Using Test Standard-Setting Methods in Educational Program Evaluation: Addressing the Issue of How Good is Good Enough

Paul R. Brandon

School districts in the United States and elsewhere commonly use standard setting to assign value to student test and assessment scores. That is, they set standards to show "how good is good enough." This paper presents a summary of the empirical findings on the most widely-studied test standard-setting method and describes what the conclusions of the summary suggest about the use of test standard-setting in educational program evaluations.

The purpose of setting test or assessment standards is to establish judgmentally the *cutscores* that show the dividing points between levels of student performance such as pass and fail, basic and proficient, proficient and advanced, and so forth. Cutscores are established with methods such as the modified Angoff method, the contrasting-groups method, the bookmark method, and several others (Cizek, 2001). As part of student and school accountability efforts, districts report to students the performance levels at which their scores fall and report to policymakers and to the public the percentages of students achieving at the various performance levels. The U. S. No Child Left Behind Act has enshrined the use of cutscores, in that schools are required to identify and report student proficiency levels and to increase the levels of students who score below proficiency.

Cutscores are set either by making judgments about test items or about examinees' performance on tests or assessments. Methods for making judgments about test items are known as *test-centered methods,* and methods for making judgments about examinee performance are known as *examinee-centered methods* (Jaeger, 1989). The test-centered method that for years was the most frequently used and that remains the most widely studied method is the modified Angoff method (Angoff, 1971), and probably the most frequently studied examinee-centered method is the contrasting-groups method. In preparation for studying how and when to use test standard-setting methods in educational program evaluations, I conducted exhaustive reviews of the literature on these two methods (Brandon, 2002, 2004).

Before districts or states set cutscores, they first must develop *performance standards*. A performance standard is a statement defining and describing the knowledge or skills that students must show at a particular performance level. Performance standards are developed before cutscores are set; cutscores are the operationalized versions of performance standards. Sometimes policy makers specify performance standards and sometimes the panels of judges that set cutscores develop them.

Under what conditions and for what purposes might it be appropriate to conduct standard setting in program evaluations? This topic has been discussed sketchily by some (e.g., Cook, Leviton, & Shadish, 1985; Rossi & Freeman, 1993; Shadish, Cook, & Leviton, 1991; Worthen, Sanders, & Fitzpatrick, 1997) and somewhat more thoroughly by a few others (e.g., Fink, Kosecoff, & Brook, 1986; Henry, McTaggart, & McMillan, 1992; Patton, 1997; Wholey, 1979). The inattention given to the topic is unfortunate, because the appropriateness of using standard-setting methods in program evaluation has not been thoroughly discussed, and the

types of evaluation instances in which using cutscores would be helpful and appropriate have not been well-established.

This article examines the use of test standard setting in educational program evaluations. It begins with a recounting of the primary findings of my review of the literature on the modified Angoff method (Brandon, 2004). I focus on this method because it has been examined empirically more than any other method. However, despite the relative abundance of research on the method, the empirical literature does not provide strong support for the validity of modified Angoff cutscores. Therefore, in this article, I am cautious about applying the method in program evaluation. I argue that it is appropriate under certain testing conditions in formative evaluation studies or when conducting preliminary summative studies of program outcomes. Studies of these types require a lesser degree of validity than summative evaluations used by policymakers to make go/no-go program decisions. Based on the results of the literature review, I discuss flaws in the methods of modified Angoff studies. I then discuss

1. the types of decisions that might be made when interpreting evaluation results in light of cutscores and the strengths of the conclusions made based on test standard setting in evaluations,
2. the program evaluation scenarios in which it is appropriate to use cutscores for interpreting evaluation results, with a focus on the stage of evaluation and the types of evaluation designs, and
3. four criteria that evaluators should address when using cutscores to help interpret evaluation results.

This article is limited by my decision to base conclusions primarily on empirical findings about the modified Angoff research. Some evaluators might wish to know

what standard-setting methods other than the modified Angoff method can be used in program evaluations. Psychometricians and researchers are continually developing new standard-setting methods (Cizek, 2001); many such as the bookmark method are proving promising, and evaluators might wish to learn from the research on them. However, the intent of this article is base conclusions on empirical research, and little sound research has been conducted methods other than the modified Angoff. For example, considerable attention has been paid to the contrasting-groups method, which for years probably was used more than any other examinee-centered approach, but little research has been conducted on it (Brandon, 2002). I base my conclusions solely on the research on the modified Angoff method because I have adopted a conservative approach to applying the standard-setting literature to program evaluation. I limit myself to the best research available; the body of modified-Angoff research may be less comprehensive than desirable, but it is broader and goes deeper than the research on other methods.

The article also is limited because it does not suggest how to apply standard setting methods for purposes other than test standard setting in program evaluation. Other than brief comments in the final paragraph of the article, I do not speculate about using the method for other purposes. Very little program evaluation research has been conducted on using standard-setting methods for purposes other than testing. (I have experimented in two evaluations with applying standard-setting methods to judging how well the evaluated programs were implemented, but the success of the efforts was mixed.) There was no research on test standard-setting methods when they were first put into wide use; I do not intend to repeat that scenario by making recommendations about using standard setting in program evaluation for purposes other than tests without an empirical basis for my suggestions. The place for extensive speculation about other uses of standard setting in program evaluation is

elsewhere.

The Methodological Soundness of the Modified Angoff Method

To learn about the soundness of test standard-setting, it is useful to discuss the modified Angoff method, not only because it is an exemplar of one of the two primary types of test standard setting, but also because more empirical research has been conducted on it than any other standard-setting method. As this section shows, the evidence for the effectiveness and validity of the method is less convincing than desirable, the literature is narrow, and many of the studies of the standard-setting method are unsound or incomplete.

The modified Angoff method includes three primary steps. The method is called *modified* because some aspects of it were developed after Angoff (1971) first proposed it. The first step is to select and train judges. The second step is to define and describe the performance level that examinees must meet—that is, to establish the performance standard. Judges can conduct this step, but often policymakers or others provide judges with the performance standard. The third step is to make *item estimates*—that is, to establish estimates of the probabilities that examinees will correctly answer the items on the test or assessment at the level of the performance standard. Usually judges conduct two or three rounds of item estimation. Between rounds, the judges review empirical information such as the difficulty level of each item and have discussions about their item estimates; then, if they wish, they revise their estimates in the next round. After the three steps are conducted the cutscore is calculated by summing the item estimates for each judge and averaging the sums across judges.

Researchers and practitioners have studied the modified Angoff method more than any other, but some of the findings on the steps are inconclusive:

Selecting and training judges. Some of the research on selecting and training judges provides conclusive findings, but other research does not. Studies suggest that the appropriate number of judges for modified Angoff studies is 10–20. The conclusions of the small number of empirical studies on this topic (Brandon, 2004) generally were within this range.

Selecting judges for their subject-matter expertise can enhance item estimation, but not all judges need have high levels of expertise. Research on this topic is inconclusive because of some of the studies that I identified had methodological flaws and because other studies examined incomplete versions of modified Angoff standard setting.

Very little research has been conducted on training judges, and no results bear summarizing here.

Defining and describing the performance standard. The findings of a small body of studies support the conclusion that definitions and descriptions of performance standards should be made using a set of prescribed steps and that performance standards should be fully explicated. Research on the topic is inconclusive because about half of the studies on it were simulations of standard-setting that did not include or fully implement all the modified Angoff steps (Brandon, 2004).

Defining and describing performance standards is a difficult step to carry out fully and validly. Developing statements of performance standards for high school graduation tests requires judges to have a full understanding of the knowledge and skills that teenagers must have upon entering the workforce or post-secondary education, and developing performance standards for earlier school grades requires judges to estimate the level of students' knowledge and skills necessary for success in the following grades. In both these standard-setting instances, judges must know

what they are setting proficiency scores *for*. That is, they must understand the purpose of the standard setting and the context that students will be in when the students use the knowledge and skills that are addressed in the examination. "To say that adequacy must be defined for some purpose has important implications for validating passing scores as well as validating performance standards. This condition is much more stringent than requiring the passing score to be consistent with the description of performance standards" (Camilli, Cizek, & Lugg, 2001, p. 459). Understanding what scores are set for is not a trivial endeavor; indeed, some would say it is impossible: "Performance standards simply cannot help us decide whether Johnny or PS 19 or Colorado has enough reading skill, because there is no sensible answer to the question, 'Enough reading skill for what?' beyond the trivial level of 'Enough reading skill to answer test question 36 correctly'" (Burton, 1978, p. 270).

There are no well-established developmental theories to guide methods for estimating what students' necessary levels of performance should be upon graduation. What students need to know and be able to do depends upon the educational or vocational paths they will follow upon graduation. The proficiency level necessary for someone to go directly into the workforce is different from level necessary for someone to enter a community college, which in turn varies from the level necessary someone entering a competitive four-year post-secondary educational institution. The minimum levels of knowledge and skills necessary to succeed in these settings, as well as the highest levels of proficiency that can be expected, vary among these settings. Similar issues apply to setting cutscores for elementary and middle school tests and assessments. Kane (2001, pp. 58, 82–83) said,

> There are generally no accepted performance standards for life after high school

> and no empirical base of information relating performance in history or science in eighth or twelfth grade to success in life (however that might be defined)… Standards seem most arbitrary when the contingencies they are designed to address are very vague and open-ended. The standards set on a high school graduation test are likely to be judgmental, because the level of skill that a graduate will need for work or life will depend on where they work and how they choose to live, and therefore there is no clear focal activity or contingency that can serve as a guide in standard setting. Standard-setting judges must know what students must be proficient for.

A comparison with standard setting in the military is informative. In military settings, training standards are established and applied in personnel decision making. Military training standards address clear external criteria such as the knowledge and skills necessary to operate equipment or perform specialized tasks. This is also more or less the case in standard setting for licensure or certification—a topic addressed in much of the standard-setting literature. It is not the case in K–12 education, where "it is highly unlikely that a teacher will have had experience in the career that his or her students eventually choose to enter. . . . Schools are relatively isolated from the world of work and the consequences of the quality of education they provide, whereas military training centers and operating units are tightly integrated" (Hanser, 1998, p. 82). If traditional K–12 standard-setting methods were used in the military, "the trainers who set the training standards could be quite divorced from field experience" (Hanser, p. 92)—a clearly unacceptable state of affairs. "Standards that are relatively context free are difficult to set and accept" (Hanser, p. 93).

Making item estimates. More research has been conducted on making item estimates than on any other modified-Angoff step. Some of the findings of this research support the conclusion that cutscores are valid, but other findings make us

question the strength of that conclusion.

The findings of research on the extent to which item estimates are correlated with item difficulty levels—a relatively common thread of research in the empirical standard-setting literature—suggest that the estimates moderately mirror item difficulty. This finding is an indication of the validity of the estimates.

Other studies have examined the effects of activities between standard-setting rounds, when judges review empirical information about items and discuss this information and their item estimates. The results of these studies suggest that judges' between-round activities affect the magnitude of cutscores. However, these results are tentative because about a third of the studies on the topic have not confirmed these findings (Brandon, 2004) .

Other results suggest that judges' between-round activities decrease item estimates' variability and increase their reliability from round to round (desirable results). However, the results about decreasing variability are inconclusive because of large standard deviations, and the results about increasing reliability are inconclusive because of the number of studies is small and the methods for calculating reliability varied among studies. Hurtz and Auerbach (2003) found that judges' discussions among themselves reduced the variability of cutscores but that reviewing empirical information did not.

Researchers also have examined the absolute value of the differences between item estimates and empirical *p*-values. Their studies address *item accuracy*. The rationale behind the studies is that there should be small differences between item estimates and the empirical *p*-values of examinees whose scores are deemed to be close to the cutscore. Although some evidence has been found that judges are able to make estimates accurately, the results of several studies suggest that item

estimation might be less valid than desirable because judges tend to underestimate the difficulty of hard items and overestimate the difficulty of easy items. Of all the findings about item estimates, these are the most troubling for the validity of modified Angoff cutscores. Indeed, Shepard (1995, p. 151) concluded that findings such as these showed that "judges were unable to maintain a consistent view of the performance they expected" and thus made judgments that were "internally inconsistent and contradictory."

Conclusions About the Modified Angoff Method and Its Literature

The findings about item accuracy and the findings about the "proficiency for what" issue lead us to be concerned about using cutscores for a wide variety of program evaluation purposes. These are not the only reasons to be cautious about using the method in program evaluations, however. There also are three flaws in the literature that throw doubt on using the method for a broad array of evaluation scenarios.

The first flaw has to do with the breadth of the literature: It is broader than the research on other standard-setting methods, but it is still narrower than desirable. Insufficient empirical research has been conducted on some steps of the modified Angoff method, particularly on selecting judges, the need for judge subject-matter expertise, judge training, and defining and describing the performance standard.

> More research has been conducted on the modified Angoff method than any other standard-setting method, but the findings of the extant research provide only the first few layers of an empirical foundation for making decisions about how to set cutscores. These layers alone cannot serve as the sole basis for deciding about how to go about setting modified Angoff cutscores; clinical guidance by experienced practitioners is also necessary.

(Brandon, 2004, p. 80)

The second flaw has to do with the reporting of studies. Many empirical modified Angoff studies have not reported full descriptions of the standard-setting methods that were used:

> The dearth of complete descriptions obfuscates the interpretation of the body of modified Angoff standard-setting literature. If the studies were described more carefully and thoroughly, patterns of interactions among the variations in methods might be discernible. As the research stands now, these patterns cannot be seen.

(Brandon, 2004, pp. 79–80)

The third flaw is methodological. Many of the findings reported in the empirical standard-setting research are from simulations in which only some of the standard-setting steps have been conducted. Research on the method that omits some of the modified Angoff steps is flawed because it does not examine all the key aspects of standard-setting; such research is akin to studying performance assessments in which students are not given instructions for conducting the assessments. Because of the omission of key steps, the findings of some studies are less generalizable than desirable to the fully implemented modified Angoff method.

The primary effect of these three flaws is that we do not have a full understanding of all of the steps of the modified Angoff method. There are not enough empirical studies to adequately examine all facets of the method, too many of the empirical studies that have been published do not explain how they conducted the steps or else do not conduct some of the steps, and too many studies are analog studies. These flaws, combined with the findings about difficulties in knowing "proficiency for what" and the findings about the difficulty in making estimates for the hardest and for the easiest items, lead me to conclude that it is questionable whether

modified-Angoff cutscores are uniformly valid for making summative, high-stakes decisions in program evaluations. Placing great weight on modified Angoff cutscores in high-stakes decisions, as occurs in K–12 education, might be more than their methodological foundation can bear, in part because some of the findings about the method are troubling and in part because the methods and reporting of many modified Angoff studies are flawed.

Evaluation Scenarios Appropriate for Developing and Using Cutscores

Program evaluators might correctly hesitate to use modified Angoff cutscores for high-stakes, summative purposes, but the findings on the validity of cutscores are not so troubling as to refrain from using them in all program evaluations. Evaluators can use them to help interpret student scores for formative-evaluation purposes or to help interpret scores for *suggesting* summative program-evaluation decisions. Cutscores do not have to be interpreted as definitive demarcations of success; "gray areas" about the cutscores can be calculated using the standard error of the mean, resulting in cutbands instead of *cutscores*. This calculation would show a band around the cutscore that would provide an accommodation to the inexactitude of standard setting. Using standard errors in this way, evaluators would have three score bands—one for students who we could reasonably state are below the desired level of performance, one for those who are more or less at the desired level of performance, and one for those who are clearly above the desired level of performance. Using this analysis, evaluators could report with a reasonable level of assurance the percentages of student scores above and below proficiency. Such descriptive reports could help evaluators understand how well programs are helping students achieve program goals without placing undue emphasis on the

cutscore itself. The reports could provide program personnel with general guidance about their programs. Formative evaluation findings and findings that are only *suggestive* of summative conclusions are not used to make go/no-go decisions about programs. When cutscores are used in ways such as these, their precision and validity are less critical than when they are used for making conclusive summative decisions about students or schools.

However, because of the limitations in the research and because of concerns about invalidity, I conclude that the modified Angoff method should be used primarily when other approaches are unavailable for interpreting student scores. That is, cutscores should be developed and used only with some kinds of evaluation designs and only in some evaluation stages. Evaluators should consider using test cutscores to help interpret test or assessment program outcome scores when no comparison or control groups are available. This scenario occurs when educational programs are implemented at all program sites, when administrators and faculty at non-program sites are unwilling to let evaluators use their sites for comparison or control groups, or, in the evaluations of small programs, when evaluation funding is too limited to have comparison or control groups. Cutscores developed when no comparison or control groups are available could help evaluators decide the extent to which children are performing at or near the desired level of performance. Cutscores might particularly be useful during the first year of an evaluation, when no year-to-year effect sizes can be calculated. Effect sizes showing annual growth are valuable for year-to-year comparisons, because they can be compared with published effect sizes about similar programs studies (Lipsey, 1990; Lynch, 1987), and because they probably are more defensible than cutscores. The two analyses together might also be useful, of course; cutscores used over several years of an evaluation can interpret how high or low program students are performing,

irrespective of the size of year-to-year effect sizes.

As long as they are interpreted with caution, cutscores might also be helpful even when comparison groups are used. They can help interpret mean scores when the differences between program and comparison groups are not statistically significant. Comparing average scores to a cutscore could help evaluators know the general levels of performance of both the program and comparison groups. Furthermore, using cutscores could help evaluators tie the interpretation of evaluation results directly to program goals. If a program's goal is, say, to have students achieve proficiency in reading knowledge or skills, evaluators could use cutscores to show the extent to which the proficiency goal had been achieved. The same kind of analysis could be conducted for other levels of student performance. Such reports are rhetorically more powerful than simply reporting whether the program group out-achieved a comparison group or surpassed a specified percentile of a norm group, because comparisons of average scores with cutscores tie evaluation results directly to descriptions of desired levels of student performance.

Criteria for Using Standard Setting in Program Evaluations

There are at least four criteria that should be addressed if evaluators use the modified Angoff method in program evaluations:

1. Standards should be set for reliable and valid tests.
2. The program for which standards are to be set should be well defined with concrete objectives that clearly show what is expected of program recipients upon completion.
3. The standard-setting judges should understand the program objectives well,

know the socioeconomic and educational context of the program, and understand the context in which program recipients will study or work after completing the program.

4. The standard setting should be feasible. The standard-setting method should not require more time and resources than the program can afford.

The necessity of the first condition should go without saying; cutscores cannot be used validly to make decisions about program success unless the test for which they are set adequately measures subject matter and produces sufficiently precise scores to make decisions about programs. The other three conditions, however, need some elaboration.

Well-defined programs. When using standard setting in program evaluations, the programs should have clear sets of concrete objectives. Clear objectives are necessary if well-defined and well-described performance standards are to be developed. Although the empirical literature on setting performance standards is not extensive, a small body of studies strongly suggests that performance standards must be thoroughly described and well understood by judges if cutscores are to be valid. Indeed, it is commonsensical that performance standards must be thoroughly explicated, because judges need to understand what students must be proficient *for*.

The "proficiency for what" issue need not be as deleterious in program evaluation standard setting as it is in K–12 accountability standard setting. K–12 public education provides a wide smorgasbord of educational services to all children. In contrast, many educational programs provide narrow, well-defined services to clearly-demarcated populations. Educational programs typically address a single subject such as reading or science or a narrow topic such as safety, drugs abuse, and so forth. Programs are designed for a single grade level or perhaps two or three

grades. They often serve subgroups of students with well-described demographic characteristics. If programs are well-designed, it is likely that their objectives will be clear and the goals more clearly defined the goals typically addressed in K–12 standard setting (i.e., advancing students to the next grade or graduating them from high school). Furthermore, judges in program evaluation standard setting can consider the social and demographic context of the schools that a program serves. Programs often serve smaller populations than entire districts. Judges can define performance standards and set cutscores while keeping in mind the population that the program serves, the wealth and the physical condition of the schools that are served, the typical longevity of teachers serving in the district, and other district demographics that evaluators can gather for judges to consider.

Judges who know the program and its context. Standard-setting judges are more likely to have reasonable expectations about student outcomes in a program if they are intimate with the program's history, aspirations, administration, line personnel, operations, and so forth. The better they know a program, the more reasonable their expectations about program outcomes will be, and the more likely it will be that they will know the answers to a number of questions, Quoting Smith (1981, p. 266), these questions are

- Has what the program is trying to do ever been done before by anyone? (If not, do not expect too much.)
- Has it ever been done the way the program is trying to do it? (Reasonable expectations are lower for innovations.)
- Is the logic which explains why this program will achieve its desired ends compelling? (The stronger the logic, the more warranted high expectations are.)

- Does the scope of this effort, in terms of time and resources, match the level of effect expected? (Real change usually requires a lot of time and effort.)
- Do contextual factors suggest that this effort might be more or less successful than previous efforts? (Higher expectations are warranted if this program is free of previous contextual constraints.)

It certainly would not be impossible to provide standard-setting judges selected from outside the program with the answers to these questions, but the standard-setting training required to address the questions fully would be onerously lengthy and expensive.

Judges are more likely to develop reasonable expectations if they are familiar with the socioeconomic and educational contexts of a program. Programs in economically disadvantaged communities or in schools lacking good equipment and facilities are less likely to show acceptable levels of performance than are programs in less-disadvantaged communities. Judges should know these contexts because of their effects on student outcomes in the program. Judges can take socioeconomic status and school conditions into account when developing performance standards and setting cutscores. Keeping in mind the mix of schools of varying socioeconomic status and of facilities with varying degrees of maintenance will help ensure that judges' standards are well-informed and reasonable.

The need for familiarity with programs and their social and demographic contexts means that standard-setting judges should be program personnel such as developers or teachers. Others might be insufficiently familiar with the program. For example, parents might not understand program expectations. Also, outside educators such as university personnel might be insufficiently familiar with the

conditions of the schools in the program. Program evaluators who are not subject to political pressures can select judges on the basis of how well they know the program and understand the school context, including both the schools themselves and the community in which they reside. It is unlikely that evaluators will find qualified personnel of this sort outside of the program setting.

Having to hire program personnel might mean selecting judges who would be inclined to set lenient program performance standards and low cutscores. Judges might establish erroneously easy performance standards and cutscores because they are loyal to the program, do not wish to see it fail, or believe that they might be under pressure to be easy on the program. This is a source of bias that evaluators should consider when developing program standards. Judges should be trained to establish performance standards that reflect the intent of the program and to set cutscores at levels that match the performance standards.

A colleague and I had teachers serve as standard-setting judges for a state-developed writing assessment that we administered during an elementary-school writing program evaluation (Brandon & Higa, 1998). After pilot-testing the standard setting in another school, all seven fourth-grade teachers in the program school set standards for their students. The teachers addressed the question, “If you instructed your students last year as well as possible, what was the best they could have done?” They answered this question for each of five dimensions of writing—meaning, voice, design, clarity, and conventions (grammar, punctuation, and so forth).

The seven teachers were deemed the only appropriate group to develop standards because other groups had insufficient knowledge about students’ achievement and educational background, writing skills, and the context within which they were

taught. The school principal did not participate because he might not have known the capabilities of the cohort of assessed students sufficiently well to have set fair standards, and parents did not participate because they knew too little about content-area knowledge or skills or about program context to arrive at fair judgments.

We were concerned that the seven teachers' estimates of how well students could perform might be lenient because they would not want the effects of their instruction to look poor. To address this concern, we examined the differences between the mean estimates for each of the five writing dimensions and the actual performance of students for which the standards were set (Brandon & Higa, 1998). If the cutscores that the teachers set had been far below student averages, it would have suggested that inappropriate methods were used or that teachers had a self-serving bias. The differences between the cutscores and the performance of the program students showed, however, that the cutscores were somewhat above students' performance, suggesting that teachers did not show a self-serving bias. Furthermore, the cutscores were not so high as to suggest inappropriate expectations. These results helped rule out claims of invalid standards.

Feasibility. Program evaluations must be feasible (Joint Committee on Standards for Educational Evaluation, 1994). Sufficient time and resources are necessary for program evaluation standard setting because good standard setting can be a labor-intensive, lengthy activity. Evaluation theoreticians and methodologists often overlook feasibility issues, but these must be addressed if practitioners are to use the methods.

In standard setting, both the development of the description of the performance standard and the setting of cutscores require sufficient time and resources.

Developing performance standards for a moderately long single-subject test can take half a day (Mills, Melican, & Ahluwalia, 1991; Livingston & Zieky, 1989). Furthermore, setting cutscores is clearly not a brief task, as should be apparent from the description presented earlier of the steps of the modified Angoff method. In modified Angoff standard setting, judges review items, make initial estimates, review empirical information about the items, hold discussions about their initial estimates, revise their estimates, and perhaps repeat the review/discussion/estimation activities for another iteration. These activities can easily last for a full day; in some instances, such as standard setting for the National Assessment of Educational progress, they take two days or more.

When setting standards for the elementary-school writing program (Brandon & Higa, 1998), we eliminated the step of having teachers prepare written descriptions of performance standards; instead, we asked them to estimate the best performance that they reasonably thought children could achieve. We eliminated the step because the rating-scale rubrics described the target level of performance for each rating-scale point. Teachers knew the rubrics well because they had used them to score student papers; they were asked to use the rubrics to substitute for performance standards. When trained in the standard-setting procedures, they simply had to review some of the materials that they had used when doing the assessments. This efficiency contributed to the feasibility of the standard setting. The standard setting method was implemented in a reasonable period of time (less than half a day). The teachers' comments, made during and immediately following the standard setting, suggested that they understood and fully used the standard-setting methods. Some teachers commented that they were unsure about the percentages to estimate for the scale points, but none resisted participation. None of the comments suggested that teachers found it difficult to apply knowledge of

the assessment to the standard-setting task.

Summary and Conclusions

Standard setting, which is widely used by school districts and states to hold students and schools accountable for their educational performance, has not been widely used by program evaluators as a means for helping decide whether a program has performed sufficiently well. Furthermore, the topic has been covered minimally in the program evaluation literature. This is unfortunate, because evaluators could use cutscores to help interpret program outcomes during the first year of an evaluation in which there are no comparison groups. They might even be useful when comparison groups are used, for they help show how high program and comparison groups are performing, irrespective of which group is performing the best.

Standard-setting consists of establishing performance standards, which are statements describing the knowledge and skills that students must attain if they are to perform at a specified performance level (basic, proficient, advanced, and so forth), and it consists of setting cutscores. The modified Angoff method is the most widely studied standard-setting method. As used in the test and assessment standard setting that schools, districts, and states conduct for accountability purposes, the modified Angoff method has three steps. Very little research has been conducted on the first step, which is to select and train the panels of judges who establish performance standards and set cutscores. Other than showing that 10–20 is an adequate range of the number of standard-setting judges, the empirical research literature is of little assistance in identifying the best mix of procedures for this step.

More research has been conducted on the second step, which is to define and

describe the performance standard (i.e., the statements describing the level of knowledge and skills that students should attain). The findings are inconclusive but commonsensically suggest that the better that performance standards are defined and explicated, the more valid cutscores are likely to be. Performance standards for educational accountability purposes are murky by nature, however, because it is impossible to know what comprises an adequate level of performance. If a performance standard is defined for graduation, should it be set for students who are going to trade schools, community colleges, state colleges, or private elite universities? What should the performance standard be for students who do not participate in any post-secondary education? If a performance standard for a particular school subject is defined for an elementary- or middle-school grade, what is the developmental or pedagogical basis for deciding what constitutes adequate performance? These questions have not been adequately addressed in the literature, and because of the epistemological complexity of the topic, are unlikely ever to be.

More research has been conducted on the third step of the modified Angoff method than on the other two steps. In this step, judges set estimates of the percentages of students who should pass each item at the level of the performance standard. During this step, judges are given empirical item *p*-values so that they know the difficulty levels of the items they are judging. The empirical research suggests that judges' discussions make a difference, but the research is not conclusive. Probably the most conclusive research about the third step has to do with the accuracy of item estimates, which is established by examining the absolute value of the differences between judges' item estimates and item *p*-values. This research suggests that judges tend to underestimate the difficulty of hard items and overestimate the difficulty of easy items. That is, the range of judges' item

estimates is less than the range of empirical *p*-values.

The research on the three steps of the modified Angoff method has not been conclusive in part because (a) the literature is more narrow than desirable, (b) some of the literature is not reported fully, and (c) the methods of the research have been of low quality. Because of problems with the methods and findings of the empirical research on standard setting, as exemplified by the research on the modified Angoff method—the most-studied of all test and assessment standard-setting methods—it might be concluded that program evaluators should avoid using the method to help make judgments about program success. However, the methods are not so unsound as to preclude their use for formative program evaluation purposes or for making *suggestive* (rather than conclusive) summative evaluation decisions. If cutscores are interpreted with caution and are considered to be suggestive of the success (or lack thereof) of a program, they can help evaluators make conclusions in evaluations that lack comparison groups.

Even though the empirical test and assessment standard-setting literature does not provide convincing evidence about the strength of standard-setting methods, it nevertheless is sufficiently thorough to help us know the conditions that should be present if evaluators use the method in program evaluations. There are at least four of these conditions. The first is that standards should be set only for valid and reliable tests. Evaluators are best advised to set standards for commercially published tests or assessments or for other carefully crafted instruments. Second, cutscores should be set only if program objectives are clearly stated. Otherwise, performance standards will be difficult to develop. Third, judges should be familiar with the program and the context within which it is taught. The task of setting performance standards for a program is conceptually less complex than the task of setting standards for a school district, because programs (at least those that well-

developed and well-run) have clear sets of methods and objectives that standard-setting judges can keep in mind when setting cutscores. This assumes that the judges know the program well and eliminates the possibility of having people outside the program serve as judges. Of course, the charge might be made that program faculty, developers, or administrators who serve as standard-setting judges might set lenient standards. However, in a trial application of standard setting in a program evaluation, it was shown that this need not be the case (Brandon & Higa, 1998). The fourth condition is that the standard setting should be feasible. Evaluators should not assume that they can set standards without proper preparation and full understanding of the mechanics and theory of the procedures. In our trial application of standard setting in a program evaluation (Brandon & Higa, 1998), we showed that it was feasible in a small school-level evaluation.

This article shows that standard setting methods have value in evaluations. They can help evaluators make decisions about program success in the first year of an evaluation that has no comparison groups. In this scenario, other means for deciding about program success are unavailable; therefore, standard setting helps address an empty slot in evaluators' methodological toolbox. The fact that there are weaknesses in the argument for using methods such as the modified Angoff method to make high-stakes decisions need not deter evaluators from using the method during programs' early years, when summative decisions are infrequent. Standard-setting methods also can help evaluators make decisions about program success in later years of evaluations that do have comparison groups. In this scenario, cutscores can help determine the extent to which both the program group and the comparison group have achieved at sufficiently high levels. In both these scenarios, cutscores should not be interpreted rigidly; they should be used to arrive at *suggestions* about program success. This use of cutscores helps make up for the

procedural weaknesses of the method. As long as (a) cutscores are set for valid and reliable tests, (b) program objectives are clear, (c) program personnel serve as standard-setting judges, and (d) there are sufficient resources to conduct the standard setting well, standard setting can contribute to evaluators' decisions.

As stated at the beginning of this article, standard-setting is a means of answering the question, How good is good enough? The conclusions about standard setting given in this article can serve as suggestions about other methods for addressing the question in evaluation studies. First, the stage of the evaluation should be considered. In the case of developing cutscores in program evaluations, the argument for using standard setting to help make evaluation decisions is the strongest in the first year of an evaluation. Other methods for deciding the quality of a program are appropriate in other phases. By way of contrast, experimental and quasi-experimental methods are appropriate when programs are mature. Second, the method for answering the question depends on the use of evaluation findings. Standard-setting methods used for deciding about program success need not be free of flaws when the decisions are formative or when the findings are used to make suggestions, as opposed to conclusive statements, about program success. Experimental and quasi-experimental approaches to evaluation are appropriate for providing conclusive findings about the quality and effectiveness of a program. Third, the context of the program should be taken into account (Smith, 1999). Evaluators using standard setting methods need to find judges who understand the context of the program, or else cutscores will not be well-informed. The importance of knowledge about context applies to all discussions about how good is good enough. Fourth, the method for answering the question must be feasible. It will not do to require, for example, that all studies use experimental or quasi-experimental designs when the setting or the resources of the evaluation do not

allow them. The current push by federal educational research funding agencies to require these designs ignores the feasibility issue—particularly since these same officials do not back up their call for experimental and quasi-experimental designs with funding for expensive evaluations. These four aspects of evaluation should be considered when developing a minimal set of guidelines that evaluators should take into account when establishing the level of performance that a program should show if it is to be considered good enough.

References

Angoff, W. H. (1971). Scales, norms and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508–600). Washington, DC: American Council on Education.

Brandon, P. R. (2002). Two versions of the contrasting-groups standard-setting method: A review. *Measurement and Evaluation in Counseling and Development, 35*, 167–181.

Brandon, P. R. (2004). Conclusions about frequently studied modified Angoff standard-setting topics. *Applied Measurement in Education, 17*, 59–88.

Brandon, P. R., and Higa, T. F. (1998, April). *Setting standards to use when judging program performance in stakeholder-assisted evaluations of small educational programs.* Paper presented at the meeting of the American Educational Research Association, San Diego, CA.

Burton, N. W. (1978). Societal standards. *Journal of Educational Measurement, 15*, 263–271.

Camilli, G., Cizek, G. J., & Lugg, C. A. (2001). Psychometric theory and the

validation of performance standards: History and future perspectives. In G. C. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 445–475). Mahwah, NJ: Lawrence Erlbaum.

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

Cook, T. D.; Leviton, L. C., & Shadish Jr., W. R. (1985). Program evaluation. In G. Lindzey and E. Aronson, *Handbook of social psychology* (3rd ed.). New York: Random House.

Fink, A. Kosecoff, J., & Brook, R. H. (1986). Setting standards of performance for program evaluations: The case of the teaching hospital general medicine group practice program. *Evaluation and Program Planning, 9*, 143–151.

Hanser, L. M. (1998). Lessons for the National Assessment of Educational Progress from military standard setting. *Applied Measurement in Education, 11*, 81–95.

Henry, G. T., McTaggart, M. J., & McMillan, J. H. (1992). Establishing benchmarks for outcome indicators: A statistical approach to developing performance standards. *Evaluation Review, 16*, 131–150.

Hurtz, G. M., & Auerbach, M. A. (2003). A meta-analysis of the effects of modifications to the Angoff method on cutoff scores and judgment consensus. *Educational and Psychological Measurement, 63*, 584–601.

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

Joint Committee on Standards for Educational Evaluation. (1994). *The program evaluation standards* (2nd ed.). Newbury Park, CA: Sage.

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

Lipsey, M. W. (1990). *Design sensitivity: Statistical power for experimental research.* Newbury Park, CA: Sage.

Livingston, S. A. & Zieky, M. J. (1989). A comparative study of standard-setting methods. *Applied Measurement in Education*, *2*, 121–141.

Lynch, K. B. (1987). The size of education effects: An analysis of programs reviewed by the Joint Dissemination Review panel. *Educational Evaluation and Policy Analysis*, *9,* 55–61.

Mills, C. N., Melican, G. J., & Ahluwalia, N. T. (1991). Defining minimal competence. *Educational Measurement: Issues and Practice, 10*(2):7–10.

Patton, M. Q. (1997) *Utilization-focused evaluation: The new century text*. 3rd ed. Newbury Park, CA: Sage.

Rossi, P. H., & Freeman, H. E. (1993). *Evaluation: A systematic approach* (5th ed.). Newbury Park, CA: Sage.

Shadish, W. R., Cook, T. D., & Leviton, L. C. (1991) *Foundations of program evaluation: Theories of practice*. Newbury Park, CA: Sage.

Shepard, L. A. (1995). Implications for standard setting of the National Academy

of Education Evaluation of the National Assessment of Educational Progress Achievement Levels. In *Joint conference on standard setting for large-scale assessments. Vol.2. Proceedings* (pp. 143–160). Washington, DC: U.S. Government Printing Office.

Smith, N. L. (1981). Constructing reasonable expectations in evaluation. *Evaluation News, 2,* 265–267.

Smith, N. L. (1999). A framework for characterizing the practice of evaluation, with application to empowerment evaluation. *Canadian Journal of Program Evaluation, Special Issue,* 39–68.

Wholey, J. S. (1979). *Evaluation: Promise and performance*. Washington, DC: Urban Institute.

Worthen, B. R., Sanders, J. R., & Fitzpatrick, J. L. (1997). *Program evaluation: Alternative approaches and practical guideline* (2nd ed.). New York: Longman.