

Standard Setters: Stand Up and Take a Stand!

Barbara S. Plake, Ph.D.

Psychometric Inquiries

2006 Career Award Address presented at the annual meeting of the National Council on
Measurement in Education, Chicago, IL, April 2007

Standard Setters: Stand Up and Take a Stand!

I have been doing standard setting studies for over 20 years. In these studies, I have worked with licensing and certification agencies in setting passing scores for various professions from Well Drilling and Pump Installers, Financial Analysts, Medical Technicians, Licensed Practical Nurses, to Poison Control Center Operators (and others). I've helped school districts set the cutscores for classifying students into performance levels categories for NCLB compliance, and worked with a province in Canada to set the passing score on the high school graduation literacy test. Across these settings I have used various judgmental methodologies with assessments that had only multiple-choice questions to ones that were a complex mixture of constructed response tasks.

Based on these experiences, I have encountered a number of decision points in planning and executing standard setting studies for which there appears to be differing opinions, but little research to inform or support the decisions. The purpose of my presentation is to point out these unresolved issues and controversies in standard setting and propose a research program to provide empirically-based information to inform the decision making process. I titled this presentation "Standard Setters: Stand Up and Take a Stand" because in my interactions with policy bodies, I have been frustrated by uninformed, but good intentioned people who want to influence the design and methodology for the standard setting study. In these situations, I wished that I had more data-based information to help guide the discussions on the design and implementation of these standard setting strategies. With more data-based information, professionals who are planning and conducting a standard setting study, who I am calling the "standard

setters”, would be better able to “stand up and take a stand” by offering research based answers to these issues in planning and implementing a standard setting study.

There are typically three principal groups involved in a standard setting study. The first is the agency or organization that decides that a standard setting study is needed. Often this group is composed of policy makers and stakeholders. The second is the group that works with the agency or organization to plan, conduct, and report the results of the standard setting study. This group is often a contractor who may be part of the test development company who has the contract to develop the test or it may be an independent contractor. For legal and political reasons it is desirable to have the group who conducts the standard setting study be independent of the agency or policy body who makes the final decisions about the cutscores based, in part, on the results of the standard setting study. A third group in the standard setting study is the participants whose judgments are gathered during the standard setting study. Although many of the issues and unresolved questions identified in my presentation are applicable to both judgmental (test-based) and empirical (examinee-based) standard setting processes, my presentation will focus on judgmental standard setting methods.

My presentation is organized around sequential events that need to happen in the design and implementation of a standard setting study. These steps are not all linearly related because some happen simultaneously, some influence the options for later steps, and some may even be independent of prior steps. But in total, these steps and their related decisions must occur in order to plan and carry out a standard setting study. For a full listing of the steps needed to conduct a standard setting study see Hambleton and Pitoniak’s chapter in the fourth edition of Educational Measurement, pages 436-437. The

steps that I identify below do not line up perfectly with those by Hambleton and Pitoniak because I am focusing on steps for which there are unresolved research issues.

A first step, determining the purpose of the standard setting study, may at first appear to be strictly a policy decision. In many instances this step may seem to be a fairly obvious and easy decision: the test scores will be used to make decisions about who will be licensed, certified, given special privilege (such as an scholarship), or placed into one of several achievement classifications. However, even in these fairly direct applications, accompanying issues and questions that have psychometric implications need to be resolved. For example, if the decision is about high-stakes situations such as licensure, certification or graduation, it must be decided about how many retake opportunities candidates who fail will have, if any. Is there a minimum period of time that must pass between administrations? If the test has multiple components, each with its own passing score, can candidates “bank” the scores on the components they pass and only concentrate on the remaining sections? What is the period of time that “banked” passing scores will be considered valid? Should the passing score be adjusted over time for retake candidates because the compilation of standard errors may, in the long run, produce more false positive decisions? In situations where a student needs to pass one (or more) examinations to qualify for promotion or high school graduation, what will be the plans for offering remediation for failing students?

There is limited research and some legal case law that can inform these decisions. With regard to retake policies for licensure and certification examinations, some programs are lenient in their retake policies, allowing multiple, and even immediate, retakes. These policies can be subject to abuse and candidates can expose an item pool or

jeopardize test security through over-exposure and theft of test questions. Another concern is that the probability of making false positive decisions will increase across multiple testing opportunities (Millman, 1989; Phillips, 2001). Testing programs and agencies need to be cautious and thoughtful as they make decisions regarding their retake policies and the consequences of those decisions.

A related early decision in planning for a standard setting study has both policy and psychometric implications. Some programs want to ensure that candidates have sufficient knowledge and skills across multiple component areas. They want to accomplish this by setting a unique cutscore for each of these components, thus requiring candidates to achieve a passing score on each component. This has great appeal to many policy makers, especially if the policy board includes specialists in these component areas. Often, the mind set is that it is essential for the candidates show sufficient expertise in each of the specialists' areas. This may be construed, in some cases as a "protection of territory": "We can't possibly license (or certify) a candidate who doesn't show adequate levels of proficiency in MY area, because it is essential to good practice".

These concerns about ensuring that successful candidates have the necessary knowledge, skills, and abilities are sincere and well intentioned. However, many of these policy makers (often practitioners) do not understand two outcomes of such policy decisions: a) the probability of passing all components and thus achieving an overall passing score reduces dramatically as the number of components needed to pass increases (this is true for virtually all conjunctive decision making models), and b) the technical quality of the component examinations may not support making these kinds of high-stakes decisions. Often, instead of these conjunctive decision models, only a

compensatory model can be supported by the levels of psychometric quality of the examination.

One extreme example of a conjunctive decision making model that illustrates the point is when policy makers want to implement “drop dead” decisions on individual questions for passing the examination. I remember working with a chiropractic licensure examination committee that wanted, as part of the examination process, candidates to read X-ray films to diagnose patient conditions. They wanted to implement a decision rule that the candidate would fail automatically if the candidate did not recognize the severity of the condition shown in one of the X-ray films. Much discussion ensued as I tried to inform them about the psychometric issues related to making a pass/fail decision based on only one (albeit crucial) data point. We ended up adding multiple questions based on that particular film to ensure that the candidate adequately addressed the diagnosis and treatment options. Although some of these questions about the use of conjunctive and compensatory models can be answered by theory, others could be informed from research on the actual impact of these policies on candidate outcomes.

Following decisions about the purpose of the standard setting process and related policies around retakes and “banking” of scores and a consideration about what decision models to use (conjunctive, compensatory, or some combination), standard setters (i.e., the persons responsible for designing and implementing the standard setting study) next need to work with the agency/organization in identifying the qualifications of persons who will serve on the standard setting panel. In my experience this typically has not been too much of an issue when working with licensure and certification agencies. They are generally amenable to the composition of the panel with professionals who represent

important constituencies in the organization. Often, areas of expertise are relevant, as well as representation geographically and demographically. Years of experience and exposure to entry-level candidates are also considered important in identifying the composition of the panelists. However, other concerns have been raised when working with state education agencies when the goal is to make decisions about student performance. Based on my recent experience, the breadth of background and experience of panelists has been increased as other stakeholders for the educational enterprise have been considered for membership on the standard setting panels. In Kane's (2001) chapter entitled, *So much remains the same: Conception and status of validation in setting standards*, he states, "Given the wide-ranging impact of the policy decisions involved in setting standards for high-stakes tests, it is important to have broad representation for groups with an interest in the stringency of the standard...It is desirable to include representatives from as many stakeholder groups as possible, even though some of these groups may not have great technical expertise (p. 65)". I can understand the interest and need to involve multiple stakeholders in the decisions about student outcomes, especially for decisions such as high school graduation. Involving stakeholder groups outside of the educational system was a key component of the Jaeger method (1982) and is currently in practice by policy for setting performance level cutscores with the National Assessment of Educational Progress (Loomis and Bourque, 2001). However, I believe that the standard setting panel is not the place to satisfy this policy issue. As Hambleton and Pitoniak (2006) state, "Seeking broad representation of constituency groups when assembling a standard-setting panel provides political advantages. Knowledge about the content is, however, a more important criterion in panel selection since standard setting

methods often require complex judgments and insights...(p. 451)". Standard setting panels are often asked to envision students who are just at the borderline for the requisite achievement level. Next, they are asked to identify the kinds of performances on the test that would be consistent with that of students whose performance is at borderline for that achievement level. This is not an easy task and, in fact, some standard setting methods (principally the Angoff, 1971 method) have been criticized as being "fundamentally flawed" due to the cognitive task that is placed on panelists when making these decisions about the probability of borderline students answering the items correctly (Pellegrino, Jones, & Mitchell, 1999). It should be noted that there is disagreement in the field about the validity of this criticism (Hambleton, Brennan, Brown, Dodd, Forsyth, Mehrens, et al, 2000). Even though the training in most standard setting studies involves a discussion of the skills of students who are at the borderline, and often panelists are given experiences with the examination, this seems to me to be an inadequate preparation for the tasks to be undertaken if the panelist is not well aware, and deeply engaged, in the educational experiences of the students. Jaeger (1989) notes that "...very little is known about the sensitivity of judges' conceptions of a minimally competent individual to the procedures used to train judges in formulating these conceptions, or to the standard-setting procedures through which judges must apply those conceptions (p. 493)". I have reviewed the reports from standard setting studies where stakeholder from outside the school system served as panelists. In many cases, the stakeholders report that they did not have a solid understanding of the target students and were challenged to understand the tasks presented to the students. Other panelists complain that these stakeholders took too much time trying to get the needed perspectives or simply were not able to participate

effectively in the standard setting workshop. Some of these stakeholders had personal biases that influenced their perspectives of the levels of outcomes expected by the target level student. In some cases these may be business representatives who felt that higher levels of performance were needed to be successful in their business environment (never mind that the purpose of the test was not to ensure that the students had the levels of achievement needed to be successful in a business environment). In other cases, these may be parents who have students with special needs and these parents were inclined to be protective of their children's eligibility to earn a diploma. My point is that the tasks that are required of the panelists are non-trivial and should not be assigned to persons who do not have the required levels of information to make meaningful and valid ratings. It is not sufficient, in my view, to give these people a brief orientation to the curriculum, student performance expectations, and the test and expect them to make valid estimates of how students who are just qualified to pass (for example) will perform on the test. The consequences of involving these unqualified people in the standard setting panel are severe: At best, their invalid responses can be removed from the data set before computing the cutscore(s); at worst, these unqualified panelists can delay, distract, and even derail the whole process. I don't mean to imply that these important stakeholders don't have a role in the policy decision about setting the passing scores on these high stakes tests. Instead, I believe these stakeholders should have other opportunities to inform the final decision about the passing score. Central to my argument is that the final decision about the passing score(s) is a policy decision that is informed by the results of the standard setting study. Research that examines the effect of including stakeholders on

the standard setting panel would help inform standard setters about the possible impact of these policies.

Another issue that is sometimes raised in conducting a standard setting study relates to the basis for evaluating examinee performance. In some applications, panelists are asked to evaluate how target examinees **WOULD** perform on the test questions or tasks. Essentially, panelists are trained to understand the target (borderline) examinee's knowledge, skills, and abilities, and then to envision how these examinees, in the test situation, would actually perform of the questions on the test. This is consistent with the underlying assumptions with most standard setting methods (the Jaeger, 1982, method is an exception). However, in some applications, panelists are asked to rate instead how these target examinees **SHOULD** perform on the test. For example, in conducting the standard setting study for the Virginia Standards of Learning Assessment, Harcourt Educational Measurement identified in their technical report that they used a modified Angoff procedure, but emphasized that the question posed to the panelists was how **SHOULD** the minimally competent candidates perform on the test questions. This is a very different expectation. The word "should" implies that they will perform without error, without any levels of test anxiety, with perfect exposure to the curriculum and test taking strategies. **SHOULD** is a very dangerous level of expectation and will most likely set passing scores that are unrealistically high. While some believe that the right decision is to have panelists make "Would" not "Should" decisions, there is not a body of literature that has examined the impact of these different approaches.

Once the panelists have been identified, the basis for the ratings decided (would instead of should), another issue involves the training of the panelists. It seems

fundamental to me that if the panelists are going to be asked to envision examinees with a certain level of knowledge, skill, achievement, and ability, then the panelists must have a deep understanding of the educational experiences and performance expectations of the examinees. This is not easily learned in a 3-hour training session (that is devoted to many more activities than merely coming to a conceptualization of the target examinees' abilities). I believe that it is critically important that these panelists have first hand knowledge and experience with examinees, including those whose skills are at the target borderline. If the panel is composed of identifiable subdivisions (such as those who provide educational experiences for candidates as opposed to those who supervise entry level candidates), it is possible to do secondary analyses to examine whether these subgroups are providing different cutscores. Plake and Impara (2000) conducted such secondary analyses to help inform the policy makers if different cutscores would have resulted from these differing panel groupings. Further, Plake and Impara (1994) found that panelists with limited first hand experience with the examinees in their area of content expertise were more influenced by the feedback data and reported less confidence in their ratings. It is important to understand what level of exposure and experience with the target candidates is required to successfully complete the rating tasks expected of the panelists.

In addition, in several judgmental standard setting methods, panelists are asked to evaluate how the target level examinees will likely perform on the tasks in the test. This is sometimes achieved by having the panelists take the examination under pseudo test conditions. In some cases, this may be appropriate, especially if the panelists have limited prior exposure to the test. However, in other situations, this may not be a good

use of the training time. In school situations, where the panelists are often teachers of the students who take the test, these teachers are often fully aware of the test and the skills and knowledge needed to complete the test. Sometimes, having the teachers take the test targeted for their students is interpreted as an insult, or at best, a waste of their time. Psychometricians in charge of planning a standard setting study should evaluate the circumstances surrounding the examination to decide whether this activity is needed and justified in the specific standard setting activity. Automatically deciding to devote hours of training time to an unnecessary activity does not, in my mind, make good sense. In some situations, these tests are administered over multiple days consuming several hours of student test time. To administer the examination under “real administrative conditions” does not seem reasonable in circumstances where the panelists are fully aware of, and closely engaged in, the administration of the test to their students. In contrast, in many licensure and certification standard setting studies, it may be that too little time is devoted to helping the panelists understand the complexities of the test. Judgments need to be made by the standard setters about what is the appropriate amount of orientation that is needed to ensure that the panelists can successfully complete the tasks of the standard setting process: being able to understand the complexities of the testing process in the eyes of the examinees for whom they are to make performance predictions. It is a research question about what degree of exposure is sufficient, and it is likely dependent on the purposes of the test (licensure/certification or assignment to performance level categories), the standard setting method used, and the training of the panelists during the standard setting study.

Another issue is the extent to which the examination used in the standard setting study needs to be the full examination or whether a subset of the examination can serve as a proxy for the full examination. Some research has addressed this topic. Using a secondary analysis from standard setting studies, Ferdous and Plake (2005b, In press) examined how close the cutscores would be if the panelists' ratings were for a mini-test that was designed to mirror the full test in content and psychometric properties. The results of these studies were quite promising; cutscores (when put of the scale for the full test) derived from the mini-tests were often within a score point from the cutscores for the full examination. However, this research was conducted using ratings that were made by the panelists when they were rating the full examination. It is unknown whether their ratings on the mini-subtest would remain stable when these items were rated alone instead of in the context of the full examination. Research is needed to address this issue. In these studies, both an Angoff and a Yes/No (Impara and Plake, 1997) standard setting methods were used. It may be that these results were also dependent on the methods used to gather the panelists' ratings and the length of these examinations. If it can be shown that ratings for a specially constructed subtest can provide equivalent results as those for the full test, savings in time, resources, and panelists' energies could be realized.

Research could also inform which components of the standard setting process are most salient in providing sound ratings from the panelists. Most standard setting procedures carefully follow a series of steps, from orientation of the panelists to the task, a presentation of performance level definitions that are used for classification into performance categories (including the performance level descriptors), a discussion of the knowledge, skills, abilities, and achievements of examinees whose skill levels are just at

the borderline for classification into the respective categories, and practice on the method that will be applied, including experience with the kinds of feedback that will be provided to the panelists between rounds of ratings. It is believed that all of these components are critically important to the success of the standard setting procedure and often serve as the foundation for procedural validity evidence to support the quality, integrity, and defensibility of the results (Kane, 2001). I am among the “true believers” that all of these components are essential, but there is a limited empirical literature base to support these perceptions. Researchers (Ferdous & Plake, 2005a; Giraud, Impara, & Plake, 2005, Skorupski & Hambleton, 2005) have studied the salience of these and other factors (such as whether the panelists set high or low round 1 ratings for student performance) on panelists’ final cutscore values. This research is only a modest beginning in a program of research to help understand the importance of the various components of the orientation and training process.

Once the panelists have engaged in the training process (which should include practice on items similar to the ones in the test using the rating forms that will be used for the operational test questions, as well as receiving experience with the kinds of information they will be provided between rounds of ratings), they are ready to begin providing their ratings for the operational test. When the standard setting method involves providing item performance estimates (such as the Angoff method), some decisions need to be made about how precise these estimates can reasonably be made by the panelists. Using the Angoff method, panelists are asked to consider a randomly selected, minimally competent candidate, and to estimate the probability that this candidate would answer the item correctly. This means that these panelists are being

asked to provide estimates between 0.00 and 1.00. In practice panelists often do not use this full probability range for making their item performance estimates. Even when they are free to do so, panelists will often provide estimates that are multiples of 5, such as .55, .60, etc. In fact, most panelists are unlikely to give estimates lower than .20 (but they often provide estimates that exceed .90). One possibility would be to provide the panelists with these preselected probability values beginning with .20 in multiples of 5 as choices for their probability values. This would have another positive benefit, as it could reduce the data entry needs between rounds as there would be fewer decimal places to enter. If panelists were given ranges of probability values (such as 0 – 9, 10 – 19, etc.) the rating could be easily accommodated using machine-scorable answer sheets.

However, as Impara (2007) pointed out, differences can occur between the “intended” and actual cutscore that results from the use of these rating scales. One approach that has promise for gaining the benefits from using the abbreviated rating scales, but may rectify the bias resulting from their use, is to only use the abbreviated rating scales during the first round of ratings, switching to the full probability scale for the second (and final) round. This approach has been studied (Jaeger, 1989; Giraud and Plake, 1998) and depending on the algorithms used, can yield results close to what would have been obtained using the full probability scale. One rationale for using the full probability scale for the second round is that the feedback between rounds 1 and 2 typically gives panelists information on item difficulty (in terms of proportion of examinees who answered the items correctly). This information uses the full range from 0.00 to 1.00; therefore, reverting to the full probability scale for round 2 would be congruent with the type of information the panelists are provided before being asked to make their round 2 ratings.

In a study designed to address the effect of switching from the abbreviated probability ratings in the first round to using the full probability scale for round 2, Giraud and Plake (1998) divided a standard setting panel into two randomly equivalent groups. One group used the full probability scale for both rounds; the second panel used the abbreviated rating scales (0-9; 10-19, etc.) for the first round and the full probability scale for the second round. The results from the two panels were within one SEE of each other. Therefore, a strategy such as this could reduce the cognitive burden on panelists during the first round and also reduce data entry demands between rounds for the staff conducting the standard setting study. More research is needed to address the use of rating scales for gathering panelists' ratings.

A related decision regards the number of rounds of ratings that should be considered. Different opinions have been expressed about the effects of "staging" the feedback provided to panelists (Hambleton and Pitoniak, 2006). The kinds of feedback provided to panelists also vary, but are typically of two kinds. One kind gives panelists norm-based information about how their ratings compare to those of the other panelists. This can be given for the overall individual panelist's cutscores and/or at the item level. The second kind of feedback is based on examinee performance, showing either the proportion of examinees who answered the items correctly (p-values) and/or the proportion of candidates who would pass (or be classified into performance categories) based on the panelists' cutscore(s). Most standard setting practitioners support the need for at least two rounds, with feedback provided between rounds. Hambleton recommends providing panelists with both kinds of feedback between rounds so that panelists can make more informed decisions in subsequent rounds (Hambleton, 2001). Some

researchers advocate for more than two rounds, with or without sequential staging of the feedback (Reckase, 2001). The rationale for more than two rounds is to reduce the variability in panelists' cutscores present at round 2. Some researchers believe that withholding impact results will reduce the focus of the panelists on that information (Hambleton and Pitoniak, 2006). However, there is limited (if any) empirical evidence to support these claims. Typically, the change in the value of the cutscore even between rounds 1 and 2 is small, with the most notable change being in the lowering variability of the panelists' ratings. How much more reduction in variability that would result from conducting a third round is open to question. This is an area where research is strongly needed; to make decisions about the number of rounds needed and the anticipated effects of differing kinds of feedback between rounds is difficult without evidence to support these opinions. Having panelists engage in additional rounds of ratings requires additional time and commitment by the panelists and increases costs to the agency/organization funding the standard setting study. If there aren't sufficient improvements in the quality of the results, it then becomes a question of whether conducting these additional rounds results in an acceptable return on the agencies' and panelists' investments of time and resources.

Another issue that needs attention occurs when multiple performance levels are to be set. In some standard setting approaches, panelists are asked to sequentially consider the performance of borderline students on the same item across the performance categories. For example, panelists are asked to first estimate the performance on an item for students who are just at the Basic level; next the panelists are asked to turn their attention, keeping the item constant, to the student who is just at the Proficient level and

to estimate how that level student would likely perform on the item. Finally, the panelist is asked to consider the student whose skills are just at the Advanced level and make yet another item performance estimate for the same item for these borderline Advanced students. It is not clear how well panelists can, in fact, adjust their focus across the three performance categories in order to make these multiple performance estimates. Would it be better to ask the panelists to first focus on the borderline Basic student and make performance estimates for all of the items in the test and then sequentially turn their attention to the other performance categories, always making the full set of performance estimates for all items in the test before moving on to the next performance category? Or even have multiple panels, each panel setting only one of the cut scores. To my knowledge there is no research to help to make these decisions. Further, if the plan is to have the panelists keep the item constant as they move from one performance category to the next, does it matter whether the panelists start at the lowest performance category and move sequentially to the next higher ones? In some settings, one of the categories may be more important, from a policy perspective, than the others. Would it be better to have the panelists start with the Proficient category, for example, and then move to either the Basic or Advanced category second? Would the rating vary if either of these approaches were employed? How can decisions be made about these issues if we don't have a sound research base to support our decisions?

It is recommended that every standard setting be accompanied by an evaluation of the process to provide procedural validity evidence (Kane, 1994; Hambleton, 2001). Although Hambleton and Pitoniak (2006) have recently provided guidance on what topics should be included in these evaluations, little research has been conducted on the

effect of the wording of these evaluation questions on the outcomes. Evaluation and survey research tells us that how these questions are asked will affect the results.

Guidance is needed to help standard setters ask evaluation questions in order to report on the quality of the procedures implemented during the process.

This address has identified a number of issues that need consideration by the field in designing and implementing standard setting studies. In some instances, I present a position that is not uniformly embraced by the field; in others I have identified areas where research is needed to help make informed decisions when designing and implementing standard setting studies. My hope is that this presentation will start a conversation about unresolved issues in standard setting and help set a research agenda to address these issues.

In my address, these are some of the research areas identified for standard setting:

1. Under what circumstances does permitting retakes and the banking scores on examination components result in the most valid cut score(s)? What are the risks of allowing multiple retakes? Do score banking rules support the continued integrity of the assessment program?
2. Under what circumstances is it appropriate to have multiple cutscores that rely on a conjunctive decision model? How can agencies that want to implement single decision points for item level performance achieve this in a psychometrically defensible manner?
3. What are the effects of including stakeholders on the standard setting panel? What are other meaningful ways that such stakeholders can be engaged in the process of setting cutscores?

4. What is the impact of using “Should” instead of “Would” on the cutscores that result from a standard setting study? Does the impact vary as a function of the purposes of the standard setting process? Do the results differ by which standard setting method(s) are used?
5. To what extent do panelists need to have first hand knowledge of the examinees and the content of the test in order for them to successfully complete the rating tasks?
6. What is the effect of having panelists take the examination in pseudo-administrative conditions? Are there some settings where this activity is not beneficial and may actually distract the panelists in their participation and evaluations of the standard setting process? Under what circumstances do panelists need full administrative experiences in order to successfully conduct the tasks of the standard setting procedure?
7. What is the effect of having panelists rate a specially constructed subtest instead of focusing on the full examination? That is, can the same results be achieved with a subset of items rather than the entire test?
8. What are the salient features of the training process? How do the various elements/components of the standard setting process vary in their impact on the final cutscore(s)? How do these effects differ across standard setting methods?
9. What are the effects of using differing levels of precision in the ratings during the rounds of a standard setting study when using Angoff procedures?

10. For settings where there are multiple performance categories, what is the impact on cutscores when panelists focus on the performance categories first and make item performance estimates for all the items in the test as opposed to making sequential performance category ratings for each item before moving on the other questions comprising the test? What is the impact on cutscores if the sequence of considering performance level categories is varied?
11. What are the effects of staging feedback data so that impact information is presented last?
12. What is the difference in results from using three rounds of ratings over two rounds?
13. How are the evaluation results affected by the phrasing of the evaluation questions? Which wordings yield the most valid information about the standard setting procedures?

Standard setting has benefited from many years of development and implementation. As new assessment types have become prevalent in high-stakes tests, new standard setting methodologies have been developed. Although research programs support many of these methods, many of the design and implementation issues that surround the process have not been well researched. The purpose of this presentation is to identify some of these design and implementation issues and then to present a program of research that could address these questions. With research to support standard setters, we will be better able to make informed decisions when designing and implementing standard setting studies.

References

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (ED.), Educational Measurement. (2nd ed., pp. 508-600). Washington, D.C.:

American Council on Education.

Giraud, G., Impara, J.C., & Plake, B.S. (2005). A qualitative examination of teachers' conception of the target examinee. Applied Measurement in Education, 18(3).

Ferdous, A.A. & Plake, B.S. (in press). Item selection strategy for reducing the number of items rated in an Angoff standard setting study. Educational and Psychological Measurement.

Ferdous, A.A. & Plake, B.S. (2005a). Understanding factors that influence panelists' decisions in an Angoff standard setting study. Applied Measurement in Education, 18(3), 257-267.

Ferdous, A.A., & Plake, B.S. (2005b). Use of test questions in an Angoff standard setting method. Educational and Psychological Measurement, 65(2), 185-201.

Hambleton, R.K. (2001). Setting performance standards on educational assessments and criteria for evaluating the process. In G.J. Cizek (Ed.), Setting performance standards: Concepts, methods, and perspectives, pp. 89 - 116. Mahwah, NJ: Lawrence Erlbaum Associates.

Hambleton, R.K., Brennan, R.L., Brown, W., Dodd, B., Forsyth, R.A., Mehrens, W.A., et al. (2000). A response to "Setting reasonable and useful performance

standards” in the National Academy of Sciences’ Grading the nation’s report card. Educational Measurement: Issues and Practice, 19(2), 5 – 14.

Hambleton, R.K., & Pitoniak, M.J. (2006). Setting performance standards. In R.L. Brennan (Ed.) Educational Measurement, (4th Ed., pp. 433-470). Washington, DC: American Council on Education.

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

Jaeger, R. M. (1982). An iterative structured judgmental process for establishing standard on competency tests: Theory and application. Educational Evaluation and Policy Analysis, 4, 461-476.

Jaeger, R.M. (1989). Certification of student competence. In R.L. Linn (Ed.) Educational Measurement (3rd Ed., pp. 485-514. Washington, D.C. : American Council on Education.

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

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. 52-88. Mahwah, NJ: Lawrence Erlbaum Associates.

Loomis, S.C., & Bourque, M.L. (2001). From tradition to innovation: Standard setting on the National Assessment of Educational Progress. In G.J. Cizek (Ed.), Setting performance standards: Concepts, methods, and perspectives, pp. 175-217. Mahwah, NJ: Lawrence Erlbaum Associates.

Millman, J. (1976). If first you don't succeed: Setting pass-rates when more than one attempt is permitted. Educational Research, 18(6), 5 – 9.

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

Phillips, S.E. (2001). GI Forum vs. Texas Educational Agency: Psychometric issues. Applied Measurement in Education, 13, 343-385. Plake, B.S., & Giraud, G. (April, 1998). Effect of a modified Angoff strategy for obtaining item performance estimates in a standard setting study. American Educational Research Association, San Diego, CA.

Plake, B.S., & Impara, J.C. (2000). 2000 Standard Setting Final Report: American Association of Poison Control Specialists, Washington, DC: AAPCC.

Plake, B.S., Impara, J.C., & Potenza, M.T. (1994). Content specificity of expert judgments in a standard setting study. Journal of Educational Measurement, 31, 339 – 348.

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. 159-174. Mahwah, NJ: Lawrence Erlbaum Associates.

Skorupski, W., & Hambleton, R.K. (2005). What are panelists thinking when they participate in standard setting studies? Applied Measurement in Education, 18, 223-255.