presuming that high stakes school tests are assessing what might be termed "essential" skills for functioning as a productive citizen. The Texas Court uses the term "essential" skills when referring to the content assessed by the TAAS:

[T]he State of Texas has determined that a set of knowledge and skills must be taught and learned in State schools. The State mandates no more than these "essential" items. (\*681)

The Texas Court might be surprised to learn that the skills needed to pass today's high stakes exit tests cannot properly be termed "essential." Policymakers and the general public, when hearing the term "essential skills," probably think of skills such as the ability to:

- read and comprehend English passages representative of widely circulated material commonly encountered in adult life,
- solve mathematical problems derived from situations commonly encountered in adult life, and
- create, in English, a written composition. (Minnesota Department of Education, 2006)

The above skills are those tested by the Minnesota Basic Skills Test (BST). It would be hard to argue against categorizing any of those three items as "essential." Not surprisingly, this test was well-accepted by the school community because there was strong consensus that the skills on the test were truly skills that everyone should possess (Yeh 2005). Although it was well accepted, this test is no longer used. It has been replaced by a series of new state tests "that help districts measure student progress toward Minnesota's academic standards and meet the requirements of No Child Left Behind [NCLB]" (Minnesota Department of Education 2005).

Any discussion of the requirements of the NCLB is far beyond the scope of this paper, other than to indicate that the passage of the NCLB in 2001 promoted a trend away from tests of basic skills in favor of tests designed to correspond to higher academic standards.

States that had basic skills tests, such as Minnesota and Texas, replaced those tests with new tests designed to test the content of the state-specific standards as expressed in the state's extensive curricular frameworks. We need only look at two contrasting items, one from the Minnesota BST and one from the standards-based TAAS, to see the difference in the skills tested.

The first mathematics problem on the 1998 BST official practice test follows. Note: the BST is *not* a timed test<sup>9</sup>:

---

<sup>9</sup> We are so accustomed to timed tests that we give little thought to the effect of such a constraint. The amount of time students are given to complete a test section should really be considered another performance variable in analyzing test results. We can assume that if given less time, many students would do less well, and vice versa.

1. If you average 50 miles per hour while driving your car, estimate the distance you would travel after 28 minutes [to be answered without a calculator].
  - A. 2 miles
  - B. 25 miles
  - C. 80 miles
  - D. 1500 miles (Minnesota Department of Children, Families and Learning, 1998)

Let's compare that question with the first mathematics question on the 2002 TAAS exit exam:

1. A minute is 0.0000019 year. How is this number expressed in scientific notation?
  - A  $1.9 \times 10^{-6}$
  - B  $1.9 \times 10^{-5}$
  - C  $1.9 \times 10^5$
  - D  $1.9 \times 10^6$  (Texas Education Agency 2002)

It would be hard to argue that the ability to express numbers in scientific notation is an "essential skill," to use the Court's term. Certainly most adults rarely, if ever, encounter such problems in everyday life.

Clearly, the TAAS is not a test of "essential skills," despite what the Court may assume. Today's high stakes exams are not tests of basic skills. Rather, they are tests of the content outlined in the state's curricular frameworks. They are properly viewed as *accountability* tests, designed to hold students and schools accountable for meeting the standards outlined in the curricular frameworks.

***Purpose of state high stakes assessments properly viewed as "accountability"***

State legislatures today often use the term "accountability" when referring to educational issues. The Massachusetts Education Reform Act of 1993 contains this description of one of the activities of the state's Board of Education:

The board shall carry out its responsibilities with a view toward increasing the accountability and effectiveness of public early childhood, elementary, secondary and vocational-technical schools and school districts for the performance of the students they serve. (Mass. Gen. Laws ch. 69, § 1B, 1993)

As this wording indicates, high stakes school testing programs

are instituted by legislators in order to hold schools "accountable" to the taxpaying public and to assure taxpayers that they are getting their money's worth. Additionally, legislators aim to raise student achievement by holding students "accountable" for achieving a certain standard in order to receive a state-certified diploma.

As we saw, in Massachusetts as well as in other states, the Department of Education is the agency that drafts the frameworks, develops the test, and sets the cut scores. Nowhere in this process does any outside agency determine if the cut score is pegged at a point that validly separates adequately educated from inadequately educated students. Rather, these high stakes tests are best viewed as internal accountability assessments, valid only for use within the system.

### **Conclusion: State high stakes tests not valid for homeschoolers**

Non-public school students, including homeschoolers, are not, by definition, students enrolled in public school. Non-public school students are not educated at public expense, nor do they receive high school diplomas from the state. Non-public school students are simply not accountable to state officials in the same way as public school students are. Therefore, the state-specific accountability tests that are valid for use in the public schools are not necessarily valid for use in non-public schools.

Policymakers who propose using such tests for homeschoolers have not considered the psychometric validity of these tests in the context of homeschooling. In fact, they do not even have a clearly articulated purpose for proposing the tests in the first place. The sponsor of a 2008 homeschool testing bill in the Nebraska Legislature is typical of legislators who propose homeschool testing. According to newspaper reports, she had no evidence that such testing would serve any useful purpose. Rather "her concern [came] from the stories she hear[d] about students who are kept out of public or private schools but receive little to no schooling" (Robb 2008). Another article reported that this same legislator "testified that she has heard anecdotes over the years about children 'who fall through the cracks' or are said to be home-schooled but show no evidence of education" (Saunders 2008).

Here, as in all the other cases of which I am aware, the rationale for legislation is ill-defined, and the evidence of need for the legislation is anecdotal. Even if legislators could provide a clear rationale and solid evidence, they would still need to face the issue of validity of the test instrument. They cannot simply assume that the measurement instrument described in

their legislation is valid for the testing job they are describing.

As we have seen, a nationally normed standardized test, the kind mentioned in existing homeschool testing policies, is not valid for use in a high stakes setting, because such tests are not designed to be used with cut points. We further saw that today's state-specific standards-based tests are not valid for use outside of the public school setting, because they are designed to measure adherence to the state's curricular frameworks. They would not be valid for non-public school students who are not required to follow those specific frameworks.

Without a valid measurement instrument, high stakes homeschool testing policies are irrevocably flawed. The lack of the proper tool is not, however, a significant policy problem, because there is also a corresponding lack of evidence that any need for such a tool exists. Most jurisdictions in the country do not mandate high stakes assessments for homeschoolers, and there is no evidence that the lack of such a requirement results in any lesser education for students in those jurisdictions. The topic of homeschool regulation is, of course, beyond the scope of this paper. Rather, based solely on psychometric considerations, I encourage policymakers to resist the urge to develop high stakes policies for homeschoolers. Without a proper tool, high stakes homeschool policies are more than just a waste of policymakers' time; they are, as I said before, a misuse of taxpayer dollars and a needless intrusion on homeschooling families.

## References

### References

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Camara, W. J. and G. Echternacht. (2000). *The SAT I and High School Grades: Utility in Predicting Success in College*. Retrieved March 2, 2008 at [http://www.collegeboard.com/research/pdf/rn10\\_10755.pdf](http://www.collegeboard.com/research/pdf/rn10_10755.pdf)
- Camilli, G., & Vargas, S. The legend of the large MCAS gains of 2000 2001. *Education Policy Analysis Archives* (14)4. Retrieved March 25, 2008 at <http://epaa.asu.edu/epaa/v14n4/v14n4.pdf>
- Cizek, G. J. (1998). *Filling in the blanks: Putting standardized tests to the test*. Washington D.C.: The Thomas B. Fordham Foundation. Retrieved March 2, 2008 at <http://www.edexcellence.net/doc/cizek.pdf>
- Commonwealth v. Roberts, 34 N.E. 402 (Mass.1893). Retrieved March 25, 2008 at <http://www.mhla.org/information/massdocuments/roberts.htm>
- Debra P. v. Turlington, 644 F.2d 397 (5<sup>th</sup> Cir. 1981).
- California Department of Education (2004). *California standardized testing and reporting (STAR). Yuba River data*: Retrieved March 25, 2008 at <http://star.cde.ca.gov/star2004/viewreport.asp?ps=true&lstCounty=29&lstDistrict=66415&lstSchool=6113138>. *Twin Ridges Elementary District data*: Retrieved March 25, 2008 at <http://star.cde.ca.gov/star2004/viewreport.asp?ps=true&lstCounty=29&lstDistrict=66415&lstSchool=>
- GI Forum v. Tex. Educ. Agency, 87 F. Supp.2d 667 (W.D. Tex. 2000).
- Glaser. R. (1963). *Instructional technology and the measurement of learning outcomes*. *American Psychologist*, 18, 519-521.

- Glass, G. V. (2003). *Standards and criteria redux*. Retrieved March 24, 2008 at <http://glass.ed.asu.edu/gene/papers/standards/>
- Hambleton, R. K. (2003). Setting passing scores on tests. Special report on education reform ten years after the Massachusetts Education Reform Act of 1993. *Education Connection (Spring 2003)*, 11-14. Retrieved March 25, 2008 at <http://www.umass.edu/education/publications/ed.connection.2003.pdf>
- Haney, W. (2002, May 6). Lake Woebeguaranteed: Misuse of test scores in Massachusetts, Part I. *Education Policy Analysis Archives*, 10(24). Retrieved March 24, 2008 at <http://epaa.asu.edu/epaa/v10n24/>
- Lillard, A., and N. Else-Quest. 2006. Evaluating Montessori education. *Science* 313 (Sept. 29):1893-1894.
- Marzano, R. J., Kendall, J. S., and Gaddy, B. B. (1999). *Essential knowledge: The debate over what American students should know*. Aurora, Colorado: McRel Institute.
- Massachusetts Department of Education (2005). *2005 MCAS Technical Report*. Retrieved November 14, 2006 at <http://iservices.measuredprogress.org/MCAS2005TechReport.pdf>
- Mass. Gen. Laws ch. 69, § 1B. Education Reform Act of 1993. Retrieved March 8, 2008 at <http://www.mass.gov/legis/laws/mgl/69-1b.htm>
- Minnesota Department of Children, Families and Learning (1998). *Minnesota 1998 basic standards practice test: Mathematics*. Retrieved March 25, 2008 at <http://education.state.mn.us/mdeprod/groups/Assessment/documents/Instruction/000395.pdf>
- Minnesota Department of Education (2005). *Assessments: MCA-II*. Retrieved March 25, 2008 at [http://education.state.mn.us/mde/Accountability\\_Programs/Assessment\\_and\\_Testing/Assessments/MCA\\_II/index.html](http://education.state.mn.us/mde/Accountability_Programs/Assessment_and_Testing/Assessments/MCA_II/index.html)
- Minnesota Department of Education (2006). *General information about the basic skills tests*. Retrieved March 24, 2008 at <http://education.state.mn.us/mdeprod/groups/Assessment/documents/FAQ/000342.pdf>

New Mexico House Bill 158. (2005). *An Act Relating to Education: Requiring Home School Students to Take Required Standards-Based Academic Performance Tests; Allowing the Public Education Department to Require Home School Students Who Do Not Demonstrate Adequate Yearly Progress to Attend a Public or Private School*. Bill introduced in New Mexico Legislature 47th legislature State of New Mexico first session 2005. Retrieved March 25, 2008 at <http://legis.state.nm.us/Sessions/05%20Regular/bills/house/HB0158.html>

New York State Department of Education. *Home Instruction in New York State*. Retrieved March 25, 2008 at <http://www.emsc.nysed.gov/nonpub/part10010.htm>

Null v. Bd. of Educ., 815 F. Supp. 937 (S.D. W. Va. 1993).

Murphy v. Arkansas, 852 F.2d 1039 (8<sup>th</sup> Cir. 1988).

Olson, L. (2004). NCLB law bestows bounty on test industry, *Education Week* 24(14): 1,18-19. Retrieved February 3, 2004 at <http://www.edweek.org/ew/articles/2004/12/01/14tests.h24.html>

Pearson Educational Measurement (2007a) *Glossary*. Retrieved March 25, 2008 at <http://www.pearsonedmeasurement.com/research/glossary.htm#grade>

Pearson Educational Measurement (2007c). *What is a stanine, and what does it mean?*. Retrieved March 24, 2008 at [http://www.pearsonedmeasurement.com/research/faq\\_2f.htm](http://www.pearsonedmeasurement.com/research/faq_2f.htm)

Phillips, S. E. (2000) *GI Forum v. Texas Education Agency: Psychometric Evidence*. *Applied Measurement in Education*, 13(4), 343-385. Retrieved March 14, 2008 at <http://marces.org/mdarch/pdf/1000024.pdf>

Robb, J. (2008, February 24). Home-school pitch pits personal choice vs. government role. *Omaha World-Herald*. Retrieved March 3, 2008 at [http://www.omaha.com/index.php?u\\_page=2798&u\\_sid=10266451](http://www.omaha.com/index.php?u_page=2798&u_sid=10266451)

Rudner, L. M. (1993). *The achievement testing provisions of the Virginia home schooling requirements*. Testimony for the Virginia State Legislature, February 4, 1993. (ERIC Document

- Reproduction Service No. ED 355267). Retrieved March 24, 2008 at <http://www.eric.ed.gov/ERICWebPortal/contentdelivery/servlet/ERICServlet?accno=ED355267>
- Sacchetti, M. (2006, June 26). Colleges question MCAS success: Many in state schools still need remedial help. *Boston Globe*. Retrieved March 24, 2008 at [http://www.boston.com/news/education/k\\_12/mcas/articles/2005/06/26/colleges\\_question\\_mcas\\_success/](http://www.boston.com/news/education/k_12/mcas/articles/2005/06/26/colleges_question_mcas_success/)
- Saunders, M. (2008, February 27). Home school families voice opposition to bill. *Omaha World-Herald*. Retrieved March 4, 2008 at [http://www.omaha.com/index.php?u\\_page=2798&u\\_sid=10268901](http://www.omaha.com/index.php?u_page=2798&u_sid=10268901)
- Texas Education Agency. (2002) Texas Assessment of Academic Skills, Exit Level released test. Retrieved March 24, 2008 at <http://www.tea.state.tx.us/student.assessment/resources/release/taas/release02/xl.pdf>
- United States Department of Education. (2003). *Standards and Assessments: Non-Regulatory Guidance*. Retrieved March 2, 2008 at <http://ed.gov/policy/elsec/guid/saaguidance03.doc>
- United States v. Delaware, 2004 U.S. Dist. LEXIS 4560 (D. Del. 2004).
- Yeh, S. S. (2005). Limiting the unintended consequences of high-stakes testing. *Education Policy Analysis Archives*, 13(43). Retrieved March 24, 2008 at <http://epaa.asu.edu/epaa/v13n43/>
- Zucker, S. (2003). *Fundamentals of Standardized Testing: Harcourt Assessment Report*. Harcourt Assessment, Inc. Retrieved March 24, 2008 at [http://harcourtassessment.com/NR/rdonlyres/20DB5E75-059E-4EB7-BFB1-F9C171ABE0C3/0/Fundamentals\\_of\\_Standardized\\_Testing\\_Final.pdf](http://harcourtassessment.com/NR/rdonlyres/20DB5E75-059E-4EB7-BFB1-F9C171ABE0C3/0/Fundamentals_of_Standardized_Testing_Final.pdf)