An NCPR Working Paper

Best Practices for Setting Placement Cut Scores in Postsecondary Education

Prepared for the NCPR Developmental Education Conference:
What Policies and Practices Work for Students?
September 23–24, 2010
Teachers College, Columbia University

Deanna L. Morgan
The College Board

National Center for Postsecondary Research
www.PostsecondaryResearch.org

The National Center for Postsecondary Education is a partnership of the Community College Research Center, Teachers College, Columbia University; MDRC; the Curry School of Education at the University of Virginia; and faculty at Harvard University.

The contents of this report were developed under a grant from the Department of Education. However, those contents do not necessarily represent the policy of the Department of Education, and you should not assume endorsement by the Federal Government.

For information about NCPR and NCPR publications, visit www.postsecondaryresearch.org.

Abstract

Cut scores are used in a variety of circumstances to aid in decision making through the establishment of a clear cut line between adjacent categories. Community colleges regularly use cut scores on placement tests to decide the appropriate course for each beginning student: the first college-level course or a developmental course, depending on student test performance. This initial placement decision sets the stage for the student's college career and must be made in the most defensible manner possible to withstand any challenges that may arise. This paper provides a broad overview of important considerations in the generic standard setting process. Specific standard setting methods are only touched upon briefly, but many commonalities in actions and processes exist among the various standard setting methods. This paper covers both actions and processes at a level that is accessible to the non-psychometric layman with only a basic understanding of assessment. The paper also provides an overview of the key elements to consider in choosing a placement test, for which the cut scores are implemented and provides case studies. Placement decisions can only be as good as the instrument and cut scores upon which they are based. Areas for potential placement testing research are also provided.

Contents

Abstract iii

1. Introduction 1

2. The Need for Placement Testing 3
Guidelines for Fair and Proper Use of Cut Scores 5
Uses of Cut Scores in Higher Education 6

3. Best Practices in Setting Cut Scores 7
The Role of Judgment in the Standard Setting Process 7
The Authoritative Body 9
Participants in a Standard Setting Study 10
Performance Level Descriptors 13
Standardization and Defensibility of the Process 16
Documentation and Evaluation of the Process 20

4. Case Studies 26
Case Study #1: Setting Cut Scores on a Diagnostics Test 26
Case Study #2: Setting the Cut Score for a Combination Placement Rule 28

5. Selection of a Placement Test: Factors to Consider 31
Validity 31
Reliability 33
Sensitivity and Bias 34

6. Conclusion and Recommendations for Further Research 36

References 38

1. Introduction

College admissions models vary widely both within and between two- and four-year colleges and universities (Rigol, 2003). At one end of the spectrum are highly selective institutions that generally rely heavily upon admissions test scores and require large amounts of evidence to choose those students they consider the best fit. At the other end of the spectrum are open admissions institutions, primarily community colleges, that operate on a more egalitarian basis, striving to serve and educate all applicants rather than just a select few (The College Board, 2002). Along the continuum towards the more selective institutions, the probability increases that scores from an admissions test, like the SAT or ACT, will be considered as one piece of information in the admissions and selection process. However, a growing number of institutions are relying less heavily on scores from admissions tests in the selection process, especially at community colleges. As reliance on such tests decreases, it becomes increasingly important to establish alternate means of assessing student knowledge and skills as a measure of their readiness for college-level coursework.

Content standards refer to the curriculum that students must master. Specifically, what skills and knowledge should students be able to demonstrate? In terms of college placement, they may be thought of as the competencies that are considered a requirement or prerequisite for enrollment in a specific course. Placement tests serve the purpose of identifying the knowledge and skill level of an examinee in the form of a test score. Placement tests should be chosen based on their alignment with the content standards for the courses in which students will be placed. Applying the test score requires the establishment of performance standards that quantify the content standards by defining how much mastery of the standards students must demonstrate to achieve a particular level of competency. For example, how much of the content standards must a student know and be able to do to be considered just sufficiently knowledgeable for placement into a college-level course? It is necessary to define the point on the score scale where the examinee has exhibited sufficient knowledge and skill to be placed into the designated course. The designated course, in the case of placement tests, is typically the college-level, credit-bearing course; the alternative consists of one or more categories representing developmental coursework. The identified score, at which anyone scoring at or above it is placed into the higher level course while anyone scoring below it is placed into the lower level course, is called the *cut score*. *Standard setting* is the generic process of determining the placement of one or more cut scores, often referred to as setting cut scores. *Standard setting methods* are specific combinations of standardized activities and processes that are used to develop cut score recommendations; examples include but are not limited to Angoff

(1971), Bookmark (Lewis, Green, Mitzel, Baum, & Patz, 1998), Body of Work (Kingston, Kahl, Sweeney, & Bay, 2001), and Contrasting Groups (Livingston & Zieky, 1982), etc.

This paper provides the following: a brief summary of the need for placement testing and guidelines for the proper use of cut scores, an overview of the best practices when setting cut scores for placement in postsecondary education, two case studies of the standard setting process, a discussion of factors to consider when selecting a placement test to use, and a conclusion that offers suggestions for future research.

2. The Need for Placement Testing

Typically, admission to a community college requires little beyond a high school diploma or certificate of equivalency. Prospective students at community colleges may or may not submit scores from, or even have ever taken, an admissions test. The removal of the requirement to submit scores from these traditional "gatekeeper" tests allows greater access to higher education for all students but also eliminates one piece of evidence in the admissions process that can assist in the prediction of how well a student can be expected to perform in the freshman year of college. As a result of such increased access, the student body may represent a wider range of socioeconomic, demographic, and academic characteristics than previously, especially when there is no standardized measure of student knowledge and skill used to set the bar for admissions by screening out low performers.

Results from a survey in winter 2010, sponsored jointly by the League for Innovation in the Community Colleges and The Campus Computing Project, with more than two thirds of the community college presidents and district chancellors responding, indicate 10 percent or greater increases in enrollment since winter 2009. Further, one third of the respondents reported increases of 15 percent or greater. This surge of enrollment—combined with open-access policies at many colleges and, in many cases, a lack of scores from traditional admissions tests—has shifted the focus on many campuses from selecting students for admission to placing students into the appropriate course to begin instruction. Many two- and four-year institutions thus require placement testing of incoming students to determine whether individual students have the knowledge and skills necessary to be successful in college-level, credit-bearing[1] courses or whether remediation, through some form of developmental or basic skills education, is required.

Use of placement testing, as opposed to other methods for deciding student admission, increases the size and diversity of the student body at many institutions. A key factor in the need for placement testing lies in the difference between the level of knowledge and skills required for high school graduation and those required to be successful in college-level, credit-bearing courses. Traditionally, high schools have focused on minimal competency for graduation with limited input from higher education or employers (Haycock, Barth, Mitchell, & Wilkins, 1999; Kirst, 2005; Venezia, Kirst, & Antonio, 2003). As a result many high school graduates have minimal life skills but are not necessarily prepared for entry into college or the skilled work force (Achieve & The Education Trust, 2008). Too often the focus in high school is distilled down into the number

[1] For the purposes of this paper, credit-bearing refers to courses taught at the college level for which the student will receive credit towards degree completion and not those which may bear credit for other purposes, such as receiving financial aid or determining full-time vs. part-time status.

of courses or credit hours needed in a subject area with little or no specification about what students should know and be able to do to be successful in postsecondary education (ACT, 2007; Haycock et al., 1998).

The need for remediation through developmental coursework extends the time to graduation and increases the cost of a college education for those students who must pay for additional non-credit-bearing courses before even beginning the credit-bearing courses required for degree completion. The National Center for Education Statistics (NCES, 2004) estimates that 95 percent of students who require remediation spend one year or less in developmental courses, resulting in a subsequent delay in expected graduation or entry into the workforce. In addition, students needing remediation have a significantly lower probability of graduating (Camara, 2003) and will incur additional tuition and related costs from the extra enrollment time. However, proper placement is important to ensure that students have the knowledge and skills to be successful without such difficult coursework that they become frustrated or fail to be sufficiently challenged such that they are bored. Both possible scenarios, frustration and boredom, can lead to decreased student motivation (Mattern & Packman, 2009).

Placement tests are used to determine the student's level of knowledge and skills so that colleges can make appropriate course placement decisions to allow the student to be successful in the course. Often a college will have developed multiple cut scores for the purpose of assigning students to varying levels of entry level or developmental education. The number of developmental courses that may be offered by a college can vary from one to as many as six levels per subject area, generally in the subjects of reading, writing, and mathematics. To complicate factors, the cut scores used for placement are typically determined locally by the individual college or system and can vary widely from one college to the next both in terms of the number of cut scores used and the placement of the cut scores along the score scale.

Placement tests are commercially available from some test vendors, such as The College Board's ACCUPLACER® and COMPANION® tests and ACT's COMPASS® and ASSET® tests. The variability between the number of developmental courses offered and the different standards that the colleges deem appropriate for placement into their local courses requires placement tests to be constructed to allow for discrimination across the entire score scale. Sufficient items of targeted difficulty that measure well-defined content and skills must be available to assess well-prepared students who are ready for the college-level, credit-bearing courses; students who will need to begin with the most basic level of developmental education and work their way up to the college-level, credit-bearing courses; and all levels of student ability between these two extremes.

Regardless of the test selected for placement decisions, cut scores will need to be established in order to actually use the test scores for placement. Accordingly, guidelines exist to encourage the fair and proper use of cut scores, and those are discussed in the next section.

Guidelines for Fair and Proper Use of Cut Scores

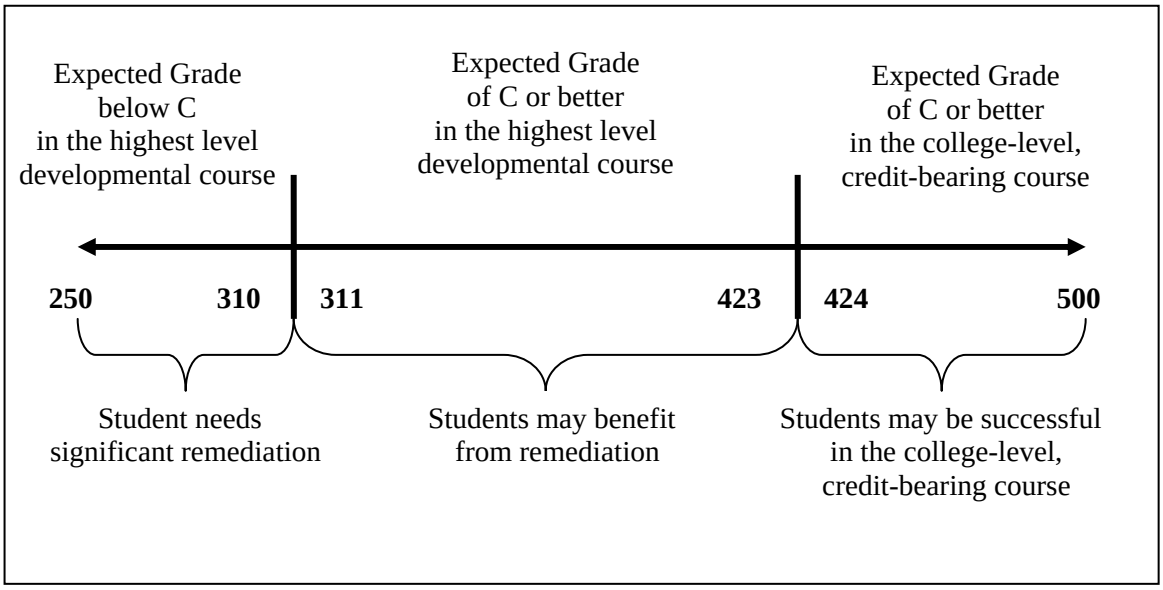
The *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education [AERA, APA, & NCME], 1999) provides guidelines for the appropriate use of tests and test scores. Standard 13.7 states that "in educational settings, a decision or characterization that will have a major impact on a student should not be made on the basis of a single test score. Other relevant information should be taken into account if it will enhance the overall validity of the decision" (p. 146). Therefore, decisions about examinees should be based on a collection of evidence and not solely on a single test score. For example, the *Guidelines on the Uses of College Board Test Scores and Related Data* (The College Board, 2010) notes that test scores should be used "in conjunction with other indicators, such as the secondary school record (grades and courses), interviews, personal statements, writing samples, portfolios, recommendations, etc." (Guideline 5.2, p. 7). Other relevant College Board guidelines suggest that test scores be used as approximate indicators rather than a fixed and exact measure of an examinee's preparation (i.e., Guideline 5.3).

Of particular relevance to the idea of using cut scores in placement decisions as approximate indicators is knowledge of the standard error of measurement (SEM) associated with a test score as a measure of the consistency of an individual person's score and how much that score could be expected to vary should the student be tested multiple times. For example, if Abbie scored 70 on an assessment and the SEM for scores on the assessment was +/- 3, then Abbie's scores on repeated testing would be expected to fall between 67 and 73 in 68 percent of retests (using +/- 1 SEM) and between 64 and 76 in 95 percent of retests (using +/- 2 SEM). Smaller values of SEM indicate greater precision in the score and are better than larger values of SEM, which indicate less precision. The SEM can be used when students score just below the cut score to evaluate whether the student may benefit from a retest given the likelihood that he or she would obtain a score at or above the cut score. If the range of scores created by using the student's score +/- 1 SEM includes the cut score then it is likely that the student, if retested, would score at or above the cut score and a retest would be worthwhile, though it is equally likely that a retest would result in a lower score. Using the SEM may help to save testing costs due to retesting large numbers of students that are unlikely to receive a passing score with a second chance.

Uses of Cut Scores in Higher Education

Making placement decisions is one of the most common uses of cut scores in higher education. College placement decisions may require that one or multiple cut scores be established depending on the number of courses in which students may be placed. An example may be the use of two cut scores to separate examinees into three groups: (1) those who should begin in the college-level, credit-bearing course, (2) those who require some remediation and should begin in the highest level developmental course, and (3) those who should be placed into a lower level developmental course for even greater remediation and basic skills development. Figure 1 provides an example of a hypothetical placement decision for placing examinees into one of three course levels.

Figure 1: Diagram of Hypothetical Placement Decisions Using Two Cut Scores[2]



For simplicity, most examples in this paper feature the use of one cut score to separate examinees into two categories for placement purposes. However, the information is applicable and easily transferable to the use of multiple cut scores for readers interested in those uses. For additional examples of setting multiple cut scores, see Morgan and Michaelides (2005) and Morgan and Hardin (2009). For information on setting cut scores for admissions, see Morgan (2006).

[2] The cut scores of 311 and 424 were arbitrarily chosen for this example and should not be interpreted as recommendations for the placement of the cut scores in placement decisions.

3. Best Practices in Setting Cut Scores

Many excellent resources are available that describe the standard setting process, standard setting methodologies, and other key facets of setting cut scores. Comprehensive treatments of standard setting can be found in Cizek (2001, 2006), Cizek and Bunch (2006), Jaeger (1989), and Hambleton and Pitoniak (2006). More user-friendly treatments with a focus on the implementation and actual applications of conducting a standard setting study can be found in Pitoniak and Morgan (in press), Morgan and Michaelides (2005), Morgan and Hardin (2009), Morgan (2006), and Zieky, Perie, and Livingston (2008). Following a brief discussion of the role of judgment in the standard setting process, this section discusses the best practices that should be observed in setting cut scores, focusing on five major components of the process: (1) The Authoritative Body or Policy-Makers, (2) Participants in a Standard Setting Study, (3) Performance Level Descriptors (PLDs), (4) Standardization and Defensibility of the Process, and (5) Documentation and Evaluation of the Process.

The Role of Judgment in the Standard Setting Process

Critics of standard setting often point to the role of judgment in the process, with the implication that results are subjective or arbitrary. Judgment and the subjectivity of the participants play a key part in the process, and it is possible that changes in the participants or process would culminate in a different result. However, the process is designed to include a carefully considered set and sequence of activities that rely heavily upon input from multiple perspectives, standardization, and replicability of the process and activities; and on informed judgment through the use of knowledgeable subject matter experts. While more subjective than many other processes encountered in the area of psychometrics where a definite right or wrong answer exists, standard setting results in cut score decisions based on student knowledge and skills but lacking the capriciousness often implied by criticisms of arbitrariness. Standard setting has been defined by Cizek (1993) as “the proper following of a prescribed, rational system of rules or procedures resulting in the assignment of a number to differentiate between two or more states of degrees of performance” (p. 100).

Another approach that has often been used to set cut scores in the higher education setting is the use of validity studies. Validity studies are based upon statistical calculations, often employing logistic regression, that use observed student data such as test scores and course grades to predict the probability of success or failure in a course based on the examinee’s test score. Validity studies differ from traditional standard setting methodologies most notably in the empirical nature of the data provided and the

consideration of that data in isolation from the content being measured and the knowledge and skills that the students can demonstrate. Though course grades are often provided as input to the process as a proxy for what students should know and be able to do, the meaning of grades across instructors, schools, etc., can vary widely so no claims can reliably be made about the amount of knowledge represented by the grade. Some teachers grade more or less strictly than others; use different weightings of factors such as homework, exams, and participation; and are vulnerable to a host of other characteristics that can contribute to students, both within and among classrooms, being assigned similar grades for what may be very different levels of subject matter mastery. However, when investigated, the analysis is unable to account for the variance in grading policies and all students receiving the same letter grade are treated the same resulting in an increase in the associated error that is a direct result of human subjectivity and judgment.

Validity studies are useful to institutions in terms of feedback to determine how well cut scores are working and identify potential placement rates should it be necessary for cut scores to be changed. It is time consuming to collect the necessary data to conduct a validity study, but the analyses are somewhat simple and often offered as a free or low-cost service by testing companies to users of their assessment products. Proponents of validity studies for setting cut scores often cite objectivity of results as a leading advantage over using a standard setting methodology. This argument is flawed, as the interpretation and ultimate use of the results from validity studies also rely upon human judgment and subjectivity. Once presented with a list of the probabilities of success and failure associated with test scores, one or more individuals must make a judgment about which is most desirable. The decision is often left to one policy-maker or a small group of policy-makers, and is based heavily upon the selection of desirable placement or pass rates without the link to content and examinee knowledge and skills that is integral to the standard setting process. The integration of content, knowledge, and skills in the standard setting process is a primary factor in the recommendation by Morgan and Michaelides (2005) that a standard setting methodology be used to set initial cut scores and that validity studies serve as tools to evaluate the results of the standard setting and how well the cut scores are functioning over time rather than as the deciding force behind the placement of initial cut scores.

The complete elimination of judgment, in either standard setting or validity studies, is unlikely. The following section on the five major components of standard setting addresses how to design the process to take advantage of the informed judgment of subject matter experts while minimizing the arbitrary nature through training and participation in a sequence of well-defined activities with the goal of capturing expert judgments. These expert judgments will then be combined to form a recommendation as to the most appropriate placement of the cut score(s) for a given context.

The Authoritative Body

The authoritative body, or policy-maker, makes many decisions through the course of the standard setting process (with the input of the facilitator), which may include which method to use, what the make-up of the panel of subject matter experts (SMEs) will be, what data will or will not be provided to the SMEs, the criteria for making final decisions, and ultimately the location of the final cut score(s). Though the authoritative body makes several decisions related to the standard setting, it should not participate in the actual standard setting session, since that could inhibit the participation of other panelists.

The authoritative body may be a single person but generally is a small group of people. The composition of the authoritative body will differ based on the policy and procedure at each institution; for example, it may include the department head, dean of admission, director of placement testing, president of the college, Board of Regents, etc. The key is to ensure that those selected have the authority and knowledge to evaluate the SME recommendations and other relevant material and then make the final decision on the location of the cut score(s).

The authoritative body should receive copies of all materials used in the standard setting study and a copy of all results provided to the panel, as well as any additional information that may be available for consideration, such as expected placement rates, a summary of the results of the panelists' evaluations, and alternative cut score recommendations that may be considered based on the standard error of measurement and the standard error of judgment. All tests are fallible, and an examinee is not likely to earn exactly the same score if he or she took a different version or tested the next day, assuming that no additional training/learning occurred overnight. Should an examinee retest many times (assuming no learning or fatigue), there would be a distribution of test scores. The mean of this hypothetical distribution of scores for an examinee is called the examinee's true score: what the examinee would get if there were no error in the measurement process. The standard error of measurement (SEM), discussed earlier in this paper, is an estimate of the standard deviation of this hypothetical distribution of scores (see Cizek and Bunch, 2006, for information on its calculation).

Even if the authoritative body believes that the panel's recommendation represents the best standard, it may still want to adjust the standard downward (or upward). An examinee whose true score is exactly at the recommended standard will fail 50 percent of the time because there is error in the measurement process. This possibility does not mean that mistakes were made in the design and administration of the test but instead reflects the fact that it is not possible to measure students' "true ability" and that test scores must be used as an estimate. Sources of measurement error are extraneous factors that are not

directly related to what colleges are intending to measure and may include the particular items that a particular form of the test contains, leniency or severity in scoring of responses to constructed-response items, conditions related to the testing environment, and examinee-related factors, such as fatigue.

The authoritative body may decide that failing a qualified examinee is a worse error than passing an unqualified examinee. In this case, the authoritative body might lower the standard by one (or two) SEM so that examinees at or slightly above the panel's recommendation are not likely to fail due to errors of measurement. Obviously, the probability that examinees with knowledge and skill levels slightly below the panel's recommendation will pass also increases, so lowering the risk of failing a qualified examinee increases the risk of passing an unqualified examinee.

After reviewing the recommended cut scores and associated information provided, the authoritative body will make the final decision about the location of the cut score(s) that will be put into use at the institution. The final version of the technical report should document the decision of the authoritative body. It should also be noted that while the facilitator consults with the authoritative body and manages the standard setting process, and the SMEs serve on the panel and make recommendations based on their content expertise and knowledge of the student population, it is the authoritative body that actually sets the final cut score that will be used.

Participants in a Standard Setting Study

Participants in a standard setting study should be chosen carefully for the role to which they will be assigned. Two key roles are those of the study facilitator and the panelists, also known as subject matter experts (SMEs). Selection of participants for the standard setting study is central to both the validity and generalizability of the standard setting results. Characteristics of and considerations for participants in each role are discussed more fully in the following sections.

The Facilitator

The facilitator serves as a link between the SMEs who will participate in the study and the authoritative body, or policy-makers, that will make the final call about the cut score location that will be used for placement. The facilitator works with the authoritative body prior to the actual study by offering informed advice regarding the standard setting process and a host of decisions that should be made in advance of the study. The facilitator assists the authoritative body in selecting an appropriate standard setting method,

identifying the selection criteria for the SMEs who will participate in the study, and communicating the logistical needs for the standard setting meeting in terms of space, equipment, timing of meals and activities, etc.

Prior to the standard setting study, the facilitator designs the plan for the standard setting meeting and oversees the development and preparation of materials that will be used in the meeting. During the meeting the facilitator ensures that appropriate procedures are followed to allow the SMEs to provide cut score recommendations within the established time frame for the meeting while following procedures that are psychometrically sound and recognized by the measurement community as fundamental to the establishment of valid cut scores. In addition, the facilitator maintains documentation of the materials used, the standard setting process, and the decisions/recommendations that are part of the standard setting meeting.

The facilitator should be a professional, with training in measurement and experience with standard setting and meeting facilitation. Knowledge of the subject matter on which cut scores are being established is not required of the facilitator; in fact, subject matter knowledge may be a disadvantage, since it can make it more difficult for the facilitator to be impartial and unbiased in their interactions with the panelists. The facilitator should be skilled in meeting management, able to train the SMEs on the standard setting task, and able to elicit participation from each SME, while ensuring that the process is not dominated by one or a group of SMEs such that others feel unheard or inhibited in providing an opposing view.

The facilitator must remain impartial and have no immediate stake in the outcome of the standard setting so as to avoid the possibility or the appearance that the panel's recommendation is not independent of the facilitator. Ideally, the facilitator would be an outside consultant or perhaps a member of the college faculty with the appropriate training and experience. Examples of college faculty include the Director of Institutional Testing at the institution, a faculty member in the Department of Educational Psychology, or a member of an affiliated research center. The facilitator could also be someone from the Placement Office, but this choice could have the appearance of bias or lead to real bias in the process and results. Whether the facilitator comes from the institution or is hired externally, it is critical that the individual has training and experience in standard setting and has no direct stake in the outcome of the study. The facilitator should not provide any personal opinions on test content or skills that may influence the judgments of the SMEs (Geisinger, 1991; Mehrens, 1986). Also of importance when using a facilitator from within the institution is ensuring that the subject matter experts do not have a reporting relationship with the facilitator that may prohibit one or more subject matter experts from speaking freely during the study due to fear of repercussion.

The Subject Matter Expert

Subject Matter Experts (SMEs) provide the judgments used to form the recommended cut score. Raymond and Reid (2001) provide an extensive discussion on the selection and role of the standard setting panel. The SMEs should be knowledgeable about the examinee population and the skills and knowledge required of examinees in relation to the decisions being made. The SMEs should be experts in the content or proficiency area under consideration, and most should be faculty members currently teaching in the subject or proficiency area at the institution(s) that will use the resulting cut score(s). The panel of SMEs should be representative of the college or institution for which the decisions are being made. SME representation should be considered in terms of gender, race/ethnicity, tenure (both veteran staff and those employed more recently), and, in cases where the cut score may be intended for multiple campuses or locations, the geographical location and campus size. For example, if the cut score will be used system-wide, then representatives from around the system, not just from the main campus, should be included, and representatives should include both two- and four-year colleges.

The more representative the panel of SMEs, the more generalizable and valid the results will be. For that reason, stakeholders other than faculty may be considered for inclusion. However, anyone chosen to serve on the panel must be knowledgeable in the content or proficiency area in which the cut score will be set. Depending on the specific intended use of the cut score and the population to which it will be applied, other stakeholder groups may need to be represented on the panel, such as those with knowledge of or expertise about students with disabilities or students for whom English is a second language. Community employers or representatives from secondary education may sometimes be appropriate for inclusion, especially in cases where it is important to have "buy-in" from these groups, but in most cases their inclusion would be limited to only one or two members of the panel.

Discussion and interaction among panelists play a central role in the standard setting process; therefore, the panel should be sufficiently large for the purposes of representation but kept to a manageable size to allow every panelist to participate and contribute. The method of standard setting chosen for use may have implications for the number of panelists that may be needed. In general, at least 10 panelists are needed, with a maximum number of approximately 25 panelists. A trained facilitator can offer guidance on the optimal number and the desirable characteristics of the panel. It is important to set targets for different types of panelists (location, expertise, etc.) to be recruited and not just assemble a "convenience" panel. The characteristics of the panel are an important source of validity evidence when evaluating the results of the standard-setting study, since the participation of non-representative or non-knowledgeable panelists may cast doubts on the

appropriateness of the cut score. Documentation of the standard setting, to be discussed more fully later, should include a detailed account of the members of the standard setting panel and their representation in terms of gender, ethnicity, experience, geography (where appropriate), and any other relevant criteria.

Performance Level Descriptors

Performance Level Descriptors (PLDs) define the rigor or expectations associated with the categories, or performance levels, into which examinees are to be classified. For example, for placement cut score(s), the PLDs may clearly delineate the difference in content and/or proficiency expectations for examinees in a developmental course versus a college-level, credit-bearing course. SMEs bring into the standard setting process a diverse set of opinions about and experiences with both examinees and courses. While this diversity increases the generalizability of the standard setting results, it may also introduce a variation in initial definitions of the examinees within a performance level. Consider, for instance, a group of faculty members teaching the same course at a college or university. The course may be the same and may use the same curriculum and materials, but it is not uncommon for the requirements necessary to earn a grade of "A" from one professor to differ slightly, or even dramatically, from the requirements necessary to earn a grade of "A" from another professor. Therefore, it is likely that when asked to think of an examinee in a given performance level each SME will picture this hypothetical person differently. As a result, it is important that the SMEs have for their use in standard setting a set of *performance level descriptors* (PLDs). The PLDs facilitate the calibration of panelists by providing each panelist with the same working definition for each performance level.

The PLDs may be created during the standard setting study or in advance; some experts argue that the latter perhaps allows for a more thorough concentration on this key component of the process. The creation of PLDs can be a very time-consuming endeavor, adding up to a full day to the process of setting cut scores. For that reason, it is sometimes preferable to convene a panel of experts expressly for the purpose of creating the PLDs prior to the cut score session, so that this critical part of the process is given full attention. Then, during the process of setting cut scores, panelists are given the prepared PLDs and provided an opportunity to discuss, edit, and refine them. However, some experts prefer including the creation of the PLDs at the standard setting meeting, since the process of developing the PLDs serves to unify and calibrate the panelists and offers them a full understanding and familiarity with the PLDs that may not be fully achieved when beginning with PLDs crafted prior to the meeting by a group that may not overlap in membership with that of the SMEs serving on the standard setting panel.

Whether developed at the standard setting meeting or in advance, the set of performance level descriptors should:

- Describe what examinees at each level should reasonably know and be able to do for placement into the category.
- Relate directly to the content or proficiency standards, i.e., course pre-requisites and course requirements.
- Distinguish clearly from one level (developmental course) to the next (college-level course).
- Be written in positive terms.
- Be written in clear and concise language without using non-measurable qualifiers such as often, seldom, thorough, frequently, limited, etc.
- Quantify knowledge and skills as much as possible.
- Focus on achievement.
- Focus on the knowledge, skills, and abilities exhibited by the examinee at the borderline or threshold of the category rather than the typical or average performance across all students placed in the category.

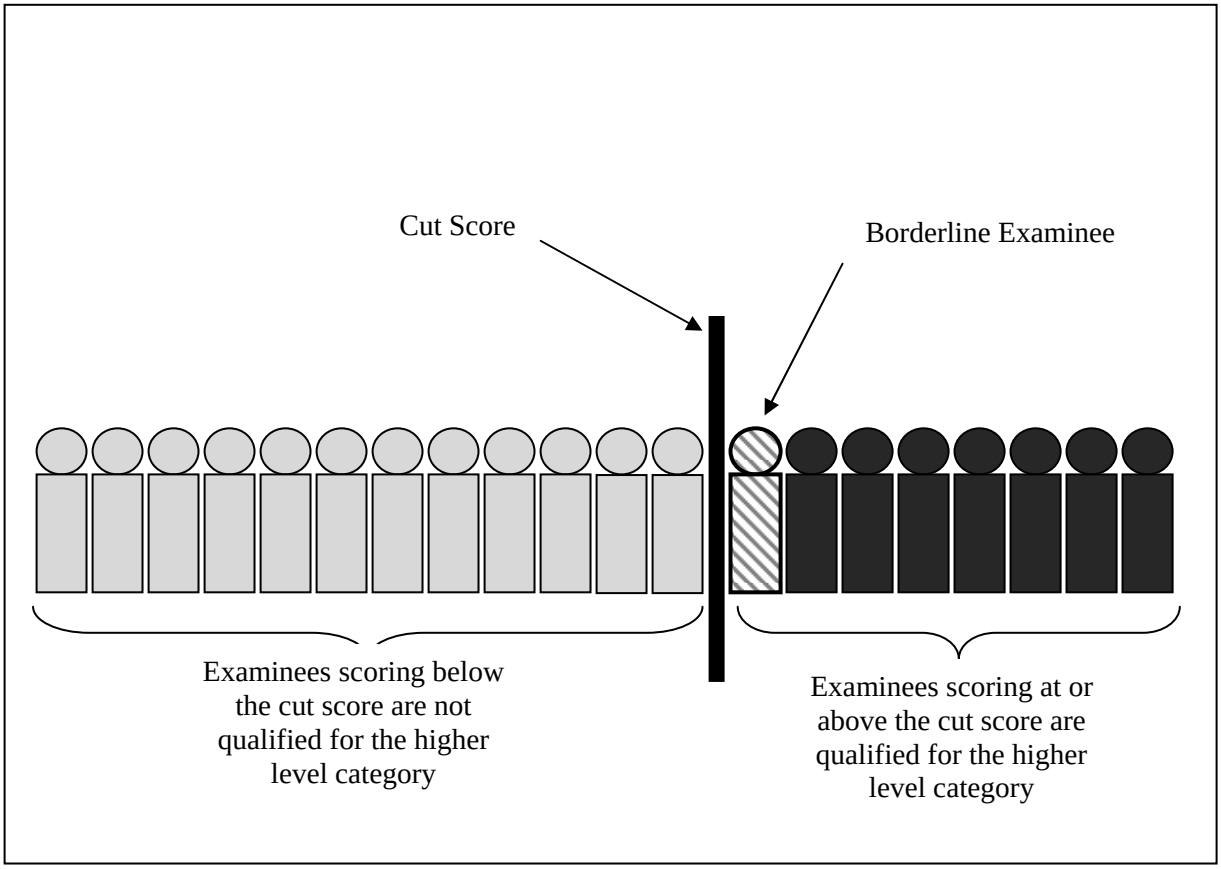
The objective of the standard setting process is to identify the point on the score scale that separates examinees who meet the specified expectations from those who do not. Generally each group (those who meet or exceed the standard and those who do not meet the standard) contains some portion of examinees that obviously belong in that specific group; for example, very low scorers or very high scorers. However, each group will also contain a number of examinees who either exhibit (1) just enough proficiency to be placed in the higher category or (2) who lack proficiency, but just barely, to keep them from being placed in the higher category. Examinees who exhibit just enough proficiency for placement into a category are known as *borderline examinees.*[3] When setting a cut score, the SMEs should make decisions with the borderline examinee in mind. Specifically, the performance of the borderline examinee on the test is important because many standard setting methods use the test score obtained by the borderline examinees as the cut score.

[3] Other terms for the borderline examinee that are used in the standard setting literature include *minimally competent examinee* and *just qualified examinee*.

Figure 2 provides an illustration of the ability continuum and the location on the continuum of the borderline examinee.

The performance level descriptors will be a major part of the documentation provided to the authoritative body and in the final report. The final set of performance level descriptors provides the meaning in words of the numeric cut score that will be set and adds to the validity of the standard setting process and the resultant cut score (Hambleton, 2001; Perie, 2008). For additional information on writing performance level descriptors and examples that have been used in standard setting studies, see Perie (2008), Cizek and Bunch (2006), Hambleton (2001), or Hansche (1998).

Figure 2: Location of the Borderline Examinee on the Ability Continuum



Standardization and Defensibility of the Process

A plethora of standard setting methods exists and the number increases steadily as new tests and item types are developed and researchers look for ways of accomplishing the standard setting task more efficiently. No single method can be determined as best, given that each method has advantages and disadvantages that will lend themselves better to some tests than to others. The choice of method must be made after careful consideration of the fit between the method and the test type along with consideration of several factors such as time required, resource requirements and availability, precedence of use, and degree of research and evaluation. This paper does not discuss specific standard setting methods and the advantages and disadvantages of each; other sources provide thorough overviews of various methods and can be consulted for additional information on specific methods (Cizek, 2001; Cizek & Bunch, 2006; Hambleton & Pitoniak, 2006; Hansche, 1998; Morgan & Michaelides, 2005; Pitoniak & Morgan, in press; Zieky et al., 2008).

Types of Standard Setting Methods

Tests used in the higher education context are comprised of various types of items. Information related to the items—such as the text of the item itself or information on its difficulty—may or may not be available to the institution, especially when commercially available tests are being used, due to concerns for test security. Initial considerations for choosing a standard setting method pertain to the characteristics of the test: Is it a multiple choice test, an essay or constructed response test, a combination of multiple choice and essay, or some other type of test construction? Some methods work well on multiple choice tests but do not lend themselves to other item types. Other methods work well with tests composed of multiple items or score points but are not adaptable to tests with a single essay question or other formats that may have a small number of score points. Identifying the type of test items used and methods that work well with that item type is a good first step in determining the method to use.

Once the types of methods that work well for the item types on the test being administered are identified, it is useful to consider the different types of information required to implement each method. Some of the methods require actual test items for panelists to rate, while others require samples of examinee test responses, i.e., the actual essay written or responses to a set of short answer questions. A select group of "examinee-centered" (Kane, 1998) methods requires that the SMEs be familiar enough with the examinees that judgments can be made about student ability beyond what is offered by test scores or samples of student responses. For incoming students undergoing placement testing, it is unlikely that college faculty would have the necessary knowledge of the

examinees to make use of an “examinee-centered” method. For cut scores intended for use in post-test or program evaluation contexts, “examinee-centered” methods may be possible. In addition to test items or examinee responses, some methods also require information on item difficulty, or examinee performance on each item and/or total test score. The availability of these different sources of information will be a determining factor in the selection of the standard setting method.

Another characteristic of tests that can have important consequences with respect to the type of method is the test administration method. Tests can be administered online, by computer, or by traditional paper-and-pencil. Traditional paper-and-pencil tests are most adaptable, since it may be possible to obtain an actual test booklet that can be used intact or copied into the necessary configuration for the chosen method. Computer or online tests can be more difficult to work with, but tests that are simply computerized versions of the traditional paper-and-pencil test have the same advantage of paper-and-pencil tests, though it may be difficult to obtain a printed version for copying and/or reordering, depending on the security protocols of the test vendor. Other tests administered online or by computer may take advantage of complex test administration algorithms that may alter the items administered to each examinee based on varying rules concerning item exposure, student performance, or branching rules designed to increase test security and/or identify the examinee’s ability in the most efficient manner possible. Because most examinees do not see the same set of test items identifying a single test form for use in the standard setting, study can be difficult and may necessitate the use of all items in the item pool or a method where the SMEs would actually take the test online or by computer, responding as they would expect the borderline examinee to respond rather than rating individual items or sorting student performance as is typical in many other methods. This would then require the availability of computers for the standard setting method, which may not be feasible in some situations.

The resources required by each standard setting method will vary in terms of the time and materials required. Most methods require that multiple rounds of ratings be made by the panelists. Some methods require that panelists provide separate ratings for every item. These methods can be quite time consuming, especially for tests of many items. Other methods require panelists to sort samples of student work into multiple categories. These sorting methods can be time consuming, as they generally are used for tests consisting entirely or primarily of essay and constructed response items that must be read by the panelists. The time needed to read each response is often lengthy and the preparation of materials can be time consuming since a complete set of multiple essays or test booklets needs to be prepared for each panelist. Additional time and resources are needed for materials preparation prior to and during the meeting, such as compilation and duplication

of materials, and for data entry and data analysis during and after the meeting. These activities may require that additional staff or special equipment be available, such as a copier or laptop. Moreover, data analysis may require special skills, depending on the method chosen, ranging from calculation of the mean, median, or percent correct to performing logistic regression or even a familiarity with item response theory.

Standard setting methods vary in their frequency of use, as well as in the extent to which they have been subjected to research and evaluation. The extent of research supporting the validity of the method and the degree to which the method is used play a role in establishing the method's legal defensibility. Ideally, research will be available on the method to assist in determining how well it has withstood both psychometric and legal scrutiny. All placement and admissions decisions have an impact on students, and some have argued that the impact of admissions decisions is greater or higher stakes than that of placement decisions. However, for the affected student, the consequences of either can be life changing. In general, the higher the stakes of the decision, the more important it is to use a method that has a strong history of use and research.

Common Characteristics of Standard Setting Methods

Certain functions and features of standard setting methods are common across most methods and play an integral role in establishing the validity and defensibility of the method. Knowledge of these common characteristics can assist institutions in evaluating standard setting methods for possible use:

1. Requires a diverse and representative set of subject matter experts (SMEs) to serve on the panel and make recommendations as to the placement of the cut score(s).

2. Requires the development/use of performance level descriptors (PLDs).

3. Requires knowledge of the test on which cut scores are being set, which includes taking the test for familiarity and a better understanding of the test difficulty.

4. Requires training on the method and includes the opportunity to practice using the method before making operational judgments.

5. Requires panelists to participate in multiple iterative rounds of independent judgment and ratings, alternating with large and/or small group discussions with feedback provided between judgment rounds regarding the location of the cut score(s) or tentative group membership from the most recent round of ratings. May or may not include impact data such as the expected pass/fail or placement rates if the current cut score(s) were to be implemented and other normative type data. Many experts argue that impact data turns the primarily content grounded process into a numbers-based process and argue against the introduction of impact data, while others feel it is a necessary reality check that should occur prior to the final round of ratings. (Cizek & Bunch, 2006; Morgan & Michaelides, 2005; Reckase, 2001; Zieky et al., 2008).

6. Collects panelist feedback about the process and panelists' level of confidence in the resulting cut score(s) at multiple points in the process. Typically evaluations are collected after training but before beginning operational judgments, so that readiness to proceed with the operational activities can be judged, and after the final round of judgments to collect panelists' confidence levels regarding the process, expected outcome, and any other factors that may be of interest for research or for planning purposes in the future, e.g., food, lodging, time needed, etc. Morgan (2006) and Morgan and Hardin (2009) provide examples of evaluation forms that may be used.

The six common characteristics described above are found in most standard setting methods, though the order, precise implementation, and frequency of data collection points may vary. These characteristics form the basis for establishing the validity of the standard setting process. Throughout the process, documentation will be produced and should be maintained as proof of the activities that occurred in case the validity of the cut scores is ever challenged. In the placement scenario, it is typically not in the student's best interest to challenge the cut score and the resultant course placement, since winning the challenge would most likely result in placement in a course for which he or she was not ready. However, an exception may be the case where a student experienced an atypical testing situation or came into the testing session at a disadvantage because of illness or emotional distress that affected performance. Most commonly, students who wish to challenge are given another opportunity to test or, in some cases, are allowed to enroll in the higher course to avoid the college's need to deal with the repercussions of insisting on the lower

course placement. Should the institution wish to stand firm on placement decisions, or perhaps have another use for cut scores that is challenged and must be defended, the composition of the SME panel, the method that was used, and all available documentation will be instrumental in showing that the cut scores were established with due diligence to detail and ensuring that a standardized and replicable process was followed. Unfortunately, when questioned about the origin of existing cut scores and the method used to establish the scores, many institutions indicate a lack of knowledge about how and when the cut scores were established or cite the adoption of cut scores in use at other institutions with little or no evidence about how they were derived and how well they fit the standards and course offerings of the institution before they were placed into use. If challenged, these institutions would be at a disadvantage in defending their cut scores and any decisions based upon those scores.

Documentation and Evaluation of the Process

Documenting all of the steps that were undertaken in designing and implementing the standard setting study is critical. A technical report should be compiled indicating how the method was chosen, how it was carried out, and providing detailed coverage of the results.

Documentation

Documentation is an integral part of the planning process. In the event that the cut score recommendations are ever challenged, the standard setting study report and other documentation are the evidence of what occurred and of what the panelists recommended. Documentation should include, but not be limited to, the following information (Pitoniak & Morgan, in press):

Documentation for the technical report:

- How the standard setting method was chosen (i.e., issues considered when making the decision).
- Information on panelist recruitment and qualifications (i.e., the target characteristics that were identified, and the degree to which they were attained).
- Agenda for study (including notes on activities that took a much shorter or longer time than planned).

- Performance level descriptions and description of process through which they were constructed.
- Detailed description of the method implemented.
- Types of feedback provided to panelists.
- Facilitator scripts, if used.
- Summaries of panelists' ratings and how they changed across rounds.
- Recommended cut scores after each round.
- Standard error of measurement.
- Standard error of judgment (which estimates the extent to which the cut score would vary if the study were replicated with many different samples of panelists).[4]
- Summary of impact data,[5] if used.
- Recommendation to the authoritative body and their subsequent decision

Documentation for the materials:

- A full set of the materials used and handed out in the study (e.g., test items, training materials, presentations, blank rating and evaluation forms, performance level descriptors, feedback information).
- Completed rating forms.
- Spreadsheet with compiled ratings and any subsequent analyses.
- Completed evaluation forms.
- Spreadsheet with summary of evaluation form responses.

It is advisable to document the process as soon after the study as possible. Having a written plan will facilitate completion of the technical report as much of the information

[4] For information on computing the SEJ, see Zieky et al. (2008) and Morgan (2006).

[5] It is helpful to provide estimates of the percentages of students in each performance category based on the cut scores +/- 2 SEJs and +/- 2 SEMs, for the total population and possibly for any subgroups of interest.

will already be in place. Describing the results of the study immediately after its completion will ensure that no important details are forgotten.

Evaluation

An integral part of the documentation is an evaluation of the execution of the study and its results. This is essential to judging the validity of the inferences to be made on the basis of the categorizations made using the cut scores. Many sources of validity evidence have been discussed in the literature. Kane (1994, 2001) has described them as falling into three categories: procedural, internal, and external (see also Hambleton & Pitoniak, 2006). It is desirable to include information from each of these sources in the technical report.

Procedural. The procedures that will be used in the standard setting study should be clearly articulated prior to the study to allow a careful review of plans by the authoritative body. Any decisions made should be spelled out as evidence that they were clearly thought out and not made in reaction to the results of the study. Examples of decisions made include whether to provide impact data, how to treat extreme ratings, whether the mean or median will be used, etc. Panelist evaluations are another key source of information about the procedural validity of the process. Information to be collected from panelists includes topics such as the efficacy of the orientation, understanding of performance category descriptors and the borderline examinee; training in the rating task; the helpfulness of discussion and feedback; and the level of confidence in the resulting standards. The questions that will be asked of panelists on evaluation forms and the number of evaluation opportunities needed should be carefully planned and documented in the standard setting plan. Results of the evaluation forms should be summarized and included in the final technical report for the study.

Internal. The *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999) indicates that "whenever feasible, an estimate should be provided of the amount of variation in performance standards that might be expected if the standard-setting procedure were replicated" (p. 60). The standard error of judgment (SEJ) can be computed as an estimate of the standard deviation of a large number of replications of the panel's recommendations. It is often advisable to calculate the recommended cut score +/- 2 SEJ for consideration by the authoritative body in determining the final cut score to be adopted. Morgan (2006) and Zieky et al. (2008) provide information on the calculation of the standard error of judgment. Other sources of internal validity evidence include inter-panelist and intra-panelist consistency. Inter-panelist consistency refers to the degree to which ratings are consistent across panelists. A very low estimate of inter-rater reliability would suggest that panelists may lack a common understanding of the performance level

definitions, the borderline examinee, or the rating task. Intra-panelist consistency refers to both (1) the relationship between the panelist's rating and item-level data, if available, and (2) to the extent to which the panelist's ratings change across rounds. A very low relationship between item difficulty and the panelist's ratings may call into the question the panelist's content knowledge. In addition, that a panelist's ratings do not change at all across rounds suggests that he or she may have an agenda and/or is not considering the feedback provided.

External. External validity may be evaluated by examining the extent to which the categorization of examinees into performance levels aligns with other information available about those examinees' knowledge, skills, and abilities. For example, an institution can compare the performance level categorization of examinees to their performance on another measure of the same content area. An example of this often occurs when an institution is investigating the use of a new placement test and is able to compare placement results with one instrument and the current operational cut scores to those with the new instrument and the newly recommended cut scores to determine how well the placement results align. Another approach, described by Morgan and Michaelides (2005), is to use logistic regression to examine the relationship between test scores and course grades through a validity study.[6] While this approach has some drawbacks as a method for setting initial cut scores for placement, its use as a method of confirming or evaluation results of the standard setting study and how well the cut scores are functioning can be valuable.

Validation of Cut Scores

It is important that cut scores be periodically evaluated to see that they are functioning as intended. It is a good idea to do this following the initial cut score study as a way of validating the new cut scores. Validity studies, though not recommended for setting initial cut scores, which use logistic regression to estimate the probability of success in a course given a specific test score, are effective ways to gather this information using only test scores and course grades. A study can consist of first administering the assessment to incoming students prior to any coursework being undertaken in the subject area. Next, students can be placed in courses using the method that was used previously by the college to determine placement and without any regard to the students' performance on the

[6] The organization that administers and scores the test may offer a service to assist with validating cut scores. These services include the Admitted Class Evaluation Service (ACES) offered by the College Board to users of its tests (http://professionals.collegeboard.com/higher-ed/validity/aces) and the Course Placement Service offered by ACT for its tests, including ACT, ASSET, or COMPASS (http://www.act.org/research/services/crsplace/). The authoritative body may want to check to see if the test publisher for the test on which cut scores were set offers this service.

placement test under consideration, which should be saved for use later in the process. When students finish the course where they were placed using whatever placement rules were already in existence at the college, their final course grade should be collected. It is important that the grades collected be for the course where they were placed and that no intervening coursework occurred between placement testing and entry into the course for which grades are being collected. The college should define what it considers to be "success" in the course; typically a grade of "B" or a grade of "C" is used. Employing a statistical procedure known as logistic regression, it is possible to predict the probability of success for students achieving each score on the score scale. This information can then be used to evaluate whether the test scores are working as intended for placement into the courses, i.e., that most students placing into the course are able to successfully complete it. In addition, the information gained about the probability of success for each score point on the test score scale can be used to inform decisions about the cut score that may be appropriate for use in placing students into the course. While not recommended for use in setting the initial cut score to use in placement testing because of the undefined relationship with content and what students should know and be able to do, this method is a good way to evaluate how well the cut scores established are working as a follow up in subsequent years to the standard setting study where the initial cuts were established.

When conducting a validity study, it is important that test scores and grades be collected for all students enrolled in the course with no intervening training occurring between testing and the start of the course of interest. Otherwise, it is possible to have skewed results due to self-selection and range restriction. Range restriction occurs when a portion of the population being examined is excluded from the sample such that the resulting sample is no longer representative of the population. This restriction of the range of population members can lead to a reduction in the correlation and a biased interpretation of the results (see Mattern & Packman, 2009). For example, if testing is not required for all students during the study period, it is likely that only a small portion of the students will participate and this portion may not be representative of the typical group of incoming students that is placed into the entry-level, credit-bearing course. Further, for colleges that use ACT or SAT as a screening device to exempt students from testing, the sample is likely to include results based only on the population of students with scores too low to allow for placement based on ACT or SAT score alone and the non-ACT or non-SAT takers. This group is typically of lower ability than those students submitting scores high enough to be used for exemption and there is, thus, a resulting decrease in representation of the higher ability students from what would be found in the total population of students entering college. Not only will a self-selected sample possibly result in skewed results for the validation study, but the restricted range will also cause the predictive power of the test scores to be underestimated. For a more thorough discussion of logistic regression, see

Agresti (1996), or see Mattern and Packman (2009) for an overview of placement validity studies and range restriction.

As mentioned earlier in this paper, “examinee-centered” standard setting methods are also effective ways of evaluating that cut scores are functioning. These “examinee-centered” methods require student test scores and faculty familiar enough with each student to make judgments about the appropriate performance level classification into which each student should be placed. One drawback to these methods is also the strength of the method—faculty know each student—because it is possible that extraneous factors such as attendance, behavior, participation, etc. will bias the judgment of the faculty member in making performance level judgments for the student. For that reason these methods are not preferred for setting initial cut scores but can be used to evaluate how well the cut scores are functioning.

In addition to evaluating cut scores following the initial study, it is recommended that cut scores be revisited periodically through a validity study or a standard setting methodology every five to seven years as a way of determining if the cut scores need adjustment. Cut scores may need to be revisited more often if the population of the institution, the institution’s curriculum, the placement test in use, or other factors change significantly such that the validity of the original study may be called into question as no longer valid and representative of the institution. All validity studies and standard setting studies, whether for validation purposes or for resetting cut scores, which are no longer valid or are not functioning as intended, should follow the guidelines provided for documentation and evaluation from start to finish in case a challenge to the cut scores is ever issued. Although validity evidence cannot establish that a cut score is appropriate, the lack of such evidence can cast doubts on its validity.

4. Case Studies

The following section provides a summary of two different standard setting scenarios and the methods that were used for each. The first describes the procedure used in a large standard setting to set multiple cut scores on diagnostics tests for a suite of commercially available placement tests. The second case study describes the combined use of methods that were employed to make cut score recommendation for a statewide policy on writing featuring an essay with a backup multiple choice exam for examinees scoring just short of the cut score required for automatic placement.

Case Study #1: Setting Cut Scores on a Diagnostics Test

A large group of approximately 60 subject matter experts were convened to set cut scores on a new set of diagnostics tests planning to implement a three-category scoring system. The anticipated scoring system would classify examinees into one of three performance categories based on the degree of strength or weakness in each test domain, or subscore, as determined by the examinees' performance on items within each domain. This particular standard setting was very large, with two cut scores being set on each of 20 test domains representing four content areas. A separate panel of approximately 15 SMEs was convened for each content area and had the task of setting two cut scores on each of five domains over a period of five days. While the content area and domains differed among each panel, the standard setting task was the same. Therefore, only one domain with two cuts will be discussed in this example.

The method used was a modification of the Bookmark Method (Lewis et al., 1998; Lewis et al., 1996; Mitzel, Lewis, Patz, & Green, 2001), a method that has been used in many contexts and is popular for use in setting cut scores on K-12 state exams. The set of SMEs participating on the panel were selected to be representative of the testing population in terms of geographical location, gender, and the mix of race/ethnicity. Both two- and four-year institutions, including public and private schools, were represented and the SMEs ranged in experience from two to more than 30 years of experience teaching in the content area at the college level. Each panel also contained at least one representative from secondary education in recognition of the use of the assessments at the high school level in addition to use at the postsecondary level.

The panelists began by taking the test and then crafting performance level descriptors (PLDs) of the three score reporting categories that would be used on the exam and then narrowing them down to the target definitions of the borderline examinee at each cut score. This task was accomplished by providing a very general description of what may

be an appropriate description of the category and allowing the SMEs to delete any parts that they did not agree with, add any information they felt was missing, and generally make any edits they liked, up to and including totally rejecting the descriptions provided and crafting new ones from a blank slate. The facilitator shepherded this process by asking the SMEs to respond to the provided descriptions and brainstorm to produce any ideas that may be appropriate. All ideas were accepted and treated as worthy of consideration during this stage. The follow-up, once the brainstorming was finished, involved a line-by-line review to identify common themes, contradictory statements, and, where needed, to word craft the list items to be more reflective of the category. This process was completed separately for each category and then the categories were reviewed together to ensure that they were complementary and did not contradict each other.

The SMEs underwent training in the use of the method, which included a chance to practice by using the method on a small set of items that would not be reviewed operationally until later in the schedule to minimize the chance that the training activity would unduly influence the judgments about items during the actual study. Once training was completed, the SMEs were asked to complete an evaluation form responding to a few procedural questions and indicating whether they felt confident in their ability to proceed with the standard setting task or if additional training was needed. Once all SMEs indicated that they were ready to proceed, then the actual standard setting task began.

The standard setting task involved the SMEs review of a booklet of test items in the operational test item pool; this was a computer adaptive exam and no one official set of items was administered to all students so it was necessary to use the item pool rather than a single test form. The set of items representing each domain was ordered from the easiest item to the most difficult item based on actual student performance data. The SMEs reviewed the set of items, making note of the specific knowledge and skills necessary to answer each item correctly. Then, using the PLDs as a reference, the SMEs individually identified the point in the ordered item booklet such that the knowledge and skills represented by items prior to that location were necessary but not sufficient for inclusion in the higher category and the knowledge and skills represented by items up to and including that spot were considered both necessary and sufficient for placement of the borderline examinee into the higher category. Once this process had been completed by all SMEs for each cut score that was needed, feedback was provided that indicated the range of possible cut scores as determined by the placement of each SMEs bookmark(s) in the ordered item booklet. The facilitator then led a discussion by prompting the SMEs to provide rationales for the bookmark placements paying particular attention to those at the extremes of the range of placements. Following the discussion, the SMEs were given an opportunity to revise their bookmark placements, should they choose to do so. The recommended cut score

was then computed as the median placement of the bookmark across panelists. Evaluations of the process and meeting logistics were collected before the SMEs were released. The recommended cut score, the PLDs, a summary of the results from the evaluation forms, and other relevant data were then provided to the test vendor for review and the final decision on the location of the cut scores for each domain.

Case Study #2: Setting the Cut Score for a Combination Placement Rule

The second case study was conducted on behalf of a statewide system to update the placement rule in response to a recent change in the scoring rubric for the essay used as part of the rule. This system used a two-part placement rule that placed students automatically into the college-level, credit-bearing course based on their achievement of a minimum score on the essay administered to each examinee. Examinees who did not qualify for automatic placement based upon the essay score were divided into two groups: those who scored too low for placement into the college-level, credit-bearing course and those that did not score high enough on the essay for automatic placement but were close enough that they may still have qualified by demonstrating proficiency in writing on a multiple-choice assessment of writing skills.

This study required a two-pronged approach to standard setting. The first was to determine the score on the essay under the new scoring rubric that would be used as the cut score for automatic placement into the college-level, credit-bearing course. The second approach was to identify the point on the multiple-choice exam that would be considered acceptable for placement when students fell just below the cut score on the essay. The cut score on the essay was established using the Analytical Judgment Method (Plake & Hambleton, 2001). The 20 Subject Matter Experts (SMEs) used in the study were the same for both approaches and were selected to be representative of the institutions, both two- and four-year, found in the statewide system. A group constituting a representative mix of gender, race/ethnicity, geographical representation, and teaching experience was convened; the individuals included representatives from secondary education, developmental, and college-level courses. Performance level descriptors (PLDs) were created in a similar fashion as discussed in Case Study #1 above, and the SMEs were trained in the standard setting method to be used in the first part of the study.

Each SME was provided with a packet of student essays written to the same essay prompt and representing each of the rubric points on the new scoring rubric. Multiple examples of each rubric point were provided, but no indicator of the score received on the essay was available to the SMEs. Using the PLDs as a reference point, the SMEs reviewed

each essay and independently sorted the essays into two stacks: those by examinees who showed sufficient knowledge and skills for placement into the college-level, credit-bearing course and those that did not. Once each SME finished the sorting task, a discussion of the relative merits of each essay, and the rationales behind each SME's decision about whether to place the essay in the higher or lower group, was conducted to allow the different perspectives and insights to be shared among the group in case the new information would make a difference in the sorting decisions by one of the SMEs. The SMEs were then asked to again review the essay sets and were given the opportunity to change the sorting of one or more essays should they desire to do so. Following this second round of sorting the facilitator identified the top three scores placed in the lower category and the bottom three scores placed into the upper category and then computed the median score within this set of essays to derive the recommended cut score on the essay.

The second part of the study was conducted on a multiple-choice assessment of writing skills administered adaptively by computer to each examinee such that most examinees would see items varying, in part or completely, from those of other examinees, depending on the ability of each examinee. The lack of consistency between the items that were administered to each examinee added increased difficulty to the standard setting task. Case Study #1 discussed one procedure for dealing with this; Case Study #2 utilized another approach, best described as a modification of the Yes/No Angoff Method (Angoff, 1971; Impara & Plake, 1997). The Yes/No Angoff method requires that SMEs review each item on the test and indicate whether the borderline examinee would be expected to answer the item correctly (Yes) or incorrectly (No). The Yes responses are given a value of 1 and the No responses are given a value of 0. The average of all Yes and No responses provided by the SME provides an estimate of the percent correct that each SME would expect of the borderline examinee. A discussion of the range of percent correct values from each SME in the context of why specific items may have been given a Yes or a No by some SMEs occurred after the round of judgments; afterwards the SMEs had the opportunity to change their judgments on any or all items as they desired. An average across the expected percent correct of all SMEs following the last round of judgments provided the recommended cut score for the assessment.

In the case of this statewide system, the collection of items seen by each examinee varied and it would be extremely cumbersome to make Yes/No judgments on every possible item combination or on all items in the item pool. As an alternative, a computer-based approach to standard setting was used. All the SMEs had use of a computer that allowed them to access the online site of the test vendor and to take the multiple-choice exam that would be used in the second part of the statewide system's placement rule. The SMEs were asked to take the exam and to respond to each item as they would expect the

borderline examinees to respond. If the borderline examinees would be expected to answer the item incorrectly, then the SME was to provide an incorrect response and if the borderline examinee would be expected to answer the item correctly, then the SME provided a correct response. Which incorrect response was provided did not make a difference as long as the response was incorrect or correct as intended by the SME.

When the SMEs reached the end of the exam, the score that each received on the exam was used to calculate a median score across all SMEs that would serve as the temporary cut score. Referencing that score, a discussion was held about the content that caused SMEs to enter an incorrect response versus a correct response so that diverse perspectives could be shared. Following the discussion, the SMEs once again took the exam, responding incorrectly or correctly as they would expect of the borderline examinees. Scores following this second round of testing were collected and the median score served as the recommended cut score for examinees not scoring high enough for automatic placement based on the essay alone. All results and documentation were provided to the statewide system for review and the statewide system served as the authoritative body that would make the final decision regarding the essay and multiple-choice scores to be used in the placement rule.

5. Selection of a Placement Test: Factors to Consider

The primary focus of this paper has been to summarize the use of and impetus for placement testing at institutions of higher education, with particular focus on the use of cut scores and best practices for establishing those scores. Before closing, it is important to discuss the factors that colleges should consider when selecting a placement test or, indeed, any exam, and the evidence needed to justify their selection decision.

At a minimum evidence of the following should be available and fully reviewed prior to selecting a placement test (Morgan, in press):

Validity

- Does the test measure content that is relevant for and aligned with the college-level, credit-bearing course in which students will be placed?
- Have the test scores been validated for the proposed purpose of the test?
- What evidence is available supporting the validity of the test for placement into the college-level, credit-bearing course?

Validity refers to the interpretation and use of assessment results and not to the test instrument being used, although the quality of the instrument does play an important role in the validity of the results. It is possible that a college may want to use the results from an assessment for multiple purposes: measuring students' knowledge and skills for proper course placement, describing students' growth from the beginning of class until the end of it, evaluating instructor performance over the length of a course or a particular instructional program/strategy. While all of these goals may be admirable, it is unlikely that one assessment has been constructed and validated to produce scores serving all of these purposes. Therefore, it is prudent to identify the primary purpose and use of the assessment results and investigate the evidence that exists to validate that use. Commercial test developers should have information available about the validity of score interpretations for specific purposes and uses. Colleges choosing to use locally developed assessments need to conduct validity studies to provide this evidence themselves. Should a placement decision or other decision be made that uses the assessment results and then fall under scrutiny, it is certain that the validity of the test scores and score interpretations will be among the first

points challenged. A lack of evidence to support the validity of using the score results will place a college on shaky ground in terms of defending itself.

An important aspect of score validity concerns the content of the assessment. Does the assessment measure what the college needs it to measure? For example, placement into a math course should be accomplished using a test that measures mathematical skills and produces a score that reflects the students' knowledge of math. A math test should not be so confounded with reading level or other content that a student with sufficient skills in math is unable to demonstrate that knowledge because of a reading difficulty; this is sometimes an issue when math tests have a preponderance of word problems and result in a score reflective of math and reading ability rather than just math ability. However, it may be that the college-level, credit-bearing mathematics course requires a student to have a particular reading level in addition to math skills to be successful and, therefore, the score reflecting both math and reading ability is appropriate for the intended use. Appropriateness of an instrument should be evaluated against the college-level, credit-bearing course content and requirements rather than assuming a math score from a particular instrument is appropriate for use in placement. The same holds true for the level of content being measured by the test: if, for example, a student will need to have an understanding of trigonometry to be successful in the college-level, credit-bearing math course, then the test should be assessing trigonometry and not solely computation or Algebra I. It is a good practice to have faculty review operational, if available, or sample test items prior to endorsing the use of scores from any placement test.

In addition to evaluating what the assessment is measuring and what knowledge and skills are being reflected by the test score, it is also important to assess the validity of the use of test scores for placement into the college-level, credit-bearing course. Typically, colleges must weigh the information they seek to obtain from an instrument against the financial resources and time available. A scarcity of either can limit the number of test questions that may be administered to a student and has implications for both the validity and reliability of test scores (Kane, 2006; Nitko, 1996; Payne, 1992; Popham, 1999).

Regardless of a college's resources, a determination of whether the test score is sufficiently valid for the purpose of placing students into the college-level, credit-bearing course is still to make. The institution would not want to waste resources on an assessment that is not serving its purpose, nor would it want to be in the position to have to defend any decisions made based on an assessment where evidence about the defensibility of the score use is missing. Thus, it is recommended that a predictive validity study be conducted to determine whether students placed into the courses are adequately prepared for success in the courses where they are placed.

Reliability

- What evidence is available supporting the reliability of the test scores? What are the reliability coefficients, standard error of measurement (SEM) for the test, and conditional standard errors of measurement (CSEMs) for each scale score and particularly for those a college intends to use for cut scores?

When thinking of the reliability of test scores, it is helpful to do so in terms of consistency. If the same student is tested multiple times with no intervening instruction or learning between test administrations, does he or she get a similar score each time? All scores contain some error; that is, fluctuations in test scores may be due to factors unrelated to student ability and not an indication that a mistake was made. The error may be systematic, meaning it has the same influence on the student in repeated testing sessions or it affects all examinees on one occasion; for example, poor lighting or a disruption in connectivity to the Internet during testing one afternoon in the testing lab would have an effect on all students in the testing lab at that time. The error may also be unsystematic, showing no consistent pattern and fluctuating from one situation to the next; for example, the student may have different levels of motivation during one or more testing sessions due to fatigue, emotional distress, illness, or other such factors. A test score is a snapshot of student performance and ability on that test on that day at that time under those circumstances. Changing one or more of these factors could result in a change in the student's score. In addition, a test form is simply a sampling of items from the domain of all possible items that could have been used. On a different set of items it is possible that a student could receive a higher, or lower, score because the item set administered was a better match to the examinee's skill set. As such, the more items on the assessment the greater the percentage of all possible items in the domain being represented. For this reason, longer tests with more items measuring the intended construct generally have higher score reliability than shorter tests with fewer items.

An exception to the positive correlation between test length and reliability may occur when one of the tests scores being compared results from a traditional paper-and-pencil exam while the other score is achieved on a computer adaptive test (CAT). In the case of the CAT, a smaller number of items can produce scores that are as reliable as the traditional paper-and-pencil test score, if not more so, due to the ability of the item selection algorithm to choose items that are specifically targeted in both content and difficulty to the ability of the student based on his or her performance on previous test items. This targeted item selection allows for an efficiency of measurement that results in shorter test lengths producing scores with the same or higher reliability. (Haertel, 2006; Hambleton, Swaminathan, & Rogers, 1991; Nitko, 1996; Payne, 1992; Popham, 1999).

Reliability estimates range from zero to one, with estimates of one indicating that the test score is free of error and estimates closer to zero indicating the score is mainly a function of error. Many ways to measure reliability exist and include test-retest, internal consistency, and parallel forms. Higher reliability estimates are desirable but the exact size of an acceptable estimate will vary depending on the intended score use and the type of reliability being estimated (Haertel, 2006). In addition to the reliability estimate, it is also advisable to consider the Standard Error of Measurement (SEM) reported for the scores. The SEM is an indicator of the consistency of an individual person's score and how much that score could be expected to vary in the event the examinee is tested multiple times. Smaller values of SEM indicate greater precision in the score and are better than larger values of SEM which indicate less precision. The precision of scores can vary at different points along the score scale due to factors such as not having sufficient numbers of items available at certain points on the score scale. For this reason, SEM values may be computed conditional upon the score of interest. These Conditional Standard Errors of Measurement (CSEMs) should be examined for the places on the score scale where a cut score may be located to ensure sufficient precision at the cut for the decisions being made.

Sensitivity and Bias

- Have test items been reviewed for bias and sensitivity?
- Have studies been conducted to investigate Differential Item Functioning (DIF) of items? Do items disadvantage any subgroup when controlled for student ability?

An important part of evaluating an assessment for potential use as a placement test is ensuring that the test results are free of bias. This process begins with an investigation of the item development practices of the test developer. How are the items developed? Who develops them, and based on what specifications? Are the items field tested prior to use operationally to ensure that only items that are working as intended are administered to a student and used to compute the student's score? Are the test items reviewed by experts for potential bias or sensitivity issues? For example, do the items contain any content that could inadvertently upset the examinee in such a way that if would affect test performance? Other factors to watch out for with test items are words or contexts that may have nothing to do with the knowledge needed to correctly answer the item but which are not familiar to the examinees and may cause them to be unsure in their response due to the lack of familiarity with the context. For example, examinees in rural Iowa may have difficulty answering an item in the context of high and low tide, but may be perfectly fine if asked to demonstrate knowledge on the same skill in the context of a snowy day or other context familiar to them.

A former student once had difficulty answering a question of the life cycle of frog because in his home country tree frogs were abundant and never passed through the tadpole stage illustrated on the assessment. A bias and sensitivity review of all items should be part of any item and test development activities. (Camilli, 2006; Camilli & Shepard, 1994; Crocker & Algina, 1986).

Another way to assess the potential bias in an assessment includes reviewing evidence for differential item functioning (DIF). A DIF analysis is a statistical procedure that compares the performance of examinees who are the same or very similar in terms of ability levels and the overall test score but who differ in subgroup membership (e.g., gender, race/ethnicity, SES, ESL), in terms of their performance on each item. The assumption is that examinees scoring the same or very similarly on the total assessment should also perform the same or very similarly on each individual item. When examinees matched in ability are separated by subgroup and performance on an individual item differs such that members of one subgroup (females or Hispanic examinees, for example) are much more likely to answer an item incorrectly than the subgroup used as a reference group (in this case, males or White examinees) then the item needs to be examined more closely to try to determine why examinees from a certain subgroup are performing differently from members of other groups on this particular item. If the contributing factor can be identified then the item may be edited and field tested again for re-use on the exam. More often, the cause is not clear and the item is removed from the exam and not re-used (Camilli, 2006; Camilli & Shepard, 1994; Holland & Wainer, 1993).

In the event that an examinee or other entity ever challenges the use of a placement test at a college, it is very likely that evidence showing that the tests are free of bias will be one of the top considerations. Therefore, it is important to use an assessment with sufficient evidence to support that scores are free of bias and to maintain a copy of that evidence for the lifetime of the college's use of the assessment.

6. Conclusion and Recommendations for Further Research

Cut scores are used in a variety of circumstances to aid in decision making through the establishment of a clear cut line between categories. This paper has attempted to provide an overview of the many facets of the standard setting process that must be taken into consideration. A key step in placement testing begins with ensuring that the instrument being used produces valid and reliable score interpretations for the intended use at the institution. Once this foundation has been established, the setting of cut scores through a defensible and standardized process of activities allows for the implementation of decision rules as part of the placement process. The suggested steps and activities that are part of this standardized process have been the primary focus of this paper. In addition to setting cut scores, many other aspects of placement testing require further consideration. The remainder of this section lists some possible areas for further research.

Research is needed to better understand the role and consequences of placement testing in higher education The use of placement tests is widespread but surprising little research exists on the outcomes of student placement in terms of the students' success in the course where placed. Long-term studies are needed that track students placed in college-level, credit-bearing courses and those placed in developmental courses to determine retention, time to graduation, comparability of curriculum taken, etc.

A recent trend shows that community colleges visit local high schools and, in some cases middle schools, to test students. This undertaking is both a means of recruitment and a pro-active approach to remediation, since it identifies student strengths and weaknesses in time to address some concerns while the student is still in high school and can obviate the need and expense of remediation for both the student and college once the student is on the college campus. Stories and anecdotal evidence of these efforts abound at meetings and conferences of community college test administrators, but strong empirical evidence is lacking on the prevalence and gains associated with such practices.

Additional related research is needed to investigate the consequences of placement testing in the high schools. One concern centers around the large range of placement test cut scores in use at colleges. Colleges are typically free to set their own cut scores and this often results in different scores being required for placement into the same college-level, credit-bearing course depending at different colleges. Anecdotal evidence suggests that some students have learned to "play the game" and essentially shop around for the best placement: At which institution will the scores earned qualify the student for placement in the college-level, credit-bearing course? Confirmation of the occurrence of this behavior

and quantification of its prevalence is needed to determine what, if anything, should be done to discourage it.

Conversely, what is the effect on the high school student who tests under the placement rules of one college that visits the student's high school campus and indicates the student is "college ready" for placement into the college-level, credit-bearing course at that college? Does the student recognize that this does not necessarily guarantee the same placement at other institutions? Does the student, having received the "college ready" designation, spend less time preparing for college, not realizing that the college he or she plans to attend could have more rigorous placement testing standards? Does the student who is told "not college ready" try harder to overcome weaknesses, or become frustrated and lack motivation when he or she may be qualified for placement into the college-level, credit-bearing course at another institution?

An even more basic question that should be addressed: What is the level of student awareness regarding the likelihood of being required to take a placement test? Could student remediation rates be lowered by increasing the student expectation that placement testing will be required and educating them while in high school on the use of such tests so that they come into testing better prepared?

The above areas all deserve further study and research on them would provide a meaningful contribution to the higher education literature base while also having the potential to inform policy decisions. Answers to some of these questions may be a first step in starting to close the gap between high school and postsecondary expectations. Organizations such as a ACHIEVE, Inc. and the Council of Chief State School Officers (CCSSO) are working on the common core standards in an attempt to begin to close the gap between high school and college expectations by identifying students as "college and career ready." However, early efforts do not seem to extend far enough to actually close the gap. Perhaps with a combination of the common core standards effort and the implementation of new programs or policies at the postsecondary level through increased student awareness, early identification of student strengths and weaknesses using college placement tests, and a better coordination between institutions regarding the standards or cut score that student's must achieve for placement into the college-level, credit-bearing course, the gap can be narrowed more quickly.

In the immediate future, ensuring cut scores in use are recent, fully documented, and defensible against challenge will improve the standing of any college test use. Additionally, knowledgeable review and use of any testing products is a must in the event that test use or any decisions resulting from test use are ever challenged.

References

Achieve, & The Education Trust. (2008, November). *Making college and career readiness the mission for high schools: A guide for state policymakers.* Retrieved from http://www.achieve.org/files/MakingCollegeandCareerReadinesstheMissionforHighSchool.pdf

ACT. (2007). *Rigor at risk: Reaffirming quality in the high school core curriculum*. Iowa City, IA: ACT.

Agresti, A. (1996). *An introduction to categorical data analysis*. New York, NY: John Wiley.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education [AERA, APA, & NCME]. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508–597). Washington, DC: American Council on Education.

Camara, W. J. (2003, March). *College persistence, graduation, and remediation* (Research Notes No. RN-19). New York, NY: The College Board.

Camilli, G. (2006). Test fairness. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 220–256). Westport, CT: American Council on Education.

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items* (Vol. 4). Thousand Oaks, CA: Sage.

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

Cizek, G. J. (2001). *Setting performance standards: Concepts, methods, and perspectives*. Mahwah, NJ: Erlbaum.

Cizek, G. J. (2006) Standard setting. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 225-258). Mahwah, NJ: Erlbaum.

Cizek, G. J., & Bunch, M. B. (2006). *Standard setting: A guide to establishing and evaluating performance standards on tests.* Thousand Oaks, CA: Sage.

The College Board (2002). *Best practices in admissions decisions: A report on the third College Board conference on admissions models.* New York, NY: Author.

The College Board. (2010). *Guidelines on the uses of College Board test scores and related data.* New York, NY: Author.

Crocker, L. A., & Algina, J. (1986). *Introduction to classical & modern test theory*. New York: Holt, Rhinehart, and Winston.

Geisinger, K. F. (1991). Using standard-setting data to establish cutoff scores. *Educational Measurement: Issues and Practice, 10*(2), 17–22.

Haertel, E. H. (2006). Reliability. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 65–110). Westport, CT: American Council on Education.

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*. Mahwah, NJ: Erlbaum.

Hambleton, R. K., & Pitoniak, M. J. (2006). Setting performance standards. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 433–470). Westport, CT: American Council on Education.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Thousand Oaks, CA: Sage.

Hansche, L. N. (1998). *Handbook for the development of performance standards: Meeting the requirements of Title I*. Washington, DC: U.S. Department of Education, Council of Chief State School Officers. Retrieved from: http://www.ccsso.org/publications/details.cfm?PublicationID=131

Haycock, K., Barth, P., Mitchell, R., & Wilkins, A. (Eds.). (1999). Ticket to nowhere: The gap between leaving high school and entering college and high-performance jobs. *Thinking K-16, 3*(2). Washington, DC: Education Trust.

Holland, P. W., & Wainer, H. (Eds.) (1993). *Differential item functioning*. Mahwah, NJ: Erlbaum.

Impara, J. C., & Plake, B. S. (1997). Standard setting: An alternative approach. *Journal of Educational Measurement, 34*, 353–366.

Jaeger, R. M. (1989). Certification of student competence. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 485–514). New York, NY: Macmillan.

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

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

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

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: American Council on Education.

Kingston, N. M., Kahl, S. R., Sweeney, K., & Bay, L. (2001). Setting performance standards using the body of work method. In G. J. Cizek (Ed.), *Standard setting: Concepts, methods, and perspectives* (pp. 219-248). Mahwah, NJ: Erlbaum.

Kirst, M. W. (2005). Rethinking admission and placement in an era of new K-12 standards. In W. J. Camara & E. W. Kimmel (Eds.), *Choosing students: Higher education admissions tools for the 21st century* (pp. 285–312). Mahwah, NJ: Erlbaum.

Lewis, D. M., Green, D. R., Mitzel, H. C., Baum, K., & Patz, R. J. (1998, April). *The bookmark standard setting procedure: Methodology and recent implementations.* Paper presented at the meeting of the 1998 National Council on Measurement in Education, San Diego, CA.

Lewis, D. M., Mitzel, H. C., & Green, D. R. (1996, June). Standard setting: A bookmark approach. In D. R. Green (Chair), *IRT-based standard setting procedures utilizing behavioral anchoring*. Symposium presented at the Council of Chief State School Officers National Conference on Large-Scale Assessment, Phoenix, AZ.

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

Mattern, K. D. & Packman, S. (2009). *Predictive validity of ACCUPLACER scores for course placement: A meta-analysis* (Report No. 2009-2). New York, NY: The College Board.

Mehrens, W. A. (1986). Measurement specialists: Motive to achieve or motive to avoid failure? *Educational Measurement: Issues and Practice, 5*(4), 5–10.

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: Erlbaum.

Morgan, D. L. (2006). *Setting local cut scores on the SAT Reasoning Test™ Writing Section for use in college placement and admissions decisions* (College Board Special Report). New York, NY: The College Board.

Morgan, D. L. (in press). College placement testing of entering students. In C. Secolsky (Ed.), *Measurement, assessment, and evaluation in higher education*. New York, NY: Routledge.

Morgan, D. L., & Hardin, E. (2009). *Setting cut scores with WritePlacer®* (College Board Special Report). New York, NY: The College Board.

Morgan, D. L., & Michaelides, M. P. (2005). *Setting cut scores for college placement* (Report No. 2005-9). New York, NY: The College Board.

National Center for Education Statistics (NCES). (2004). *Remedial education at degree-granting postsecondary institutions in fall 2000* (Report No. 2004-010). Washington, DC: U.S. Department of Education.

Nitko, A. J. (1996). *Educational assessments of students* (2nd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Payne, D. A. (1992). *Measuring and evaluating educational outcomes*. New York, NY: Macmillan.

Perie, M. (2008). A guide to understanding and developing performance-level descriptors. *Educational Measurement: Issues and Practice, 27*(4), 15–29.

Pitoniak, M. J., & Morgan, D. L. (in press). Setting and validating cut scores for tests. In C. Secolsky (Ed.), *Measurement, assessment, and evaluation in higher education*. New York, NY: Routledge.

Plake, B. S., & Hambleton, R. K. (2001). The analytic judgment method for setting standards on complex performance assessments. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp.283–312). Mahwah, NJ: Erlbaum.

Popham, W. J. (1999). *Classroom assessment: What teachers need to know* (2nd ed.). Needham Heights, MA: Allyn & Bacon.

Raymond, M. R., & Reid, J. B. (2001). Who made thee a judge? Selecting and training participants for standard setting. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 119–157). Mahwah, NJ: Erlbaum.

Reckase, M. D. (2001). Innovative methods for helping standard-setting participants to perform their task: The role of feedback regarding consistency, accuracy, and impact. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 159–173). Mahwah, NJ: Erlbaum.

Rigol, G. W. (2003). *Admissions decision-making models: How U.S. institutions of higher education select undergraduate students*. New York, NY: The College Board.

Venezia, A., Kirst, M. W., & Antonio, A. L. (2003). *Betraying the college dream: How disconnected K-12 and postsecondary education systems undermine student aspirations*. Retrieved from Stanford University, Institute for Higher Education Research, Bridge Project website: http://www.stanford.edu/group/bridgeproject/embargoed/embargoed_policybrief.pdf

Zieky, M. J., Perie, M., & Livingston, S. A. (2008). *Cutscores: A manual for setting standards of performance on educational and occupational tests.* Princeton, NJ: Educational Testing Service.