DOCUMENT RESUME

ED 401 318 TM 025 891

AUTHOR Green, Bert F.
TITLE Setting Performance Standards: Content, Goals, and Individual Differences.
INSTITUTION Educational Testing Service, Princeton, NJ. Policy Information Center.
PUB DATE 6 Nov 95
NOTE 23p.; Paper presented at the annual William H. Angoff Memorial Lecture (2nd, Princeton, NJ, November 6, 1995).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Descriptive (141) -- Viewpoints (Opinion/Position Papers, Essays, etc.) (120)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Certification; *Course Content; *Cutting Scores; Elementary Secondary Education; *Performance Based Assessment; *Prediction; *Psychometrics; Standards; Student Motivation; Test Content
IDENTIFIERS *Standard Setting

ABSTRACT

Setting performance standards is an area that different constituencies see quite differently. The choices of elements for a particular standard depend to a large extent on the purposes the standard is intended to serve. Standards can be used in certification, as predictors, as descriptors, and as motivators. While performance standards indicate how much of a domain has been mastered, content standards define the extent of the domain to be tested. The bridge from one to the other is of central importance in validating performance standards. Performance standards must reflect content standards. The psychometric problem of determining just where a cut-point should be placed on a scale is important, but deciding what to test and how to test it are more important. In prediction, placing the standard on the right scale is important, while for description and motivation, the placement of the points is less important than having enough points to be descriptions and goals for the full range evaluated. Finding a way to map content standards onto performance standards is an extremely important challenge in standard setting. (Contains 3 figures, 3 tables, and 23 references.) (SLD)

WILLIAM H. ANGOFF
MEMORIAL
LECTURE Series

SETTING PERFORMANCE STANDARDS: CONTENT, GOALS, AND INDIVIDUAL DIFFERENCES

BY
BERT F. GREEN

William H. Angoff
1919 - 1993

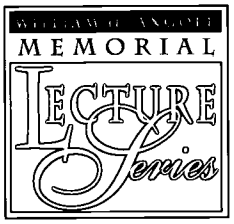
William H. Angoff was a distinguished research scientist at ETS for more than forty years. During that time, he made many major contributions to educational measurement and authored some of the classic publications on psychometrics, including the definitive text "Scales, Norms, and Equivalent Scores," which appeared in Robert L. Thorndike's Educational Measurement. Dr. Angoff was noted not only for his commitment to the highest technical standards but also for his rare ability to make complex issues widely accessible.

The Memorial Lecture Series established in 1994 honors Dr. Angoff's legacy by encouraging and supporting the discussion of public interest issues related to educational measurement. The annual lectures are jointly sponsored by ETS and an endowment fund that was established in Dr. Angoff's memory.

The William H. Angoff Lecture Series reports are published by the Policy Information Center, which was established by the ETS Board of Trustees in 1987 and charged with serving as an influential and balanced voice in American education.

Setting Performance Standards:

Content, Goals, and Individual Differences



The second annual William H. Angoff Memorial Lecture was presented at Educational Testing Service, Princeton, New Jersey, on November 6, 1995.

Bert F. Green
Johns Hopkins University

Policy Information Center
Educational Testing Service
Princeton, NJ 08541-0001

Preface

The William H. Angoff Lecture Series Reports are the most recent addition to the roster of Policy Information Center publications.

In the 10 years since the National Council of Teachers of Mathematics (NCTM) published the oft cited *Curriculum and Evaluation Standards for School Mathematics,* much of the discussion and debate in educational reform has stemmed from one word: standards. While professional and product standards are widely accepted, educational standards are still in their infancy. Educators, parents, and policymakers across the U.S. are struggling to achieve consensus on what schools should teach and what students should know and be able to do. The quest for meaningful standards can be seen everywhere — from the burgeoning documents from various professional associations following suit with the NCTM to state-level initiatives in Kentucky and California to the national efforts to establish skills standards.

In the current report, Dr. Bert F. Green, professor of psychology at Johns Hopkins University, turns our attention to the key issues associated with the daunting task of setting performance standards. Dr. Green observes that standards are poorly understood and that the measurement community faces numerous challenges in identifying methods for standard setting. However, he argues that the primary policy problem is not how to set standards, but how many standards should be set, and on what measures should standards be set.

Drawing on the rich experiences of his work in the fields of psychometrics, statistics, cognitive psychology, artificial intelligence, and computerized adaptive testing and insights from his work on numerous advisory panels and professional associations, such as the National Assessment of Educational Progress and the American Psychological Association, Dr. Green examines the measurement issues of standard setting in terms of the broader purposes that standards serve.

I would like to thank the following individuals for their contribution to this publication: Ric Bruce designed the report; Carla Cooper provided desktop publishing services; Jim Chewning coordinated production; and Shilpi Niyogi was the editor.

Paul Barton
Director, Policy Information Center

Preamble

I want to thank Henry Braun, and Educational Testing Service, for inviting me to deliver the annual Angoff Memorial Lecture. Choosing a relevant topic was not difficult. Almost any topic in educational measurement would reflect Bill Angoff's work, since he had a hand in so many of the technical problems of measurement. Indeed, most topics in psychology or education would be appropriate because of Bill's service and devotion to the American Psychological Association and the science it represents. Bill and I served together on the APA Council of Representatives, as representatives of the Division of Evaluation, Measurement, and Statistics. We often shook our heads at some of the antics of our colleagues. Despite our scientific conservatism, we shared an interest in developing techniques to address the new problems that measurement always faces.

Introduction

In the inaugural Angoff memorial lecture, Bob Linn (1995) spoke about some aspects of educational standards. Today I want to talk with you about setting performance standards, an arena to which Bill Angoff contributed more than he expected. It is an area that various constituencies see quite differently, and one that causes heartburn among measurement experts. I shall also comment on content standards, and the tenuous link between content standards and performance standards, an arena that has scarcely been addressed by psychometricians. I have not contributed directly to the study of standards, but have been observing it closely, as a technical advisor to the National Assessment of Educational Progress, and as a technical advisor to the Maryland State Performance Assessment Project. I learned a lot at a joint conference on large scale assessments in 1994, especially from Michael Zieky, Samuel Livingston, and Samuel Messick.

Educational standards are always of concern to educational policymakers. There is perennial pressure for improvement. Colleges complain that the incoming students are not well-prepared. Employers complain that some high school graduates have only a slim grasp of the basics. In the 1970's, there was an upsurge of interest in doing something about standards. Most people felt that no one should be given a high school diploma without exhibiting at least minimum competency in the basic subjects of reading writing, math, and perhaps social studies. The minimum competency movement swept the country. (Jaeger, 1989). Tests of the basic skills were devised, and given to high school seniors. Cut points, or minimum competency standards were placed on the scales. To many, the demands seemed minimal indeed. Nevertheless, some students failed.

At first, the onus was placed on the students. Then students and their families complained that they had not been given an adequate opportunity to learn the required skills. Some states, such as Florida, had introduced the standards abruptly, rather than phasing them in over a span of several years. The court permitted Florida to go ahead, while the education vocabulary gained a new phrase and acronym, Opportunity To Learn (OTL) (Citron, 1983). At present, minimum competency tests are widely used and accepted. In Maryland, the tests are given as computer-based adaptive tests, and many students pass them before they enter high school, although the tests are not required until high school graduation.

Policymakers felt that the minimum competency tests would stimulate teachers and learners alike. The policy may have worked. The performance of the worst students was improved, without noticeable harm to the better students. Still, as Linn (1995) noted, there was general agreement that the minimum competency criterion led to more emphasis on the basics, and less emphasis on more advanced topics in the curriculum.

High Stakes Assessments

A new round of concern developed in the 1980's. John Cannell (1988) reported that standardized tests of achievement showed that most states were above average. Cannell named the effect after Garrison Keillor's fictional Lake Wobegon,

where all the men are strong, all the women are good-looking, and all the children are above average.

Alas, we can't all be above average. A variety of problems were identified, including out-of-date norms, teachers teaching to the tests, and in some cases teaching the actual test items themselves. It began to dawn on educational policymakers that when stakes are high, any empirical evaluative criteria, like test scores, are vulnerable to manipulation. Sometimes, this concept is slow to be recognized. In the State of Maryland, nearly all students in the public schools were being promoted every year. The phenomenon was not the result of remarkable education, nor above-average students, but seems to stem from the practice of allotting state money to school districts on the basis of the promotion rate in the schools. That formula has recently been changed.

Just last month, the Maryland State Department of Education reported that their annual Performance Assessment in the elementary schools had been compromised. Baltimore schools' Student Assessment Service published a teachers' guide to the assessment that included a few of the actual questions from the forthcoming assessment, with a slight change of wording. Dr. Steve Ferrara, whose State office was responsible for the assessment called it a significant breach of security. Dr. Amprey, Superintendent of Baltimore Schools, called it "poor judgment." (*The Baltimore Sun*. October 21, 1995.) Similar kinds of problems have plagued educators across the country since the assessment stakes have been raised.

The current wave of excitement about improving the quality of education is built to some extent on the notion that, "If you can't beat them, join them." The strategy seems to be to build a test that represents what the students should know, so that teaching to the test becomes teaching the curriculum that is central to student achievement. If standards are set on relevant assessments, the teachers will scramble to prepare students to do well on the assessments.

Criterion-Referenced Test Scores

Cannell's jibe about too many above average students called attention to the fact that a normative scale is not really what is needed in assessing achievement. Norm-referenced tests tell us whether Susie from a rich suburban school knows more than Johnny from a poor city school. That is not really the point. What we really want to know is "What do the students know and what can they do?" This kind of question implies an absolute scale, rather than a relative scale. In psychometrics, such a scale is called criterion-referenced (Berk, 1976). Achievement assessment calls for criterion-referenced scales. In fact, these names, norm-referenced and criterion-referenced, are not totally appropriate. The difference between the two is not only in the referents for the test scales (distributions of scores for norm reference vs. content coverage for criterion reference) but in the test construction itself — performance differentiation is the main factor driving the construction of norm-referenced tests, whereas content coverage is the main factor driving the construction of criterion-referenced tests.

It sometimes seems that policymakers do not understand the distinction between criterion-referenced and norm-referenced tests. In the good old

tradition of competition, the Secretary of Education prepared a wall chart showing, among other things, average SAT scores for each of several states. Technically, this is an outrageous misuse of SAT scores (Wainer, Holland, Swinton & Wang, 1985; Wainer, 1986). The SAT is designed to be norm-referenced. Its purpose is to differentiate among students bound for those colleges that require SAT scores. The SAT is in no sense an achievement test, even though it measures verbal and quantitative abilities that have been developed and expanded in school.

The National Assessment of Educational Progress (NAEP) improved upon the wall chart by conducting a state-by-state assessment of mathematics achievement, using a criterion-referenced measure. This provided a more reasonable way of obtaining comparisons between states. Still, comparing achievement in one state with that in another state does not seem to be very enlightening. States should be interested in what their students know, not in whether their students know more than the students in other states. The recent rash of international assessments seems likewise designed to show that the United States students are not number one. At the local level, school administrators tell me that they observe great interest in school improvement on the part of real estate agents, who are probably seeking a comparative advantage.

Elements of Performance Standards

The term "standards" is used in so many ways in education and testing that it sometimes seems that we aren't even talking about the same concept. In order to sort out this melange, it may be helpful first to consider the elements that go into a performance standard, and then to consider how those elements apply to performance standards that are used for different purposes. The discussion will be confined to standards applied to people, or groups of people, rather than standards for products like electric wiring, or vacuum cleaners.

In setting performance standards for students, one must first have a scale on which to set it. Perhaps standards could be set without a scale, but that seldom happens in education. The only alternative is a list. Standards for products are generally lists. One such list is the *Standards for Educational and Psychological Testing and Assessment*, prepared jointly by the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education (1985). Tests are expected to live up to each one of those standards. Sometimes shortcomings in one area might be ignored because of good performance in other areas, but that is not the intent of the standards. There are, of course, no sanctions for tests that don't meet the standards.

But, except for lists, performance standards are set on score scales. The choice of the test or tests is critical. Is the test to be built mainly for differentiation, or is it to be a criterion-referenced test, built mainly to represent the content domain? Nearly always, a criterion-referenced test is needed. There is then the question of how to specify the content

9

that the test is to cover. This involves content standards, which will be considered later.

Next, is there to be one standard on a scale, as in minimum competency, or should other levels of competency be recognized? That is, should the grade be pass-fail, or should it be A, B, C, D. There are many circumstances in which several levels would be useful.

Is one scale enough, or is there so much variety of content that several scales should be used? If there are several scales, is the standard to be an overall standard, obtained from some combination of scale scores, or should there be a standard on each scale, and some rule for an overall decision. In high school tests of minimum competency, for example, there is a minimum standard for each of several key subjects. The student usually must pass each subject.

Finally, how shall the cut-point actually be determined? There are some sophisticated methods for setting standards, the most popular of which is the so-called Angoff method. The term "so-called" is appropriate because Bill described that method offhandedly, and mainly in a footnote, in the classic, *Scales, Norms, and Equivalent Scores* (Angoff, 1971). In this procedure, each of a set of expert judges is asked to imagine a person whose skill is just at the borderline between acceptable and unacceptable, and then to mark the exam as that person would have marked it. The average score of the papers marked by all the experts is then taken as the cut-point or performance standard. The main problem is getting judges to agree about what is meant by a person at the borderline.

The Angoff method has some variants. Each judge can be asked to indicate, for each test item, how likely it is that the borderline person would give the correct answer. Critics say that this amounts to asking people to judge probabilities, a task that they can't do very well. But the responses can be rescaled for maximum interjudge agreement; the actual probabilities don't have to be believed for the method to work.

Instead of imagining a marginally competent person, Michael Zieky and Samuel Livingston of ETS point out that it would also be possible to assemble a group of real people who were marginally competent, and to find out how they did on the test. Another scheme is called the contrasting group method. A group of well qualified professionals, and a group of aspirants who are clearly not qualified, are both given the exam. The cut point is set so as to best discriminate between the two groups.

There are also some not-so-sophisticated methods in regular use. For example, norms are sometimes used as standards. That may seem arbitrary, but it happens. Course grades, for example. Course grades in school and college are a species of performance standard. Teachers sometimes grade on the curve (norm referenced), or else they assign grades based on their own and their schools' standards (criterion-referenced). It is not always easy to know what grades mean. I shall always remember a Shakespeare course I took at Yale, when I briefly thought I would major in English. I did not understand very much of the seminar discussion, but I did OK on the biweekly papers, or so I thought until I showed one of my graded papers to a classmate who exclaimed, "B+ ! That's the lowest grade anyone has gotten on a paper this semester." I quickly converted to a math major.

Purposes Of Performance Standards

The choices for the elements of a particular standard depend to a large extent on the purposes that standards are intended to serve. They can be used in certification, such as when minimum competency tests are used to certify a high school graduate. Standards can also be used as predictors, such as when standards are set on entrance exams for college or for employment. Standards may serve merely as descriptors. And finally, standards can be used as motivators.

Standards as Certifiers

Minimum competency tests and the many professional certification exams — like the actuarial exams, the certified public accountant exams, bar exams for attorneys, and medical board exams — are designed to establish competence. These tests are intended to be criterion-referenced, in the sense that their main purpose is to represent the relevant content. There might also some consideration of testing method. An architect should be asked to design something. A surgeon would seem to need to display actual hands-on performance, as well as job knowledge, but perhaps a portfolio of patients could be submitted.

One cut-point would seem enough. A person is, or is not, certifiable, as in the case of a minimum-competency exam. But often several scales are involved. The high school student must demonstrate competence in several areas. Almost all professional certification exams involve tests in several areas. The main question for certification examinations is how many different scales are to be assessed, and how the results are to be combined. Sometimes the different areas are sufficiently correlated that a single overall score can be obtained by some kind of weighted or unweighted average. A history achievement test might include American history and world history in some proportion, and vast knowledge of one can compensate for little knowledge of the other. More often, separate standards are set on each scale. A C.P.A. needs to be adequately adept at business law and ethics, auditing, business accounting, and non-business accounting. Each of four exams must be passed, within three years. The actuaries had, when I started taking the series, eight exams, which all had to be passed, but over an extended time span.

In practice, certification standards are often set without benefit of psychometrics. One certification exam has a mandated passing grade of 70% of the items. However, adjustments are made so that 30% of the candidates pass. Similar methods are used in a number of specialty exams. Passing grades on bar exams are frequently set by this percentage method, but the percentage varies a little, from year to year, depending on how many new lawyers are needed that year.

Standards as Predictors

Standards are often used for college entrance or employment tests. For example, a college academic department might believe that a student who scores less than 500 on any section of the Graduate Record Examination General Test is unlikely to make the grade in graduate school. In such situations, it is

11

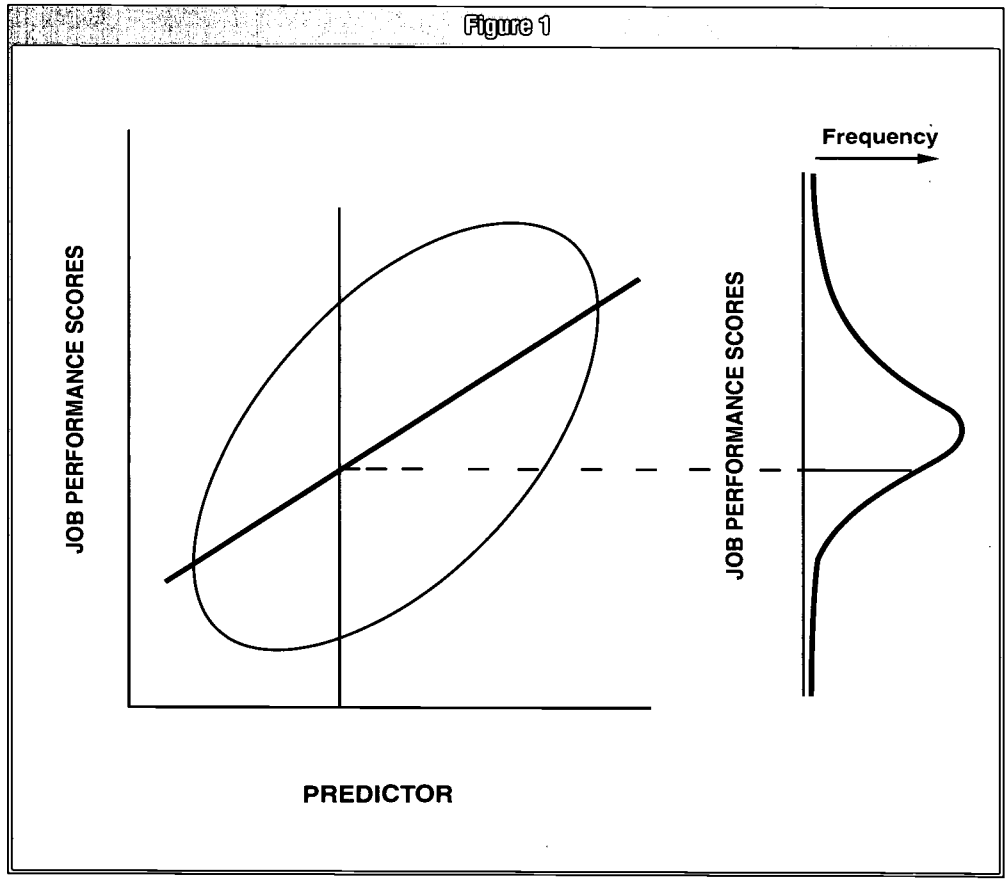
important to set the standard on the right scale. Sometimes the Angoff method is used to set cut-offs on the predictor of success, the test. Judges are asked to estimate the probability that a successful worker, or a successful student, would correctly answer each item on the selection test. That is an incorrect application of the methodology. The judgments are on the wrong continuum.

In a prediction situation, the standard should first be set on the criterion. When predicting job performance, it is job performance that should be dichotomized. In predicting academic success in college, it is college grade point average (g.p.a.) on which standards should be set. First we locate the criterion point that separates the successes from the failures, then we need to locate the cut point, i.e., the value that predicts that criterion cut point. For example, the cut point on the predictor could simply be the score for which the person is predicted to have a 50-50 chance of success. Figure 1 depicts this situation. It is important to notice that when the implied cut-point on the predictor is then applied, the result is not a sharp break in the criterion distribution, because of the errors in prediction, as shown in Figure 2.

The choice of elements for the standards then applies to the criterion, not to the tests or other predictors. There is usually only one criterion. When there are several criteria, one criterion is usually primary. In college entrance that is academic success. For example, at Johns Hopkins, being a good lacrosse player counts for a lot, but the applicant still must be able to avoid academic probation.

Sometimes, success is predicted from a combination of several test scores and other indicators,

Figure 1



12

and there is no reasonable way to use the Angoff method on a regression composite.

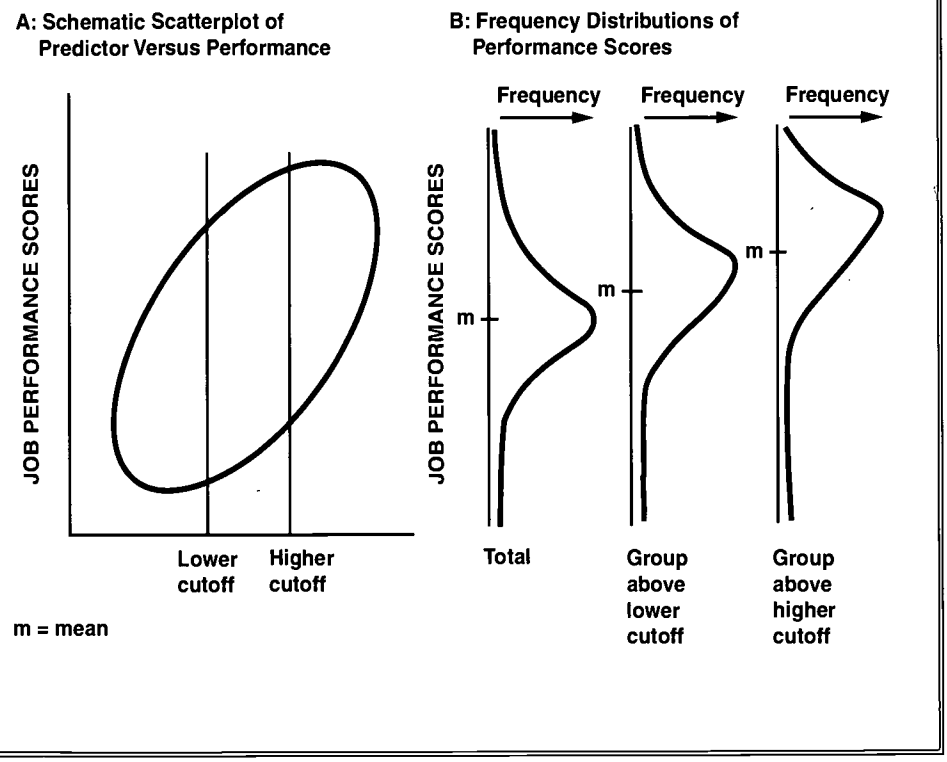
Once the cut-point is set on the criterion, say freshman g.p.a., then it is possible to determine what value of the [composite] predictor yields the requisite probability of success. But then, no cut points are implied on the individual predictors, since in a regression, the predictors are combined in a compensatory way.

One example of a regression composite can be found in the process the National Collegiate Athletic Association (NCAA) uses to set a minimum standard for participation in college athletics (NCES, 1995). This year, the standard for eligibility for participation in college athletics programs applies not only to the entrance tests but also to the high school g.p.a. Moreover, the description of the decision rule is compensatory: if you have lower test scores, you need a higher g.p.a.; if you have a lower g.p.a., you need higher test scores. The standard can be represented as a line in the graph of the predictors, and can also be represented by a table, as shown in Table 1.

For the past decade, I have been working on various personnel testing projects for the U.S. armed forces. In the U. S. Military, applicants must pass a cut point on the Armed Forces Qualification Test (AFQT) to qualify for entrance. Then for each potential military job, there is a second, job-specific standard on some combination of the 10 tests in the Armed Services Vocational Aptitude Battery (ASVAB). Persons who can just barely qualify for entrance often can qualify for only a few of the hundreds of military specialties.

Between 1982 and 1992, the U. S. Military conducted a long project to establish stan-

Figure 2



13

dards based on actual hands-on performance (Wigdor & Green, 1991). Rather than attending to how well the incumbents had done in training school, the standards were based on how well they could do their job. Moreover, in some cases it was possible to get judgments from supervisors of the value of various performance levels. In some instances, poor performance while not valuable, is also not threatening, whereas in other jobs, poor performance can be dangerous to the incumbent; in some jobs, poor performance can pose a serious danger to others. Poor performance by a cook may result in disgruntled colleagues; poor performance by an air traffic controller may result in dead colleagues.

As one looks at the problem of recruiting service personnel and placing them in jobs, the issues of performance value and performance cost must be weighed. Not only are highly talented individuals in short supply, but they are in demand. Civilian companies want to hire them too. Most of them would go to college if they could afford to, so the Army has instituted a signing bonus: in return for a full tour of duty, the Army will provide money for college tuition. This is an expensive program, and there have been attempts to consider just how much good performance the military can afford. There are costs of various sorts associated with both good and poor performance. Obtaining good performance means paying bonuses for signing up; poor performance can be dangerous, but what mounts up is the cost of recruiting someone who can't make it through specialty school, or who resigns, or is fired before they contribute much to the organization.

The U. S. Military now has an econometric model that computes the costs involved in setting particular standards (Green & Mavor, 1994). But

Table 1

Minimum Standards for Eligibility for Participation in Division I College Athletics (abridged from NCES, 1995)

*Core GPA**	*SAT*	*or*	*ACT*
2.5 or Above	700		17
2.4	740		18
2.3	780		19
2.2	820		20
2.1	860		21
2.0	900		21

*Core GPA has a detailed definition in terms of course requirements.

14

standards are not the main focus of the model. In a large organization like the U. S. Military, many other jobs also need recruits. The jobs are in competition with each other for the available talent, and talented applicants are in short supply. If some jobs got many of the recruits with great promise, the other jobs would have to settle for second best. In order to try to parcel the talent equitably, each job, or each class of jobs establishes a goal of expected performance, expressed as a distribution. It is openly recognized that all new incumbents are not going to develop into experts, but there is plenty of work for journeymen. Even if most eventually became superior workers, there is always movement into the job from new hires, and exodus due to promotion and attrition, which in the case of Military, includes finishing a contracted tour of duty with no interest by the employer or the employee in signing a new contract. So for all these reasons, as well as the inevitable individual differences, the quality of job performance among incumbents at any moment is best described as a distribution. That distribution becomes a goal for the current recruiting effort. If it is decided to raise the goal, by raising the mean of the expected performance distribution, then there will be cost implications. More effort will have to go into recruiting. Moreover, raising the goal in one job means draining talent from other jobs, unless the means of their desired performance distributions are also raised. So a system of balancing has to be used. The econometric model recognizes the need for balance. The model can also address the "down-sizing" economic question: How will performance suffer if the recruiting budget is lowered?

Note that our standards for prediction have changed into performance goals, and that the goals are expressed as distributions. Setting cut points on the entrance tests has been done indirectly by setting performance standards. Each job has an entrance standard, but it is of marginal interest. The main interest is the whole distribution, which could be specified by assigning percentages above various points ordered along the performance continuum. Some of us tried unsuccessfully to promote the use of such points, to distinguish categories of job performance (novice, apprentice, journeyman, master, and expert).

Standards as Descriptors

The simplest use of standards is to clarify the meaning of a scale. That was the point of suggesting the categories of job performance. When a scale is being used for some kind of individual or system-wide evaluation, the meaning of various points on the scale need clarification. An excellent example is the NAEP scales, as described by the NAEP proficiency levels, also known as anchor points, developed in the mid 1980's (Beaton & Zwick, 1992).

NAEP assesses several achievement areas: reading, writing, math, science, and occasionally some other areas. There is a NAEP scale for each area. Each scale is centered at 250, and ranges from 0 to 500. Some points were selected as anchor points, also called proficiency levels, and descriptive phrases were developed in order to characterize the skills represented by each level. As an example, Table 2 shows the mathematics proficiency descriptors. These descriptors were developed by examining the items

that were generally answered correctly by most of the persons below the level, and by few of the persons above that level.

The NAEP scales are developmental, in the sense that one scale is used to describe the progress of students through the first 12 years of school. There are discussions about whether the scale is really measuring the same thing at lower score levels as it does at the higher levels, but that is not the issue today. The point here is that the proficiency levels are descriptors. Standards could be set for the system by aiming for a distribution of 17 year olds as 100% at level 200, 90 % at or above 250, 60% at or above 300, and 20% at or above 350. These would be goals to aspire to, not really likely to be met soon, without incredible change in the educational system. Nevertheless, they would constitute standards for the system. It would be almost as good to set the standard in terms of the mean and standard deviation of the proficiency distribution, but such a specification would not be so easily understood by the general public.

Since the levels are being used as descriptors, it doesn't much matter which points are chosen, nor how many. There should be enough points, well-spaced along the scale, but four or six would have done about as well. The important methodological point is that when standards are used as descriptors, one can pick the points first, and then find appropriate descriptions by examining the item information.

There has been a recent uproar about NAEP standards because of a change called for by a new National Assessment Governing Board (NAGB). This new board was appointed a few years ago, in the midst of a national cry for educational improvement. Employers moaned that high school graduates

Table 2

Percentages of Students Performing at or Above Mathematics Proficiency Levels: 1986

Level	*Description*	*Age 9*	*Age 13*	*Age 17*
350	Can solve multi-step problems and use basic algebra	0.0	0.4	6.4
300	Can compute with decimals, fractions, and percents; recognize geometric figures; and solve simple equations; and use moderately complex reasoning	0.6	15.9	51.1
250	Can add, subtract, multiply and divide using whole numbers, and solve one-step problems	20.8	73.1	96.0
200	Can add and subtract two-digit numbers and recognize relationships among coins	73.9	98.5	99.9
150	Knows some basic addition and subtraction facts	97.8	100.0	100.0

couldn't do much, and scholars interested in cross-cultural comparisons pointed out that United States students were not Number 1; far from it. So NAGB decided to set standards in terms of a new set of generic levels on the scales, which were called achievement levels, and were named Basic, Proficient, and Advanced (NAEP, 1993; Lissitz & Borque, 1995). That is, rather than picking the points first, and then characterizing them, they picked the names first, and then tried to find out where their new labels belonged on the scales. That has turned out to have been ill-advised. My colleague, Warren Torgerson, says that asking someone to place "Proficient" on a test score scale is like asking someone draw a "moderately long line" on a sheet of paper. There are bound to be different notions of what the labels mean. Moreover, any method that is used to find locations for the labels on the scale had better do all the labels at once, rather than one at a time, lest proficient turns out to be placed higher than advanced. That actually happened once or twice, according to hearsay. Moreover, the implication that Basic in mathematics means roughly the same as Basic on writing is no more than a suggestion.

Standards as Motivators

The main reason for setting high standards for educational achievement is partly for assessment, and partly for motivation. Teachers are being encouraged to raise the achievement distribution of their students. The standards are goals again. Those who set the goals must recognize that some students will be better than others. Individual differences are inevitable. The standards should provide goals for all students. The NAGB did so when it chose to have three standards, leading to four regions of the achievement dimension: Unsatisfactory, Basic, Proficient, and Advanced. Perhaps everyone should be expected to reach at least a basic level, but there will always be some for whom Basic is in itself a notable achievement. Others could easily become proficient, and some should be expected to reach expert status and beyond. (This seems not very different from the classic A, B, C, D).

In the context of goals, almost any levels would serve. The goal for the system could just as well be specified in terms of the proportions of students expected to be in the various categories. In that case, roughly any categories would do—they are only being used as descriptors. In particular, the NAGB could have used the NAEP anchor points. There was no need to shift to another set of points, and no need at all to try to locate those arbitrary points empirically. Of course, NAGB and NAEP are aiming only at the system level. Some states are aiming at the student level—i.e., using an assessment that is long enough to provide students some feedback about the quality of their individual performance. At the individual level, accessible goals are needed for all students.

Content Standards

As noted above, there is an important distinction between content standards, which define the extent of the domain to be tested, and performance standards, which indicate how much of the domain has been mastered. Messick (1994) has recently argued that the bridge from the one to the other is of central importance in validating performance

standards. Methods of setting performance standards assume that the test adequately reflects the domain of knowledge on which the test-takers are being evaluated. A prior step would seem to be to define the domain.

The performance standards have to reflect the content standards. The bridge from the content standards to the performance standards depends on the test specifications, the item writers and test editors, and on the resulting performance measurement scale. Logically, it would seem preferable for the judges to set standards first on the content domain. They could identify what parts of the domain are basic, what parts go with proficient persons, and what parts would mainly be mastered by advanced students. It is not at all clear how to do this, but a way might be found. Judges might also be useful in evaluating the bridge from content to performance. This would seem a more straightforward task than imagining the test behavior of marginally competent test-takers.

NAEP has used what they called frameworks to delineate the assessment content. Forsyth (1991) criticized the NAEP anchor point descriptions primarily because the link back to the frameworks was unclear. He felt that even if the anchor points were adequately described by certain items, it was still necessary to show that the items adequately represented their framework. The NAEP frameworks are cross-classified by subject knowledge and levels of understanding. For example, the Reading framework (Table 3) is a matrix. The number of items representing each cell can be counted. The framework for the new mathematics assessment is more complex, consisting of five broad content strands, three mathematical abilities, and three unifying themes (Figure 3). An item might represent more than one aspect of this multi-way frame

Table 3

NAEP Reading Proficiency Levels

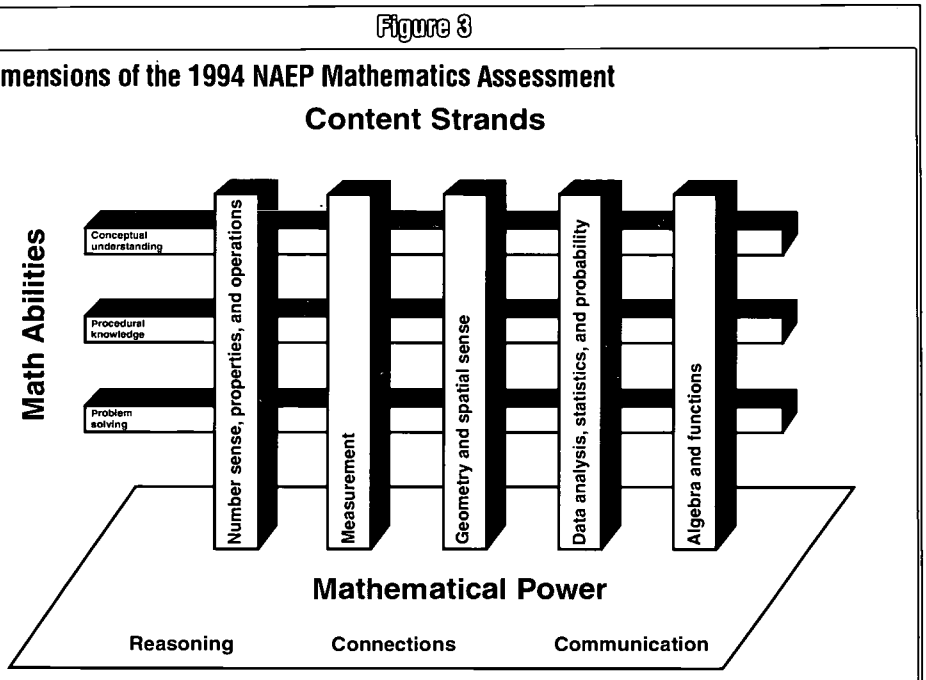
350	Can synthesize and learn from specialized reading materials
300	Can find, understand, summarize, and explain relatively complicated information
250	Can search for specific information, interrelate ideas, and make generalizations
200	Can comprehend specific or sequentially related material
150	Can carry out simple, discrete reading tasks

The bridge from such content standards to performance standards is complex (Lane, 1993).

As many of you know, various groups of educators have been devising a new round of content specifications for what students should know about their given area. The mathematics standards have been published (National Research Council Mathematics Sciences Education Board, 1993.), and some science standards are on the way. These standards are not designed to facilitate measurement. They are standards for the curriculum, not the assessment. Indeed some of these groups do not place a high value on tests. The mathematics standards opens with a quote from an Iowa farmer, "You can't fatten a hog by weighing it." One possible reply would be, "You can't tell about the feed without weighing the hog." Despite this brilliant repartee, we can be sure that the content standards are not stated in a manner that leads in any direct way to content specs for tests. Moreover, the current wave of content standards appear to be focused on expert performance. The goal is deep thought, and expert problem solving. Not everyone can be an expert. Everyone can be good at something, but most of us get along without being proficient in everything.

Figure 3

Dimensions of the 1994 NAEP Mathematics Assessment



SOURCE: U.S. Department of Education, National Center for Education Statistics, National Assessment Governing Board. *Mathematics Framework for the National Assessment of Educational Progress.* Washington, DC: 1994.

19

Conclusion

In summary, the psychometric problem of determining just where a cut-point should be placed on a scale seems not to be a central feature of standard setting. Cut points are important in certification, but so are deciding what to test and how to test it. In prediction, placing the standard on the right scale is important. For description and for motivation, the placement of the points is less important than having enough points to serve as descriptions and goals for the full range. And finally, finding a way to map content standards onto performance standards is a challenge.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

Angoff, W. H. (1971). Scales, norms and equivalent scores. In R. L. Thorndike (Ed.), *Educational Measurement.* Second Edition. Washington, D.C.: American Council on Education

City schools misuse state test questions. (1995, October 21). *The Baltimore Sun*, p. B1.

Beaton, A. E., & Zwick, R. (1992). Overview of the National Assessment of Educational Progress. *Journal of Educational Statistics*, 17, 95-110.

Berk, R. A. (1976). Determination of optimal cutting scores in criterion-referenced measurement. *Journal of Experimental Education*, 45, 4-9.

Cannell, J. J. (1988). The Lake Wobegon effect revisited. *Educational Measurement: Issues and Practice*, 7 (No. 4) 12-15.

Citron, C. H. (1983). Courts provide insight on content validity requirements. *Educational Measurement: Issues and Practice*. 2, 6-7.

Forsyth, R. A. (1991) Do NAEP scales yield valid criterion-referenced interpretations? *Educational Measurement: Issues and Practice*. 10(3) 3-9.

Green, B. F. & Mavor, A. S. (1994). Modeling cost and performance for military enlistment standards. Washington, DC : National Academy Press.

Jaeger, R. M. (1989). Certification of student competency. In Linn, R. L., (ed). *Educational Measurement*. Third Edition. New York: American Council on Education/MacMillan.

Lane, S. (1993) The conceptual framework of the development of a mathematics performance assessment instrument. *Educational Measurement: Issues and Practice*, 12(2) 16-23.

Linn, R. L. (1995). *Assessment-based reform: challenges to educational measurement.* Princeton, N.J: Educational Testing Service.

Lissitz, R. W. & Bourque, M. L. (1995). Reporting NAEP results using standards. *Educational Measurement: Issues and Practice*, 14(2), 14-23.

Livingston, S. A. (1994, October). Standards for Reporting the Educational Achievement of Groups. Paper presented at the Joint Conference on Standard-Setting for large-scale Assessments, Washington, DC.

Livingston, S. A. & Zeiky, M. J. (1982) Passing scores: A manual for setting standards of performance on educational and occupational tests. Princeton, NJ: Educational Testing Service.

National Research Council Mathematics Sciences Education Board. (1993). *Measuring What Counts: A Conceptual Guide for Mathematics Assessment.* National Academy Press, Washington, DC.

Messick, S. (1994). Standards-based score interpretation: establishing valid grounds for valid inferences. *Journal of Educational Measurement* (in press).

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

National Center for Education Statistics. (1995). *Who Can Play?* Report NCES 95-763. Washington, D. C.: U.S. Department of Education.

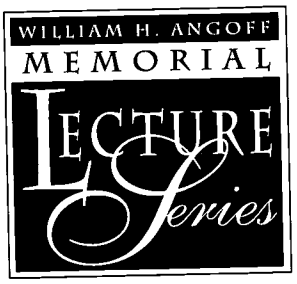
Wainer, H., Holland, P. W., Swinton, S., & Wang, M. H. (1985). On state education statistics. *Journal of Educational Statistics*, 10, 293-325.

Wainer, H. (1986). Five pitfalls encountered when trying to compare states on their SAT scores. *Journal of Educational Measurement*, 23, 69-81.

Wigdor, A. & Green, B.F. (Eds.). (1991). Performance assessment for the workplace. Committee on the Performance of Military Personnel, Commission on Behavioral and Social Science and Education, National

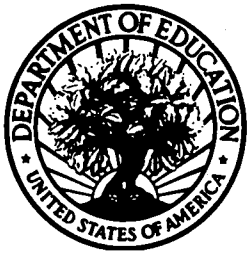
Research Council. Washington, D.C.: National Academy Press.

ZEIKY, M.J. (1994, October). *Historical perspective on standard setting for large-scale assessments.* Paper presented at the Joint Conference on Standard-Setting for Large-Scale Assessments, Washington, D.C.



POLICY INFORMATION CENTER
Educational Testing Service
Princeton, New Jersey 08541-0001

04206-13515 • Y56M3 • 204860 • Printed in the U.S.A.



U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement (OERI)
Educational Resources Information Center (ERIC)

NOTICE

REPRODUCTION BASIS

☒ This document is covered by a signed "Reproduction Release (Blanket)" form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.

☐ This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").