DOCUMENT RESUME

ED 478 077 TM 035 018

AUTHOR	Wang, LihShing; Pan, Wei; Austin, James T.
TITLE	Standards-Setting Procedures in Accountability Research: Impacts of Conceptual Frameworks and Mapping Procedures on Passing Rates.
SPONS AGENCY	Department of Education, Washington, DC.; Ohio State Dept. of Education, Columbus.
PUB DATE	2003-04-00
NOTE	60p.; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, IL, April 21-25, 2003).
CONTRACT	OSP02153
PUB TYPE	Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE	EDRS Price MF01/PC03 Plus Postage.
DESCRIPTORS	*Accountability; *Cutting Scores; Educational Research; *Pass Fail Grading; *Standard Setting (Scoring)
IDENTIFIERS	*Mapping

ABSTRACT

Standard-setting research has yielded a rich array of more than 50 standard-setting procedures, but practitioners are likely to be confused about which to use. By synthesizing the accumulated research on standard setting and progress monitoring, this study developed a three-dimensional taxonomy for conceptualizing and operationalizing the various procedures: outcome versus growth assessment, theory-driven versus data-driven approach, and observed scale versus latent scale mapping.An empirical study is reported to illustrate how these various approaches can be implemented to meet the accountability challence in the No Child Left Behind era. Consistency analysis of 12 standard-setting procedures reveals vastly disparate pass/fail decisions among different procedures, even within the same conceptual framework or mapping operation. Particularly disturbing is the finding that the passing rate may jump from as low as 29% to as high as 79%, depending on whether the standard is mapped to the observed-score scale or the latent-score scale. Implications and future directions for policy makers, school officials, and psychometricians are discussed. (Contains 9 tables and 107 references.) (Author/SLD)

STANDARDS-SETTING PROCEDURES IN ACCOUNTABILITY RESEARCH:
Impacts of Conceptual Frameworks and Mapping Procedures on Passing Rates

LihShing Wang
University of Cincinnati

Wei Pan
University of Cincinnati

James T. Austin
Ohio State University

Paper presented at the annual meeting of American Educational Research Association, Chicago, April, 2003. This study is jointly supported by research grants from US Department of Education and Ohio Department of Education (Contract No. OSP02153). The authors wish to thank the sponsoring agencies for their support, but the conclusions drawn from this study do not necessarily reflect their views or opinions. The authors are also indebted to the following project team members (in alphabetical order) who have contributed to this study in various ways: Mary Boat, Rhonda Brown, Dawn Denno, Lawrence Johnson, Catherine Maltbie, Sandy Miller, David Sacks, Jerry Walker, and Debra Zorn. Correspondence should be addressed to LihShing Wang, Educational Studies, University of Cincinnati, Cincinnati, OH 45221, USA, email: Leigh.Wang@uc.edu.

Abstract

Three decades of standard-setting research has yielded a rich array of more than fifty standard-setting procedures. When turning to this wealth of literature for methodological guidance, however, practitioners are likely to be confused as to which procedure to follow. By synthesizing the accumulated research on standard setting and progress monitoring, this study proposes a three-dimensional taxonomy for conceptualizing and operationalizing the various procedures: *outcome* vs. *growth assessment*, *theory-driven* vs. *data-driven approach*, and *observed-scale* vs. *latent-scale mapping*. An empirical study is reported to illustrate how these various approaches can be implemented to meet the accountability challenge in the *No Child Left Behind* era. Consistency analysis of twelve illustrative standard-setting procedures reveals vastly disparate pass/fail decisions among different procedures, even within the same conceptual framework or mapping operation. Particularly disturbing is the finding that the passing rate may jump from as low as 29% to as high as 79%, depending on whether the standard is mapped to the observed-score scale or the latent-score scale. Implications and future directions for policy makers, school officials, and psychometricians are discussed.

Standards-Based Accountability Reform

Historical Perspective on Education Reform

In the U.S. history of education, the ideology of *education reform*, *standards*, and *accountability* goes a long way back. Starting as early as the *Committee of Ten* in the late 1800s, the National Defense Education Act in the 1950s, the *Elementary and Secondary Education Act* of 1965, to the more recent *A Nation at Risk* of 1983, the *America 2000* of 1991, the *Goals 2000* of 1994, all the way down to the *No Child Left Behind* Act (NCLB) of 2001, the American education history is replete with admirable pursuits for better education for all (Horn, Jr., 2002; Vinovskis, 1996).

What makes the NCLB Act an unprecedented reform endeavor, however, is its forceful provisions on using state-mandated testing to hold schools accountable for helping their students attain prescribed standards and attach high-stakes consequences to the outcomes of closing achievement gaps (The White House, 2001). According to Secretary of Education Rod Paige (2003), "this is a very tough law.... There are consequences for inaction, and visibility for performance" (Education USA, 2003).

Another unique feature of this NCLB Act is its ambitious goal to bring all children, irregardless of their race, class, or disability status, to the same rigorous standards within a given period of time. In the Executive Summary of the NCLB Act, it states, "These systems must be based on challenging State standards in reading and mathematics, annual testing for all students in grades 3-8, and annual statewide progress objectives ensuring that all groups of students reach proficiency within 12 years."

Although these reform endeavors were mostly contextualized in elementary and secondary education, they have exerted far-reaching impacts on both preschool and higher education. To illustrate, the Ohio Legislature (1997) passed a bill that mandated early childhood programs administered by the State to use a standardized assessment system as a common tool for comparing and tracking student achievement (cited in Brown, Johnson, Pretti-Frontczak, & Kowalski, 2001, p. 1). Another estimated fourteen States and the District of Columbia have implemented pre-kindergarten standards and accountability systems (Costello & Zarowin, 2002). Such evaluation efforts, starting as early as the preschool level, are critical not only for improving program quality (Gilliam & Zigler, 2001) but also ensuring that school children are provided with a solid head start in their life-long learning journey (Shepard, Kagan, & Wurtz , 1998). The cumulative impacts of the various reform movements have also reached higher education as well as non-education disciplines such as medicine (Searle, 2000) and environmental study (Jasanoff, 1998).

Central to this most recent wave of accountability movement is the endorsement of rigorous research on providing credible evidence of program accountability in helping all children meet "challenging content and performance standards" and in making reward/sanction decisions based on "adequate yearly progress (AYP)" (US Department of Education, 2002). According to this law, the States must develop at least three levels of exit performance standards toward the target content standards: *partially proficient*, *proficient*, *advanced*; and the ultimate goal is to help all children attain at least the *proficient* level by the time they exit their programs. In order to reach that ultimate goal, the States must monitor the yearly progress of low-performing students in order to ensure that they are making *continuous* and *substantial* progress toward the exit performance standards within an *appropriate* time frame.

Current Issues of the New Accountability Movement

Although stakeholders generally applaud the noble goal of *leaving no child behind* in their schooling journey, the implementation of the NCLB Act has met with much confusion, skepticism, disgruntlement, and even opposition (e.g., Amrein & Berliner, 2002; Burd, 2002; Center on Education Policy, 2003; Kucerick, 2002; McNeil, 2000; Olson, 2002; Shepard, 2002-2003). Some of the debates center on the consequential validity of the standards-based reform. For example, Darling-Hammond (1997) argued that standards should be used as "guideposts" to mobilize system resources, rather than "straitjackets" to punish students and schools (p. 213). Others questioned the validity of the testing instruments used to assess student achievement and school effectiveness. As Sternberg (2003) puts it, "… the current emphasis on narrow accountability may inadvertently straitjacket teachers and schools in what and how they teach, and implicitly devalue wisdom and the responsibility that is concomitant with it because these values are not measured on the tests" (p. 5). Still others confront the ideological dimensions of searching for social justice by holding students of all races and classes to the same rigorous standards. Leonardo (2003), for example, questioned whether the reform process is "democratic in nature" by holding schools accountable without hearing voices from disadvantaged groups or addressing larger structural issues (p. 40).

Summarizing the past five decades of accumulated evidence on high-stakes testing, Linn (2000) concludes that “the unintended negative consequences of the high-stakes accountability uses often outweigh the intended positive effects” (p.14), a view shared and reinforced by a more recent and comprehensive study by Jones, Jones, and Hargrove (2003, p. 172). The American Evaluation Association (2002) even issued a position statement that unequivocally downplays “the use of tests as the sole or primary criterion for making decisions with serious negative consequences for students, educators, and schools.” This position statement clearly runs counter to the NCLB Act because the mandated state assessment is being used as the “primary criterion” for making decisions that may carry “serious negative effects” for all constituents involved.

While accountability-driven assessment continues to be a dominant theme in the public dialogue among policy makers and evaluation researchers, state officials and school administrators are facing the formidable task of solving the puzzle on their own. At least three issues related to standard setting have been raised: *lack of methodological guidance*, *controversies in standard-setting procedures*, and *narrow focus on student outcome assessment.*

Lack of methodological guidance. Much of the chaos in the accountability movement is associated with the lack of methodological guidance on how to set passing performance standards, both for the content knowledge proficiency at the student level, and for the Adequate Yearly Progress at the school level as mandated by the Act.

To answer the question: “How good is good enough?” (National Council on Educational Standards and Testing, 1992, cited in Shriner & Ysseldyke, 1994), over two decades of standard-setting research has yielded more than fifty standard-setting procedures. An up-to-date synthesis on the different standard-setting procedures available can be found in Zieky (2001). Included in this body of literature are the original Angoff method (Angoff, 1977) and its numerous variants, the Contrasting-Groups approach (Livingston & Zieky, 1982), the Briefing-Book method (Haertel, 2002), the Bookmark procedure (Mitzel, Lewis, Patz, & Green, 2001), the Generic Eclectic method (Berk, 1996), the Behavior-Anchoring approach (American College Testing, 1993), which was later elaborated into the Domain-Score approach (Schulz, Kolen, & Nicewander, 1999). As Kane (1998) aptly puts it, "we have an embarrassment of riches" in the large number of available standard-setting procedures (p. 137).

Unfortunately, other than that the standards must be *criterion-referenced*, which is consistent with the criterion-referenced measurement movement in the 1970s and 1980s (Berk, 1980) and still largely embraced by standard-setting researchers well into the 21st century (Cizek, 2001, p. ix), the States are left on their own to figure out the complex task of setting performance standards. Existing guidelines on state assessment (e.g., Malcom, 1993; Shepard, Kagan, & Wurtz, 1998; US Department of Education, 2003) only provide general principles on setting standards without prescribing which method to use or detailing how a standard should be set.

This lack of methodological prescription should not be surprising because numerous researchers have warned of the non-existence of a "true" standard or the "best" standard-setting procedure. Even the *Standards for Educational and Psychological Testing* only delineates that the standard-setting procedure be "clearly documented" (*Standard 4.19*, AERA, APA, & NCME, 1999, p. 59).

Controversies in standard-setting procedures. This lack of methodological guidance for setting standards is further compounded by the many controversies in standard-setting procedures. The National Assessment of Educational Progress (NAEP), employing perhaps the most sophisticated standard-setting procedure to date, reports student performance in three achievement categories: *Basic*, *Proficient*, and *Advanced* (Hambleton, 1998a; Reckase, 1998). However, its standard-setting methodology was criticized and widely publicized by the National Academy of Sciences (NAS) as "fundamentally flawed" because it failed to identify cut scores that produce consistent results and communicate useful information and reasonable expectation about student achievement (Pellegrino, Jones, & Mitchell, 1999). This criticism was quickly refuted by Hambleton et al. (2000) as one that is "not only just unwarranted—it is simply wrong" (p. 13) and refuting the NAS review as "clearly failed to produce a trust-worthy and credible evaluation" (p. 13). Aside from the technical complexity involved in the debate between these two reputable parties, the sheer sophistication of the scaling procedure used for mapping the performance standards to the score scales would have proved a big puzzle for practitioners to follow.

Adding another layer of controversy to this debate is Linn, Baker, and Betebenner's (2002) study that reported a disturbing discrepancy between state assessment results and NAEP categories in seven out of eleven States. Similar studies (e.g., Koretz, Barron, Mitchell, & Stecher, 1996) have found large increases in state assessments over time and small changes in national assessments such as NAEP. Whether such changes in state assessment scores represent real or superficial changes in student achievement is of course an intriguing question related to validity (e.g., Mehrens, Popham, & Ryan, 1998), but given the dissimilar nature of state vs. national assessment, it is generally recognized that a defensible external criterion for establishing the validity of a state assessment may be difficult to come by (Lissitz & Bourque, 1995, p.21), and therefore caution is needed to consider "the match in the content between the state and national tests, which can vary substantially over grades" (Schwarz, Yen, & Schafer, 2001, p. 28) . As Tindal (2000) aptly puts it, "such an emphasis on concurrent validity results in an endless spiral with no eventual resolution" (p. 3).

Another controversy in standard-setting literature is how high the bar should be set. The National Council on Education Standards and Testing (1992) argued for rigorous standards because "in the absence of demanding content and performance standards, the United States has gravitated toward having a de facto minimal skills curriculum. ... Such low expectations shortchange students and ill-serve the country" (p. 12). But how high is high enough? Some educational measurement literature endorses 80% as the content mastery cutoff (Mislevy & Bock, 1990, p.1-23). Ebel (1965) noted that some "percent of 'perfection' usually between 60 and 75 percent was ordinarily regarded as the minimum passing score" (p. 406). According to Linn, Baker, and Betebenner (2002), the original House and Senate versions of NCLB set the school-level AYP targets at one percentage of increase in the number of students achieving the *Proficient* level (p. 7). This seemingly modest annual gain was found to be not only realistically unattainable but also statistically unreliable (p. 12). Using the Kentucky state assessment data and the standard deviation as guidepost, Kifer (2001) reported that it would take 17 to 47 years for all the schools to reach the target (p. 86)!

Where to set this cutoff standard is a highly arbitrary and politically-charged issue. According to Wiliam (1996), any attempt to define a complex set of relationships between many variables as a single point necessarily entails a degree of arbitrariness, requiring the resolution of conflicting priorities. Kane (1994) also warned us that the tradition of requiring 70% correct is especially arbitrary because we can easily manipulate item difficulty to raise test scores. According to a disturbing but unconfirmed report by Morse (2002), "Ohio 'refined' its criteria for calculating low-performing schools; afterward, the number receiving Fs fell from 760 to 200" (p. 426)!

Perhaps the most disturbing controversy in standard-setting practice is the consistent finding that each standard-setting procedure produces a different passing standard, which in turn translates to a different student passing rate (Berk, 1996) and a different school rating (Olson, 1998). Studies comparing different standard-setting procedures applied to the same test data invariably yielded incomparable passing standards and inconsistent classification results (Jaeger, 1989). For example, in a study that compared three different growth standards, Linn and Baker (1999) found very different results in reaching the standard. "The most consistent finding from the research literature on standard setting is that different methods lead to different results" (National Academy of Education, 1993, p. 24).

Some researchers justified the disparate impact data by claiming that because different methods place emphasis on different aspects of performance, the discrepancies arise naturally out of the manifold interpretations that are placed on test results (e.g., Hambleton, 1980). Others expressed a more pessimistic view on the lack of consistency among different standard-setting procedures. As Kifer (2001) puts it, the

States have "idiosyncratic assessments and arbitrary standards to use for growth targets" and "arbitrary proficiency standards may have been turned into capricious growth targets" (p. 85).

Narrow focus on student outcome assessment. The above issue of unrealistic expectation for annual progress brings us to the distinction between two types of assessment: *outcome* (or *status*) assessment and *growth* (or *progress*) assessment (Kifer, 2001, p. 57). Outcome assessment affixes a score that purports to reflect a person's knowledge at the time of assessment. Growth assessment monitors differences in achievement over time and describes what has been learned in the process.

Unfortunately, the methodology for studying growth has received very little attention in the standard-setting literature, probably because the NCLB Act focuses on content mastery rather than individual growth. The AYP refers to the progress of schools in increasing the number of students reaching the static outcome standard. As such, the AYP is only a progress measure of school effectiveness capitalizing on the outcome measure of student achievement. Some researchers even go to the extreme of arguing for the focus on outcomes rather than inputs or processes (Musick, 1997, cited in Kifer, 2001, p. 30). Most studies that do look at growth are cast at the school level (i.e., percentages of students achieving the passing standards) rather than the student level (e.g., Guskey, 1994; Linn, Baker, & Betebenner, 2002; Millman, 1997 Schwarz, Yen, & Schafer, 2001).

Many researchers have argued against such narrow focus on student outcome assessment. Camilli, Cizek, and Lugg (2001) contends that "a broader conceptualization of performance standards validation suggests assessments that are even more focused, more individualized, and more consequential for individual students" (p. 470). Popham (1999) also expressed concerns over the preoccupation of large-scale assessment with accountability and too little attention on instruction. He argued convincingly that

> According to Aristotle, the most repugnant metaphysical evil is "unactualized potency"—that is, the failure of an entity to achieve its inherent potential. Because I am certain that our children can be learning more effectively, I am metaphysically miffed that large scale assessment is doing so little to foster improved educational quality" (Popham, 1999, p. 14).

This redirection of standards-based reforms toward individual student growth has found many voices in recent accountability literature. Yen and Henderson (2002) argued for the use of large-scale state assessments in making instructional decisions for individual students. Buly and Valencia (2002) found distinctive and multifaceted patterns of student's reading abilities that require dramatically different instructional emphases. Darling-Hammond (1997) pointed out the importance of building the *feedback* mechanism to identify needs and progress and to provide internal self-correctives that support continual improvement (p. 222 and p. 245). Shepard, Kagan, and Wurtz, (1998) recommended that procedures be put in place to monitor student progress using instructionally relevant assessments (p. 31). Linn, Baker, and

Betenenner (2002) proposed using "index scores" to monitor individual student progress that would recognize all students' progress rather than only those who are at the borderline of passing the standard (p. 15). In fact, an entire session in the 2003 AERA/NCME annual meeting was devoted to "value-added approach to school accountability" that owes much of its development to Sanders, Saxton, & Horn's study (1997) which describes changing performance of the same individuals over time after adjusting for differences in background variables such as social-economic status and school resources.

In summary, when one turns to the standard-setting literature for guidance, he or she is likely to be appalled by the large array of standard-setting methods, confused by the continuous debate on the "golden standard" for standard-setting, and yet at the same disappointed by the narrow conception of outcome performance. Most standard-setting methods to date operate within the criterion-referenced framework using the number-correct raw score or linearly derived normative score without taking into consideration either the standard error of measurement or the learning progress from entry to exit. A systematic investigation of the various standard-setting methods cast in the enlarged perspective of "educative assessment" (Wiggins, 1998) and how their conceptual frameworks as well as operational procedures affect the consequences data seems to be in order.

Purpose and Significance of Study

The purpose of this study is two-fold: (a) to propose a new paradigm for categorizing the various standard-setting methods that encompasses a broadened perspective of accountability assessment, and (b) to compare different conceptual frameworks and standard-mapping procedures on the resulting passing rates for empirical state assessment data.

To accomplish Purpose (a), this study revisits the standard-setting literature, summarizes the major findings to date, proposes new methods to complement existing ones, and describes in detail how each method is operationalized. To accomplish Purpose (b), we use a potentially high-stakes state assessment dataset to illustrate how these conceptual frameworks and mapping procedures can be implemented, how different standard-setting methods affect passing rates, and how to select the procedures that are most consistent with the intended interpretation as dictated by the policy and most appropriate as recommended by modern psychometric theory.

The proposed taxonomy of standard-setting procedures operates in three dimensions: *outcome assessment* vs. *growth assessment*; *theory-driven approach* vs. *data-driven approach*; and *observed-score mapping* vs. *latent-score mapping*. Within the general framework of *growth assessment*, four different conceptualizations of growth are explored: *growth referenced on total domain*, *growth referenced on entry score*, *growth referenced on individual gain,* and *growth referenced on normative gain.* Furthermore, operational procedures for mapping the performance standard to the *observed-score scale* and the *latent-score scale* for each conceptual framework are described and compared. For the *latent-score mapping* procedure, three mapping procedures are investigated: *latent-scale range*, *cubic regression*, and *domain score.* Empirical evidence is provided on how these different conceptualizations and operationalizations of standard setting impact the passing rates of Early Math, Language and Literacy, and Social Development at the preschool level. The merits and limitations of each standard-setting procedure are examined, and the implications for meeting the NCLB accountability standards are also discussed.

This study provides the much-needed methodological guidance on how to select the appropriate conceptual framework and mapping procedure for standard setting. It goes beyond the current narrow focus on criterion-referenced outcome assessment to include a variety of growth assessments and norm-referenced assessments. The reader shall be amazed by the richness of the different conceptualizations and operationalizations available to them, and yet given helpful guidance on how to make the most sensible selection given the context and purpose of assessment.

This paper revisits the long-standing issue of sensitivity of different standard-setting methods and the well-documented inconsistency in the impact data. Although findings of inconsistency are to be expected, particularly when different conceptual frameworks are involved, the lack of consistency even among different operationalizations of the same conceptual definition still presents a potential threat to any validity arguments in high-stakes testing (Haertel, 1999). It should not be surprising, therefore, that among the set of criteria for evaluating standard-setting methods, "potential for replicability of the cutscore decision rule" (Plake, 1995, p.8) is still considered an important evaluation criterion. Hambleton (1980) even goes so far as to assert that decision consistency is the "acid test" of the worth of the standard setting procedure.

Toward a Taxonomy of Standard-Setting Procedures

What is Standard Setting?

Before we proceed any further, a brief statement on some of the key terms used in this study would be helpful. The standard-setting literature makes the distinction between *content standard* and *performance standard* (e.g., Hambleton, 1998b). A content standard refers to the strands of knowledge and skills that students are expected to know and do. A performance standard refers to the level of performance that students must demonstrate in relation to the content standards. Although standard setting may refer to setting content or performance standards, it is customary to reserve this term for setting performance standards only. This study also follows this convention.

A popular definition of setting performance standards is the identification of one or more cut points on the score scale representing varying degree of content mastery and attach a particular performance interpretation to the resulting scores (e.g., Wiliam, 1996). This definition is consistent with the NAEP's "achievement levels" defined by the National Assessment Governing Board (see Hambleton, 1998b). However, this definition focuses on content mastery to the exclusion of other conceptions of performance standards. Another more generic definition is offered by Cizek (1993) as "the proper following of prescribed, rational system of rules or procedures resulting in the assignment of a number to differentiate two or more conceivable states or degrees of performance." This definition affords us a broader conceptual framework for setting performance setting.

Kane (1994) made the distinction between setting a *performance standard* and setting a *cut score*. A performance standard, defined as the "minimally adequate level of performance for some purpose" (p. 425), is a written description of the knowledge or content that students must demonstrate to show that they meet a specified level of performance. A cut score, defined as "a point on the score scale" (p. 425) that corresponds to the performance standard, categorizes students into two groups above and below the cut point. The former refers to the conceptualization of the performance standard, and the latter to the operationalization of mapping that standard to the score scale. Although this view is not universally shared by measurement researchers, we find this distinction helpful in our study because our purpose is to illustrate how performance standards can be affected by the conceptual definitions and mapping operations involved in the standard-setting process. According to Kane (2001), "most standard-setting studies have focused on determining the cut scores rather than on defining the performance standards" (p. 66). This study examines both components--defining standards and determining cut scores--in the standard-setting process.

Classification Schemes

To categorize this large array of different standard-setting procedures, Wiliam (1996) proposed a two-dimensional classification scheme: one on *test-centered* vs. *examinee-centered* continuum, the other on *meaning-oriented* vs. *consequence-oriented* continuum. The test-centered vs. examinee-centered distinction is similar to the *criterion-referenced* vs. *norm-referenced* distinction and the *judgmental* vs. *empirical* distinction popular in the measurement field, but as Jaeger (1989), Hambleton (1980) and others have forcefully argued, all criterion-referenced assessments have a normative element and all standard-setting procedures are essentially judgmental. This study uses the *theory-driven* vs. *data-driven* distinction because it is consistent with the long-standing opposing strategies in modeling research, but it is similar to the test-centered vs. examinee-centered distinction popular in the standard-setting literature.

The meaning-oriented method adjusts the cut score to fit the items, and the consequence-oriented method adjusts the items to fit the cut score. We did not include this continuum in our classification scheme because our assumption was that all items have been carefully developed according to a pre-determined test specification table, content validity has been established for this carefully assembled test, and therefore only adjusting the cut score to fit the items would make logical and practical sense.

Kane (2001) suggested adding another dimension *holistic* vs. *analytical* to the classification scheme (p. 62). The holistic method sets standards based on an overall assessment of complete performances; whereas the analytical method sets standards based on small chunks of the overall performance as indicators of achievement. Again, we did not include this continuum in our study because we approached the taxonomy from a broader perspective and assumed that a certain proficiency level (e.g., 70% outcome mastery) has been established using one of the procedures recommended in the literature.

Summarizing and building on the above standard-setting procedures and classification schemes in the literature, we propose a new taxonomy that envisions a broader perspective on the various conceptual frameworks for defining performance standards and the different operational procedures for translating the performance standard to a cut point on the score scale. In addition to the theory-driven vs. data-driven dichotomy, we also propose two other dimensions: *outcome* vs. *growth assessment*, and *observed-score* vs. *latent-score mapping*. This set of three-dimensional dichotomies represents a multi-faceted cube that encompasses all major standard-setting procedures in accountability research. Each dimension and the pros and cons associated with each dichotomy are described below.

Conceptual Frameworks: ***Outcome vs. Growth***

Outcome standard. An outcome standard involves status assessment of content mastery. It is an *absolute* criterion against which each student's performance at a particular point in time is compared and evaluated. An outcome standard describes the level of performance required for a student to enter or exit a program without taking into consideration *process* variables such as time. Most familiar standard-setting procedures such as the Angoff method fall into this category of outcome assessment. Other less known procedures such as latent-class modeling approach (see below) that do not take into consideration demographic variables such as ethnicity or contextual variables such as school resources are also considered outcome assessment. A notable exception is the *Value-Added Assessment* proposed by Sanders, Saxton, and Horn (1997), which computes individual outcome scores by statistically adjusting for various *Opportunity-To-Learn* (OTL) variables such as ethnicity and social-economic status.

Whether outcome assessment should adjust for OTL is a highly debated and politically charged issue in large-scale assessment. On one hand, holding schools accountable for reaching the same standards without taking into consideration pre-existing differences or contextual variables seems to violate the notion of *fair assessment*. On the other hand, adjusting for OTL variables, which is equivalent to adjusting the outcome standard, defies the very purpose of the NCLB Act to bring all children to the same level of achievement. A few states, e.g., California (Rogosa, 2000), have implemented such an adjustment scheme but have been challenged for its concession to bringing all students to the same standards and helping them compete in a global economy (Thum & Bryk, 1997). Alternatively, allowing disadvantaged schools a longer period of time to reach the same outcome standards appears to be a reasonable compromise (Linn, Baker, Betebenner, 2002). This brings us to the following topic on growth standard.

Growth standard. *Growth* (or *gain score*) is defined as the difference between the *exit score* and the *entry score*, where entry score refers to the pre-test taken before program intervention (e.g., beginning of a school year) and exit score refers to the post-test taken after program intervention (e.g., end of a school year) within a specified period of time (e.g., a school year). Note that exit score does not necessarily refer to a terminal test of program intervention (such as a high school graduation test), but it could refer to any reasonable midpoint assessment during program intervention (such as a ninth-grade proficiency test). The entry and exit scores used to compute growth can be observed scores or latent scores, with or without statistical adjustment for OTL variables.

The methodology for the study of change has been extensively researched in many disciplines (e.g., Bourque, 1999; Collins & Sayer, 2001; Hess, 2000). As early as the 1950s, researchers had been promoting the importance of studying individual differences in their developmental growth (e.g., Rao, 1958). Numerous studies have since been produced that expanded on the basic growth curve to include more elaborate models (see, e.g., Duncan, Duncan, Strycker, Li, & Alpert, 1999; Gottman, 1995; Wang & Chou, 1996). Ranging from the simplest difference score to the complicated *latent growth modeling* coefficient, these studies promote the representation of *intraindividual variability* over time in which each individual's developmental trajectory is described and analyzed (Moskowitz & Hershberger, 2002).

According to the NCLB Act, by Year 2014 all schools must bring all (100%) children to the *proficient* level. Although this ambitious goal has been demonstrated to be highly unrealistic, it does accentuate the importance of *growth* within a prescribed period of time. The Act targets at school-level growth, but we must set growth standards on individual students in order to monitor their yearly progress. This study describes four conceptual definitions for individual growth: *growth referenced on target domain*, *growth referenced on entry score*, *growth referenced on individual gain*, and *growth referenced on normative gain*.

Growth Referenced on Target Domain. Each individual student's growth is compared to the target domain, which may be either the total domain (e.g., 100% content mastery) or the outcome standard (e.g., 70% content mastery). This index $(Exit - Entry)/Target$ computes the percentage of learned knowledge over the target domain. Note that the NCLB Act sets the target at the outcome standard. This method compares each individual growth to a fixed target, irregardless of where the student started. As such, it monitors each student's growth against this fixed target and holds everyone to the same standard. This method does not recognize individual differences at the entry point and may frustrate those who have made significant progress relative to where they started but still fall short of meeting the outcome standard.

Growth Referenced on Entry Score. Each individual student's growth is compared to the entry score: $(Exit - Entry)/Entry$. This method evaluates each individual growth against his or her starting point, irregardless of where they are supposed to end. As such, it monitors each student's growth against a variable baseline and adjusts to the varying levels where they started. This method does recognize individual differences at the entry point and is particularly suitable for encouraging low-achieving students to grow, but it conveys little information about how they fare in terms of the final target, and it unfairly penalizes those already high-achieving students who have little room for growth. An added issue in operation is that when the entry score is zero, we need to get around the "division-by-zero" problem.

Growth Referenced on Individual Gain. A compromise between the above two methods is to evaluate observed growth against individual target gain: $(Exit - Entry)/(Individual\ Target\ Gain)$. This method evaluates each individual growth against his or her target gain, taking into consideration both the starting point and the end target. *Individual target gain* can be operationalized as either $(Target - Entry)$ or $(Target - Entry)/(Years\ to\ Target)$, where *Years to Target* refers to the number of years the student is expected to reach the target, assuming a yearly assessment program. For example, if a 4th grader's entry score is 30, exit score is 40, target score is 48, and he is expected to reach the target when he exits the 6th grade. His growth index would be computed as $(40-30)/[(48-30)/(6-4)]=1.10$. A growth index of greater than 1.0 can be considered passing because it means the student is making expected progress to reach the target standard. It should be noted that this latter definition most closely parallels the NCLB Act's objective of 100% students reaching proficiency before a given year. This method monitors each student's growth against a growth target that is adjusted to the individual starting point. It does not unfairly penalize low-achievers or high-achievers. An added advantage is that no division-by-zero problem would be incurred. Note, however, that this method is only appropriate for students whose exit scores have not reached the target passing standard.

Growth Referenced on Normative Gain. In sharp contrast to the above three growth assessment methods, which are criterion-referenced in nature, a fourth growth assessment method is to evaluate individual growth against the normative gain, computed as the expected gain score based on the regression of exit on entry. As such, this method is norm-referenced in nature, and how a student performs is compared to how his or her peers perform in terms of growth. The *residual score* (i.e., observed exit score minus predicted exit score) is computed for each student, and generally a positive residual is considered above the normative expectation and therefore passing, whereas a negative residual is considered below the normative expectation and therefore failing.

This above discussion on criterion-referenced vs. norm-referenced assessment leads us to the second dimension in our classification scheme: theory-driven vs. data-driven.

Operational Approaches: Theory-Driven vs. Data-Driven

There are two general approaches used to search for the "best" performance standard: *theory-driven* and *data-driven*. This distinction is well recognized in the model-fitting literature, particularly the structural equation modeling literature (e.g., Byrne, 1998), but is less widely embraced by the standard-setting field (see, however, Bergan, Bergan, & Feld, 2002).

Theory-driven approach. A theory-driven approach sets the standard based on a preconceived or rationally-derived judgment of what performance characteristics must be present in order to be judged as *passing* the standard. As such, a theory-driven approach is also called *judgmental, rational,* or *criterion-referenced* approach. As indicated before, most, if not all, of the standard-setting methods documented in the literature are predominantly theory-driven. Although normative data on item statistics are often provided to assist in the standard-setting process, the fundamental principle is still criterion-referenced.

Data-driven approach. A data-driven strategy sets the standard based on statistical evidence of where the cut score should be without human intervention. Two notable examples of the data-driven strategy are *latent-class modeling* and *residual-score analysis*. Latent-class modeling is similar to cluster analysis in that they assume the data come from two or more different populations (e.g., masters vs. nonmasters), and the task is to find a linear combination of the multiple measures that best differentiates the groups while minimizing classification error (e.g., Bergan, 1983; Hagenaars, & McCutcheon, 2002). As such, latent-class models are mathematical models used to assign students or schools to categories without human judgment. This method has not found wide acceptance in the standard-setting area because its purely *data-driven* approach is at odds with the criterion-referenced nature imbedded in the standards-based accountability movement.

Residual-gain analysis computes residual scores by subtracting the predicted score from the obtained score ($e = Y - \hat{Y}$). In essence, what this amounts to is that each student's gain score is being compared with the norm's average gain. If a student grows faster than the norm's growth rate, the residual score is positive ($Y - \hat{Y} > 0$) and therefore considered *pass*; if slower then negative ($Y - \hat{Y} < 0$) and *fail*.

In sharp contrast to the theory-driven approach above, these data-driven approaches do not require any human intervention or value judgment. The *empirical* and *norm-referenced* nature of the data-driven strategy clearly raises some concerns. For example, the very purpose of standards-based accountability is to ensure that all students can reach the predesignated standards. Put in this perspective, students should be evaluated on their growth toward the achievement criteria, rather than the average growth of their peers as in the residual-score analysis. The empirical search for a cut score that statistically reduces classification error in latent-class modeling also fails to recognize the criterion-referenced purpose of assessment.

In sum, standards-based accountability assessment calls for a theory-driven, criterion-referenced approach in the search for the passing standard. A data-driven, norm-referenced approach is helpful in providing supplementary normative data and reducing classification error, but it should never replace the theory-driven approach as the primary search strategy.

Mapping Procedures: Observed Score vs. Latent Score

As the modern psychometric theory--*Item Response Theory (IRT)*—is gaining popularity in the measurement field (e.g., Embretson & Reise, 2000; Hambleton & Swaminathan, 1985), any standard-setting practice must take into consideration the potential impact of mapping the performance standard to the observed-score scale vs. the latent-score scale. By calibrating latent ability scores based on observed item responses, this mapping procedure has been found to be superior than others when reference must be made to the content domain rather than the specific test items (Bock, Thissen, & Zimowski, 1996; Hambleton, 1998; Reckase, 1998).

Mapping to the observed-raw-score scale is very straightforward: one simply multiplies the standard expressed in a percentage to the total score. This applies to both outcome standard and growth standard. Mapping to a derived score scale, whether it is linearly or nonlinearly transformed, observed or latent, is a bit tricky, because the "total score" can be hard to define. To illustrate, on a linearly transformed scale such as $X' = 500 + 50[(X - \overline{X})/S_X]$, should the total score be defined as three standard deviations above and below the mean? Or four? If we defined total score as three standard deviations above and below the mean, the total score would be $[500 + (50 \times 3)] - [500 - (50 \times 3)] = 650 - 350 = 300$. If, however, we defined total score as four standard deviations $[500 + (50 \times 4)] - [500 - (50 \times 4)] = 700 - 300 = 400$.

The same problem applies to the IRT latent score, which is further compounded by the nonlinear transformation involved in the scaling procedure and the *latent* (i.e., unobserved) nature of the score scale. In latent-scale mapping, the concept of content mastery is no longer referenced on the total number of items in a test, but on the abstract notion of *predicted domain mastery*. Three latent-scale mapping approaches are explored in this study: (a) *latent-score-range mapping* (i.e., multiplying the mastery standard to the total latent-score range and adding to the minimum score, e.g., $[(700-300)\times 70\%]+300=580$); (b) *regression-score mapping* (finding the predicted latent score by regressing latent score on raw score using nonlinear cubic-curve fitting; and (c) *domain-score mapping* (computing the average of item probability functions or *Test Characteristic Curve,* TCC). Of these three methods, the domain-score mapping is probably the best candidate because it is most consistent with the theoretical and mathematical assumptions in IRT scaling.

The above discussion serves to illustrate that when mapping the performance standard to the score scale in order to find the cutoff point, we need to be concerned with not only what score scale to use (observed vs. latent), but also what mapping procedure to use for finding the cut score on the latent scale. Different mapping procedures for the same performance standard operating within the same conceptual framework using the same operational approach may result in different passing rates and different accountability interpretations. Before we turn to an empirical study to illustrate this point, the proposed three-dimensional taxonomy of standard-setting procedures is summarized in Table 1:

Table 1

Taxonomy of Standard-Setting Procedures: Conceptual Frameworks, Operational Approaches, and Mapping Procedures

	Theory-Driven		Data-Driven	
	Observed Scale	Latent Scale	Observed Scale	Latent Scale
Outcome	- Passing standard × Total score	- Latent-score range - Cubic regression - Domain score	- Value-added score	- Latent-class modeling
Growth	- Referenced on target domain - Referenced on entry score - Referenced on Individual gain	- Referenced on target domain - Referenced on entry score - Referenced on individual gain	- Referenced on normative gain - Value-added score	- Referenced on normative gain

An Empirical Study

Assessment Data

This study used a longitudinal panel design in which preschool children were observed at two time points: Autumn 2000 and Spring 2001. Teacher observations of student performance in Early Math, Language and Literacy, and Social Development were collected. About 11,000 children in the Autumn of 2000 (the entry point) and 8,600 in the Spring of 2001 (the exit point) participated in the study, but only non-IEP (Individualized Educational Program or students with special needs) children aged between four and six with complete entry and exit data were included in the analysis. The final sample size is 5,636 for Early Math, 5,658 for Language and Literacy, and 5,496 for Social Development. Although this subgroup of the original sample may not be representative of either the sample or the population, this possible non-representativeness was deemed inconsequential because the purpose of this empirical study is to illustrate the various standard-setting procedures and compare their impact on the passing rates. No substantive interpretation on the passing rates per se (e.g., a passing rate of 30% associated with the 70% outcome standard) should be used to infer how well or how poorly prepared the preschoolers were for subsequently schooling.

The instrument used for data collection is a state-mandated electronic assessment system in which teachers entered their observational evaluations of student performance in a checklist (*Learned/Not Learned*) format. There are 68 items in Early Math, 68 in Language and Literature, and 65 in Social Development. Both internal and external evaluations of the technical quality of the instrument have shown high internal consistency, but validation studies have yielded inconclusive findings (Wang et al., 2002). Nevertheless, the external review concluded that the dataset was "adequate for preliminary investigation of program accountability, standard setting, and educational equity, but no high-stakes decisions should be based on the [data] alone" (p. i).

Standard-Setting Procedures

This study compared twelve standard-setting procedures derived from the three-dimensional taxonomy described in Table 1. Five conceptual frameworks were explored: *outcome*, *growth referenced on total*, *growth referenced on entry*, *growth referenced on individual annual gain*, and *growth referenced on normative gain*. Each of the conceptual framework is mapped onto two scales: *raw score* and *latent score*, but three mapping procedures were used for mapping the outcome standard to the latent scale: *latent-score range*, *cubic regression*, and *domain score*, resulting in altogether twelve standard-setting procedures.

Each standard-setting procedure is examined under the hypothetical outcome standard of 70% for content mastery at the exit point and the growth standard of 30% for the growth referenced on total, growth referenced on entry, and growth referenced on individual annual gain. A residual standard of zero is used for the growth referenced on normative gain. It is beyond this paper's scope to discuss where and how these cutoff standards should be set. These hypothetical standards were chosen to illustrate how different conceptual frameworks and mapping procedures may affect the dichotomous outcomes, while holding the conception of "satisfactory performance" constant.

Although the *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 1999) mandates that the standard error of measurement (SEM) around the cutoff score be reported (Standard 2.14), Phillips (2002) argues that when multiple attempts are allowed, "lowering the passing standard to adjust for possible error of measurement has little effect on the negligible number of erroneous diploma denials, but substantially increases the already significant number of students who pass without having actually met the required standard" (p. 115). Besides, since substantive interpretation is discouraged for the purpose of this study, we do not think it would be necessary to compute the SEM and further complicate the already complicated evaluation scheme in this study.

Each of the twelve standard-setting procedures are briefly described below:

Standard 1: The total number of items in Early Math is 68, therefore, 68 ×70% = 46.

Standard 2: The latent-score scale has a mean of 500 and standard deviation 50, yielding a latent-score range of $500 \pm 4(50) = 700 - 300 = 400$ and a cut score of $400 \times 70\% + 300 = 580$.

Standard 3: The cubic regression equation is $Y = 369.01 + 9.22X + (-.176)X^2 + .0016X^3$. With an observed cut score of 46, the mapped cut score is $369.01 + 9.22(46) + (-.176)(46)^2 + .0016(46)^3 = 583$. See Figure 1 for the cubic regression plot:

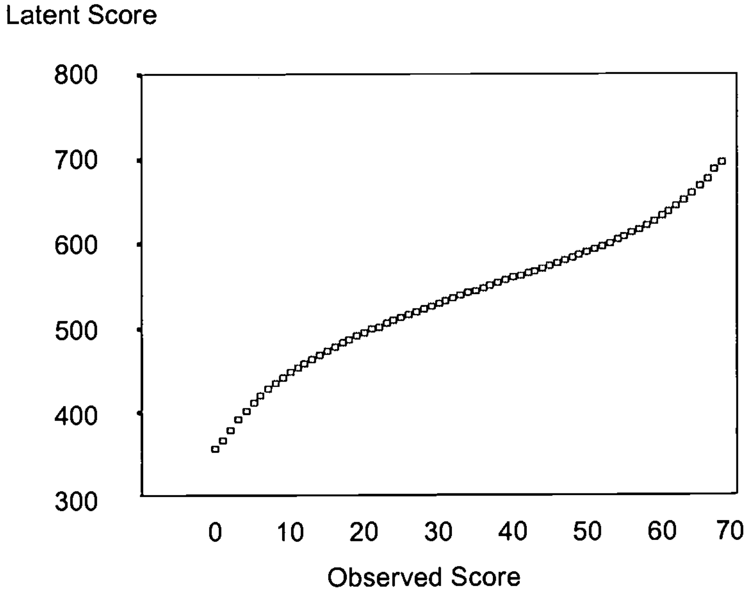

Figure 1. Cubic regression plot of latent score on observed score for Early Math.

Standard 4: The Test Characteristic Curve that maps the domain score (expressed as probability) to the latent standard score is depicted in Figure 2.

The latent standard score corresponding to the .70 domain score is .500, which translates to 500+50(.5) = 525 on the latent transformed scale.

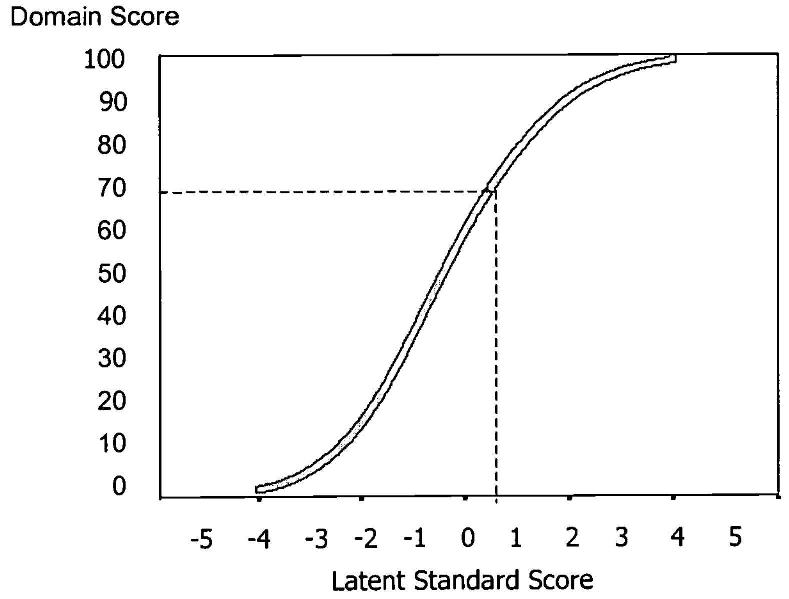

Figure 2. Test Characteristics Curve for mapping domain-score outcome standard to latent standard score scale for Early Math.

Standard 5: Each student's growth (*Observed Exit – Observed Entry*) is compared to the *Observed Entry*, and if the growth rate is greater than or equal to the standard rate (30%) then Pass.

Standard 6: Each student's growth (*Observed Exit – Observed Entry*) is compared to the *Observed Total*, and if the growth rate is greater than or equal to the standard rate (30%) then Pass.

Standard 7: Each student's growth (*Observed Exit – Observed Entry*) is compared to the expected annual *Individual Gain*, operationalized as (*Observed Outcome Standard – Entry*) / *(Years to Target)*, where *Observed Outcome Standard* is 68 × 70% = 48 and *Years to Target* refers to the number of years the student is expected to reach the target. For example, if a student is currently age 4 and expected to reach the outcome standard of 48 by age 6, then *Years to Target* is: *Terminal Grade – Current Grade* = 6 – 4 = 2. A growth index of greater than 1.0 (or 100%) is considered Pass because it means the student is making expected progress to reach the target outcome standard.

Standard 8: Each student's gain score is being compared with the norm's average gain. If a student grows faster than the norm's growth rate, the residual score is positive and therefore considered Pass; if slower then negative and therefore Fail.

Standards 9-12: Same as Standards 5-8 in that order, except that *Observed* is replaced by *Latent*.

Table 2

Syntax Rules for Mapping Conceptual Standards and Computing Passing Rates

No.	Conceptual Framework	Reference	Mapped Scale	Cut Score	Syntax
1	Outcome	Total score	Observed	68 × 70% = 46	IF Obs. Exit >= 46 THEN Pass
2	Outcome	Latent range	Latent	400 × 70% + 300 = 580	IF Latent Exit >= 580 THEN Pass
3	Outcome	Cubic reg.	Latent	$369.01+9.22(46)+(-.176)(46)^2 + .0016(46)^3 = 583$	IF Latent Exit >= 583 THEN Pass
4	Outcome	Domain score	Latent	500+50(.5) = 525	IF Latent Exit >= 525 THEN Pass
5	Growth	Entry	Observed	30% growth rate	IF (Obs. Exit – Obs. Entry)/Obs. Entry >= .30 THEN Pass
6	Growth	Total	Observed	30% growth rate	IF (Obs. Exit – Obs. Entry) /68 >= .30 THEN Pass
7	Growth	Individual gain	Observed	100% annual target gain	IF (Obs. Exit – Obs. Entry)/[(48 – Obs. Exit)/(6-Age)] >= 1.0 THEN Pass
8	Growth	Normative gain	Observed	Residual gain = 0	IF Obs. Residual >= 0 THEN Pass
9	Growth	Entry	Latent	30% growth rate	IF (Latent Exit – Latent Entry)/Latent Entry >= .30 THEN Pass
10	Growth	Total	Latent	30% growth rate	IF (Latent Exit – Latent Entry/400 >= .30 THEN Pass

11	Growth	Individual Gain	Latent	100% annual target gain	IF (Latent Exit – Latent Entry)/[(580 – Latent Exit)/(6-Age)] >= 1.0 THEN Pass
12	Growth	Normative gain	Latent	Residual gain = 0	IF Latent Residual >= 0 THEN Pass

Impact and Consistency Analysis

Two major analyses were performed: *impact analysis*, which computes the passing rate associated with each standard-passing procedure; and *consistency analysis*, which computes three consistency indices: percentage of agreement, Cohen's Kappa coefficient, and Cochran's Q test.

According to *The Program Evaluation Standards* (Joint Committee on Standards for Educational Evaluation, 1994), a common error in evaluation research is "interpreting reliability coefficients for measures of continuous variables as evidence of the reliability of dichotomous decisions (e.g., pass-fail, mastery-nonmastery) based on these measures" (Standard A6 Reliability Information, p. 155). Although the popular Cronbach's α is applicable to dichotomous data, it has been shown to be sensitive to number of items (i.e., procedures in our case) and therefore not appropriate for our study.

Alternatively, we decided to report the following three consistency indices: *percentage of agreement*, Cohen's *kappa* coefficient and its extension to three or more ratings, and Cochran's *Q*-test. The percent-agreement index, which is simply the ratio of number of students who received the same ratings across different standards (e.g., *Pass-Pass* or *Fail-Fail*) over the total sample, is widely used in the standard-setting literature. Such simple measures are easily understood and have much to recommend as long as their limitations are borne in mind (House, House, & Campbell, 1981).

The Cohen's kappa coefficient (1960) is an improvement over the simple percentage of agreement by correcting for chance error in agreement, and the *overall kappa* can be used for three or more standards (Landis & Koch, 1977). However, several studies have addressed the problems associated with the kappa coefficient, such as how *chance* should be estimated depends on whether the marginals are free to vary or fixed (Brennan & Prediger, 1981) and that it fails to take into consideration the underlying continuous distribution of the nominally dichotomous categories (Hoehler, 2000).

We therefore also report Cochran's Q-test, which test the hypothesis that several related dichotomous variables have the same mean (i.e., marginal percentages or passing rates). For example, Method 1 yields 75% passing, Method 2 yields 60%, and Method 3 yields 50%, and the Q statistic tests the hypothesis that all three methods yield the same passing rates. The Q-index is a nonparametric test, therefore it is more suitable for dichotomous data, but by focusing on marginal percentages rather than individual cases, it is a crude index of consistency and easily yields very large values and highly significant tests when the sample size is large.

The Kappa coefficient is interpreted the same way as the simple percentage of agreement, i.e., the larger the coefficient, the more consistent the different standard-setting procedures. The Q statistic is interpreted as a significance test on residuals—the larger the Q, the smaller the *p*, and the less consistent the procedures.

Results and Discussion

Multivariate Repeated-Measure Analysis

A cursory look at the descriptive statistics reported in Table 3 shows that the students generally improved from entry to exit, and they performed better on Social Development than on Language and Literacy, which is in turn better than Early Math.

Table 3

Descriptive Statistics of Entry and Exit scores

Subject	Time	Observed Score			Latent Score		
		N	*Mean*	*SD*	*N*	*Mean*	*SD*
Early Math	Entry	5636	21.14	12.43	5636	490.89	51.40
	Exit	5636	39.51	13.25	5636	558.88	47.17
Language and Literacy	Entry	5658	27.57	13.94	5658	503.28	47.48
	Exit	5658	46.45	12.74	5658	565.83	45.10
Social Development	Entry	5496	32.44	15.44	5496	467.77	77.09
	Exit	5496	50.29	12.12	5496	563.40	71.71

Note. Number of items in Early Math = 68, Language and Literacy = 68, Social Development = 65.

A closer look at the multivariate repeated-measure analysis generally confirms the above observations (Table 4): the improvement between entry and exit is both statistically significant ($p < .000$) and substantively significant ($\eta^2 = .786$ for observed scale and .776 for latent scale); the performance on the three subject areas also differs significantly both statistically ($p < .000$) and substantively ($\eta^2 = .611$ for observed scale and .224 for latent scale). The strong interaction effect between Time and Test for the latent scale ($\eta^2 = .287$) suggests that the improvement between entry and exit may be different for the three subject areas.

Table 4

Multivariate Repeated-Measure Analysis of Variance

Source	*df*	Wilks' Λ	*F*	*P*	η^2
		Observed Scale			
TIME	1, 4798	.214	17574.540	.000	.786
TEST	2, 4797	.389	3773.305	.000	.611
TIME × TEST	2, 4797	.987	32.577	.000	.013
		Latent Scale			
TIME	1, 4798	.224	16647.539	.000	.776
TEST	2, 4797	.776	692.526	.000	.224
TIME × TEST	2, 4797	.713	965.038	.000	.287

Note. N = 4799; TIME = Entry vs. Exit; TEST = Early Math, Language and Literacy, Social Development.

Follow-up tests of contrasts (Table 5) also reveal that pair-wise comparisons among the three subject areas are both statistically (p < .000) and substantively (η^2 ranging from . 129 to .479) significant.

Table 5

Repeated-Measure Tests of Contrasts between Subject Areas

Source	Test	*df*	*F*	*p*	η^2
		Observed Scale			
TEST	EM vs. LL	1	4417.757	.000	.479
	LL vs. SD	1	1285.248	.000	.211
Error	EM vs. LL	4798	(100.820)		
	LL vs. SD	4798	(133.986)		
		Latent Scale			
TEST	EM vs. LL	1	713.681	.000	.129
	LL vs. SD	1	880.921	.000	.155
Error	EM vs. LL	4798	(1341.302)		
	LL vs. SD	4798	(3902.378)		

Note. N = 4799. EM = Early Math, LL = Language and Literacy, SD = Social Development. Values in parentheses Represent mean square errors.

Cut Scores and Passing Rates

Using the standard-mapping procedures described in Table 2, we computed the cut scores and passing rates as shown in Table 6:

Table 6

Cut Scores and Passing Rates for Twelve Standard-Setting Procedures

Standard (Reference, Mapped Scale)	Early Math		Language & Literature		Social Development	
	Cut	Pass	Cut	Pass	Cut	Pass
1. Outcome (Total, Observed)	48	29%	48	52%	46	70%
2. Outcome (Total, Latent)	580	32%	580	36%	580	45%
3. Outcome (Reg., Latent)	583	29%	553	61%	520	72%
4. Outcome (Domain Score, Latent)	525	79%	496	95%	471	89%
5. Growth (Entry, Observed)	.30	81%	.30	77%	.30	66%
6. Growth (Total, Observed)	.30	40%	.30	40%	.30	36%
7. Growth (Individual Gain, Observed)	1	23%	1	54%	1	36%
8. Growth (Normative Gain, Observed)	0	48%	0	50%	0	51%
9. Growth (Entry, Latent)	.30	8%	.30	5%	.30	26%
10. Growth (Total, Latent)	.30	12%	.30	8%	.30	31%
11. Growth (Individual Gain, Latent)	1	26%	1	49%	1	26%
12. Growth (Normative Gain, Latent)	0	46%	0	45%	0	47%

The impact data show that very different cut scores and passing rates were produced depending on how the standard was conceptualized and operationalized. The passing rates ranged from 8% to 81% for Early Math, 5% to 95% for Language and Literature, and 26% to 89% for Social Development. In particular, different latent-scale mapping procedures for the same conceptual standard of *Outcome Referenced on Total* (Standards 2, 3, 4) also produced vastly different passing rates, ranging from 29% to 79% for EM, 36% to 95% for LL, and 45% to 89% for SD. Likewise, the same conceptual standard also produced vastly different passing rates, depending on whether the standard was mapped to the observed-score or the latent-score scale. Take *Growth Referenced on Entry* for example. The raw-score mapping (Standard 5) for EM produced 81%, but the latent-score mapping (Standard 9), only 8%. Similar disparities were found for LL and SD. The only less disturbing comparisons were *Growth Referenced on Individual Gain* (Standards 7 and 11), ranging from 23% to 26% for EM, 54% to 49% for LL, and 36% to 26% for SD, and *Growth Referenced on Normative Growth* (Standards 8 and 12), ranging from 48% to 46% for EM, 50% to 45% for LL, and 51% to 47% for SD.

Consistency Analysis

Obviously an omnibus consistency analysis comparing all twelve standards is of little interest here because very different conceptual definitions and mapping procedures were involved. We therefore conducted four sets of consistency analysis, each serving a special purpose for illustration.

The first one compared three totally different conceptual standards: *Outcome*, *Growth Referenced on Total*, which is theory-driven in nature, and *Growth Referenced on Normative Gain*, which is data-driven in nature. By computing the consistency indices separately for the observed and latent scores, this analysis controls for the mapping scale effect. The results are reported in Table 7.

Table 7

Consistency Measures Among Different Conceptual Standards: Outcome, Growth Referenced on Total, and Growth Referenced on Normative Gain

Subject Area	Scale	Standards	Percent	Kappa	Q
Early Mathematics	Observed	1, 6, 11	62%	.46	754.47*
	Latent	2, 9, 12	55%	.29	2147.28*
Language & Literature	Observed	1, 6, 11	59%	.45	332.80*
	Latent	2, 9, 12	54%	.26	2677.24*
Social Development	Observed	1, 6, 11	51%	.35	1909.26*
	Latent	2, 9, 12	62%	.47	625.57*

* *p* < .0005

Not surprisingly, different conceptualizations of standards yielded inconsistent passing decisions. Although the percentages of agreement range from 51% to 62%, they become much smaller after corrected for chance (from .26 to .47). All Q statistics indicate that the inconsistency is significant at p < .0005.

The second set compared three theory-driven growth standards: *Growth Referenced on Entry, Growth Referenced on Total*, and *Growth Referenced on Individual Annual Gain.* By focusing only on theory-driven growth standards and by computing the consistency indices separately for the observed and latent scores, this analysis controls for conceptual frameworks and mapping procedure effects. The results are reported in Table 8.

Table 8

Consistency Measures Among Theory-Driven Growth Standards: Growth Referenced on Entry, Growth Referenced on Total, and Growth Referenced on Individual Annual Gain.

Subject Area	Scale	Standards	Percent	Kappa	Q
Early Mathematics	Observed	5, 6, 7	30%	.06	4345.77*
	Latent	9, 10, 11	72%	.29	1035.75*
Language & Literature	Observed	5, 6, 7	49%	.30	2242.34*
	Latent	9, 10, 11	54%	.05	4310.14*
Social Development	Observed	5, 6, 7	49%	.32	1905.80*
	Latent	9, 10, 11	66%	.43	91.65*

* $p < .0005$

Somewhat surprisingly, comparing only theory-driven growth standards did not improve the consistency in passing decisions. The percentages of agreement are generally even lower, and some kappa coefficients even dropped to less than 10% of agreement. The Q statistics are all statistically significant and large in value.

The third set compared the same conceptual standards mapped to different score scales *Observed Scale* vs. *Latent Scale.* This analysis controls for conceptualization of standards and should in theory yield higher consistency (see Table 9).

Table 9

Consistency Measures Between Observed-Score and Latent-Score Mapping for Each Conceptual Standard

Subject Area	Standards	Percent	Kappa	Q
Early Mathematics	1, 2	97%	.94	143*
	1, 3	100%	1.00	0
	1, 4	50%	.20	2818*
	5, 9	27%	.04	4093*
	6, 10	72%	.35	1528.1*
	7, 11	96%	.88	163.35*
	8, 12	93%	.86	24.81*
Language & Literature	1, 2	84%	.69	899*
	1, 3	92%	.83	472*
	1, 4	58%	.12	2404*
	5, 9	28%	.03	4062*
	6, 10	68%	.24	1802*
	7, 11	85%	.71	129.00*
	8, 12	91%	.82	144.93*
Social Development	1, 2	75%	.52	1371.00*
	1, 3	98%	.95	122.98*
	1, 4	81%	.46	1026.00*
	5, 9	60%	.31	2199.00*
	6, 10	91%	.79	170.30*
	7, 11	75%	.43	208.38*
	8, 12	95%	.89	186.61*

* *p* < .0005

Table 9 generally shows improved consistency when the conceptual definitions are fixed at a particular framework. However, some consistency indices are still disturbingly low, particularly comparisons between Standards 1, 4 and 5, 9. Standards 1 and 4 both refer to outcome standards of 70% content mastery, but Standard 1 is referenced on the observed scale and Standard 4 on the latent scale using the domain-score mapping. What is particularly disturbing is that domain-score mapping is generally regarded as the best approach to mapping an outcome standard to a latent domain. However, a consistency measure of *Kappa* ranging from .12 for Language and Literacy and .46 for Social Development raises some very serious legal challenges to school systems that choose one score scale over the other. Referring back to Table 6, the passing rate for Early Math when mapped to the observed scale is only 29%, but 79% when mapped to the latent scale! In other words, the same school would be rated as *Failing* when they adopted the traditional observed raw score scale, but they could easily boost their rating to *Passing* if they decided to go with the modern latent score scale, without sweating any reform efforts!

Even if every school decided to map the standards to the latent scale, another thorny issue remains: different observed-to-latent mapping procedures still yield vastly inconsistent passing decisions. The last set compared three different latent-scale mapping methods: *Latent Range*, *Cubic Regression*, and *Domain Score* (Table 10).

Table 10

Consistency Measures Among Different Mapping Procedures for Outcome Standard

Subject Area	Percent	Kappa	Q
Early Mathematics	50%	.33	5364.51*
Language and Literature	42%	.16	5002.14*
Social Development	56%	.32	3668.36*

* $p < .0005$

We can see that the picture does not look any brighter here. Different latent-scale mapping procedures yielded very inconsistent passing decisions. The Kappa coefficients reported in Table 10 are generally even worse than those in comparing observed vs. latent standards (Table 9).

Conclusions and Challenge

Inconsistency of Passing Rates

The current findings confirm previous studies that have shown inconsistency of passing rates associated with different standards. In Jaeger's (1989) study that summarized 28 studies on differences in cut scores set by different methods, he found ratios of the highest standard to the lowest standard to range from 1.00 to 52.00, with a median value of 1.46 and a average value of 5.30. Hambleton (1978) tried to alleviate concerns by noting that the different methods of setting cut scores defined minimal competency in different ways, so differences in results were to be expected. Shepard (1983), however, pointed out that even though the various methods obviously use different "operational definitions" (p. 63) of minimal competency, "they do not have correspondingly different conceptual definitions" (p. 63). Supporting Shepard's contention, this study has found inconsistency in the same conceptual standard mapped to different score scales (observed vs. latent) and also when different latent-score mapping procedures were used for operationalizing the same conceptual standard.

Although this lack of consistency may not be readily interpreted as *unreliability* in the traditional psychometric sense, such inconsistency still raises serious concerns in standard-setting research, particularly when high-stakes consequences are to be attached to the assessment results. As Kane (2001) puts it, "an examination of the coherence of the decision process, and in particular, the consistency among the purpose and context of the decision, the conception of the achievement being evaluated, the assessment methods, and the approach to standard setting, does not establish the validity of the results. But a lack of coherence among the different elements in the decision process undermines confidence in the decision process as a whole" (p. 63).

Where Do We Go from Here?

Although there are numerous sources of measurement error that may explain the inconsistent passing rates, many have attributed this inconsistency to the value-laden process of standard setting. As pointed out by Keller and Zanetti (2000), the standard is "a reflection of what policy makers consider to be important, and as such cannot exist in a positivist framework" (p. 8). Kane (1998) also argued that, "to set a standard is to set a policy, and policies are evaluated in terms of their appropriateness, reasonableness, consistency, rather than accuracy" (p. 129). This view is also shared by Cizek (2001): "Standard setting is perhaps the branch of psychometrics that blends more artistic, political, and cultural ingredients into the mix of its products than any other" and see it "much less of a technical challenge and much more of a policy endeavor" (p. 5).

Selecting a conceptual framework that is most consistent with the intended interpretation of the results, therefore, ultimately lies in the stakeholders and policy makers. Each conceptual framework for standard setting has its own merits and limitations, and users of those standards should be well informed as to when to use which for whom and for what purpose.

Psychometricians, on the other hand, can play a critical role in assisting with the selection of an operational procedure for mapping the conceptual standard to the score scale. This study has shown how mapping to the observed-score scale and mapping to the latent-score scale can exert profound impact on the passing rates. With some states using the observed-score scale and others using the latent-score scale, such inconsistency can make across-the-state comparison extremely unfair and may even be legally controversial. Even a small passing-rate difference can exert high-stakes consequences on the individuals or programs being evaluated.

When large-scale standardized achievement testing is largely embraced by stakeholders both in the US and in the rest of the world (Phelps, 1998, 2000), the No Child Left Behind Act stands an unprecedented chance of reforming our education as so many previous reform advocates had tried and failed. The success of this wave of accountability movement, however, depends to a large extent on how policy makers, stakeholders, and psychometricians work together to "standardized" the standard-setting procedure during this chaotic flux of change so that no child or school is unfairly penalized by the undue process involved when reaching for the worthy goal of "No Child Left Behind."

References

American College Testing. (1993). *Setting achievement levels on the 1992 National Assessment of Educational Progress in mathematics, reading, and writing.* Iowa City, IA: Author.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards fro educational and psychological testing.* Washington, DC: The Authors.

American Evaluation Association (2002). *American Evaluation Association position statement on high-stakes testing in PreK-12 education.* Retrieved May 6, 2002 from www.eval.org/hst3.htm

Amrein, A.L., & Berliner, D.C. (2002). High-stakes testing, uncertainty, and student learning [Electronic version]. *Educational policy Analysis Archives, 10*(18). Retrieved April 1, 2003 from hattp://epaa.asu.edu/epaa/v10n18

Angoff, W.H. (1977). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (pp. 508-600). Washington, DC: American Council on Education.

Bergan, J.R. (1983). Latent class models in educational research. In E.W. Gordon (Ed.), *Review of Research in Education*, 10 (pp. 305-360). Washington, DC: American Educational Research Association.

Bergan, J.R., Bergan, J.B., & Feld, J. (2002). *Learning outcomes for The Ohio Indicators of Success Project Program Year 2000-2001.* Tucson, AZ: Assessment Technology Incorporated.

Berk, R.A. (1980). *Criterion-referenced measurement: The state of the art.* Baltimore, MD: Johns Hopkins University Press.

Berk, R.A. (1996). Standard setting: The next generation (Where few psychometricians have gone before!). *Applied Measurement in Education, 9*(3), 215-235.

Bock, R.D., & Thissen, D. (1997). IRT estimation of domain scores. *Journal of Educational Measurement, 34*(3), 197-211.

Bourque, M.L. (1999). Assessment of educational change: A review of selected threats to validity. In A.C. Tuijnman & T.N. Postlethwaite (Eds.), *Monitoring the standards of education.* Oxford, England: Pergamon.

Brennan, R.L., & Prediger, D.J. (1981). Coefficient Kappa: Some uses, misuses, and alternatives. *Educational and Psychological Measurement, 41*, 687-699.

Brown, R.D., Johnson, L.J., Pretti-Frontczak, K., & Kowalski, K. (2001). *Early Childhood Indicators of Success Year 3 Evaluation Report: Stakeholder Survey and Concordance Study.* Cincinnati, OH: Evaluation Services Center, University of Cincinnati.

Buly, M.R., & Valencia, S.W. (2002). Below the bar: Profiles of students who fail state reading assessment. *Educational Evaluation and Policy Analysis, 24*(3), 219-239.

Burd, S. (2002). Accountability of meddling? *Chronicle of Higher Education, 49*(4), 23-25.

Byrne, B.M. (1998). *Structural equation modeling with LISREL, PRELIS, and SIMPLIS: Basic concepts, applications, and programming.* Mahwah, NJ: Lawrence Erlbaum.

Byrk, A.S., & Raudenbush, S. (1992). *Hierarchical linear modeling*. Chicago: Sage.

Camilli, G., Cizek, G.J., & Lugg, C.A. (2001). Psychometric theory and the validation of performance standards: History and future perspectives. In G.J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp.445-475). Mahwah, NJ: Lawrence Erlbaum.

Center on Education Policy. (2003). *From the capital to the classroom: State and federal efforts to implement the No Child Left Behind Act*. Washington, DC: Author. Retrieved April 20, 2003 from http://www.ctredpol.org/pubs/nclb_full_report_jan2003/nclb_full_report_jan2003.htm

Cizek, G. J. (1993). Reconsidering standards and criteria. *Journal of Educational Measurement, 30*(2), 93-106.

Cizek, G.J. (2001). Conjectures on the rise and call of standard setting: An introduction to context and practice. In G.J. Cizek (Ed.), *Setting performance standards Concepts, methods, and perspectives* (pp. 3-17). Mahwah, NJ: Lawrence Erlbaum.

Cizek, G. J. (2001). *Setting performance standards*. Hillsdale, NJ: Lawrence Erlbaum.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37-46.

Collins, L.M., & Sayer, A.G. (Eds.). (2001). *New methods for the analysis of change.* Washington, DC: American Psychological Association.

Costello, K., & Zarowin, D. (2002). Technology connects assessment, accountability standards in early childhood education. *THE Journal, 30*(4), 48-49.

Darling-Hammond, L. (1997). *The right to learn*. San Francisco, CA: Jossey-Bass.

Duncan, T.E., Duncan, S.C., Strycker, L.A., Li, F., & Alpert, A. (1999). *An introduction to latent variable growth curve modeling.* Mahwah, NJ: Lawrence Erlbaum.

Ebel, R.L. (1965). *Measuring educational achievement.* Englewood Cliffs, NJ: Prentice-Hall.

Education USA. (2003). ED: All 50 States submitted accountability plans in time. *Education USA, 45*(2), 7.

Embretson, S.E., & Reise, S.P. (2000). *Item response theory for psychologists.* Mahwah, NJ: Lawrence Erlbaum.

Gilliam, W.S., & Zigler, E.F. (2001). A critical meta-analysis of all evaluations of state-funded preschool from 1977 to 1998: Implications for policy, service delivery, and program evaluation. *Early Childhood Research Quarterly, 15*, 441-473.

Gottman, J. M. (1995). *The analysis of change.* Mahwah, NJ: Lawrence Erlbaum.

Guskey, T.R. (Ed.). (1994). *High stakes performance assessment: Perspectives on Kentucky's educational reform.* Thousand Oaks, CA: Corwin.

Haertel, E. H. (1999). Validity arguments for high-stakes testing: In search of the evidence. *Educational Measurement: Issues and Practice, 18*(4), 5-9.

Haertel, E. H. (2002). Standard setting as a participatory process: Implications for validation of standards-based high-stakes assessment. *Educational Measurement: Issues and Practice, 21*(1), 16-22.

Hagenaars, J.A., & McCutcheon, R.L. (Eds.). (2002). *Applied latent class analysis.* Cambridge, England: Cambridge University Press.

Hambleton, R.K. (1978). On the use of cut-off scores with criterion-referenced tests in instructional settings. *Journal of Educational Measurement, 15*, 277-290.

Hambleton, R.K. (1980). Test score validity and standard-setting methods. In R.A. Berk (Ed.), *Criterion-referenced measurement: The state of the art* (pp. 80-123). Baltimore, MD: Johns Hopkins University Press.

Hambleton, R.K. (1998a). Enhancing the validity of NAEP achievement level score reporting. *Proceedings of Achievement Levels Workshop* (pp. 77-98). Washington, DC: National Assessment Governing Board.

Hambleton, R.K. (1998b). Setting performance standards on achievement tests: Meeting the requirements of Title I. In L. Hansche (Ed.), *Handbook for the development of performance standards* (pp. 87-114). Washington, DC: U.S. Department of Education and the Council of Chief State School Officers.

Hambleton, R.K., Brennan, R.L., Dodd, B., Forsyth, R.A., Mehrens, W.A., Nellhaus, J., Reckase, M., Rindone, D., van der Linden, W.J., & Zwick, R. (2000). A response to "Setting reasonable and useful performance standards" in the National Academy of Sciences' Grading the Nation's Report Card. *Educational Measurement: Issues and Practice, 19*(2), 5-15.

Hambleton, R.K., & Swaminathan, H. (1985). *Item response theory: Principles and applications.* Boston: Kluwer.

Hess, B. (2000). Assessing program impact using latent growth modeling: A primer for the evaluator. *Evaluation and Program Planning, 23,* 419-428.

Hoehler., F. K. (2000). Bias and prevalence effects on kappa viewed in terms of sensitivity and specificity. *Journal of Clinical Epidemiology*, 53, 499-503.

Horn, Jr., R. A. (2002). *Understanding educational reform.* Santa Barbara, CA: ABC-CLIO.

House, A.E., House, B.J., & Campbell, M.B. (1981). Measures of interobserver agreement: Calculation formulas and distribution effects. Journal of Behavioral Assessment, 3(1), 37-57.

Jaeger, R. M. (1989). Certifictaion of student competence. In R. L. Linn (Ed.), *Educational measurement* (pp. 485-514). Washington, DC: American Council on Education/MacMillan).

Jasanoff, S. (1998). Science and judgment in environmental standard setting. Applied *Measurement in Education, 11*(1), 107-120.

Joint Committee on Standards for Educational Evaluation (1994). *The Program Evaluation Standards* (2nd ed.). Thousand Oaks, CA: Sage.

Jones, M.G., Jones, B.D., & Hargrove, T.Y. (2003). *The unintended consequences of high-stakes testing*. Lanham, MD: Rowman & Littlefield.

Kane, M. T. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research, 64*(3), 425-461.

Kane, M. T. (1995). Examinee-centered vs. task-centered standard setting. In M.L. Bourque (Ed.), *Proceedings of the Joint Conference on Standard Setting for Large-Scale Assessments* (Vol.2, pp. 119-141). Washington, DC: National Assessment Governing Board and National Center for Education Statistics.

Kane, M. T. (1998). Choosing between examinee-centered and test-centered standard-setting methods. *Educational Assessment, 5*(3), 129-145.

Kane, M. T. (2001). So much remains the same: Conception and status of validation in setting standards. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 53-88). Mahwah, NJ: Lawrence Erlbaum.

Kelly, L.A., & Zanetti, M. L. (2000). *Validity issues in standard setting.* Amherst, MA: University of Massachusetts, Laboratory of Psychometric and Evaluative Research.

Kifer, E. (2001). *Large-scale assessment: Dimensions, dilemmas, and policy.* Thousand Oaks, CA; Corwin.

Koretz, D. M., Barron, S., Mitchell, K., & Stecher, B. (1996). *Perceived effects of the Kentucky Instructional Results Information System (KIRIS).* Santa Monica, CA: RAND.

Kramer, A., Muijtjens, A., Jansen, K., D ü sman, H., Tan, L., & van der Vleuten, C. (2003). Comparison of a rational and an empirical standard setting procedure for an OSCE. *Medical Education, 37*(2), 132-139.

Kucerick, E. (2002). The No Child Left Behind Act of 2001: Will it live up to its promise? *Georgetown Journal on Poverty Law and Policy, 9*(2), 479-487.

Landis, J.R., & Koch, G.G. (1977). The Measurement of observer agreement for categorical data. *Biometrics*, 33, 159-174.

Leonardo, Z. (2003). The agony of school reform: Race, class, and the elusive search for social justice. *Educational Researcher, 32*(3), 37-43.

Linn, R.L. (2000). Assessment and accountability. *Educational Researcher, 29*(2), 4-14.

Linn, R.L., & Baker, E.L. (1999). *Absolutes, wishful thinking, and norms: The CRESS line.* Los Angeles, CA: University of California at Los Angeles, Center for Research on Evaluation, Standards, and Student Testing.

Linn, R.L., Baker, E.L., & Betebenner, D.W. (2002). Accountability system: Implications of requirements of the No Child Left Behind Act of 2001. *Educational Researcher, 31*(6), 3-16.

Lissitz, R. W., & Bourque, M. L. (1995). Reporting NAEP results using standards. *Educational measurement: Issues and practice*, 14(2), 14-31.

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.

Malcom, S.M. (1993). *Promises to keep: Creating high standards for American student*. Washington, DC: National Education Goals Panel. Retrieved April 1, 2003 from http://www.negp.gov/page9-3.htm#Std

McNeil, L.M. (2000). *Contradictions of school reform: Educational costs of standardized testing*. New York: Routledge.

Mehren, W.A. (1995). Methodological issues in standard setting for educational exams. In M.L. Bourque (Ed.), *Proceedings of the Joint Conference on Standard Setting for Large-Scale Assessments* (pp. 221-263). Washington, DC: National Assessment Governing Board, National Center for Education Statistics.

Mehrens, W. A., Popham, W. J., & Ryan, J. M. (1998). How to prepare students for performance assessments. *Educational Measurement: Issues and Practice, 17*(1), 18-22.

Millman, J. (Ed.). (1997). *Grading teachers, grading schools: Is student achievement a valid evaluation measure?* Thousand Oaks, CA: Corwin.

Mislevy, R.J., & Bock, R.D. (1990). *BILOG 3: Item analysis and test scoring with binary logistic models* (2nd ed.). Chicago, IL: Scientific Software International.

Mitzel, H.C., Lewis, D.M., Patz, R.J., & Green, D.R. (2001). The bookmark method: Psychological perspectives. In Cizek, G.J. (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 249-281). Mahwah, NJ: Lawrence Erlbaum.

Morse, J. (2002, September 23). Anything to avoid an F. *Time*, 22.

Moskowitz, D.S., & Hershberger, S.L. (Eds.). (2002). *Modeling intraindividual variability with repeated measures data*. Mahwah, NJ: Lawrence Erlbaum.

Musick, M.D. (1997). *Accountability in the 1990s: Holding schools responsible for student achievement.* Atlanta, GA: Southern Regional Accreditation Board.

National Academy of Education. (1993). *Setting performance standards for student achievement*. Stanford, CA: Author.

National Council on Education Standards and Testing. (1992). *Raising standards for American education*. Washington, DC: US Government Printing Office.

Norcini, J. J., & Shea, J.A. (1997). The credibility and comparability of standards. *Applied Measurement in Education, 70* 39-59.

Ohio Legislature. (1997). *The Amended Substitute House Bill 215*. Retrieved March 28, 2003 from http://www.legislature.state.oh.us/BillText122/122_HB_215_1_N.htm

Olson, L. (1998). An "A" or a "D": State rankings differ widely. *Education Week, 17*(31), 1-2.

Olson, L. (2002). States anxious for federal guidance on yearly progress. *Education Week, 22*(13), 1-3.

Pellegrino, J.W., Jones, L.R., & Mitchell, K.J. (Eds.) (1999). *Grading the nation's report card: Evaluating NAEP and transforming the assessment of educational progress*. Washington, DC: National Academy Press.

Pellegrino, J.W., Wise, L., & Raju, N. (Eds.). (1998). Setting consensus goals for academic achievement. *Applied Measurement in Education, Special Issue, 11*(1), 1-120.

Phelps, R.P. (1998). The demand for standardized student testing. *Educational Measurement: Issues and Practice, 17*(3), 5-23.

Phelps, R.P. (2000). Trends in large-scale testing outside the United States. *Educational Measurement: Issues and Practice, 19*(1), 11-21.

Phillips, S.E. (2002). Legal issues affecting special populations in large-scale testing programs. In G. Tindal and T.M. Haladyna (Eds.), *Large-scale assessment programs for all students* (pp. 109-148). Mahwah, NJ: Lawrence Erlbaum.

Plake, B.S. (1995). The performance domain and structureof the decision space. *Applied Measurement in Education, 8*(1), 3-14.

Popham, W.J. (1978). As always, provocative. *Journal of Educational Measurement, 15*, 297-300.

Popham, W.J. (1999). Where large scale educational assessment is heading and why it shouldn't. *Educational Measurement: Issues and Practice, 18*(3), 13-17.

Rao, C. R. (1958). Some statistical methods for the comparison of growth curves. *Biometrics, 14*, 1-17.

Reckase, M.D. (1998). Converting boundaries between National Assessment Governing Board performance categories to points on the National Assessment of Educational Progress score scale: The 1996 science NAEP process. *Applied Measurement in Education, 11*, 9-21.

Rogosa, D. (2000). *Interpretative notes for the Academic Performance Index.* Sacramento, CA: California Department of Education. Retrieved May 16, 2003 at http://www.cde.ca.gov/psaa/apiresearch.htm

Sanders, W., Saxton, A., & Horn, S. (1997). *The Tennessee value-added assessment system: A quantitative, outcomes-based approach to educational assessment. In J. Millman (Ed.), Grading teachers, grading schools* (pp. 137-162). Thousand Oaks, CA: Corwin.

Schulz, E.M., Kolen, M.J., & Nicewander, W.A. (1999). A rationale for defining achievement levels using IRT-estimated domain scores. *Applied Psychological Measurement, 23*(4), 347-362.

Schwarz, R.D., Yen, W.M., & Schafer, W.D. (2001). The challenge and attainability of goals for Adequate Yearly Progress. *Educational Measurement: Issues and Practice, 20*(3), 26-33.

Searle, J. (2000). Defining competency—the role of standard setting. *Medical education, 34*(5), 363-366.

Shepard, L. (1984). Setting performance standards. In R.A. Berk (Ed.), *A guide to criterion-referenced test construction* (pp. 169-198). Baltimore, MD: Johns Hopkins University Press.

Shepard, L.A. (1983). Standards for placement and certification. In S.B. Anderson and J.S. Helmick (Eds), *On educational testing* (pp. 61-90). Washington, DC: Jossey –Bass.

Shepard, L. A. (2002-2003). The hazards of high-stakes testing. *Issues in Science and Technology, 19*(2), 53-58.

Shepard, L.A., Kagan, S. L., Wurtz, E. (Eds.). (1998). *Principles and recommendations for early childhood assessment*. Washington, DC: The National Education Goals Panel. Retrieved March 28, 2003 from http://www.negp.gov/page9-3.htm

Shriner, J. G., & Ysseldyke, J. E. (1994). Standards for all American students. *Focus on exceptional children, 26*(5), 1-19.

Sternberg, R.J. (2003, March). Responsibility: One of the other three Rs. *Monitor on Psychology*, 5.

The White House. (2001). *No Child Left Behind*. Washington, DC: Author. Available at http://www.ed.gov/offices/OESE/esea/index.html

Thum, Y.M., & Bryk, A.S. (1997). Value-added productivity indicators: The Dallas system. In J. Millman (Ed.), *Grading teachers, grading schools: Is student achievement a valid evaluation measure?* Thousand Oaks, CA: Corwin.

Tindal, G. (2002). Large-scale assessments for all student: Issues and options. In G. Tindal and T.M. Haladyna (Eds.), *Large-scale assessment programs for all* *students: Validity, technical adequacy, and implementation* (pp. 1-24). Mahwah, NJ: Lawrence Erlbaum.

US Department of Education. (2003). *Guidance on standards, assessments, and accountability*. Retrieved March 28, 2003 from http://www.ed.gov/offices/OESE/StandardsAssessment/overview.html

Vinovskis, M.A. (1996). An analysis of the concept and uses of systemic educational reform. *American Educational Research Journal, 33,* 53-85.

Wang, L., & Chou, T.F. (1996). IRT item linking design in the construction of ability growth curves. *Annals of Psychological Testing, 43,* 53-65.

Wang, L., Johnson, L.J., Boat, M., Brown, R., Pan, W., Zorn, D., & Austin, J.T. (2002). *Ohio Early Childhood Indicators of Success Year IV evaluation study*. Cincinnaci, OH: University of Cincinnati, Evaluation Services Center.

Wiggins, G. (1998). *Educative assessment: Designing Assessments to Inform and Improve Student Performance*. San Francisco: Jossey-Bass.

Wiliam, D. (1996). *Meanings and consequences in standard setting. Assessment in education: Principles, Policy & Practice, 3*(3), 287-307.

Yen, W.M., & Henderson, D.L. (2002). Professional standards related to using large-scale state assessments in decisions for individual students. *Measurement and Evaluation in Counseling and Development, 35*, 132-143.

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. 19-51).

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

ERIC
Educational Resources Information Center

REPRODUCTION RELEASE

(Specific Document)

TM035018

I. DOCUMENT IDENTIFICATION:

Title: Standard-setting Procedures in Accountability Research: Impacts of Conceptual Frameworks and Mapping Procedures on passing Rate	
Author(s): Lihshing Wang, Wei Pan, and James T. Austin	
Corporate Source: U of Cincinnati	Publication Date: Apr 2003

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents	The sample sticker shown below will be affixed to all Level 2A documents	The sample sticker shown below will be affixed to all Level 2B documents
PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 1	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2A	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2B
Level 1 ☒	Level 2A □	Level 2B □
Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.	Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only	Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here, please

Signature:	Printed Name/Position/Title: Lihshing Wang, Assistant Professor	
Organization/Address: U of Cincinnati,	Telephone: (513) 556 3628	FAX: (513) 556-3535
	E-Mail Address: Leigh.Wang@uc.edu	Date: 6/2/03

2624 Clifton Ave, Cincinnati,
OH 24221

(Over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, *or*, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:

Address:

Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:

Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

ERIC CLEARINGHOUSE ON ASSESSMENT AND EVALUATION
UNIVERSITY OF MARYLAND
1129 SHRIVER LAB
COLLEGE PARK, MD 20742-5701
ATTN: ACQUISITIONS

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
4483-A Forbes Boulevard
Lanham, Maryland 20706

Telephone: 301-552-4200
Toll Free: 800-799-3742
FAX: 301-552-4700
e-mail: ericfac@inet.ed.gov
WWW: http://ericfacility.org