

Running head: CHECKS AND BALANCES

Checks And Balances: Why a Federal Agency Should Monitor State-Administered Standardized

Testing Processes

Jade Caines
Emory University
Fall, 2006

The [No Child Left Behind Act \(NCLB\)](#)¹, passed in 2002, made history by mandating that public schools across the nation assess students regularly and systematically. Each state is required to implement an accountability system that tests children annually in reading and mathematics from grades three through eight and once in high school. Due to this federal mandate, as well as time restrictions and financial limitations, most states have implemented high-stakes standardized testing. There lacks investigation, however, into state testing accountability methods and procedures. States are required to submit an accountability plan to the [United States Department of Education](#) (which is either rejected or accepted²), but the public is unaware of the criteria used in this process. How do states decide passing scores on these high stakes tests? What steps are taken to minimize the unintended consequences of high stakes standardized tests? This research paper will address these questions and recommend that a bipartisan federal agency be created in order to protect public school students from the inappropriate uses of standard setting methods that state accountability programs may employ.

I. Standards and Standard Setting

Politicians and lawmakers often report that America's public school system needs higher standards and more accountability. But what are *standards*? How do states define standards and how should our country increase standards? Before improvements can be made to our education system these important questions must be addressed. In this section I will explain why the development and measurement of standards is a policy problem. I will also define standards and

¹ United States Department of Education. (n.d.). Elementary and Secondary Education Act (The No Child Left Behind Act of 2001). Retrieved December 1, 2006, from <http://www.ed.gov/policy/elsec/leg/esea02/index.html>

² There are actually several categories: They include three approval categories (full approval, full approval with recommendations, approval expected), three approval pending categories (approval pending levels 1, 2, and 3, where level 3 is a withholding of funds), and one not approved category. [Phillips, S.E. (2006, September). Legal corner: NCLB peer review update. National Council on Measurement in Education newsletter, 14(3), 4-6.

explain how they are utilized in standardized testing. Last, I will outline important elements to the standard setting process.

A. A Policy Problem

In order to ensure academic success on a national level, public policy has had to contend with the issue of standardization. It is hard to evaluate public school performance nationwide without some level of consistent analysis. [NCLB](#), for example, has required states to administer standardized assessments. In this new standardized era, however, millions of students are being assessed in ways that may not be most appropriate. Therefore, standardized assessment models need to be critiqued in order to ensure accurate measures of student mastery.

An article in an [American Psychological Association](#) (APA) publication stated that there is concern that many schools use standardized test results in inappropriate ways and that tests themselves fail to adjust for America's diverse student population (Smith, 2001). In addition to APA, other national organizations have called for a deeper look into educational testing. The [National Board on Educational Testing and Public Policy](#), for example, has proposed that research on testing become a part of a national agenda on education reform (Clarke, Madaus, Pedulla, & Shore, 2000). American policymakers and state education legislators, however, are often just as mystified about the process of creating and applying standardized tests as the average taxpayer. Currently, the [Department of Education](#) holds states accountable for compliance to [NCLB](#), but there is no one to ensure that steps taken to gain compliance are fair. We must move beyond compliance; a federal monitoring agency is needed in order to ensure states fairly and accurately implement standardized tests.

B. Performance Standards

In order to understand why a federal agency that monitors issues related to standardized testing is needed, we must first understand the great potential for negative consequences. To illustrate this point, I will discuss the performance standard setting process as it relates to standardized testing. By highlighting this aspect of accountability programs, I will demonstrate why a federal monitoring agency is not only beneficial, but imperative.

1. Defining Performance Standards

Since the standards-based movement began in the late 1980s, accountability has been a well-used buzz word. We must hold states, districts, schools, principals, teachers, and students *accountable*. This notion of accountability is based on the standards that students must reach to be deemed “educated.” There are, however, different layers of standards that are discussed within the education arena. *Content* standards classify the subject material that teachers are expected to teach and students are expected to learn (Linn, 2002). This is different from *performance* standards that identify the levels of performance that students can achieve on a given assessment; performance standards allow states to categorize students based on mastery of content knowledge. Examples of commonly used performance levels include *Below Basic*, *Basic*, *Proficient/Meets the Standard*, and *Advanced/Exceeds the Standard* (Linn, 2002). Figure 1 shows an example of one state’s performance standard definitions.

Figure 1: Performance Level Definitions

Louisiana Educational Assessment Program for the 21 Century Achievement Ratings

Advanced: Superior performance beyond the proficient level of mastery

Mastery: Competency over challenging subject matter

Basic: Demonstrates only fundamental knowledge and skills

Approaching Basic: Partially demonstrates fundamental knowledge and skills

Unsatisfactory: Has not demonstrated fundamental knowledge and skills

Source: Louisiana Department of Education, Standards Assessment, and Accountability (<http://www.doe.state.la.us/lde/saa/2273.html>)

2. How Performance Standards Become Cut Scores

Although performance standards are often referred to as the aforementioned levels, they become operational when they are translated into passing scores on an assessment (Linn, 2002). These passing scores, also referred to as cut scores, are the points on a score scale that separate one performance standard from another (Horn, Ramos, Blumer, & Madaus, 2000). States are responsible for setting these cut scores and then submitting them to the [United States Department of Education](#) for approval. These scores are critical because they determine how demanding the standards and assessments are for any given student in a state’s public school system ([Education Sector](#), 2006). Figure 2 shows an example of performance levels and corresponding test scores on an average 50-question multiple-choice exam. Although these numbers will vary depending on the assessment, each performance level will always have a specific range of correct responses (Horn et al., 2000)

Figure 2: Cut Scores and Corresponding Percentage of Items Correct

Performance Level	Number of Items Correct	Percent Correct
Advanced	40-50	80-100
Mastery	30-39	60-79
Basic	20-29	40-59
Approaching Basic	10-19	20-39
Unsatisfactory	0-9	0-19

Source: Adapted from “Cut Scores: Results May Vary” by Horn, Ramos, Blumer, & Madaus (<http://www.bc.edu/research/nbetpp/reports.html>)

3. *Variation of Performance Standards*

Unfortunately, as a result of [NCLB](#) mandates, there is extreme variation among states in regards to performance standards. States are under severe pressure to make adequate yearly progress by any means necessary. The negative consequences associated with failing schools, (e.g. families opting for private schools, decreased property values, and negative labeling) leave states with few options. [Ryan](#) (2004) states that schools have four options to avoid failure: (1) they can try their hardest to genuinely improve achievement, (2) they can postpone large annual increases in school goals and hope that by 2014 federal legislation will change³, (3) states can simply refuse federal monies and ignore [NCLB](#) in its entirety, or (4) states can make the standardized tests easier by lowering the cut scores. Many states have opted for the latter option and have been cited for intentionally setting low cut scores to ensure high passing rates (Ryan, 2004). [Ryan](#) argues that [NCLB](#) implicitly encourages this by requiring states to set high standards that must be met yearly, while allowing states to set their own standards and cut scores. If students do not pass the standardized tests, schools can be classified as “in need of improvement.”⁴ High cut scores make it harder for students to pass, thereby encouraging states to set them low in order to meet adequate yearly progress. Although the [Department of Education](#) has to approve each state’s accountability system individually, including cut scores, national consistency is tough to establish because limited oversight exists.

Several states have set low cut scores in order to achieve high passing rates. Louisiana, for example, used to divide performance levels into three categories: basic, proficient, and advanced. When only 17% of eighth graders scored proficient or advanced on the Louisiana

³ A goal of NCLB is that by 2014 all students in public school will be proficient in all subjects.

⁴ The U.S. Department of Education categorizes school success in two ways: (1) has met adequate yearly progress and (2) in need of improvement (NI). An *NI* label, according to the federal government, is not synonymous with *failing*. However, as newspapers nation-wide publish test scores, these *NI* and *failing* labels are used interchangeably in almost all popular venues.

state English test, and only 5% performed at those levels on the mathematics exam, the state simply designated those students that scored basic as “proficient” in order to meet [NCLB](#) mandates (Ryan, 2004). Colorado and Connecticut also increased their passing rate by redefining score categories in order to satisfy [NCLB](#) (Ryan, 2004). State administrators in Texas also felt pressured as a result of the federal legislation; they lowered the number of correct answers needed to meet proficiency on the third grade reading test (Ryan, 2004). It is clear that many states feel pressure to meet the standards and may aim low when establishing state cut scores in order to achieve adequate yearly progress.

C. The Process of Standard Setting

It is unclear how many educators, lawmakers, parents, and community members are aware of the process in which cut scores are established. Although some may assume that it is a highly complicated process that requires a certain level of knowledge to understand, it is important to dispel that myth. Standard setting in *any* field, including professional, medical, and education arenas, is a highly subjective process that is prone to error, just like any other process that involves human beings. In this section, I will explain two key aspects that create an effective standard setting process: (1) the selection and training of judges and (2) the chosen standard setting method.

1. The Selection of Judges⁵

Although, a variety of standard setting methods exist, there are general steps that all processes must include. Selection and training of standard setting participants, for example, is central to a fair and valid determination of cut scores.

⁵ I will use the terms *judge* and *participants* interchangeably.

There are several purposes undergirding the selection of participants. First, the selection process strives to identify people with the characteristics necessary to acquire the knowledge and skills taught in the training sessions. In other words, effective training cannot happen without effective selection. Therefore, the selection of participants is interrelated to the training that will occur throughout the standard setting process (Raymond & Reid, 2001).

It is helpful to choose judges that can learn quickly, but it is more effective to select participants that do not need training. Hence, another purpose of participant selection is to minimize the amount of training necessary by locating people who already possess the necessary knowledge and skills needed for a successful determination of cut scores. Although this may be hard, the selection process must occur with the understanding that training is quick and time is limited. Therefore, knowledgeable participants, with a capacity to learn quickly, are ideal (Raymond & Reid, 2001). Figure 3 shows a sample of tasks necessary to conduct an effective selection and training process. Column one identifies what judges would have to do throughout, and by the end of, the process. Column two addresses the skills necessary to accomplish the standard setting goals. Column three outlines factors to consider during the selection process. Finally, column four discusses activities used in the training of judges.

There are six criteria that should be used when choosing judges: (1) subject matter expertise, (2) an understanding of examinee population, (3) the ability to estimate item difficulty, (4) knowledge of instructional environment, (5) an appreciation of the consequences involved with standard setting, and (6) a fair representation of all involved stakeholders (Raymond & Reid, 2001). I will give a brief description of each criterion and the challenges involved.

a. Subject matter expertise.

In order to have a fair determination of cut scores, each participant must be knowledgeable about the subject content the assessment measures. The more specialized the assessment, the more difficult the task of finding subject matter experts. The goal, however, is to identify participants with ample expertise in assessment domains (Raymond & Reid, 2001).

For example, [The National Assessment of Educational Progress](#) (NAEP), also considered the nation’s report card, convenes standard setting panels of 30 individuals, 70% of which are educators that possess subject matter expertise. It is important that derived cut scores have input from those professionals that understand key concepts that students must learn in a given discipline (Raymond & Reid, 2001).

Figure 3: Sample Task Analysis for a Specific Standard Setting Method

<i>Major Standard Setting Tasks</i>	<i>Sample Knowledge and Skills Requirement</i>	<i>Sample Selection Factors</i>	<i>Sample Training Activities</i>
1) Acquire understanding of the context of the standard-setting activity and the environment to which the standard will be applied.	<ul style="list-style-type: none"> ◆ purpose of the exam ◆ Test specifications and test development process ◆ Rationale for, and consequences of, standard setting 	<ul style="list-style-type: none"> ◆ Ability to recognize benefits and limitations of testing ◆ Ability to appreciate consequences of applying a standard ◆ Knowledge of instructional environment 	<ul style="list-style-type: none"> ◆ Compare and contrast purpose of the test to other possible purposes. ◆ Explain test development and item writing procedures.
2) Develop definition of borderline examinee performance.	<ul style="list-style-type: none"> ◆ Characteristics of examinee population ◆ Education and training experiences of examinee population ◆ Examination performance data (item performance and examinee performance) 	<ul style="list-style-type: none"> ◆ Experience or contact with population of interest ◆ Knowledge of levels of proficiency in examinee population 	<ul style="list-style-type: none"> ◆ Discuss rationale for standard setting. ◆ Describe cognitive characteristics of examinees.
3) Estimate minimum passing levels for each item. <ul style="list-style-type: none"> a. Read each item and evaluate the correct answer. 	<ul style="list-style-type: none"> ◆ Detailed knowledge of the domain being assessed 	<ul style="list-style-type: none"> ◆ Ability to read at the level required by the exam 	<ul style="list-style-type: none"> ◆ Evaluate levels of examinee proficiency on the exam and criterion of

b. Evaluate the relative difficulty of the item.	<ul style="list-style-type: none"> ◆ Item characteristics that influence difficulty ◆ Examinee characteristics that influence item difficulty 	<ul style="list-style-type: none"> ◆ Analytic skills (written comprehension; reasoning; speed of closure; problem sensitivity) 	<ul style="list-style-type: none"> ◆ interest. ◆ Review educational preparation of examinees. ◆ Present charts depicting exam statistics and discuss varying proficiency levels.
c. Estimate the proportion of borderline examinees that will provide a correct response.	<ul style="list-style-type: none"> ◆ Basic understanding of probability 	<ul style="list-style-type: none"> ◆ Number facility and related skills 	<ul style="list-style-type: none"> ◆ Practice estimating item difficulty with feedback and discussion.
d. Repeat step 3 for each item on the test.		<ul style="list-style-type: none"> ◆ Ability to concentrate for long periods of time; persistence 	<ul style="list-style-type: none"> ◆ Present concept of measurement error associated with each items.

Source: Adapted from Raymond, M.R. and Reid, JB. (1991). Who made thee a judge?: Selecting and training participants for standard setting. In G.J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 119-157). Mahwah, NJ: Erlbaum.

b. Understanding the examinee population.

Before establishing cut scores for a given exam, standard setting judges should have a thorough understanding of the target population. In other words, judges should grasp an idea of the students taking the exam in question. An appreciation of the psychological factors related to learning and academic achievement is important to the process of standard setting. Also, judges should consider the range of individual differences that exist in a specified student population. Although the amount of understanding necessary is unclear and undefined, it still remains an important criterion for selecting participants (Raymond & Reid, 2001).

c. Estimating item difficulty.

Standard setting panelists need to be familiar with the varying factors that influence the difficulty of test items. An exam's format, choice of wording, and item order, for example, can all influence student test performance (Raymond & Reid, 2001). Developed cognitive abilities are necessary to understand the interaction between a student and a particular test item. At the

same time, judges must consider the factors that will influence the outcome of that interaction, thereby influencing an item's difficulty (Raymond & Reid, 2001).

d. Knowledge of instructional environment.

Some standard setting processes include student performance data on given assessments that may influence cut score derivations. Data shown can play an important role in final performance standard determinations; therefore, it is important that participants understand the design of instructional environments, including type of instruction, instructional context, and variation in the quality of instruction (Raymond & Reid, 2001). When performance data is introduced, panelists will understand contextual factors that may have affected students' opportunities to learn the assessed content. Otherwise, cut score decisions will be uninformed and irrelevant.

e. An appreciation of standard setting consequences.

Panelists do not make judgments in a vacuum. In this era of high-stakes standardized accountability, cut scores must be derived with sensitivity to potential negative consequences. A given panel may be responsible for determining cut scores for a high school graduation test, for example, where students who fail will not receive a high school diploma. Decisions, therefore, cannot be taken lightly. Judges must fully comprehend the purpose of a given examination and the rationale involved for setting cut scores (Raymond & Reid, 2001). Panelists must consider the benefits to society in establishing cut scores for a given exam, the social and individual consequences for students that fail, and the potential social consequences of judgment errors within the process. There are research tools, including questionnaires and interviews, that can be used in the selection process to weed out judges who are unable to appreciate the importance of the process and the possible consequences involved.

f. Fair representation of involved stakeholders.

A standard setting panel should include a collective representation of all relevant communities of interest in the process. In other words, those affected by the outcome of the standard setting process have a right to be involved in that process (Raymond & Reid, 2001). Those involved in K-12 education represent a large group, making this criterion difficult to meet. Also, it may be difficult to train a high school math teacher, a university professor, and a concerned parent simultaneously; each representative may hold strong convictions about what students should know in a given domain. Figure 4 shows one breakdown of a standard setting panel.

2. Standard Setting Methods

There are five basic steps that all standard setting processes include. In addition to the participant selection step, judges must also define “borderline” knowledge and skills for a given assessment. These two steps are the same for all standard setting methods. The remaining steps include participant training, collection of judgments, and final cut score determination. These steps differ depending on the methodology used.

Standard setting methods are numerous.⁶ They date back to the 1950s and, since that era, multiple methods have been researched and published. Two of the more popular methods, Angoff and Bookmarking, are worthy of analysis.

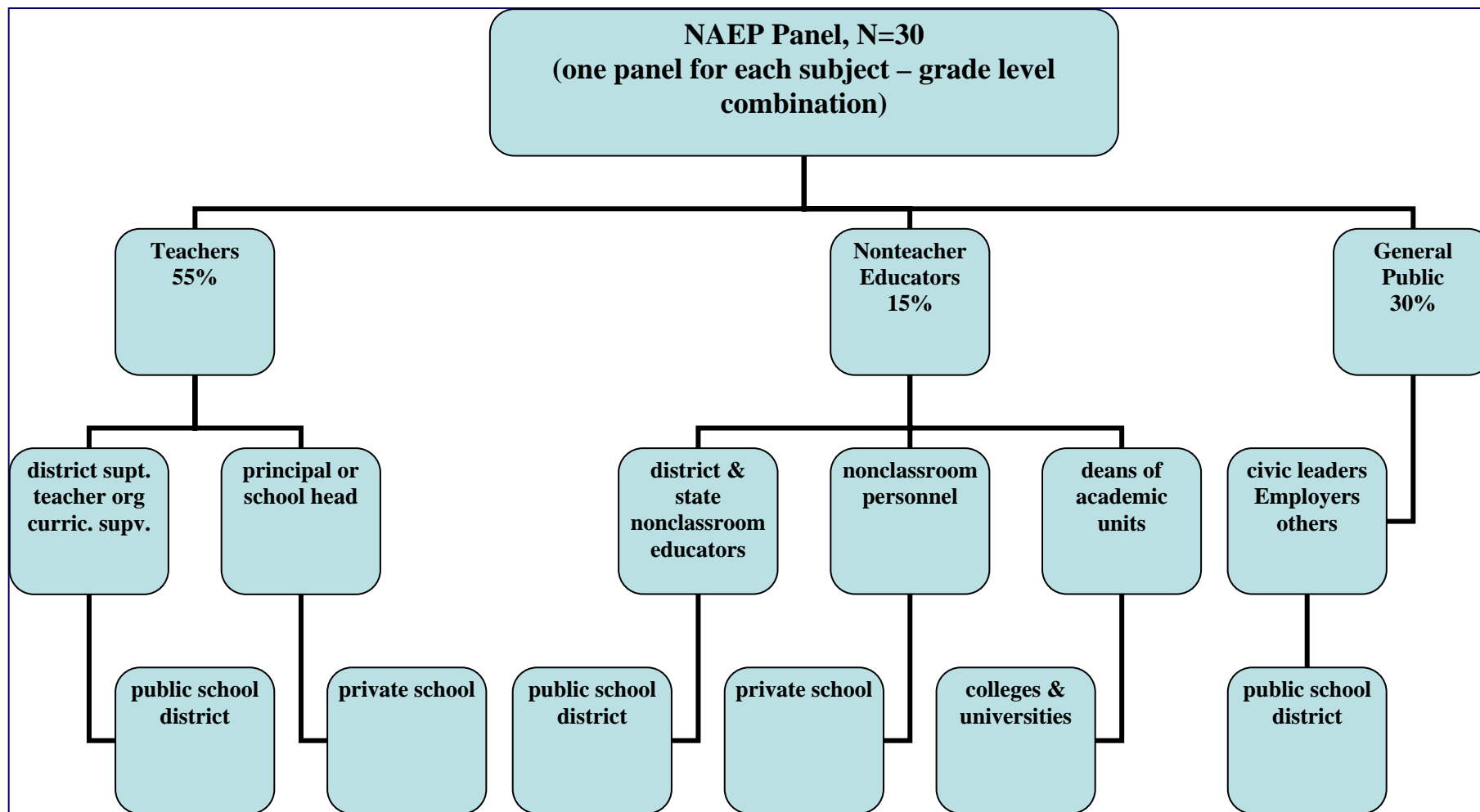
a. Angoff method.

The original Angoff method (1971) requires participants to determine the probability of a borderline examinee answering a specific test question correctly (Livingston & Zieky, 1982;

⁶ There are a variety of standard setting methods that have been used over the past 50 years. They include Nedelsky’s method, Angoff’s method, Ebel’s method, the Borderline Group method, the Contrasting Groups method, modified Angoff methods, the Bookmark method, and the Body of Work method. Only a few methods will be discussed in this section.

Horn et al., 2000; Zieky, 2001). It is similar to the first published standard setting method (Nedelsky's method) except that it can be used with tests that are not only multiple-choice.

Figure 4: NAEPS's standard setting participant breakdown



Source: Adapted from Raymond, M.R. and Reid, JB. (1991). Who made thee a judge?: Selecting and training participants for Standard Setting. In G.J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 119-157). Mahwah, NJ: Erlbaum.

Also, the Angoff method, unlike Nedelsky's method, does not require the participant to consider each possible wrong answer separately. Instead, panelists consider test items as a whole and make judgments about the probability that a borderline test taker would answer the question correctly (Livingston & Zieky, 1982). Each judge creates a probability score between 0.0 and 1.0 for each question, which is then averaged to create a minimum competency cut score. Then participants convene as a group and average each judge's probability score. The final number is the cut score for the examination (see figure 5).

Over the years a Modified Angoff method, and other extensions of the original method, have emerged. The key difference is that the original Angoff allowed judges to determine probabilities on their own, while the Modified Angoff restricts the probabilities judges can choose to eight choices (.5, .20, .40, .60, .75, .90, .95, and "do not know").

Figure 5: An Example of Angoff's Method

Calculations		Setting the Cut Score	
Question	Probability of Correct Answer	Judge	Cut Score
1	.95	1	5.80
2	.80	2	6.00
3	.90	3	6.00
4	.60	4	5.40
5	.75	5	5.00
6	.40	6	5.30
7	.50	7	5.50
8	.25	8	4.80
9	.25	9	6.10
10	.40	10	5.50
Sum=5.80			Sum=5.54

Source: Adapted from Livingston S. and Zieky, M. (1982). Methods based on judgments about test questions. In *Passing Scores: A Manual for Setting Standards of Performance on Educational and Occupational Tests*. Princeton, NJ: Educational Testing Service.

**FINAL
CUT
SCORE**

b. The bookmark procedure.

This method was developed in 1996 by three research scientists at [CTB/McGraw-Hill](#) (Lewis, Mitzel, & Green). It was developed to address the limitations of other methods, i.e. the need to create multiple cut scores in a single exam, the ability to accommodate multiple-choice *and* constructed response items, and the need to simplify the process for participants (Karantonis & Sireci, 2006). Using a statistical technique called item response theory, test questions (also referred to as items) are ordered along a scale from easy to hard. In the first round of the process, participants decide where cut scores will be set along that scale by placing a “bookmark” at the designated point (Horn et al., 2000). Then judges confer about their scores to mitigate discrepancies. Round two involves participants repeating the bookmark placement process to make necessary adjustments (after evaluating what they think students should know at each level *along with* consideration of group discussion). A third round allows judges an opportunity to defend (and/or amend) their final cut score determinations.

A survey conducted in 2000 investigated standard setting methods used by state assessment systems. It found that the Bookmark Procedure was used in 18 states, by far the most widely used method in America at that time (Karantonis & Sireci, 2006). A 2001 report stated that it had been used in as many as 28 states. Lastly, in 2005, it was reported that 31 states have used the Bookmark method (Karantonis & Sireci, 2006). This method has been widely used for several reasons: (1) it requires less data entry than other methods, (2) it takes less time for judges to complete, and (3) it can be used with all types of questions (Zieky & Perie, 2006).

II. Legal Precedents

In a legal context, standard setting processes have become quite important in determining fairness and equity issues within assessment contexts. In addition to education, legislation has also occurred in the professional certification and licensure field. Although the focus of this

paper is to address measurement issues in public education, non-education related examples only strengthen the argument for the creation of a federal test-regulatory agency. Several cases, inside *and* outside of the education arena, serve as cornerstones in understanding the crucial role cut scores play in determining students' futures.

A. Richardson v. Lamar County Board of Education [729 F. Supp. 806 (M.D. Ala. 1990)]

In 1986, Alice Richardson brought a lawsuit against Lamar County Board of Education in Alabama. She claimed that, according to Title VII of the Civil Rights Act of 1964, she was wrongfully refused to renew her teaching contract. Richardson claimed disparate impact, arguing that the teacher certification exams disproportionately discriminated against African American teachers (*Richardson v. Lamar County Bd. of Educ.*, 1990).

The Alabama District Court found the test development process to be outside of the realm of professionalism due to the cumulative effect of several serious errors committed by the test developers when it formulated the 45 exams in 1981 and 1982 (*Richardson v. Lamar County Bd. of Educ.*, 1990). First, the teacher certification exams (Alabama Initial Teacher Certification Test and early childhood and elementary education exams) were found invalid. In assessing overall validity, the Court addressed cut score validity.

The Court outlined two reasons why the Alabama teacher certification exams were cut score invalid: (1) “the cut scores bear no rational relationship to competence as that construct was defined by Alabama educators” and (2) “evidence revealed a cut score methodology so riddled with errors, that it can only be characterized as capricious and arbitrary” (*Richardson v. Lamar County Bd. of Educ.*, 1990). Overall, both rationales support the creation of a federal monitoring testing agency.

B. Guardians Ass'n of New York City Police Dept., Inc. v. Civil Service Commission of the City of New York [630 F.2d 79 (C.A.N.Y., 1980)]

Exam No. 8155 was a test used to screen potential police officers in New York City. The Court of Appeals held that the exam had a disparate racial impact and that it lacked cut score validity (*Guardians Ass'n of New York City Police Dept., Inc. v. Civil Service Commission of the City of New York*, 1980).

Although cut score validity is not the only type of validity an exam must possess, it is the cut score that ultimately determines whether a person passes or fails. A cut score is invalid if it leads to the failure of examinees that have mastered the content assessed. The converse is also true; exams that yield passing scores for candidates that have not mastered the content are cut score invalid (*Guardians Ass'n of New York City Police Dept., Inc. v. Civil Service Commission of the City of New York*, 1980).

The City of New York had several options in establishing cut score validity: (1) they could have used a professional estimate of the required ability levels or (2) analyze the test results to locate a logical “break point” in the distribution of scores (*Guardians Ass'n of New York City Police Dept., Inc. v. Civil Service Commission of the City of New York*, 1980). The Court of Appeals found that the defendant did not use either process. Instead, they selected as many candidates as they needed and set the cut score in order for the remaining candidates to fail. Therefore, the police officer’s entrance examination was improper and invalid because the cut score was not achieved using appropriate testing guidelines and procedures. The examination violated Title VII and had a significant disparate racial impact.

C. Cureton v. National Collegiate Athletic Ass'n [37 F.Supp.2d at 687 (E.D.Pa., 1990)]

This class action lawsuit was brought by four African American college athletes that challenged an NCAA rule requiring a minimum passing score on either of two standardized tests in order to participate in sports their freshmen year (*Cureton v. National Collegiate Athletic Ass'n*, 1990). The Court held that the NCAA had to follow Title VI and that requiring a minimum score on a standardized test had an unjustified disparate impact against African Americans.

III. What Can We Learn

There is a lot to be learned from a thorough investigation of standard setting processes and judicial responses to abusive cut score methodology. A few salient themes include stakeholder participation, standard testing practices, and the unintended consequences of standardized testing. The abuse, negligence, and/or dismissal of these concepts demonstrate the need for a federal agency that will monitor adherence to testing procedures.

A. It Takes a Village: The Notion of Consensus in Setting Performance Standards

Standardized testing is highly value-laden. Who should set acceptable passing scores on standardized tests? Who decides how much mastery students need to demonstrate in a particular subject matter? These questions are often answered by policymakers and state education legislatures with minimal input from concerned community members. As previously mentioned, standard setting processes should involve all community members with concerned interest. NAEP standard setting committees (see figure 4) have a fair representation of all stakeholders. Unfortunately, however, not every standardized assessment has such a representative participant sample. *Richardson v. Lamar County* showed how test publishing companies can abuse their power and dismiss input from valid experts.

Currently, [Harcourt Educational Measurement](#) and [CTB/McGraw-Hill](#) each have control of 40 percent of the test-design market, while [Riverside Publishing](#) controls the last 20 percent (PBS, *Frontline*). Should we trust corporate America with the development of assessments that can make or break our youth? What amount of standard setting for state-administered examinations has input from teachers, parents, and other concerned constituents? Federal oversight could answer these questions.

1. Test Secrecy

In 1974, Congress enacted the [Family Educational Rights and Privacy Act](#) (FERPA), also known as the Buckley Amendment (Looney, 2004). A key component of this act ensured that parents, legal guardians, and students themselves (depending on the circumstances) would have the right to inspect and review a student's records in a timely fashion. This act establishes appropriate access to student records as a fundamental right within educational contexts. Students should be allowed to view their graded standardized tests. Furthermore, students, parents, teachers, and concerned citizens should be aware of the processes involved with setting performance levels.

Unfortunately, testing industries have historically been clandestine. For decades, testing companies would not release student test booklets after scoring was complete. The rationale was based on testing companies reusing items from one administration to the next. Obviously, if they release the actual graded test to the student, they would be unable to re-circulate questions back into the test item bank. Furthermore, test companies are unwilling to dissect student exams in order to identify standards that need to be re-taught; it would be quite costly. Despite these concerns, however, *not* informing students of missed items/content standards contradicts the

purpose of assessment. Wrong answers imply a lack of knowledge on given domains. If students do not know what items they missed, they are unable to relearn un-mastered content.

Although progress has been made in some areas of testing, the transparency of state performance standards remains elusive throughout America (see figure 6). As previously discussed, the cut score process is based on judgment. Although it uses methodology that has been researched and widely accepted, it is still a process centered on judgment. Concerned stakeholders, especially parents, have the right to be involved in the process or, at the very least, be aware of its components. If [FERPA](#) can ensure families access to information that concerns students, surely a federal monitoring agency can ensure transparency of a process that can make or break students' lives.

2. Testing Community Addresses Unintended Consequences

Although it may not be common knowledge, the standardized testing industry is held to industry standards. The American Education Research Association, American Psychological Association, and the National Council for Measurement in Education (1995) co-authored *Standards for Educational and Psychological Testing* in an effort to govern test practices.

Figure 6: State Transparency Ratings: Is Information About Cut Scores Available on the Internet?

Transparent (Information Available on the Web)			Non-Transparent (Very Difficult To Locate/Unavailable On The Web)	
Alaska	Maine	Ohio	Alabama	Nebraska
Arizona	Maryland	Oregon	Florida	Nevada
Arkansas	Massachusetts	Pennsylvania	Georgia	North Carolina
California	Minnesota	South Carolina	Hawaii	North Dakota
Colorado	Mississippi	South Dakota	Idaho	Oklahoma
Connecticut	Montana	Tennessee	Illinois	Rhode Island
Delaware	New	Texas	Iowa	Utah
Indiana	New Jersey	Virginia	Kansas	Vermont
Kentucky	New Mexico	Washington	Michigan	West Virginia
Louisiana	New York	Wisconsin	Missouri	Wyoming

Source: State Department of Education Web sites; adapted from Education Sector. (2006, July). [Making the cut: How states set passing scores on standardized tests](#). Washington, DC: A.J. Rotherham.

As a result of the *Standards*, the same joint committee on testing practices convened to establish the [Code of Fair Testing Practices in Education](#) ([American Psychological Association](#), 2004).

The purpose of the code is to discuss the obligations of test developers in creating fair tests for different educational contexts and to address those issues that affect the proper use of examinations. Several points are worthy of examination.

The *Code* encourages test developers and users to avoid unintended consequences at all costs. They instruct developers to ensure test appropriateness for examinees of different ethnic and linguistic backgrounds. Meanwhile, they state that test users should ascertain test appropriateness by deciphering the test content and norm group. In regards to cut scores, the *Code* says that developers should provide information that will be helpful in the standard setting workshops and users should explain cut score determinations to the general public ([American Psychological Association](#), 2004). The [Joint Committee on Testing Practices](#) recognizes that it takes a village to properly educate, and assess, today's students.

B. A Federal Task Force is Needed

An introduction to standards and standard setting, as well as several legal examples of standard setting processes gone wrong, have both served as platforms to promote the notion of federal monitoring of test practices. Some education organizations have lobbied for clearly defined and reasonably structured federal objectives. I propose that an increase of federal oversight is needed to limit the disparate racial impacts of cut score processes.

The [Center on Education Policy](#) (2001) has reported that although federal contributions to a state's education are minor, [NCLB](#) now guides much of the decisions made about today's classrooms. Unfortunately, the new mandates force states to meet approval deadlines that

encourage inappropriate behavior (e.g. arbitrary cut scores). Successful development and implementation is compromised for expediency. Our nation cannot afford large-scale mistakes in public school accountability programs. Therefore we need a bipartisan, federal agency that monitors state testing procedures.

The [Center on Education Policy](#) (2001) also outlines four core principles for an improved federal role in education. Principle one states that “the federal government should continue to encourage high academic standards, but should also demand meaningful accountability from the states for increased student achievement *and accept national responsibility to help in the proper use of tests*” [emphasis added] ([Center on Education Policy](#), 2001). The report suggests several ways for an improved federal role in education. One suggestion is to establish a bipartisan agreement around education issues. In other words, Democrats and Republicans should agree on educational problems and best practice solutions. Political affiliation should be irrelevant when deciding what is best for America’s public school students. Each state will have to compromise, but the winners will be millions of children ([Center on Education Policy](#), 2001).

This report proposes a plan that would lay the groundwork for this federal taskforce. It suggests that an objective national group work with states to determine performance levels and identify appropriate test measures. This objective group could be “contracted” by the federal government, along with psychometricians, statisticians, subject matter experts, and school administrators. The process depends on bipartisan support, with input from many stakeholders.

Education legislation prior to [NCLB](#) included this kind of task force.⁷ The [National Education Standards and Improvement Council](#) (NESIC) was established as a part of the [Goals 2000: Educate America Act](#) (www.ed.gov). The central purpose of NESIC was to identify areas

⁷ In 1994 President Bill Clinton passed Goals 2000: Educate America. See www.ed.gov

where specific content standards needed to be developed. States had a financial incentive to adopt the recommended content standards, thereby creating the *standards movement* (Mulcahy, 1994). This council proved that the creation of a bipartisan council regarding education issues could be successful. NESIC encouraged most states to implement defined content standards in key subject areas. Unfortunately, NESIC was dismantled after [NCLB](#) replaced Goals 2000. Although testing issues (such as alignment with standards and assessments) had been a part of the council's goals, they were unable to thoroughly address the issue due to the death of the legislation.

C. Learning from the Past

The groundwork has been laid to show why a federal taskforce is not only beneficial, but necessary. Concerns, however, may exist about federalism. Does Congress have the enumerated power to create such a task force? What previous legislation serves as a model for a proposed federal test monitoring agency? In this section, I will answer these questions.

1. Federalism Concerns

Federalism is the division of government between the national, state, and local levels. Some critics may claim that the creation of a federal agency that monitors education testing might violate federalism. Constitutional restraints require the federal government to carefully consider whether federal intervention is within proper authority. Several legislation, including the national truth-in-testing law, the commerce clause, and the Belmont Report, quiet criticism about the creation of a federal taskforce. The notion of checks and balances is also addressed.

a. [National truth-in-testing legislation.](#)

In 1980, Representative Ted Weiss (NY), Shirley Chisholm (NY), and George Miller (CA) sponsored The Educational Testing Act, also called national truth-in-testing legislation. The purpose was to encourage greater accountability to the public and facilitate accessibility to and innovation in the testing process (Weiss, 1980). Specifically, the bill would have compelled testing organizations to provide more information on testing procedures.

Weiss listed several reasons why a national truth-in-testing legislation was needed. One reason was the unintended consequences of cut scores. It is possible that a federal testing taskforce could bypass federalism concerns by requiring testing companies, not state departments of education, to disclose the standard setting process. State departments of education convene standard setting panels, but psychometricians from test publishing companies are consultants that usually organize the panels. Once the panel has decided on a final cut score for the given assessment, the report is then delivered to the state department of education for final approval. Under the national truth-in-testing legislation, the federal government would mandate testing companies for full disclosure, not state agencies, thereby addressing potential federalism concerns.

b. *[The Belmont Report](#)*.

The National Research Act (Pub. L. 93-348) was signed into law in 1974. As a result, it created the National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. The primary charge of the commission was “to identify the basic ethical principles that should underlie the conduct of biomedical and behavior research involving human subjects and to develop guidelines which should be followed to assure that such research is conducted in accordance with those principles” (www.hhs.gov/ohrp/humansubjects/guidance/belmont.htm). In other words, entities receiving

monies from the federal government are required to follow basic principles when doing certain research involving humans. [*The Belmont Report*](#) was a result of the commission's findings and is interwoven throughout the act.

Similarly, a federal agency that monitors testing practices would identify the basic fairness principles that underlie educational testing involving students. The taskforce would develop testing guidelines that would address standard setting, along with other state-administered testing procedures (similar to the Belmont Report and the Standards for Educational and Psychological Testing).

[*The Belmont Report*](#) addresses (1) the boundaries between biomedical and behavioral research and the accepted practice of medicine, (2) the role of assessment in risk-benefit criteria in determining appropriateness of research involving human subjects, (3) appropriate guidelines for selection of human subjects for research participation, and (4) the nature and definition of informed consent in different research settings. There is a long, detailed process researchers must complete in order to gain approval for proposed projects (called the Institutional Review Board process). Once approval is gained from the IRB, researchers are bound to follow the Belmont principles; if they violate the principles in any way, the IRB immediately discontinues the research.

The federal agency for the monitoring of testing practices would model [*The Belmont Report*](#). The federal taskforce would address (1) the boundaries between teaching and assessment, (2) the appropriateness of standard setting methods (3) the appropriate selection of testing instruments, and (4) the nature and definition of unintended consequences in standardized testing. This model has clearly worked in the research context. It is certainly worthwhile to try it in the standardized testing context.

2. Checks and balances vs. local control

Standard setting is simply one aspect of standardized testing that is prone to mistakes. Every year there are countless examples of errors in testing due to creation, dissemination, and scoring (see www.fairtest.org). A system of checks and balances must be in place to ensure that our nation's public school students receive fair opportunities to demonstrate their knowledge. A federal agency that monitors processes involved in standardized testing is one step in the right direction.

Critics, however, may argue that a federal testing taskforce would undermine states' authority over public education. The 10th Amendment states that powers not delegated to the United States by the Constitution, nor prohibited by it to the States, are reserved to the States respectively, or to the people (<http://caselaw.lp.findlaw.com/data/constitution/amendment10/index.html>). Therefore, education is under state, not federal control.

Legislation over the past 15 years, however, shows an increasing federal role in public education. [The Improving America's Schools Act of 1994](#) (IASA) amended the [Elementary and Secondary Act of 1965](#) (Pub. L. 103-382). Essentially, it offered financial incentives for states to institute more rigorous content standards. Most of the Act's mandates were voluntary, thereby limiting federal involvement. The [No Child Left Behind Act of 2001](#) increased federal involvement by requiring states to assess children regularly, make annual achievement gains, and offer families a way out of failing schools.⁸ States that do not comply will lose federal funding.

Although a federal taskforce that monitors state-administered testing processes may seem to violate the 10th Amendment, in actuality it is following in the footsteps of previous legislation.

⁸ United States Department of Education. (n.d.). Elementary and Secondary Education Act (The No Child Left Behind Act of 2001). Retrieved December 11, 2006, from <http://www.ed.gov/policy/elsec/leg/esea02/index.html>

Most states have higher academic standards because of the [National Education Standards and Improvement Council](#) (created as a result of IASA). States were not performing well on national assessments and the federal government stepped in to improve the situation. Currently, some states are performing well on state-administered assessments but poorly on national assessments, while other states are failing state-administered exams but scoring high on national tests. It is clear that cut scores need to be adjusted for positive, *and negative*, inflation.

Critics may also claim that the federal government cannot monitor test processes because it is outside of their expertise. This argument is certainly true. Therefore, the federal taskforce would only be *commissioned* by the [United States Department of Education](#). The group of “monitors” would include experts in the field of education, administration, and testing. Psychometricians and subject matter experts would comprise the majority of the group.

Additionally, critics may claim that there are not enough financial resources to implement such a taskforce. Education budgets are already overspent at the federal and state levels. Expenditures such as taskforce members’ salaries, travel expenses, and incidentals would require monetary allocations that are simply too much to bear. A federal taskforce/agency is not a wise financial investment.

Proponents of the testing taskforce could give several responses to the financial challenge. First, [NCLB](#) has been under-funded from its inception. More federal monies are needed in order to ensure a fair, equitable, and successful education system. Second, federal education evaluators could add taskforce-related responsibilities to their already existent contract. Currently, the federal government must approve states’ accountability system. The employees involved in that process can simply shift foci to examine standardized testing processes. Finally, expenses can be limited if test publishing companies were required to submit

standard setting processes to the state *and* federal departments of education. Members would not have to travel, significantly reducing fact-finding trips to state departments of education.

Presently, psychometricians (on behalf of standard setting panels) only submit final recommended cut scores to the state. In order to make the process more transparent, they should be required to submit recommendations to the federal government, thereby decreasing taskforce expenditures.

The last criticism is that standardized testing processes are not flawed enough to warrant federal action. Opponents may think that standard setting is a judgment-free process that does not need monitoring. On the other hand, critics may agree that standard setting is a process centered on judgment, but completely trust the people chosen to make those tough decisions.

Unfortunately, cynics would disagree with both arguments. First, the experts in the field of psychometrics all agree that standard setting is a process based on judgments. Although great effort has been invested in trying to standardize the process, it still must rely on human judgment. Second, with human judgment at the center of this high-stakes process, error is inevitable. The only way to substantially lower the risk of error is to monitor the process. It is not enough to “trust” the experts, because even the most notable people can make erroneous decisions.

Furthermore, education is not the only field that standard setting affects. There are countless lawsuits that have challenged professional licensure exams that affect teachers, police officers/firefighters, and medical personnel.⁹ In addition, university entrance examinations (including SAT, ACT, and LSAT) have also been challenged on ground of disparate racial impact. Although examinees affected by inappropriate standard setting processes cover the

⁹ This paper explored a few cases. For more information visit www.westlaw.com

spectrum, it is always possible for teachers to get another job or potential law students to choose another career path. Children, however, in America's public elementary and secondary schools are left with no options once they are failed by a standard setting process gone wrong.

Children fail high-stakes examinations for numerous reasons. The federal government offers children with low academic ability tutoring services. They attempt to address low teacher quality by mandating minimum teacher qualifications. They also attempt to offer children that attend deteriorating schools with low resources an opportunity to attend a better school. The federal government is clearly willing to address reasons for student failure. Standard setting processes, although not always unsuccessful, have inherent flaws that can fail students. Creating a federal monitoring taskforce that monitors standardized testing procedures is the best way to combat unintended consequences of this high-stakes process.

References

- American Psychological Association. (2004). *Code of Fair Testing Practices in Education*. Washington, DC: Joint Committee on Testing Practices.
- Carpenter, S. (2001). The high stakes of educational testing [Electronic version]. *APA Monitor* 32, 5.
- Center on Education Policy. (2001, February). *An education agenda for the congress and the new administration*. Washington, DC: J. Jennings.
- Cizek, G.J., Bunch, M.B., Koons, H. (Winter 2004). An NCME instructional module on setting performance standards: Contemporary methods. *Educational Measurement: Issues and Practice*.
- Clarke, M., Madaus G., Pedulla J., & Shore, A. (2000). *An agenda for research on educational testing* (NBETPP Statements Series Vol. 1, No. 1). National Board on Educational Testing and Public Policy: Boston College, Lynch School of Education.
- Cureton v. National Collegiate Athletic Ass'n. 37 F.Supp.2d at 687 (E.D.Pa., 1990).
- Education Sector. (2006, July). *Making the cut: How states set passing scores on standardized tests*. Washington, DC: A.J. Rotherham.
- FairTest Examiner. (Spring 2002). *Arizona, New York tests made open*. Retrieved June 25, 2005, from www.fairtest.org/examarts/Summer%2002/Foul%20Ups.html
- FairTest Examiner. (Summer 2003). *Testing errors continue*. Retrieved June 25, 2005, from www.fairtest.org/examarts/fall%2003/More%20Testing%20Errors%20.html
- Guardians Ass'n of New York City Police Dept., [Inc.](#) v. Civil Service Commission of the City of New York. 630 F.2d 79 (C.A.N.Y., 1980).

- Horn, C., Ramos, M., Blumer, I., & Madaus, G. (2000). *Cut scores: Results may vary* (NBETPP Monograph Vol. 1, No. 1). National Board on Educational Testing and Public Policy: Boston College, Lynch School of Education.
- Karantonis, A. & Sireci, S.G. (2006). The bookmark standard-setting method: A literature review. *Educational Measurement: Issues and Practice*, 25(1), 4-12.
- Linn, R.L (2002). Validity of the uses and interpretations of results of state assessment and accountability systems. In G. Tindal & T.M. Haladyna (Eds.), *Large scale assessment programs for all students: Validity, technical adequacy, and implementation* (pp.27-48). Mahwah, NJ: Erlbaum.
- Linn, R.L., & Miller, M.D. (2005). *Measurement and assessment in teaching* (9th ed.). Upper Saddle River, NJ: Pearson.
- Livingston, S.A. & Zieky, M.J. (1982). *Passing scores: A manual for setting standards of performance on educational and occupational tests*. Princeton, NJ: Educational Testing Service.
- Looney, S.D. (2004). *Education and the legal system: A guide to understanding the law*. Upper Saddle River, NJ: Pearson.
- Mulcahy, D.G. (1994, November). *Goals 2000 and the role of the national education standards and improvement council*. Paper presented at the annual conference of the American Educational Studies Association, Chapel Hill, NC.
- Phillips, S.E. (2000). *GI Forum v. Texas Education Agency: Psychometric evidence*. *Applied Measurement in Education*, 13(4), 343-385.
- PBS, Frontline. (n.d.). *The testing industry's big four*. Retrieved December 5, 2005, from www.pbs.org/wgbh/pages/frontline/shows/schools/testing/companies.html

- Quillin, M., & Kurtz, M. (1997, August 1). Johnston Schools Sued over Testing. *The Raleigh News and Observer*, B-1.
- Raymond, M.R. & Reid, J.B. (2001). Who made thee a judge? Selecting and training participants for standard setting. In G.J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives*. (pp.119-158). Mahwah, NJ: Erlbaum.
- Richardson v. Lamar County Bd. Of Educ. 729 F. Supp. 806 (M.D. Ala. 1990).
- Ryan, J.E. (2004). The perverse incentives of the No Child Left Behind Act. *New York University Law Review*, 79(3), 932-989.
- Shore, A., Pedulla, J., & Clarke, M. (2001). *The building blocks of state testing programs* (Statements Series Vol. 2, No. 4). National Board on Educational Testing and Public Policy: Boston College, Lynch School of Education.
- Smith, D. (2001). Is too much riding on high-stakes tests? [electronic version]. *American Psychological Association Monitor*, vol. 32, no. 11.
- Weiss, T. (1980). National truth-in-testing legislation. *The Journal of Negro Education*, 49(3), 233-237.
- Wu, E. (2005, October). U.S. falls in education rank compared to other countries. *The Kapio Newspress*. Retrieved December 5, 2005, from <http://kapio.kcc.hawaii.edu/upload/fullnews.php>
- Zieky, M.J. (2001). So much has changed: How the setting of cutscores has evolved since the 1980s. In G.J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives*. (pp.19-52). Mahwah, NJ: Erlbaum.
- Zieky, M.J. & Perie, M. (2006). *A primer on setting cut scores on tests of educational achievement*. Princeton, NJ: Educational Testing Service.