Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Copyright is retained by the first or sole author, who grants right of first publication to the *Practical Assessment, Research & Evaluation.* Permission is granted to distribute this article for nonprofit, educational purposes if it is copied in its entirety and the journal is credited.

Volume 15, Number 6, May, 2010 ISSN 1531-7714

Recommending cut scores with a subset of items: An empirical illustration

Chad W. Buckendahl
Alpine Testing Solutions

Abdullah A. Ferdous
American Institutes for Research

Jack Gerrow
Dalhousie University and National Dental Examining Board of Canada

Many testing programs face the practical challenge of having limited resources to conduct comprehensive standard setting studies. Some researchers have suggested that replicating a group's recommended cut score on a full-length test may be possible by using a subset of the items. However, these studies were based on simulated data. This study describes a standard setting application using two independent panels providing judgments on a 300-item licensure test. Specifically, one panel provided judgments on all 300 items; whereas the second panel made judgments on a randomly-selected subset of 150 items. Both panels also participated in an alternate standard setting method to evaluate panel comparability. Results suggest caution for practitioners considering using subsets of items for standard setting studies.

Few testing programs operate without scores being used to make some decision about individuals or groups. Although we may characterize some of these decisions in the context of the consequences associated with the decision (e.g., high stakes, low stakes), the decisions are meaningful for some intended use or interpretation. The ability to make these decisions is facilitated by the use of performance standards that become operational through the application of cut scores. Thus, establishing cut scores is the link between the information we collect about an examinee's abilities and the decisions we make about those abilities represented as test scores.

There are a number of methods in the literature for systematic processes to recommend cut scores to policymakers. Hambleton and Pitoniak (2006) provide a summary of methods that are currently being used in the field. However, in addition to concerns about psychometric integrity, many testing programs face the practical challenge of having limited resources to conduct comprehensive standard setting studies that contribute to the validity evidence for their program. To begin to address this challenge, some researchers have suggested that replicating a group's recommended cut score on a longer, full-length test may be possible by using a subset of the items.

The purpose of this study was to empirically investigate whether the cut score recommendation for a randomly selected subset of items from a 300-item test would converge with the cut score recommendation on the full-length test. In this study, two independent panels provided judgments on a 300-item licensure test. One panel provided judgments on all 300 items; whereas the second panel made judgments only on a randomly selected subset of 150 items of the full test. Our design for this study was informed by previous and concurrent work on the topic.

Other theoretical research has suggested that it may be feasible to replicate a panel's recommended cut score on a full-length test with a subset of the items. Specifically, Ferdous and Plake (2005) examined different random samples of items (e.g., 50%, 70%) from a full-length test to evaluate whether a panel's recommended cut score using a modified Angoff (1971) method could be replicated from these shorter length versions. Results suggested that recovering the full-length cut score could be reliably accomplished, potentially increasing the efficiency of the standard setting process. For a lengthy test, this strategy may allow a test sponsor to reduce the number of days and the related financial and logistical resources needed for a traditional standard setting study.

Similarly, Smith and Ferdous (2008) examined alternative sampling methods for selecting a subset of items for standard setting studies that used a modified Angoff (1971) method. The proposed method used an automated strategy for selecting items that could replicate the recommended cut score for the full-length test. In this study, non-random strategies that reflected different meaningful strata for selecting items were proposed to evaluate whether particular item characteristics allowed replication of the full-length cut score to be more efficient than a random selection. Results of these analyses also suggested that recommended cut scores on the full-length test could be recovered through this systematic sampling approach.

Beyond conceptual proposals and simulation studies, operational standard setting methods have also essentially employed a subset of items from either an item pool or from selecting items from multiple test forms to represent the interpretative scale. Specifically, because the Bookmark method (Mitzel, Lewis, Patz, & Green, 2001) is commonly used in educational settings, Ferdous (2007) also proposed a strategy for examining subsets of items for these applications. In the Bookmark method, panelists do not make judgments for individual items, but rather they make judgments about a collection of items and a target examinee's likely (typically assigned with a response probability criterion) performance on that collection. The Ordered Item Booklet (OIB) that is constructed for the Bookmark method often has items that have very similar item locations. Thus, from one perspective the use of a subset of items has the potential to proactively address some of the perceived item dis-ordinality that occurs during these studies. However, the selection criteria for which items to retain or omit may raise concerns about whether the content of the test is fully represented by the OIB.

Though the results of these studies appeared promising, Hambleton (2007) expressed concerns about the potential use of only a subset of items for standard setting judgments as minimizing the importance of this process for testing programs. More important, these studies were conceptual discussions and simulations that were unable to consider how panelists might respond to only providing judgments on a subset of items in an applied study. Therefore, to evaluate some of these concerns, we engaged in the applied, empirical study described here.

Using Kane's (2001) framework for evaluating standard setting studies, we collected evidence to evaluate the procedural, internal, and external validity of the recommendations. For procedural evidence we focused on the selection of panelists, training procedures, and the independent application of the Angoff (1971) method as described in Impara and Plake (1997) for both panels. The internal evidence we evaluated was collected from within panel and between panel judgments.

Collecting external evidence to evaluate the recommended cut score often proves to be the most difficult component of Kane's (2001) framework. Hambleton and Pitoniak (2006) suggest that "results from other standard setting methods, information from other sources and evidence of the reasonableness of the outcomes (p. 461)." Cizek and Bunch (2007), Green, Trimble, and Lewis (2003), and Jaeger (1989) also suggested that the use of multiple standard setting methods may provide additional sources of external evidence that can be used to inform the policy decision.

For this study, both panels also independently participated in a modification of the Bookmark method on the full length test as described by Buckendahl, Smith, Impara, and Plake (2002). The second methodology was incorporated to provide a source of external evidence and information about the comparability between the panels that could influence the interpretation of the results from two variations of the same method.

Information about the Written Examination

The National Dental Examining Board (NDEB) of Canada's Written Examination is one of two tests required as part of the licensure process for dentists. It is intended to provide information on the extent that candidates for initial dental licensure have attained the entry-level knowledge, skills, and abilities necessary for safe, independent practice. An objective of administering this test is to classify candidates into two categories: 1) candidates who are incompetent and not eligible for licensure, and 2) candidates who are at least minimally competent and eligible for licensure.

The Written Examination contains 300 multiple-choice items that measure elements of the dental content domain as defined by NDEB through the *Competencies for a Beginning Dental Practitioner in Canada.* All items are scored dichotomously and each item counts one point. For the 2007 Written Examination the total number of points available was 300.

METHODS

Variations of two commonly-used standard setting methods were used in this study. These methods were: a) modified Angoff (1971) method (Impara & Plake, 1997); b) a modified Bookmark method (Buckendahl, et al., 2002); and c) a modified Angoff method using a randomly selected subset of items (Ferdous & Plake, 2005). Each of these methods is described briefly below.

Angoff Method

The Angoff (1971) method entails using subject matter experts (SMEs) to examine each test item and estimate how a typical Minimally Competent Candidate (MCC) will perform on that item. We applied the Yes/No modification of Angoff's (1971) original methodology as described by Impara and Plake (1997) to a full length version and a subset version of the exam. The subset version represented a random selection of 50% of the items in the exam as suggested by Ferdous and Plake (2005). Although Smith and Ferdous (2007) suggested alternative item selection approaches, the design for this study had already been previously proposed and approved by the sponsor's policy body prior to the availability of this additional research to inform the design. Therefore, items were randomly selected using a random number generator without consideration of item difficulty or content representation of the random subset of items.

Bookmark Method

The Bookmark method (Mitzel, Lewis, Patz, & Green, 2001) also uses expert judges to examine content represented by items on the test and estimate how a typical MCC will perform. Items are ordered empirically from least difficult to most difficult and compiled into a booklet. Item ordering is often accomplished using multi-parameter item response theory (IRT) methods recommended in Mitzel et al., 2001. However, other models for item ordering have been used. For example, Wang (2003) suggested that item ordering using a single parameter IRT model would also be appropriate. Similarly, Buckendahl, et al. (2002) illustrated that for the multiple choice items in their study, classical test theory p-values correlated highly with the item ordering produced by a three-parameter logistic model with guessing factored out.

PROCEDURES

For this study, 35 panelists were selected by NDEB to represent three specific stakeholder groups within the dental profession in Canada. The first stakeholder group was dentists who were recently licensed (i.e. within the last 5-7 years). The second stakeholder group was more experienced dentists (i.e. generally more than 7 years experience). The third stakeholder group was educators from dental schools in Canada. These characteristics were the primary selection criteria as these specifications help to ensure that the panelists have both content knowledge and some familiarity with the abilities of the population of candidates who would be taking the Written Examination. A final selection criterion was that panelists needed to be representative of the diversity of Canada's dental population (i.e. geography, gender, language). To create the independent panels that would participate in the two variations of the Angoff method, the panel was divided into two groups, Group A (Full Angoff; n=18) and Group B (Subset Angoff; n=17) to be representative of the stakeholder groups and selection criteria noted above.

In the training, panelists were informed that they would be participating in multiple methods with a variation between the groups in the application of the Angoff method. This means that the Group A knew they would be making judgments on all 300 items and that Group B knew that they would be making judgments on a randomly selected subset of 150 items. The sequence of the methods was the same for both groups. Specifically, the two variations of the Angoff method were applied first followed by the Bookmark method on the full length test. Because the same methodology was applied first for both groups, there was a potential order effect.

Following an initial orientation and discussion of the content contained in the *Competencies*, the full panel engaged in a discussion of the minimally competent candidate (MCC). To assist panelists in their conceptualization of the MCC, NDEB provided a Performance Level Descriptor (PLD) that was shared with the panelists as the broad policy definition and a starting point for further discussion. This PLD characterized the MCC as, "The candidate has the entry-level knowledge, skills, abilities and judgments necessary to begin safe, independent practice." The NDEB has developed more specific descriptors of what entry-level dentists need to know and be able to do. These descriptors are contained in the Board's document entitled, *Competencies for a Beginning Dental Practitioner in Canada*[1] and were provided to panelists as supplemental information.

For the Written Examination after engaging in training activities, panelists were asked to conceptualize a specific MCC with whom they were familiar. Keeping this candidate in mind, they were directed to begin with the easiest item (first page of the booklet) and move through the booklet until they found the place where their MCC would have at least a 0.67 probability (see Huynh 1998; 2006 for discussion of RP67 criterion) of answering the collection of items up to that point correctly. Once they reached an item that they thought the MCC would have less than a 0.67 probability of answering the item correctly, panelists were asked to continue in the booklet for 3-4 more items to evaluate whether these items would also be answered correctly by the MCC less than the RP criterion. If these additional items confirmed their initial judgment (i.e., also appeared to be difficult for the MCC candidate), they went back to the first item that they identified and placed their bookmark at that point. If the MCC would have at least a 0.67 probability of answering these additional items correctly, the panelist continued on in the booklet until they reached their respective bookmark location. Davis, Buckendahl, and Gerrow (2008) provide detailed information about the application of the Bookmark method in a cross-method comparison.

Group A (Full Angoff)

In Group A, panelists began by making their first round ratings for each of the 300 multiple choice items on the Written Examination and occurred on the first of two days of the study.

Following data entry, panelists' rating forms were returned to them on the second day with an explanation of the results of the first round of the Angoff procedure. As feedback, panelists received their respective Round 1 recommendation, the group's average (mean and median) recommendation, and variability (standard deviation, minimum, and maximum) for the group. Panelists also received actual performance data (p-values) from candidates' March 2007 exam performance (n=550). After responding to questions about Round 1 judgments, panelists then received the impact of the panel's initial recommendation based on a cumulative percent distribution of actual candidate performance. The final source of information that the panelists received after their initial judgments was the answer key for the items to assist them in their judgments of how difficult the items were for the candidates. Panelists then made their final recommendations on each of the 300 items. Following their Round 2 judgments on the Angoff method, the panelists then participated in the Bookmark method as described above.

Group B (Subset Angoff)

To develop the test booklet for Group B, we randomly selected 50% of the 300 items from the Written Examination for the study. This was a simple random sample that did not stratify content representation or item difficulty in the selection of the items. Because simulation studies (e.g., Ferdous & Plake, 2005) suggested that panelists' judgments could be reproduced with different randomly selected proportions, we sought to test this finding empirically. This variation of the Angoff method was conducted both as external validity evidence and also in response to ongoing discussions about the appropriateness of using subsets of items in operational standard setting studies (Ferdous & Plake, 2005; Hambleton, 2007).

The same sources of feedback data that were provided to panelists in Group A were also provided to panelists in Group B. However, there was one modification. Specifically, to communicate the impact data of the initial round judgments on the pass/fail rates of candidates, the panelists' recommended values were doubled to reflect the impact of their initial recommendation on the full length examination. After panelists completed their second round Angoff ratings, they conducted the Bookmark method as described above and completed a process evaluation form that provided feedback on the activities of the study.

[1] For more information about these *Competencies*, see www.ndeb.ca.

RESULTS

The results in Table 1 illustrate the psychometric characteristics of the full length form of the test that was used in Group A and the subset version of the test that was used in Group B. From these data we can see that there are some small differences in the mean and median that would yield an expected, approximate difference in recommended cut scores of two raw score points. Note also that the shapes of the distributions and the estimated internal consistency estimates suggested differences in the characteristics of the two versions of the exam that has the potential to mitigate interpretations of the recommended results of the study.

Table 1. Comparison of psychometric characteristics for full and subset length tests.

	Full Length Test (n=300)	Subset Length Test (n=150)
Mean	240.34	118.63
Median	242.00	120.00
Standard Deviation	20.12	10.56
St. Error of the Median	1.04	0.55
Skewness	-0.93	-0.76
Kurtosis	1.92	1.32
Coefficient alpha	0.90	0.82
St. Error of Measurement	6.36	4.54

Table 2. Results from Angoff method for Rounds 1 and 2.

Round 1

	Group A (Full Angoff)	Group B (Subset Angoff)
Mean	216.6	204.0
Median	214.0	206.0
Standard Deviation	18.6	17.6
St. Error of the Median	5.49	5.35
Impact (% below)	13.1%	5.6%

Round 2

	Group A (Full Angoff)	Group B (Subset Angoff)
Mean	214.1	204.4
Median	217.0	206.0
Standard Deviation	16.9	15.9
St. Error of the Median	4.99	4.83
Impact (% below)	10.5%	5.6%

Note that values for the subset Angoff group were doubled to place Group B's results on the same scale as Group A for comparability.

The recommended cut scores from each round for each group are shown in Table 2. Providing feedback data between rounds one and two had some influence on the panelists as the second round mean cut score recommendation for Group A (Full Angoff) dropped by two points. In comparison, the mean cut score recommendation for Group B (Subset Angoff) remained essentially unchanged. The variation in cut score recommendations for both groups decreased from Round 1 to Round 2 (18.6 to 16.9 for Group A; 17.6 to 15.9 for Group B). This reduction in variance suggests that panelists' judgments converged slightly in Round 2. Note that the results for Group B were doubled to place them on the same scale for greater ease of comparison.

Bookmark Method

The panelists' recommendations from the Bookmark standard setting method are shown in Table 3. Providing feedback data between rounds one and two also had some influence on the panelists as the second round cut scores for both Groups A and B decreased by approximately nine points. The variation in cut scores also decreased from Round 1 to Round 2 for both groups. However, the variation for Group A was higher than for Group B. Although the variation was greater, the recommended mean values were similar. The median values, though, suggested that Group B's recommendations were approximately one standard error of the median lower than Group A's.

Table 3. Results from Bookmark method for Rounds 1 and 2.

Round 1

	Group A	Group B
Mean	225.2	226.7
Median	224.0	223.0
Standard Deviation	32.1	13.4
St. Error of the Median	9.48	4.07
Impact (% below)	21.8%	25.3%

Round 2

	Group A	Group B
Mean	216.3	218.4
Median	218.0	213.0
Standard Deviation	20.1	12.5
St. Error of the Median	5.94	3.80
Impact (% below)	12.4%	13.6%

Evaluation

The process evaluation form that panelists completed at the end of their operational ratings on the variations of the Angoff and Bookmark method consisted of eight parts. Specifically, Part 1 focused on the orientation and training; Part 2 focused on Round 1 of the Angoff ratings; Part 3 focused on Feedback data for Angoff; Part 4 focused on Round 2 of the Angoff ratings; Part 5 focused on Round 1 of the Bookmark ratings; Part 6 focused on Feedback data for Bookmark; Part 7 focused on Round 2 of the Bookmark ratings; and Part 8 focused on panelists overall evaluation of the study. Panelists' evaluation ratings of these activities and processes were generally positive across these sections with higher values reflecting this assertion. Selected panelists' responses by group are provided in Table 4.

Table 4. Selected comments from panelists for Groups A and B

Group A comments
• One thing I would have liked to discuss was the pass rate that is aimed for and how that can be modified to reflect the reality of candidates MCC.
• I believe both methods are efficient and having the 2 rounds makes it more valid and realistic.
• I felt very comfortable with my test scores in the end (Angoff) as I am involved with the curriculum development & I am a fairly recent grad. . . .
Group B comments
• I feel my standards for an MCC may be too high and I am really thinking of the "Average" candidate rather than the weakest possible passing candidate. As a result, I may have set the bar too high for a passing mark.
• Explain how the p-values were used in the past by the people involved in such a process.
• I question the validity of this work process. The process is good to identify a passing score based on content, but the final passing score must be determined using other factors than the content.

Although they were informed that they were only providing judgments on half of the full length test, Group B's panelists did not express concerns or comments regarding these reduced judgments. Most comments from both groups were focused on the comparison across the Angoff and Bookmark methods (See Davis, et al. 2008) and are not reported here given the scope of this study.

DISCUSSION

Conducting standard setting can be a costly component of the test development and validation process, particularly for licensure and certification testing programs that rely on highly compensated subject matter experts to provide judgments. Evaluating strategies that improve the efficiency of the process without threatening the validity of the interpretations are at the core of this line of research. In discussing the results of simulated studies on this topic, Hambleton (2007) expressed concerns about the potential use of only a subset of items as minimizing the importance of the standard setting process for testing programs. Because previous studies were unable to consider how panelists might respond to only providing judgments on a subset of items in an empirical study, we sought to further explore the question. Thus, the purpose of this study was to empirically investigate whether the cut score recommendation on a randomly selected subset of items from a full-length test would converge with the cut score recommendation on the full length test.

In this study, two independent panels provided judgments that were applied to a full-length licensure tests. One panel provided Angoff Yes/No judgments (Impara & Plake, 1997) on all items; whereas the second panel made these item-level judgments on a randomly selected subset of 50% of the items. Results suggested that there were meaningful differences in the second round recommendations of these independent panels that warrant further study to evaluate whether the promising findings from simulation studies (e.g., Ferdous & Plake, 2005; Ferdous, 2007; Ferdous & Smith, 2008) can be replicated in operational settings.

Applying Kane's (2001) suggested framework, we discuss the procedural, internal, and external validity evidence we used to evaluate these results.

Procedural evidence

The procedural evidence for the study appeared to be strong. First, panelists were selected to represent primary stakeholder groups in the licensure testing process (i.e. recently licensed practitioners, experienced practitioners, and dental faculty). These panelists were also randomly assigned to Group A or Group B within their respective strata. Second, the panelists for both groups participated in a common, systematic training procedure to orient them to the performance level descriptor (PLD), content, skills of the target candidate population, and the methods that they would use in the study. Third, both groups independently completed methods that were consistent with published standard setting literature, receiving feedback data between two rounds of judgments about item level difficulty and likely candidate performance. Finally, the panelists completed a written process evaluation form that documented their confidence and comfort with the various aspects of training and operational ratings. The results of this evaluation suggested that panelists had a positive experience with the methods applied in the study.

There were two notable limitations in the procedural evidence. One limitation rests within the methodology chosen to select items for the Subset Angoff variation. Because Ferdous and Plake (2005) proposed differing levels of random selection, we chose to evaluate this methodology empirically in this study. Alternative item selection methodologies could have been chosen. For example, items for the subset could have been selected to stratify content and difficulty to approximate the full length version (Ferdous & Smith, 2008). Although the randomly selected subset of items had similar psychometric characteristics to the full length test, they were not identical which could have had an impact on the comparability between the panels' recommendations.

A second limitation was the order in which the methodologies were applied. In both instances, the panelists made their Angoff judgments on their respective version (full or subset) before making their Bookmark judgments on the full length version of the test. This order may have had differential impact on the groups because one group had seen the full length version during their Angoff judgments whereas the other group would have seen only half the items.

Internal evidence

Evaluating the internal validity evidence is related to the design of the standard setting methodology. For example, when there is more than one round of judgments or extended group discussion between rounds of judgments, the variation of panelists' responses may converge. However, this convergence may be an artificial estimate of consensus and more influenced by social desirability to converge on the central tendency of the group.

For this study, the variation in the recommendations of the groups was lower in the second rounds than in the first suggesting that panelists were in greater agreement with the resultant recommended cut score. The reduction in variation was also accompanied by a reduction in the recommended cut score. This suggested that panelists were influenced by the feedback data, most likely the impact of the groups' initial recommendations. Our evaluation of the internal evidence generally supported the results of the methods that were applied in this study.

External evidence

Because external validity evidence is often the most difficult to collect, particularly for licensure and certification programs, this study was designed to prioritize multiple sources that would inform our evaluation. These sources were the a) use of independent panels, b) variations of the same method, and c) an alternative standard setting method. An additional source of external evidence considered by the policymaking body was the historical pass rate for the program.

In this study, two equivalent panels were trained using common methods, but then separated to provide independent judgments using multiple methods to cross-validate their recommendations. These panels both participated in variations of the Angoff standard setting method as the second source of external evidence. The results from these methods suggested that the panels yielded recommended cut scores that differed by greater than two standard errors of the median for the respective Angoff method. With the initial difference in form difficulty, we may have expected a difference in recommended values of two raw score points that would still suggest divergent results. An evaluation of these results could compare the experimental method with the traditional method, but without additional, empirical evidence of the panels' comparability, this evaluation is incomplete.

The third source of external evidence that we evaluated was the panelists' recommendations with a second standard setting methodology. Because both panels completed the same application of the Bookmark (Mitzel, et. al 2001) method in the same order, it served as a check on the equivalence of the panels. Using the median, the results of this methodology suggested the panelists in Group B generally recommended a lower cut score than Group A. The median results were different, however, this time within approximately one standard error of the median. If this difference between groups is systematic, the observed differences in the variations of the Angoff method may not be as different as they appear in isolation, particularly if combined with the approximate two raw score point difference in the difficulty of the versions of the test.

Considering these external data collectively, the results suggested that there were differences between the variations of the Angoff method that may require policymakers to underweight the results of the experimental method. However, results for this study may not be as divergent as observed in this study. Group B's recommendations were lower on both methodologies suggesting that they had a slightly lower expectation for the MCC as a group than did Group A. If these observations were better controlled, the results of the study may be different. Thus, the results diverged from what we expected based on Ferdous and Plake (2005) based on a likely combination of factors that included the study design, order of the methods, panel variability, and group dynamic of the respective panels. Further study is warranted to evaluate whether controlling these conditions would yield more convergent results.

Utility evidence

An additional element to consider as an extension of Kane's (2001) framework is the utility of using a subset methodology as a source of validity evidence in the policymaking process. Some of the elements that may be considered in an evaluation of utility evidence include panelist fatigue, stakes of the program, availability of qualified panelists, and resources. For this study, the potential reduction of time from two days to one day would have substantively reduced costs to the program for panelists' honoraria, travel expenses, and meeting logistics. Specific costs for these components of the study could be estimated based on a number of dependent factors such as the respective profession (e.g., dentist versus dental assistant) and meeting location (e.g., San Francisco versus Omaha).

One of the advantages of using a subset of items when conducting an Angoff study is to reduce the amount of time needed for gathering judgments for recommending a cut score to policymakers. This reduction of time has implications for many of the practical considerations noted above. However, the intended uses of scores within a testing program should weigh more heavily in the decision of the standard setting design.

In this application, the intended use of scores was to distinguish candidates who were at least minimally competent from those who were not in a dental licensure setting. Given the greater risk to the public of candidates who should not be licensed (i.e., Type I errors), asking panelists to make a recommendation on a subset of information may reduce the credibility of the process if challenged. Although we have some information about the comparability of the groups through the Bookmark method that both groups conducted, there is insufficient information to know how Group B would have reacted to seeing the full set of items. Until additional studies can further evaluate this question, there is not a compelling reason to readily accept the results of the subset variation of the Angoff methodology as being appropriate as a stand-alone method given the external validity evidence described above. However, this is not to suggest that this program of research is without merit.

Further empirical studies are needed to inform the appropriateness of the concept. For example, a fully counterbalanced design could be used to replicate the study described in this article in which panels participate in both variations of the method. In addition, alternative item selection approaches could also be used that may place specific content and difficulty constraints as suggested by Ferdous and Smith (2008) or randomly select items that meet specific blueprint and psychometric characteristics for each person. By engaging in this work, the observed differences from this study may be further explained to better inform practice.

CONCLUSIONS

Many testing programs face the practical challenge of having limited resources to conduct comprehensive standard setting studies that inform policymakers' decisions and contribute to the validity evidence for their program. To address this challenge, some researchers have suggested that replicating a group's intended, recommended cut score on a full-length test may be possible by using a subset of the items. Although an evaluation of the validity evidence in this study suggested that the panels' recommended values were somewhat different, it should not discourage researchers from considering additional explorations of variations on this methodological concept given the limitations of generalizing the results this study. Practitioners are also encouraged to consider utility as another source of evidence in their judgments about the standard setting design and their evaluation of other sources of validity evidence, particularly as it relates to the intended use and interpretation of the scores.

REFERENCES

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R.L. Thorndike (Ed.), *Educational Measurement,* (2nd ed., pp. 508-560), Washington, D.C.: American Council on Education.

Buckendahl, C. W., Smith, R. W., Impara, J. C., & Plake, B. S. (2002). A comparison of Angoff and Bookmark standard setting methods. *Journal of Educational Measurement, 39*(3), 253-263.

Cizek, G. J. & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests.* Thousand Oaks, CA: Sage.

Davis, S. L., Buckendahl, C. W., & Gerrow, J. (2008). A comparison of Angoff and Bookmark standard setting methods for an international licensure test. Paper presented at the annual meeting of the National Council on Measurement in Education. New York, NY.

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

Ferdous, A. (2007, April). The use of subsets of test items in a bookmark standard setting method. Paper presented at the annual meeting of the National Council on Measurement in Education. Chicago, IL.

Ferdous, A. & Smith, R. W. (2008, March). Considerations for using subsets of items for standard setting. Paper presented at the annual meeting of the National Council on Measurement in Education. New York, NY.

Green, D. R., Trimble, C. S., & Lewis, D. M. (2003). Interpreting the results of three different standard setting procedures. *Educational Measurement: Issues and Practice, 22*(1), 22-32.

Hambleton, R. K. (2007). Symposium on standard setting. Discussant presentation at the annual meeting of the National Council on Measurement in Education. Chicago, IL.

Hambleton, R. K. & Pitoniak, M. J. (2006). Setting performance standards. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 433-470). Westport, CT: Praeger.

Huynh, H. (1998). On score locations of binary and partial credit items and their applications to item mapping and criterion-referenced interpretation. *Journal of Educational and Behavioral Statistics, 23*, 35-56.

Huynh, H. (2006). A clarification on the response probability criterion RP67 for standard settings based on bookmark and item mapping. *Educational Measurement: Issues and Practice, 25*(2), 19-20.

Impara, J. C. & Plake, B. S. (1997). Standard setting: An alternative approach. *Journal of Educational Measurement, 34*, 355-368.

Jaeger, R. M. (1989). Certification of student competence. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 485-514). Englewood Cliffs, NJ: Prentice-Hall.

Jaeger, R. M. (1995). Setting performance standard through two-stage judgmental policy capturing. *Applied Measurement in Education, 8,* 15-40.

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. 53-88). Mahwah, NJ: Erlbaum.

Smith, R. W. & Ferdous, A. (2007, April). Using a subset of items for a modified Angoff standard setting study – An automated item selection procedure. Paper presented at the annual meeting of the National Council on Measurement in Education. Chicago, IL.

Wang, N. (2003). Use of the Rasch IRT model in standard setting: An item-mapping method. *Journal of Educational Measurement, 40,* 231-252.

Citation

Buckendahl, Chad W., Ferdous, Abdullah A. & Gerrow, Jack (2010). Recommending cut scores with a subset of items: An empirical illustration. *Practical Assessment, Research & Evaluation*, 15(6). Available online: http://pareonline.net/getvn.asp?v=15&n=6.

Note:

This research was sponsored by the National Dental Examining Board of Canada (NDEB). The authors are grateful to the NDEB and board members for their support.

Corresponding Author:

Chad W. Buckendahl
Alpine Testing Solutions
2467 Cordoba Bluff Ct.
Las Vegas, NV 89135
Email: chad.buckendahl [at] alpinetesting.com
Office: (702) 586-7386