

EDUCATION POLICY ANALYSIS ARCHIVES

A peer-reviewed scholarly
journal
Editor: Gene V Glass
College of Education
Arizona State University

[home](#) [abstracts](#) [complete](#) [editors](#) [submit](#) [comment](#) [notices](#) [search](#)

Copyright is retained by the first or sole author, who grants right of first publication to the **EDUCATION POLICY ANALYSIS ARCHIVES**. EPAA is a project of the [Education Policy Studies Laboratory](#).

Articles appearing in **EPAA** are abstracted in the *Current Index to Journals in Education* by the [ERIC Clearinghouse on Assessment and Evaluation](#) and are permanently archived in *Resources in Education*.

PDF file of this article available: [PDF](#)

This article has been retrieved **1703** times since September 1, 2003

Volume 11 Number 31 September 1, 2003

ISSN 1068-2341

Performance Standards: Utility for Different Uses of Assessments

Robert L. Linn
University of Colorado at Boulder
&
**National Center for Research on Evaluation,
Standards, and Student Testing**

Citation: Linn, R. L. (2003, September 1). Performance standards: Utility for different uses of assessments. *Education Policy Analysis Archives*, 11(31). Retrieved [Date] from <http://epaa.asu.edu/epaa/v11n31/>.

Abstract

Performance standards are arguably one of the most controversial topics in educational measurement. There are uses of assessments such as licensure and certification where performance standards are essential. There are many other uses, however, where performance standards have been mandated or become the preferred method of reporting assessment results where the standards are not essential to the use. Distinctions between essential and nonessential uses of performance standards are discussed. It is argued that the insistence

on reporting in terms of performance standards in situations where they are not essential has been more harmful than helpful. Variability in the definitions of proficient academic achievement by states for purposes of the No Child Left Behind Act of 2001 is discussed and it is argued that the variability is so great that characterizing achievement is meaningless. Illustrations of the great uncertainty in standards are provided.

Measurement specialists generally conceive of achievement as a continuum and they prefer to work with scale scores with many gradations rather than with a small number of categorical scores. It is recognized that there are a number of purposes for which the scores need to be lumped into a small number of categories that require the identification of one or more cut scores. Some leading measurement specialists, however, have suggested that it is best to avoid setting performance standards and associated cut scores if possible. For example, Shepard (1979) advised that it is best to “avoid setting standards whenever possible” (p. 67) and Green (1981) concluded that “fixed cutting scores should be avoided whenever possible” (Green, 1981, p. 1005).

There obviously are some purposes where the identification of one or more cut scores cannot be avoided because they are essential to the use of a test. Tests used to make licensure and certification decisions, must have a cut score identified that will collapse the score scale into only two categories – pass and fail. In other situations, the scores are collapsed into 4 or 5 categories. The College Board, for example, converts a weighted combination of scores on the multiple-choice and the constructed-response sections of their Advanced Placement (AP) Examinations to a final grade that is reported on a five-point scale:

- 5 – extremely well qualified
- 4 – well qualified
- 3 – qualified
- 2 – possibly qualified
- 1 – no recommendation.

The pass-fail dichotomy is required for the decision to be made in the case of a licensure or certification test. The use of 5 categories on AP examinations, on the other hand, also supports dichotomous decisions about whether or not a student will receive college credit based on his or her AP grade, but allows colleges and universities to determine the grade required to be awarded credit.

Certification tests and AP examinations are just two of many situations where the primary use of test scores is to determine whether the test taker has met a performance standard. The performance of the person who barely met the standard is more similar to that of the person who barely failed to meet the standard than to that of someone who exceeded the standard by a comfortable margin. Indeed, due to measurement error, there is a substantial probability that the person who barely met the standard may be a false positive while the person who barely failed may be a false negative. The two could easily switch

places if they took an alternate form of the test. Although such classification errors are of concern and should be minimized to the extent possible, they cannot be avoided altogether and there are legitimate practical reasons that require that a decision be made. Thus, Shepard's (1979) and Green's (1981) advice to avoid the use of a fixed standard or cut score cannot always be followed because standards are an essential element of the use of the test results. In the last 10 or 15 years, however, performance standards have been mandated and used with increasing frequency in situations where they are not essential to the use being made of test results.

Nonessential Uses of Performance Standards

Performance standards began to be introduced in uses of tests where they were not really essential as the result of the criterion-referenced testing movement that was spawned by Glaser's (1963) classic article. Ironically, Glaser's conceptualization of criterion-referenced measurement did not require the establishment of a fixed standard or cut score, but the use of cut scores to determine that a student either did or did not meet a performance standard became associated with criterion-referenced tests (Hambleton & Rogers, 1991; Linn, 1994). Glaser's later discussions of criterion-referenced testing (e.g. Glaser & Nitko, 1971) recognized the use of criterion in the sense of a fixed standard noting that "[a] second prevalent interpretation of the term criterion in achievement measurement concerns the imposition of an acceptable score magnitude as an index of achievement" (p. 653). Nonetheless, the setting of a fixed standard or cut score is not an essential to criterion-referenced measurement.

There are a number of reasons to question the wisdom of setting a performance standard for a test if the standard is not essential to the use of the test results. Green's (1981) desire to avoid setting fixed performance standards whenever possible was based on the recognition that a single item provides very little information by itself but "one item may well make the difference between passing and failing" (p. 1005). Others who have been critical of the use of performance standards have focused on limitations of the standards. Based on his review of standard setting methods, Glass (1978), for example, concluded that standards setters "cannot determine 'criterion levels' or standards other than arbitrarily. The consequences of arbitrary decisions are so varied that it is necessary either to reduce the arbitrariness, and hence the unpredictability or to abandon the search for criterion levels altogether in favor of ways of using test data that are less arbitrary, and hence safer" (p. 237). Although respondents to Glass's article noted that although standards are arbitrary in the sense that they are set judgmentally, they need not be capricious (Hambleton, 1978; Popham, 1978).

For a licensure test there is a clear context for standard setters who have a responsibility for thinking about the minimal skills required to protect the public from incompetent practitioners. The broader good of requiring some minimal level of performance on a test to be certified and therefore allowed to practice justifies the judgmentally established cut score. Moreover, the need to make

certification decisions provides judges with a clear context for setting the standard. Standard setters may also have a clear sense of the proportion of candidates who have passed the examination in the past. Similarly, for an AP test there is a clear link between college grades assigned in courses for which credit may be awarded. Judgments regarding minimal acceptable performance when standards are set on tests that must be passed to graduate from high school or for promotion to the next grade have a more generalized context than the one that applies to setting standards for licensure examinations. The need to establish a cut score is essential to the use of a test as a high school graduation requirement and at least some of the consequences of the use are known to the standard setters who can weigh the potential benefits (e.g., restoring the meaning of a high school diploma or motivating greater effort by teachers and students, Mehrens & Cizek, 2001, p. 480) against the potential negative consequences (e.g., that the minimums will become the maximum or that more students will drop out of school, Mehrens & Cizek, 2001, p. 481).

In many instances the standards that are set are not used to make any pre-specified decisions about individual students. Instead the performance standards are used for reporting the performance of groups of students and for tracking progress of achievement for schools, states, or the nation. Examples include the setting of performance standards for the National Assessment of Educational Progress (NAEP) or a state assessment. The context that standard setters have had for setting a cut score to determine proficient, basic, or advanced performance for NAEP or a state assessment has, until recently, lacked any clear context other than a sense of an aspiration for high levels of student achievement.

Standards and Aspirations

Six broad educational goals, two of which concerned student achievement, were agreed to by Governors and the President at the Education Summit held in Charlottesville, Virginia in 1989. The National Educational Goals Panel was created and given the responsibility of monitoring progress toward the goals set at the Education Summit. Five years after the Charlottesville summit, the Goals 2000: Educate America Act of 1994 was signed into law (Public Law 103-227). Goals 2000 encouraged standards-based reporting of student achievement. As defined by a technical planning group for the Goals Panel, “performance standards specify ‘how good is good enough’” (National Educational Goals Panel, 1993, p. 22). Unanswered is the question: good enough for what? It is clear that the performance standards are expected to be absolute rather than normative. (Although normative comparisons have been eschewed by many proponents of criterion-referenced tests and standards-based assessments, norms have considerable utility for providing comparisons of relative performance across content areas and even for a single measure are often more readily interpreted than are criterion-referenced reports or reports of results in terms of performance standards (see, for example, Hoover, 2003).) In keeping with the zeitgeist of the time, it is also clear that they were expected to be set at “world class” levels.

The high aspirations of 1989 Education Summit which were encouraged by Goals 2000 and the Goals Panel provided the primary context for the setting of standards on NAEP and on many state assessments. As might have been expected the performance standards set on NAEP on a number of state assessments were set at high levels. In support of the judgment that the performance standards were set at ambitious levels consider the fact that in 1990, the first year that NAEP results were reported in terms of achievement levels, the proficient level on the mathematics assessment corresponded to the 87th percentile for 4th grade students, the 85th percentile for 8th grade students, and 88th percentile for 12th grade students (see, for example, Braswell, Lutkus, Grigg, Santapau, Tay-Lim & Johnson, 2001). Linkages of NAEP to the Third International Mathematics and Science Study (TIMSS) grade 8 mathematics results which revealed that no country is anywhere close to having all of their students scoring at the proficient level or higher (Linn, 2000) also attest to conclusion that the NAEP achievement levels are set at very ambitious levels.

Performance standards set by a number of states for their state assessments during the past decade have also been set at quite high levels in many cases. The high levels that were set by many states are evident from the linkage of state assessments to NAEP (e.g., McLaughlin & Bandeira de Mello; 2002, see also Linn, in press, for a discussion of the McLaughlin & Bandiera de Mello results). It is not unusual for a state to have the proficient performance standard set at the 70th percentile or even higher. Since performance standards on NAEP and on a number of state assessments have, in most cases in the past, had no real consequences for students and there is no requirement for actually achieving the aspiration of having all students at the proficient level or above, one might conclude that there is no harm in having high standards. There may be reason for concern that reports that, say, less than half the students are proficient may paint an unduly negative picture for the public, but many would argue that it is good to have ambitious goals even if they are never achieved. "If you reach for the stars, you may not quite get them, but you won't come up with a handful of mud either" (Samuel Butler, as quoted by Applewhite, Evans, and Frothingham, 1992, p. 22).

It is one thing to set performance standards at the height of stars when there are no requirements of achieving them, but it is another matter altogether when there are serious sanctions for falling short such as those that have recently been put in place by the No Child Left Behind (NCLB) Act of 2001 (Public Law 107-110). NCLB sets the goal of having all children at the proficient level or higher in both reading/language arts and mathematics by 2013-2014. It also requires schools, districts, and states to meet adequate yearly progress (AYP) targets in intermediate years that would assure that they are track to having 100% of the students at the proficient level or above by 2013-2014. There are severe sanctions for schools that fall short of AYP targets for two or more years in a row.

The percentage of students who are at the proficient level or above for purposes of NCLB is to be determined by state assessments using performance

standards established by each state. “Each State shall demonstrate that the State has adopted challenging academic content standards and challenging student academic achievement standards that will be used by the State” (P. L. 107-110, Section 1111(b)(1)(A)).

All states had to submit plans explaining how they were going to meet the accountability requirements of NCLB by January 31, 2003. States that were in process of introducing new assessments or that had not yet set performance standards will be setting standards in quite a different context than existed prior to the enactment of NCLB. In light of the new context provided by NCLB, it is reasonable to expect that they will set the standards at less ambitious levels than they would have been set a couple of years earlier. The standards recently set by Texas for their new assessment, the Texas Assessment of Knowledge and Skills (TAKS), are consistent with the expectation that states may set their sights a little lower in the context of NCLB.

States that already had their assessments and performance standards in place prior to the enactment of NCLB faced a dilemma. They confronted the question of whether they should stay the course, recognizing that their performance standards were set at levels that are unrealistic for all children to achieve within the next 12 years. Or should they lower their performance standards and risk being accused of dumbing down their standards? Some states, e.g., Colorado and Louisiana redefined their performance levels for purposes on NCLB. Colorado, for example, has reported results on the Colorado Student Assessment Program (CSAP) in terms of four levels: unsatisfactory, partially proficient, proficient, and advanced. Colorado will continue to use all four levels for reporting to parents, schools, and districts. For purposes of NCLB, however, Colorado has collapsed the partially proficient and proficient levels into one level called proficient.

NCLB Starting Points for States

In order to track their AYP toward the goal of 100% proficient or above by 2013-2014, states have to define percentage proficient starting points. The starting point for each subject (reading/language arts and mathematics) is defined to be equal to the higher of the following two values: (1) the percentage of students in the lowest scoring subgroup who achieve at the proficient level or above and (2) “the school at the 20th percentile in the State, based on enrollment, among all schools ranked by the percentage of students at the proficient level” (P.L. 107-110, Sec. 111 (b)(2)(E)(ii)). In most cases the latter value will be the higher one and define the starting point.

Because states have their own assessments and set their own performance standards it should not be at all surprising that state NCLB starting points are quite variable. Some states are yet to define their performance standards and/or starting points and some states have expressed their starting points in terms of scale scores that make comparisons difficult. Percentage proficient or above for reading/language arts starting points are available for 34 states at grades 4 and 8 (Olson, 2003). At grade 4, the starting percentages range from a low (i.e.,

most stringent) of 13.6% for California to a high (i.e., most lenient) of 77.5% for Colorado, with a median of 49.35%. At grade 8, the starting points range from a low of 13.6% to a high of 74.6% with a median of 46.2%. As at grade 4, California and Colorado define the extremes at grade 8. At grade 4, eight states have starting point percentages of 34% or less and eight states have starting points 64% or more. The corresponding percentages at grade 8 are 35% and 60%. State NAEP results indicate that states do vary in terms of student achievement, but not nearly enough to explain the huge variability in NCLB percentage proficient starting points. For the 43 states that participated in the 2002 NAEP 4th grade reading assessment, for example, the percentage of students who were at the proficient level or above ranged from a low of 14% in Mississippi to a high of 47% in Massachusetts (Grigg, Daane, Jin & Campbell, 2003).

The variability in the starting points is of similar magnitude for mathematics at grades 4 and 8 as that found for reading/language arts. The range for mathematics at grade 4 is from 8.3% in Missouri to 79.5% in Colorado and at grade 8 the range is from 7% in Arizona to 74.6% in North Carolina. On the 2000 NAEP mathematics assessment, North Carolina students did perform somewhat better than Arizona students. Thirty percent of the North Carolina students were at the proficient level or above on the grade 8 mathematics compared to 21% in Arizona (Braswell, Lutkus, Grigg, Santapau, Tay-Lim & Johnson, 2001). The grade 8 mathematics achievement of students in Arizona and North Carolina appears much more similar on NAEP, however, than is suggested by the starting points of 7% and 74.6%.

Controversy Regarding Performance Standards

Performance standards have been the subject of considerable controversy. The performance standards called achievement levels set on the NAEP assessments, for example, have been subjected to harsh criticism. Reviews by panels of both the National Academy of Education (NAE) (Shepard, Glaser, Linn, & Bohrnstedt, 1993) and the National Research Council (NRC) (Pellegrino, Jones, & Mitchell, 1998) concluded that the procedure used to set the achievement levels was "fundamentally flawed" (Shepard, et al., 1993, p. xxii; Pellegrino, et al., 1998, p. 182). The conclusions of the NAE and NRC panels were controversial and several highly-regarded measurement experts have defended the procedure used to set the NAEP achievement levels as well as the resulting levels (e.g., Cizek, 1993; Kane, 1993, 1995; Mehrens, 1995).

There is an abundance of methods for setting standards, but there is no agreed upon best method. This point was made repeatedly at the Joint Conference on Standard Setting for Large-Scale Assessments sponsored by the National Assessment Governing Board and the National Center for Education Statistics held October 5-7, 1994 in Washington, DC. The Joint Conference included 18 presentations by scholars representing multiple perspectives. The papers dealt with a variety of issues ranging from technical to policy and legal issues. Crocker and Zieky (1995) prepared an executive summary of the conference which included the following summary conclusion.

“Even though controversies and disagreements abounded at the conference, there were some areas of general agreement. Authors agreed that setting standards was a difficult, judgmental task and that procedures used were likely to disagree with one another. There was clear agreement that the judges employed in the process must be well trained and knowledgeable, represent diverse perspectives, and that their work should be well documented” (p. ES-13).

Variability in Standards

As was indicated by Crocker and Zieky, there is a broad consensus in the field that different methods of setting standards will yield different standards. This consensus is consistent with Jaeger’s (1989) summary of the literature on the comparability of standard setting methods. “Different standard-setting procedures generally produce markedly different standards when applied to the same test, either by the same judges or by randomly parallel samples of judges” (Jaeger, 1989, p. 497). It is also agreed that different groups of judges will set different standards when using the same method, especially when the groups represent different constituencies (e.g., teachers, administrators, parents, the business community, or the general public). Moreover, there is general agreement that “... there is NO ‘true’ standard that the application of the right method, in the right way, with enough people, will find” (Zieky, 1994, p. 29).

Given that there is no “true standard or “best” method for setting a standard, it is reasonable to ask what should be made of the variability in results as the function of choice of methods or choice of judges. If one wants to make generalizations across methods or groups of judges then it would seem reasonable to treat the variability in results as error variance. In doing so, we would at least acknowledge that there is a high degree of uncertainty associated with any performance standard.

Variability Due to Judges

Attempts are often made to estimate the error variability due to judges as part of the standard setting process. A difficulty that is encountered, however, is that standard setting methods usually involve group discussion following an initial set of judgments which may be made independently. Group discussion obviously makes judgments in subsequent rounds dependent which makes it impossible to estimate the error variability due to judges in the final round of judgments. Furthermore, there are good reasons to believe that the person who leads the standard setting exercise may have an important influence on the outcome. Thus, what would be desired is something akin to duplicate-construction experiment that Cronbach (1971) proposed as a way to evaluate the content validity of a test. The duplicate-construction experiment would require that two teams of “equally competent writers and reviewers” (p. 456) independently construct alternate tests. The parallel in standard setting would involve the use of independent panels of comparably qualified judges set the standards under the direction of equally competent leaders using the same method and instructions. The variance in the standards for the two panels would provide an

estimate of the amount of error due to the panel of judges and standard setting leader.

Since the bigger sources of variability in standards is apt to come from the way in which judges are identified and the method that is used to set standard, even the parallel of Cronbach's duplicate-construction experiment would greatly underestimate the real degree of uncertainty in the standards. Some idea of the degree of variability due to the identification of judges is provided by results of a study conducted by Jaeger, Cole, Irwin, and Pratto (1980). Jaeger and his colleagues had three panels, consisting of samples of teachers, school administrators, and counselors, respectively, independently set passing standards on a North Carolina test. The differences in the standard set by the different panels can be gauged by the magnitude of the differences in the proportion of students who would have failed the test according to the different groups of judges. On the reading test the proportion who would have failed ranged from a low of 9% to a high of 30%. The variability in failure rates was even greater for the mathematics test, ranging from a low of 14.4% to a high of 71.1%.

Variability Due to Method

Variability due to choice of method can be evaluated based on results of several different studies that were reviewed by Jaeger (1989). One of those studies where multiple methods were used, for example, was conducted by Poggio, Glassnapp and Eros (1981). They had independent samples of teachers set standards using one of four different standard setting methods: the Angoff (1971) method, the Ebel (1972) method, the Nedelsky (1954) method, or the contrasting groups (see, for example, Jaeger, 1989) method. Teachers set standard for tests at grades 2, 4, 6, 8, and 11. There was substantial variability in the standards set by the four different methods at every grade. At grade 8, for example, the four different methods would set the minimum passing score on a 60 item reading test at 28, 39, 43, and 48 items correct. If the most lenient standard had been used, just over 2% of the students would have failed whereas approximately 29% would have failed if the most stringent standard had been used.

In his summary of 32 contrasts of standards set by different methods Jaeger (1989) found that the ratios of the percentages of examinees who would fail range from a low of 1.00 to a high of 29.75 with a median of 2.74. That is, the typical consequence of using a different method to set standards would be to alter the failure rate by a factor of almost 3. As Jaeger (1989) concluded the "choice of a standard setting method is critical" (p. 500). He went on to endorse earlier suggestions by Hambleton (1980), Koffler (1980) and Shepard (1980; 1984) that "it might be prudent to use several methods in any given study and then consider all the results, together with extra-statistical factors, when determining a final cutoff score" (Jaeger, 1989, p. 500).

Because of the practical cost considerations, the use of multiple methods as input to a final standard setting decision is rare in operational practice, but that

was the approach recently taken in Kentucky for the assessments introduced in the state in 2000. As was recently reported by Green, Trimble and Lewis (2003), the Kentucky Department of Education used multiple methods as input to their final standard setting when the state introduced a new testing system, the Kentucky Core Content Test (KCCT) in 2000. First, three different methods, the bookmark procedure (Lewis, Green, Mitzel, Baum & Patz, 1998; Mitzel, Lewis, Patz, & Green 2001), the Jaeger-Mills procedure (Jaeger & Mills, 2001), and the contrasting group (see, for example, Jaeger, 1989) were used to set cut scores to distinguish four levels of performance (novice, apprentice, proficient and distinguished) on each of 18 different tests used in the KCCT system for various grade levels and content areas. The results of the bookmark, the Jaeger-Mills, and the contrasting groups standards setting efforts were input to a synthesis process where the results were considered by teacher committees that recommended cut scores to the Kentucky State Board of Education (Green, Trimble, & Lewis, 2003; see also CTB/McGraw-Hill, 2001; and Kentucky Department of Education, 2001, for more detailed descriptions).

Table 1 displays the percentage of students at the proficient level or above on each of six KCCT subject area tests administered at elementary school grades according to the three standard setting methods. Also shown are summary statistics across the three methods, mean, standard error, minimum, maximum, and range as well as the percentage proficient or above according to the standard set by the synthesis panel. The standard error is simply the standard deviation of the percentages for the three different standard setting methods since the standard deviation may be interpreted as a standard error if the goal is to generalize across standard setting methods. As can be seen, the standard errors are quite large, indicating that there is considerable uncertainty about the percentage of students proficient or above due to standard setting method.

Method or Statistic	Subject						Mean
	Reading	Mathematics	Science	Social Studies	Arts & Humanities	Practical Living/Vocational Studies	
Bookmark	56.5%	35.2%	35.4%	48.4%	15.3%	44.8%	39.3%
Jaeger-Mills	15.3	20.7	4.7	4.5	11.0	16.4	12.1
Contrasting Groups	29.4	19.5	24.5	27.2	24.8	24.7	25.0
Methods Mean	33.7	25.1	21.5	26.7	17.0	28.6	25.5
Standard Error	20.94	8.74	15.56	21.95	7.06	14.60	14.81

Maximum	56.5	35.2	35.4	48.4	24.8	44.8	40.85
Minimum	15.3	19.5	4.7	4.5	11.0	16.4	11.9
Range	41.2	15.7	30.7	43.9	13.8	28.4	29.0
Synthesis Standard	57.2	31.2	35.9	39.8	13.3	45.4	37.1

Tables 2 and 3 display results parallel to those in Table 1 for the KCCT tests administered at the middle school and high school grades, respectively. The results for the 12 subject area by grade combinations shown in Tables 2 and 3 are similar to those shown in Table 1 for the tests administered at the elementary school grades. The mean standard error across the 18 subject area by grade combinations in Tables 1 through 3 is 10.82 and they range from a low of 4.26 for the middle school mathematics test to a high of 26.35 for the middle school reading test. Even in the best case there is a good deal of uncertainty about the percentage of students who are at the proficient level or above as the consequence of the method used to set the performance standards. A standard error of even 4 points is large relative to the annual change in percentage proficient or above that is likely to be required to meet the AYP target for NCLB. A standard error of 26 points is gigantic in that same context.

Table 2 Percentage of Students at the Proficient Level or Above on KCCT For Middle School Grade Tests According to Standard Setting Method							
Method or Statistic	Subject						Mean
	Reading	Mathematics	Science	Social Studies	Arts & Humanities	Practical Living/Vocational Studies	
Bookmark	61.0%	19.8%	50.6%	28.6%	41.6%	40.6%	40.4%
Jaeger-Mills	10.5	17.7	10.4	12.8	14.6	16.2	13.7
Contrasting Groups	22.7	25.9	23.7	22.8	23.9	30.9	25.0
Methods Mean	31.4	21.1	28.2	21.4	26.7	29.2	26.4
Standard Error	26.35	4.26	20.48	7.99	13.72	12.29	14.18
Maximum	61.0	25.9	50.6	28.6	41.6	40.6	41.4
Minimum	10.5	17.7	10.4	12.8	14.6	16.2	13.7
Range	50.5	8.2	40.2	15.8	27.0	24.4	27.7
Synthesis Standard	51.0	25.2	27.3	28.3	35.9	35.3	33.8
Table 3 Percentage of Students at the Proficient Level or Above on KCCT For High School Grade Tests According to Standard Setting Method							

Method or Statistic	Subject						Mean
	Reading	Mathematics	Science	Social Studies	Arts & Humanities	Practical Living/ Vocational Studies	
Bookmark	27.1%	10.9%	19.0%	22.5%	21.0%	47.8%	30.7%
Jaeger-Mills	12.0	10.9	2.6	14.4	11.0	17.5	11.4
Contrasting Groups	26.3	25.9	30.9	24.0	24.5	33.5	27.5
Methods Mean	21.8	15.9	17.5	20.3	18.8	32.9	21.2
Standard Error	8.50	8.66	14.21	5.16	7.00	15.16	9.78
Maximum	27.1	25.9	30.9	24.0	24.5	47.8	30.3
Minimum	12.0	10.9	2.6	14.4	11.0	17.5	11.4
Range	15.1	15.0	28.3	9.6	13.5	30.3	18.6
Synthesis Standard	27.5	26.3	27.3	24.0	19.5	48.4	28.8

Of course Kentucky did not use any of the three methods to set the performance standards for operational use. Rather the results of the three methods were used as input to the synthesis panels that provided the final recommendations to the State Board of Education. Hence, it might be argued that the variability due to method is not relevant to judging the uncertainty of the final performance standards. Thus, a reasonable question is what is the degree of uncertainty for operational standards in Kentucky or any other state? Results recently reported by Hoover (2003) are relevant to this question. Hoover compared of the percentage of students labeled proficient or advanced according to three national test batteries and NAEP. Hoover's results show that performance standards that are finally adopted after much care and expense by national test publishers for their tests or by the National Assessment Governing Board for NAEP also have a great deal of uncertainty.

According to the three national tests the percentage of grade 4 students who are proficient or above in reading is 24% according to one test publisher, 40% according to another publisher, and 55% according to the third publisher. According to NAEP the figure is 31%. For grade 4 mathematics, the corresponding four numbers are 15%, 34%, 44%, and 26% (Hoover, 2003, p. 11). Hoover's comparisons also show that, whereas there is apparently a substantial decline in the percentage of students who are proficient or above from grade 8 to 9 (33% vs. 11%) according to one publisher, there is no such decline according to the other two publishers. While one publisher's performance standards show a fairly steady decline in percentage of students who are proficient or advanced in mathematics from grades 1 through 12 (from 41% to 5%) another publisher's performance standards show that slightly more than twice as many students (27% vs. 12%) are proficient or advanced at grade

12 than at grade 1.

Conclusion

The variability in the percentage of students who are labeled proficient or above due to the context in which the standards are set, the choice of judges, and the choice of method to set the standards is, in each instance, so large that the term proficient becomes meaningless. Insistence on standards-based reporting of achievement test results where such reporting serves no essential purpose is more harmful than helpful. This is particularly true in the context of NCLB where schools, districts, and states are subject to substantial sanctions based on the progress that is made against arbitrary performance standards that lack any semblance of comparability from state to state.

Several years ago I (Linn, 1995) described four uses of performance standards: exhortation, exemplification of goals, accountability, and the certification of student achievement. Performance standards and associated cut scores are essential only for the fourth use. Although reporting results in terms of performance standards is often done to exhort teachers and students to exert greater effort standard-based reporting is not essential to that use. Nor are performance standards essential for the purpose of exemplifying goals. NCLB and a number of state accountability systems depend on the performance standards, but that would not have to be the case.

One of the purposes of introducing performance standards is to provide a means of reporting results in a way that is more meaningful than a scale score. Certainly, reporting that a student performed at the proficient level appears more understandable than saying that the student has a scale score of 215. Parents familiar with student performance in school in terms of grades of, say, A, B, C, D and F might naturally assume that proficient is like a grade of B or C. Given the huge inconsistencies in definitions of proficient achievement and in the associated stringency of cut scores, however, it seems clear that attaching such an interpretation to performance levels cannot mean the same thing across states where standards vary so radically in their stringency. It would be better to find another way of dealing with these non-essential uses of performance standards and cut scores.

There obviously is a legitimate interest in being able to measure progress in student achievement. There are many ways of measuring progress and setting AYP targets that do not depend upon the reporting of results in terms of performance standards. Effect-size statistics that would measure the year-to-year difference in mean achievement scores in terms of the standard deviation of scores in the base year is one obvious way that progress could be measured. Holland's (2002) proposal to measure progress in student achievement by comparing cumulative distribution functions is another approach. Comparisons of cumulative distribution functions provide a means of monitoring changes in student performance throughout the range of performance. Changes in the percentage of student exceeding any selected score level can be readily determined rather than just focusing on one arbitrary cut score that corresponds

to the proficient performance standard.

Comparisons might also be made to norms for a base or reference year. If improvement in performance in State A was large enough that three quarters of the students in 2013-2014 performed above the median level in 2002-2003 that would represent a large improvement in student achievement. It would also be readily understood that that students in state A generally had better achievement than students in state B where 55% of the students in 2013-2014 scored above the 2002-2003 median. Furthermore, the norm-based results for states A and B would be much more interpretable than a statement that three-fourths of the students in State C were proficient or above in 2013-2014 compared to only 55% in State D, because the meaning of proficient might be radically different for States C and D. Indeed, if the stringency of the proficient performance standard were as variable from state to state as it is now, it might well be that the achievement of students in state D was actually better than that in state C.

Finally, if it is decided that the best way to track progress is in terms of the percentage of students scoring above a fixed cut score, sometimes referred to as PAC for percent above cut, then it would be better to pick the cut score based on norms in a base year than to use an arbitrary definition of “proficient” performance that bears little similarity to the definition of proficient performance in another state. The median or some other percentile rank in a base year might be used as the cut score. This would provide a clear and consistent meaning that does not seem possible to achieve for the proficient performance standard. Using PAC statistics to track progress would provide a reasonable alternative to tracking progress in achievement for different states in terms of percent proficient or above. PAC statistics based on a cut score at, say, the median achievement level in a base year would also be much more interpretable than percentages proficient or above when comparisons are made across states.

Reports of individual student assessment results in terms of norms have more consistent meaning across different assessments than reports in terms of proficiency levels based on uncertain standards. Furthermore criterion-referenced reports of results can be provided by illustrating the types of items that the student can answer correctly and the types that they cannot. A fixed cut score is not essential to criterion-referenced interpretations of achievement.

Postscript

For the reasons discussed above, I believe it would be desirable to shift away from standards-based reporting for uses where performance standards are not an essential part of the test use. I recognize, however, that existing state and federal laws require the setting of performance standards and the reporting of results in terms of those standards. Thus, at least until the laws are changed, there is no choice but to work to make performance standards as reasonable as possible. Assuring that judges on standard setting panels understand the context in which the standards will be used is a minimal requirement for obtaining reasonable performance standards. Normative information needs to

be made part of the process for judges to anchor their absolute judgments with an understanding of current levels of performance of students and likely consequences. As Zieky (2001) has noted, considering both absolute and normative information “in setting a cutscore can help avoid the establishment of unreasonably high or low values” (p. 38). In addition to knowing the percentile rank corresponding to particular cut scores, it would also be desirable to have some means of providing judges with comparative information about the relative stringency of their standards in comparison to standards set in other states before judgments are finalized. Normative information would be one way of making comparisons to standards in other states.

It is critical that the context in which the standards will be used be made as clear as possible to panels of judges who set the standards. The uses of the standards for purposes of NCLB with its expectation that all students will be at the proficient level or higher by 2014 and sanctions for schools that do not meet AYP targets is an important part of the current context that standard setters need to consider.

Finally, while there is no agreed upon best method for setting standards, the literature does provide useful indications of the differences among different methods in their relative stringency and ease of use. Jaeger’s (1989) advice that multiple standard setting methods be used and the results of the different methods be “considered together with extrastatistical factors when determining the cutoff score” (p. 500) seems as sound now as it in 1989. The experience in Kentucky with the use of multiple methods as input to a synthesis panel provides an excellent example where that advice was taken seriously in practice.

A number of authors have suggested helpful criteria to consider in selecting a method. Hambleton (2001), for example, identifies 20 criteria that he presents in the form of questions to be considered in evaluating a standard-setting process. Raymond and Reid (2001) provide useful advice on the selection and training of judges, Kane (2001) provides a good discussion of considerations in the validation of performance standards and cut scores, and Bond (1995) has identified five principles intended to help ensure fairness. Although such considerations cannot eliminate the arbitrariness, they can help make the standards and cut scores more reasonable and more defensible.

Acknowledgment

Based on a paper presented at the College Board, New York City, July 16, 2003. Work on the paper was partially supported under the Educational Research and Development Center Program PR/ Award #R305B60002, as administered by the Institute of Education Sciences, U.S. Department of Education. The findings and opinions expressed in this paper do not reflect the position or policies of the National Center for Education Research, the Institute of Education Sciences, or the U.S. Department of Education.

References

- Angoff, W. H. (1971). Scales, norms and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508-600). Washington, DC: American Council on Education.
- Applewhite, A., Evens, W. R. III, & Frothingham, A. (1992). *And I quote*. New York: St. Martin's Press.
- Bond, L. (1995). Ensuring fairness in the setting of performance standards. In *Proceedings of the Joint Conference on Standard Setting for Large-Scale Assessments*. (pp. 311-324). Washington, DC: National Assessment Governing Board and National Center for Education Statistics.
- Braswell, J. S., Lutkus, A. D., Grigg, W. S., Santapau, S. L., Tay-lim, B.S.-H., & Johnson, M. S. (2001). *The nation's report card: Mathematics 2000*. Washington, DC: National Center for Education Statistics.
- Cizek, G. J. (1993). *Reactions to National Academy of Education report, Setting performance standards for student achievement*. Washington, DC: National Assessment Governing Board.
- Crocker, L. & Zieky, M. (1995). Executive Summary: Joint conference on setting standards for large-scale assessments. In *Proceedings of the Joint Conference on Standard Setting for Large-Scale Assessments*. (pp. ES-1 – ES-17). Washington, DC: National Assessment Governing Board and National Center for Education Statistics.
- Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443-507).
- CTB/McGraw-Hill. (2001), *Kentucky standard setting technical report*. Monterey, CA: Author.
- Ebel, R. L. (1972). *Essentials of educational measurement*. Englewood Cliffs, NJ: Prentice-Hall.
- Glaser, R. (1963). Instructional technology and the measurement of learning outcomes. *American Psychologist*, 18, 519-521.
- Glaser R. & Nitko, A. J. (1971). Measurement in learning and instruction. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 625-670). Washington, DC: American Council on Education.
- Glass, G. V. (1978). Standards and criteria. *Journal of Educational Measurement*, 15, 237-261.
- Goals 2000: Educate America Act of 1994, Public Law 103-227, Sec. 1 et seq. 108 Stat. 125 (1994).
- Green, B. F. (1981). A primer of testing. *American Psychologist*, 36, 1001-1011.
- Green, D. R., Trimble, C. Scott, & Lewis, D. M. (2003). Interpreting the results of three different standard-setting procedures. *Educational Measurement: Issues and Practice*, 22, No. 1, 22-32.
- Grigg, W. S., Daane, M. C., Jin, Y. & Campbell, J. R. (2003). *The nation's report card: Reading 2002*. Washington, DC: National Center for Education Statistics.
- 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 testing: Motives, models, measures and consequences* (pp. 80-123). Baltimore, MD: Johns Hopkins University Press.
- Hambleton R K (2001) Setting performance standards on educational assessments and criteria for

- evaluating the process. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods and perspectives* (pp. 89-116). Mahwah, NJ: Lawrence Erlbaum.
- Hambleton, R. K. & Rogers, J. H. (1991). Advances in criterion-referenced measurement. In R. K. Hambleton & J. N. Zall (Eds.), *Advances in educational and psychological testing* (pp. 3-44). Boston: Kluwer Academic.
- Hoover, H. D. (2003). Some common misconceptions about testing. *Educational Measurement: Issues and Practice*, 22, No. 1, 5-14.
- Holland, P. (2002). Measuring progress in student achievement: Changes in scores and score gaps over time. *Report of the Ad Hoc Committee on Confirming Test Results*. Washington, DC: National Assessment Governing Board.
- Jaeger, R. M. (1989). Certification of student competence. In R. L. Linn (ed.), *Educational Measurement* (3rd ed., pp. 485-514). New York: Macmillan.
- Jaeger, R. M., Cole, J., Irwin, D. M., and Pratto, D. J. (1980). *An interactive structure judgment process for setting passing scores on competency tests applied to the North Carolina high school competency tests in reading and mathematics*. Greensboro, NC: Center for Education Research and Evaluation, University of North Carolina at Greensboro.
- Jaeger, R. M. and Mills, C. (2001). An integrated judgment procedure for setting standards on complex, large-scale assessments. In G. J. Cizek (Ed.), *Setting performance standards; Concepts, methods, and perspectives* (pp. 313-318). Mahwah, NJ: Lawrence Erlbaum.
- Kane, M. (1993). *Comments on the NAE evaluation of NAGB achievement levels*. Washington, DC: National Assessment Governing Board.
- Kane, M. (1995). Examinee-centered and task-centered standard setting. In *Proceedings of the Joint Conference on Standard Setting for Large-Scale Assessments*. (pp. 119-141). Washington, DC: National Assessment Governing Board and National Center for Education Statistics.
- 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.
- Kentucky Department of Education. (2001). *Standard setting: Synthesis of three procedures – procedures and findings*. Frankfort, KY: Kentucky Department of Education. Available at www.kde.state.ky.us.
- Koffler, S.L. (1980). A comparison of approaches for setting proficiency standards. *Journal of Educational Measurement*, 17, 167-178.
- Lewis, D. M., Green, D. R., Mitzel, H. C., Baum, K., & Patz, R. J. (1998). *The Bookmark standard setting procedure: Methodology and recent implementations*. Paper presented at the 1998 annual meeting of the National Council on Measurement, San Diego, CA.
- Linn, R. L. (1994). Criterion-referenced measurement: A valuable perspective clouded by surplus meaning. *Educational Measurement: Issues and Practice*, 13 (4), 12-14.
- Linn, R. L. (1995). The likely impact of performance standards as a function of uses: From rhetoric to sanction. In *Proceedings of the Joint Conference on Standard Setting for Large-Scale Assessments*. (pp. 267-276). Washington, DC: National Assessment Governing Board and National Center for Education Statistics.

- Linn, R. L. (2000). Assessments and accountability. *Educational Researcher*, 29(2), 4-16.
- Linn, R. L. (in press). Accountability: Responsibility and reasonable expectations. *Educational Researcher*.
- McLaughlin, D. & Bandeira de Mello, V. (2002). *Comparison of state elementary school mathematics achievement standards using NAEP 2000*. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA, April.
- Mehrens, W. A. (1995). Methodological issues in standard setting for educational exam. In *Proceedings of the Joint Conference on Standard Setting for Large-Scale Assessments*. (pp. 221-263). Washington, DC: National Assessment Governing Board and National Center for Education Statistics.
- Mehrens, W. A. & Cizek, G. J. (2001). Standard setting and the public good: Benefits accrued and anticipated. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods and perspectives* (pp. 477-485). Mahwah, NJ: Lawrence Erlbaum.
- Mitzel, H. C., Lewis, D. M., Patz, R. J., & Green, D. R. (2001). The Bookmark procedure: Psychological perspectives. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods and perspectives* (pp. 249-281). Mahwah, NJ: Lawrence Erlbaum.
- National Educational Goals Panel. (1993). *Report of the Goals 3 and 4 Technical Planning Group on the Review of Education Standards*. Washington, DC: Author.
- No Child Left Behind Act of 2001. Public Law 107-110.
- Olson, L. (2003). "Approved" is relative term for Ed. Dep.: 11 states fully meet ESEA requirements. *Education Week*, Vol. XXII, No. 43. August 6, pp. 1, 34-36.
- 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.
- Poggio, J. P., Glasnapp, D. R., & Eros, D. S. (1981). *An empirical investigation of the Angoff, Ebel, and Nedelsky standard setting methods*. Paper presented at the annual meeting of the American Educational Research Association, Los Angeles, April.
- Popham, W. J. (1978). As always, provocative. *Journal of Educational Measurement*, 15, 297-300.
- Raymond, M. R. & Reid, J. B. (2001). Who made thee judge? Selecting and training participants for standard setting. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods and perspectives* (pp. 119-157). Mahwah, NJ: Lawrence Erlbaum.
- Shepard, L. A. (1979). Setting standards. In M. A. Bunda & R. Sanders (Eds.), *Practices and problems in competency-based measurement* (pp. 72-88). Washington, DC: National Council on Measurement in Education.
- Shepard, L. A. (1980). Technical issues in minimum competency testing. In D. C. Berliner (Ed.), *Review of research in education: Vol. 8* (pp. 30-82). Washington, DC: American Educational Research Association.
- Shepard, L. A. (1984). Setting performance standards. In R. A. Berk (Ed.), *Criterion-referenced testing: Motives, models, measures and consequences* (pp. 169-198). Baltimore, MD: Johns Hopkins University Press.
- Shepard, L., Glaser, R., Linn, R. & Bohrnstedt, G. (1993). *Setting performance standards for student*

achievement. Stanford, CA: National Academy of Education.

Zieky, M. J. (1995). A historical perspective on setting standards. In *Proceedings of the Joint Conference on Standard Setting for Large-Scale Assessments n Statistics*. (pp. 1-38). Washington, DC: National Assessment Governing Board and National Center for Education Statistics.

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). Mahwah, NJ: Lawrence Erlbaum.

About the Author

Robert L. Linn

School of Education
University of Colorado at Boulder
Email: Robert.linn@colorado.edu

Robert L. Linn is Distinguished Professor of Education at the University of Colorado at Boulder and Co-Director of the National Center for Research on Evaluation, Standards, and Student Testing. He has published over 200 journal articles and chapters in books dealing with a wide range of theoretical and applied issues in educational measurement and has received several awards for his contributions to the field, including the ETS Award for Distinguished Service to Measurement, the E.L Thorndike Award, the E.F. Lindquist Award, the National Council on Measurement in Education Career Award, and the American Educational Research Association Award for Distinguished Contributions to Educational Research. He is past president of the American Educational Research Association, past president of the National Council on Measurement in Education, past president of the Evaluation and Measurement Division of the American Psychological Association, and past vice-president for the Research and Measurement Division of the American Educational Research Association. He is a member of the National Academy of Education, a Lifetime National Associate of the National Academies, and serves on two Boards of the National Academy of Sciences.

The World Wide Web address for the *Education Policy Analysis Archives* is
epaa.asu.edu

Editor: Gene V Glass, Arizona State University

Production Assistant: Chris Murrell, Arizona State University

General questions about appropriateness of topics or particular articles may be addressed to the Editor, Gene V Glass, glass@asu.edu or reach him at College of Education, Arizona State University, Tempe, AZ 85287-2411. The Commentary Editor is Casey D. Cobb: casey.cobb@unh.edu.

EPAA Editorial Board

Michael W. Apple University of Wisconsin	David C. Berliner Arizona State University
Greg Camilli Rutgers University	Linda Darling-Hammond Stanford University
Sherman Dorn University of South Florida	Mark E. Fetler California Commission on Teacher Credentialing
Gustavo E. Fischman Arizona State University	Richard Garlikov Birmingham, Alabama
Thomas F. Green Syracuse University	Aimee Howley Ohio University
Craig B. Howley Appalachia Educational Laboratory	William Hunter University of Ontario Institute of Technology
Patricia Fey Jarvis Seattle, Washington	Daniel Kallós Umeå University
Benjamin Levin University of Manitoba	Thomas Mauhs-Pugh Green Mountain College
Les McLean University of Toronto	Heinrich Mintrop University of California, Los Angeles
Michele Moses Arizona State University	Gary Orfield Harvard University
Anthony G. Rud Jr. Purdue University	Jay Paredes Scribner University of Missouri
Michael Scriven University of Auckland	Lorrie A. Shepard University of Colorado, Boulder
Robert E. Stake University of Illinois—UC	Kevin Welner University of Colorado, Boulder
Terrence G. Wiley Arizona State University	John Willinsky University of British Columbia

EPAA Spanish Language Editorial Board

Associate Editor for Spanish Language
[Roberto Rodríguez Gómez](#)
[Universidad Nacional Autónoma de México](#)

roberto@servidor.unam.mx

Adrián Acosta (México) Universidad de Guadalajara adrianacosta@compuserve.com	J. Félix Angulo Rasco (Spain) Universidad de Cádiz felix.angulo@uca.es
Teresa Bracho (México) Centro de Investigación y Docencia Económica-CIDE bracho dis1.cide.mx	Alejandro Canales (México) Universidad Nacional Autónoma de México canalesa@servidor.unam.mx
Ursula Casanova (U.S.A.)	José Contreras Domingo

Arizona State University
casanova@asu.edu

[Erwin Epstein \(U.S.A.\)](#)
Loyola University of Chicago
Eepstein@luc.edu

[Rollin Kent \(México\)](#)
Universidad Autónoma de Puebla
rkent@puebla.megared.net.mx

Javier Mendoza Rojas (México)
Universidad Nacional Autónoma de México
javiermr@servidor.unam.mx

[Humberto Muñoz García \(México\)](#)
Universidad Nacional Autónoma de México
humberto@servidor.unam.mx

[Daniel Schugurensky \(Argentina-
Canadá\)](#)
OISE/UT, Canada
dschugurensky@oise.utoronto.ca

[Jurjo Torres Santomé \(Spain\)](#)
Universidad de A Coruña
jurjo@udc.es

Universitat de Barcelona
Jose.Contreras@doe.d5.ub.es

[Josué González \(U.S.A.\)](#)
Arizona State University
josue@asu.edu

[María Beatriz Luce \(Brazil\)](#)
Universidad Federal de Rio Grande do Sul-UFRGS
lucemb@orion.ufrgs.br

Marcela Mollis (Argentina)
Universidad de Buenos Aires
mmollis@filo.uba.ar

Angel Ignacio Pérez Gómez (Spain)
Universidad de Málaga
aiperez@uma.es

[Simon Schwartzman \(Brazil\)](#)
American Institutes for Resesarch–Brazil
(AIRBrasil)
simon@sman.com.br

[Carlos Alberto Torres \(U.S.A.\)](#)
University of California, Los Angeles
torres@gseisucla.edu

[home](#) [abstracts](#) [complete](#) [editors](#) [submit](#) [comment](#) [notices](#) [search](#)

EPAA is published by the Education Policy Studies
Laboratory, Arizona State University