

Exploring a New Methodology For Setting Performance Level Standards with
Computerized Adaptive Tests

Susan L. Davis

Buros Center for Testing

June 22, 2005

Paper presented as part of the “Computer Adaptive Testing in State Assessment” session at the 35th Annual National Conference on Large-Scale Assessment sponsored by the Chief Council of State School Officers, San Antonio, TX.

Correspondence concerning this paper should be addressed to:

Susan Davis
Buros Center for Testing,
21 Teachers College Hall
University of Nebraska-Lincoln
Lincoln, NE, 68588-0352
sdavis7@unl.edu

Abstract

Innovations in testing practices are catalysts that require testing professionals to continually update and improve their methodologies and procedures for assisting with testing programs. One such innovation, Computerized Adaptive Testing (CAT), provides many useful advantages to users including shortened testing time. However, the introduction of CAT to the testing world has required some researchers to think of alternative ways to assist states (or organizations) with their testing programs. Specifically, researchers have been rethinking existing methodologies for activities such as standard setting to make them compatible with the format of a computerized adaptive test. This paper details the demonstration of a standard setting procedure that closely parallels the well-known Angoff method (1971). Included in the demonstration was an evaluation of the participants' perception of the methodology.

Introduction

The innovation of Computerized Adaptive Testing (CAT) is one example of how the world of testing is changing due to advances in technology and psychometrics.

Advances such as adaptive testing allow test creators to construct more efficient tests that minimize testing time and provide accurate estimates of examinee performance. In turn, such innovations have challenged those in the testing profession to advance the science of testing at the same pace. The purpose of this paper was to explore one method in which a current standard setting method could be implemented with an adaptive test while accommodating the unique characteristics of an adaptive test.

The most notable advantage of CATs is the reduced testing time for students due to individually tailored tests created by an adaptive item selection algorithm. This algorithm selects items for an examinee from a large item pool based on the examinee's performance, which creates individual tests specifically matched to an examinee's ability. As a result, each examinee is likely to see a different form of the test as many item selection algorithms not only choose items that are matched by difficulty to the examinees' ability but also selects items from a large pool in a way that minimizes item exposure. Given this, one issue for testing professionals is to find ways to help states set performance standards on such tests that can be different for every student (Gushta, 2003). Sireci and Clauser (2001) state the problem for measurement professionals is that CATs do not require the creation of separate standard setting methods but rather existing methodologies must be modified to accommodate the unique characteristics of the CAT.

Commonly used standard setting methodologies such as Angoff (1971) or Bookmark (Mitzel, Lewis, Patz, & Green, 2001) require panelists to view either the entire

set of test items or a representative subset. As noted previously, in a CAT many examinees will see very different forms of the test as items are drawn from a large item pool, which would require either (1) standard setting panelists to view the entire item pool or (2) a representative sample of items must be selected. The first option would require significant time and effort from the panelists and as the focus of much standard setting methodology is on reducing the cognitive complexity for the panelists (e.g., Impara & Plake, 1997; Mitzel et al., 2001), this does not appear to be a realistic option. The second option would require a sample to be chosen that provides adequate coverage of both difficulty level and content. Currently, the process by which one would ensure adequate coverage of both difficulty and content is unclear (Sireci & Clauser, 2001). The method described in the following section is one proposed method that would afford setting performance standards on a CAT without having to view the entire item bank or manually select a representative sample from the bank.

Proposed Method for setting standards with Computerized Adaptive Tests

The method demonstrated in this study has origins with methods in the field of measurement many are very familiar with. In his famous chapter, Angoff (1971) described a method for setting passing standards that is used today in many types of testing programs. The now appropriately named Angoff method requires panelists to estimate the performance of a group of examinees at the item level by indicating what proportion of these examinees would respond correctly to each item. Years later, Impara and Plake (1997) proposed an alternative version to the Angoff method with two notable modifications to the existing method. First, panelists would be asked to focus on the expected performance of one student (instead of a group) whose ability characterized the

borderline between performance categories (target student). This modification allowed panelists to focus on one student with whom they were familiar, as prior research by Impara and Plake had suggested that teachers were not very accurate in estimating the performance of a group of students. The second modification eliminated the need to estimate the proportion of examinees that would respond to each item correctly and replaced it with a 'yes/no' decision criteria pertaining only to the target student. This modification was actually part of the original Angoff method (1971).

Recently, these same standard setting principles have been applied to adaptive testing. Sireci and Clauser (2001) described what they called "The Wainer Method" (as it was based on personal communication with Howard Wainer). In this method, panelists take the adaptive test as if they are the target student (also referred to as the 'borderline' or the 'barely proficient' student) by responding to each item as they believe the target student would. Similar to the version of the Angoff method described by Impara and Plake (1997), panelists are asked to use a 'yes/no' judgment method and determine if the target student would answer the question correctly and respond in the way they believe the target student would. As a panelist progresses through the test, the adaptive algorithm estimates the ability demonstrated by their performance. The resultant test score is the recommendation for the cut score. After each panelist completes the exam, their scores are averaged to derive the recommended cut score.

Recent work by O'Neill, Tannenbaum, and Tiffen (under review) is the first noted application of this methodology. In this work, the above described method was credited to ideas noted in the Sireci and Clauser (2001) chapter as well as personal discussions with Mary Lunz in the mid 1990s. O'Neill and his colleagues used this method to set a

passing standard for the NCLEX-RN[®]. Overall, the authors felt the method was quite successful given the closeness of the existing passing standard to the passing standard set by other groups for the same exam.

In the current study, this method was tested using an 8th grade mathematics CAT exam used by school districts in Nebraska. The purpose of this activity was to demonstrate this method in an educational setting and gather panelists' perceptions of their experience and their comfort of the method. In turn, it was expected that this would be a first step in estimating the feasibility of this method in an educational setting. The following sections detail the process used for applying this method, the results, and interpretations of these results for future use.

Method

Participants

Panelists were invited to attend a workshop on standard setting methodology. Because panelists self-selected into this workshop, this panel was not considered to be an appropriate representation of subject matter experts that would be needed if the goal of this workshop was to set an operational standard. However, the purpose of this activity was to demonstrate the method and obtain the reaction of the panelists, and the panel assembled was deemed acceptable for this purpose.

Training

As a precursor to introducing the proposed method for standard setting with a CAT, panelists were given a tutorial covering different standard setting methodologies. During this informational session, panelists were introduced to the concept of the target student. The target student was described to panelists as one who has the minimal skills

needed to be considered ‘proficient’ (or ‘passing’ depending on the wording of the competency categories for a test, organization, or school district). Panelists were then given the opportunity to consider this student in their own school districts/classrooms. Specifically, panelists were asked to discuss the target student with respect to their school district’s proficient category for the 8th grade math exam. Panelists first completed this activity in small groups (3-4 individuals per group) by listing tasks that would be easy for the target student and those that would be more difficult. After each group had completed the discussion, the workshop facilitator lead a discussion with the full group where each small group shared their ideas and discussed the ideas of others. This discussion resulted in a description of the target student compiled by the entire group, which defined those tasks that would be easy for the target student and those that would be more difficult. The panelists were able to refer to this list during the operational demonstration of the standard setting method.

Demonstration of method

Panelists were then introduced to the standard setting procedure they would be using. Put simply, they were instructed to take the exam as if they were the target student by responding to each question in the way they felt the target student would respond. For the demonstration of this method, panelists took the 8th grade adaptive mathematics exam used by school districts in which they worked to test students’ math skills as compared to state standards. The test is 52 items for all examinees; the stopping rule is not based on standard errors. The test algorithm was designed to give each student a sampling of the content area (a minimum of 6 items per content area). Most panelists had some degree of familiarity with the test as they either worked with students who took the test or were

administrators in schools in which the test was used. Panelists were administered the test in the same computer lab where students took the exam and before starting the test, panelists read the same test instructions students are given. Upon completion of the test, panelists' scores were recorded and panelists completed an evaluation of the process.

Results

Estimates of student ability from the 8th grade mathematics exam are reported as a scale score and students' scores typically range from 140 to 300. The average score derived by panelists using this method was 244.9 (median=244.5) and panelists' scores ranged from 222 to 260 (standard deviation = 9.28) indicating low variance in participants' estimates of the target student's performance. Although the panel who derived this recommended standard was not considered to be the ideal group of subject matter experts, it does speak well of the method that the average cut score (244) was close to the average cut score used by other states for their 'proficient' classification (238) as reported by the test publisher.

Given that this was a demonstration of this new method for standard setting, the focus of this paper is on the success of the method as determined by the panelists' evaluation of their experience. In the selected response portion of the evaluation most panelists indicated that they were confident (or somewhat confident) and comfortable (or somewhat comfortable) with their judgments on the performance of the target student. Interestingly, panelists only exuded a moderate level of confidence that the passing score derived from this exercise should be an appropriate passing standard for the use of the test in their district. This is likely due to their range of backgrounds (mix of administrators and teachers) and the fact that many of these panelists did not work with

8th grade mathematics students on a day-to-day basis. This type of information (confidence in recommendation and process) is important to collect from panelists as it is useful for policy makers to consider when setting a standard for performance based on the results of a standard setting exercise.

In addition, panelists were asked two open-ended questions. The purpose of these questions was to understand the thought process used by panelists as they progressed through the test and determine if the method was working as intended. First, panelists were asked to describe the process they used to determine if the target student would respond correctly to each question. Most panelists indicated that they referred to the description of the target student created by the panel earlier in the day, which listed the skills that would be easy or difficult for the target student. Panelists also indicated they relied on their experience teaching or working with students, considered the difficulty and complexity of the item, and considered the state's definition of the proficient performance level as compared to the description of the performance level below proficient. Overall, these responses indicated that panelists were completing the exam as instructed and the method was working as expected.

Second, panelists were asked to describe, if they believed the target student would not answer the item correctly, which distracter the target student would select. Most panelists indicated they based their judgment on common errors that students made. Others said they looked for a response that would be a good guess but not the correct answer or looked for similarities between the item stem and response options.

Finally, panelists were given the opportunity to share any additional comments they had about the process. Several of these comments are noteworthy as they provided

information important to evaluating the process. For example, one panelist noted that he/she felt his/her performance was “better than the barely proficient student” and another noted that it was “hard to remember to respond as the barely proficient student.” Although the evaluations were anonymous, panelists were asked to note their exam scores on their evaluations. The two panelists who indicated they may not have had the right mindset both had scores more than one standard deviation higher than the average score. This result would suggest that in an operational standard setting using this method, panelists should be asked how well they were able to stay in the mindset of the target student. In turn, a facilitator may consider excluding data from any panelist who indicated having a problem following the method.

Discussion

Overall, this demonstration was very successful. The panelists experienced a hands-on opportunity to learn the method and apply it to a familiar test. The facilitators were able to gauge the feasibility of explaining this process and using it in an educational setting. In addition, the panelists’ evaluations of the method provide invaluable information as to how the panelists viewed the method, the thought processes they used, and some minimal problems that can be used to improve upon the method. For example, two of the 22 panelists indicated that they performed better than the target student. In turn, their scores were substantially higher than the scores of the other panelists. Therefore, in future use of this method, panelists should be specifically asked if they felt they were able to stay in the mindset of the target student. In addition, there are several ways in which this problem could be avoided with during the process. For example, breaks could be built into the testing session (e.g., after 10 minutes, 20 minutes) where

panelists stopped the testing process and engaged in discussion about the lists of skills (that would be easy or hard for the target student) they drafted earlier in the process. Second, given the flexibility of the testing software, prompts to respond as the target student could be built into the display. For example, in the one documented use of this method, O'Neill et al., (under review) displayed a prompt to the panelists between each question that reminded them to respond in the way the target student would.

Upon completion of this demonstration there are several apparent advantages and disadvantages to this method that are noteworthy. The major apparent advantage of this method is that it uses the test itself as the format for setting the standard. Therefore, panelists get the same experience as the test takers and standard setting facilitators do not have to prepare item booklets or packets in preparation for this activity. By using the existing test, panelists only have to view and respond to a minimal number of items (not the full item pool). Also, the scoring algorithm provides the data to be used in calculating the recommendation for the performance standard, which also minimizes the work of the facilitator.

There are also several disadvantages to this method. First, one must have access to a computer testing lab and the test itself at the time of the standard setting exercise. This may require special arrangements with a testing contractor or school. Second, because panelists are getting the real testing experience they may, as was noticed in this demonstration, slip out of the mindset of the target student and focus on answering the questions correctly. As noted earlier, there are several possible ways in which to combat this problem that should be considered in future applications. Third, because panelists are taking the test as a student would, they are not provided the correct answer and this

method makes the assumption that panelists know the correct answer to all items they encounter.

As with any standard setting method, further research must be conducted to determine the feasibility and success of this method in different situations and with different types of tests. Specifically, future research should compare the results of this method to the results of other methods and explore any substantial differences that arise. In addition, the use of a second round should be considered, as well as the impact of providing panelists feedback between rounds regarding performance of student groups in relation to their recommended standard.

As more testing programs become computerized and incorporate adaptive testing, testing professionals are challenged to acclimatize current methods to accommodate these new testing characteristics. Currently, only minimal information exists about setting performance standards with adaptive tests. This study served as a first look at a new method for setting standards with adaptive tests in a way that appears to be efficient, straight forward, and understandable by panelists. Future research will explore variations on this method and determine how it compares to other innovative methods for setting standards with adaptive tests.

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.
- Gushta, M.M. (2003, May). *Standard-setting Issues in Computerized-Adaptive Testing*. Paper presented at the Annual conference of the Canadian Society for Studies in Education, Halifax, Nova Scotia.
- Impara, J.C., & Plake, B.S. (1997). Standard setting: An alternative approach. *Journal of Educational Measurement*, 34(4), p.353-366.
- Mitzel, H.C., Lewis, D.M., Patz, D.M., & Green, D.R. (2001). The bookmark procedure: Psychological perspectives. In G.J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 249-281). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
- O'Neill, T.R., Tannenbaum, R., & Tiffen, J. (2005). Recommending a minimum English proficiency standard for entry-level nursing. *Journal of Nursing Measurement*.
- Sireci, S.G., & Clauser, B.E. (2001). Practical issues in setting standards on computerized adaptive tests. In G.J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 355-369). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.