

**Running Head: An Adaptive Procedure for Standard Setting**

An Adaptive Procedure for Standard Setting and A Comparison with Traditional Approaches

Robert W. Lissitz & Marc H. Kroopnick

University of Maryland, College Park

As described in Zieky (2001), standard setting procedures have evolved over the past 60 years or so. Early on, in the 1950s, cutscores were based on accepted percentages correct somewhere between 60 and 75 percent (Zieky, 2001). Conceptually, as Ebel (1965) notes and Zieky (2001) highlights these percentages correct can be considered the proportion of perfection a minimally proficient examinee, for example, must achieve. Since then, however, there have been many systematic methods proposed and implemented that involve experts making decisions or judgments based on the items included on the exam or the examinees actually taking the exam. One of the most popular early procedures is the Angoff (1971) procedure. Rather than relying on what experts simply think is a reasonable percentage correct to justify, for example, minimal proficiency, this method requires a standard setting panelist to judge each item individually as to whether or not he/she would expect a minimally proficient examinee to answer it correctly. This classical test theory based procedure is described in more detail later in this paper, but generally it is the aggregation of these item judgments that result in a cut score. As measurement models become more sophisticated, so do many standard setting methods. These methods would include the Bookmark (Lewis, Mitzel, and Green; 1996) and the Item Sorting Method (Sireci et al., 2000); both are based on Item Response Theory and capitalize on having item difficulty and person ability on the same latent dimension. To varying degrees, these methods require individual item judgment just as in the Angoff procedure.

Standard setting is taken very seriously by the National Council on Measurement in Education, the American Educational Research Association, and the American Psychological Association. Accordingly, they have outlined standards for the establishment of cutscores in *Standards for Education and Psychological Testing* (1999). While standards relating to

cutscores exist throughout the book, cutscores are explicitly discussed in standards 4.19 through 4.21. Generally, they require documentation for the rationale and procedures for establishing cutscores, the use of sound data to define and justify the appropriateness of different categories, and the use of procedures that allow judges to appropriately apply their knowledge and experience (AERA, APA, NCME, 1999). Clearly the establishment and application of cutscores have come a long way since the 1950s when a universally accepted percent correct was simply used.

Standard setting can be a very time consuming and expensive endeavor, but it is necessary in today's educational climate especially in the context of No Child Left Behind (NCLB). As such, it is important to develop and implement standard setting procedures that are both justifiable and efficient. This study compares the results of two very common and arguably the most popular standard setting procedures, the Angoff (1971) and the Bookmark Procedures (Lewis, Mitzel, and Green; 1996), with a new adaptive approach. The context for this comparison of these three techniques is a fixed form, multiple choice, high school mathematics exam that was actually used in a state assessment program approximately 17 years ago. The purpose of this study is to examine whether the new adaptive approach has the potential to work as well as or even better than the other procedures, whether it can be more efficient (i.e. a cutscore is agreed on more quickly with equal or greater consistency), and if the panelists' satisfaction with each procedure suggests that the new approach is worthy of more systematic development.

## Description of Procedures

*Angoff Procedure*

The Angoff procedure (1971) is a classical test theory based approach that has several variations. One variation of Angoff's original idea has been implemented where each panelist decides whether or not a Minimally Proficient Examinee (MPE) would be expected to answer each item correctly; the count of the items that the panelist expects the MPE to answer correctly would be the cutscore. This variation of Angoff's original idea will be employed in this study; Impara and Plake (1997) indicate that this approach yields similar results to the probability estimation implementation and is easier for the panelists. This procedure can also allow panelists to discuss their results and make revisions.

*Bookmark Procedure*

The Bookmark procedure, as described by Lewis, Mitzel, and Green (1996), is an IRT based approach where items on a test are ordered into a booklet by location on the latent continuum. Panelists are required to place a "bookmark" between the most difficult item a MPE would be expected to answer correctly and the easiest item a MPE would be expected to answer incorrectly. Typically, panelists work with items from an operational exam. When a 1PL model is employed, as will be the case here, a .5 probability of success rate can be used (Wang, 2003), although that is a subject of some debate (see, for example, Huynh 2006). This procedure is iterative and allows for revision and discussion. Since item difficulties are on the same scale as person ability, the location of the bookmark can be easily translated to a cutscore on the score reporting scale (Lewis et al., 1996).

*The “New” Adaptive Procedure*

This approach is also IRT based and we will assume a 1PL model for this trial. Each panelist will evaluate an initial item as to whether or not he or she expects a MPE to answer it correctly. Picking an initial item is a decision that can and should be researched further, but for this study we used one with a p-value of approximately .7. We chose this value because we believed it to be in the region of likely choice for the MPE cut-off. In a very general sense, this procedure will iterate like a computer adaptive test (CAT). If the panelist indicates that he or she expects the MPE to answer the item correctly, the panelist will be given an item with a higher difficulty parameter than the initial item. If he or she does not expect the MPE to answer the item correctly, an item with a lower difficulty parameter will be given. In the present trial, the procedure will be operationalized by human assistance. An optimal realization of this approach would involve using a computerized system for item administration.

The procedure will continue until the panelist converges, based on some stopping rule or some tolerance level, on a single difficulty parameter. This procedure can also allow for discussion amongst the panelists and an opportunity for revision. The motivation for this procedure centers on efficiently using the panelists' time. When considering just a single performance standard, it does not make sense for the panelists to evaluate items that will not be discriminating for examinees near the associated cutscore. This procedure allows the panelist to focus attention on the items near the cutscore and since this procedure is implemented in an IRT framework, the item's location on the ability scale can be easily translated to a total correct score. Given that standard setting procedures are typically time consuming and expensive, this method may represent a way to limit both factors. Note that a procedure similar to this one was suggested independently by both Howard Wainer in personal communication with Sireci and

Clauser (2001) and Walter Way in personal communication with Zieky (2001) in the context of computer adaptive testing. Our application of this approach, however, is to fixed length, non-adaptive testing. Also, all three approaches could utilize impact data presented during an interim discussion period, but the nature of this trial does not permit such a step. We would expect that doing so would improve the stability of the standard setting process, although this is an empirical question.

### Implementation

A single standard of proficient or not proficient was set for the multiple choice test described below using each of the three procedures described above. Each group of panelists participated in a discussion of what is required of a MPE. The panelists then broke into groups according to their assigned procedure for specific instructions on conducting the standard setting according to their procedure. Each standard setting procedure included two iterations. The second iteration commenced after the panelists discussed their results from the first round. No impact data were provided. At the conclusion of each procedure, the panelists completed a survey that allowed them to reflect on their experiences and to express their level of satisfaction with the standard setting process and their assigned procedure.

#### *Panelists*

Sixty-one students from two introductory classroom assessment courses served as the panelists and were randomly assigned to the procedure groups. Class one included 34 students and class two included 27 students. The vast majority of students were junior and senior female majors in early childhood and elementary education. Specifically, the enrollment in both classes was over 90% female and over 65% education majors. All 61 students had at least junior level

class standing. Please note that class one enrolled 8 graduate students and class two enrolled 1 graduate student. The majority of these graduate students are part of a graduate program in measurement. The students had been enrolled in the assessment course for approximately nine weeks and the course precedes the practicum experience in the schools for the undergraduate education majors. This study was conducted independently with each class. Note that each class may agree on a different definition of minimally proficient, so comparing results across classes must be approached cautiously.

#### *Test Data*

Dichotomous response data from the Maryland Functional Math Test (MSDE, 1990) was calibrated according to the 1PL model. These items are at a level such that the student panelists should feel competent to evaluate the items.

#### *Facilitators*

Nine graduate students in the measurement (EDMS) program assisted in this experiment. There were male and female students involved and they were trained by the second author in two sessions prior to the trial in each class.

### Administration of Standard Setting Procedures

Please find below the instructions given to the standard setting facilitators and explained to the panelists for each procedure. These instructions were discussed with all facilitators at their training sessions prior to conducting the study.

#### *Angoff Procedure*

- Panelists will be presented a test form and make pass/fail decisions about each item.

- Each panelist will score each item with a 1 if he or she expects a MPE to answer the item correctly and score it 0 otherwise.
- The cut score determined by each panelist will be the sum of the item scores.
- Panelists will discuss their cut scores and share their reasoning. After this discussion panelists will engage in a second round of judgments.
- The sum of these final item scores will then be converted to a theta score using a test characteristic curve.
- The final cutscore will be the median of the panelists' scores.

#### *Bookmark Procedure*

- Panelists will be presented a booklet with item order based on the estimated 1PL IRT difficulty parameters.
- Panelists will then be instructed to read through the entire exam and place a bookmark just below the hardest item they expect a MPE to answer correctly with at least .5 probability. We will use the point where this item lies on the latent continuum as the cutscore determined by the panelist.
- Panelists will discuss the reasoning behind their bookmark decisions and revise.
- After each panelist has placed their final bookmark, the median difficulty parameter across all panelists will be taken as the final cut score for the group.

### Adaptive Procedure

- Each panelist will be asked to evaluate the same item (with approximately .7 p-value) as to whether or not he or she thinks a minimally proficient student will have more than a 50% chance of answering it correctly.
- If the answer is yes, the next item that the student will receive will be 5 items harder than the previous one. Count up to the fifth item from that first one.
- If the answer is no, the next item that the student will receive will be 5 items easier than the previous one. Count down to the fifth item from that first one.
- If the panelist's decision to the second item is the same as the first, advance five more items in the appropriate direction (if possible). It could be the case that due to the finite number of items, it will not be possible to advance five items. So, just advance as much as possible in that case. Keep doing this until there is a disagreement between the new decision and the previous decision. If a panelist maintains the same decision across all items presented, then the item parameter associated with that panelist will be at one of the extremes of the test.
- Once there is a discrepancy between the panelist's decisions, hopefully we've found an approximate area on the latent continuum where that panelist thinks the minimally proficient student lies. So, the next item given will be as follows based on their answer to the previous item:
  - If the panelist says "NO", provide an easier item 4 away from the most recent item.
  - If the panelist says "YES", provide a harder item 4 away from the most recent item.

FOR THE NEXT ITEM

- If the panelist says “NO”, provide an easier item 3 away from the most recent item.
- If the panelist says “YES”, provide a harder item 3 away from the most recent item.

FOR THE NEXT ITEM

- If the panelist says “NO”, provide an easier item 2 away from the most recent item.
- If the panelist says “YES”, provide a harder item 2 away from the most recent item.

FOR THE NEXT ITEM

- If the panelist says “NO”, provide an easier item 1 away from the most recent item.
- If the panelist says “YES”, provide a harder item 1 away from the most recent item.
  
- If you are ever in a circumstance where the directions indicates that you should provide an item that the panelist has already seen, provide the next closest “never been seen before item” moving in the appropriate direction away from the most recent item.

- After this procedure is carried out once, there will be a group discussion. After the group discussion, we will repeat this procedure using a different subset of items (starting with an item that represents the median ending item parameter estimate based on the group's decisions from the first iteration of this procedure).

## Results

### *Initial reactions and defining the standard*

When presented with the task, the panelists in both classes were enthusiastic about participating. The introduction to the task involved a brief description of the importance of standard setting, the specific test under consideration, performance standards, and cutscores. After the introduction, the panelists were given a copy of the test for which they would be determining the cutscore for the minimally proficient performance standard. Once the panelists, had a chance to review the test, a discussion of the definition for the minimally proficient performance standard ensued. The goal of this discussion was to yield a consensus of what would be required of a minimally proficient high school math student. Despite the facilitators, introductory presentation, and population from which the panelists were drawn (both introductory classroom assessment classes serve the same population of students) remaining constant in both classes, the definition for minimally proficient differed to a larger extent than expected. The authors could have forced a common consensus on definitions but did not. In a real, rather than simulated, state standard setting activity, the leaders would usually force the participants (e.g., in different grades) to a common consensus.

The first group (class 1) determined their definition of a minimally proficient student rather quickly and with very little contention among the panelists. However, they focused on

specific tasks rather than using general words such as “functions at a basic level.” They decided the following would be required of a minimally proficient student:

- Addition, subtraction, multiplication, division of whole numbers
- Addition, subtraction, multiplication, division of decimals
- Read and answer questions from graphs and charts
- Handle money and make change
- Rank numbers from least to greatest
- Tell time; understand time

The group also discussed the following, but did not agree to include it in their definition of minimally proficient:

- Understanding word problems; particularly mapping to appropriate operation
- Shop, calculate tips, sale prices (percentages off)

The second group (class 2) had a more difficult time generating and agreeing on a definition for minimally proficient. They focused on cutscores rather than the performance standard and spent most of their time discussing numbers instead of the words to describe those numbers. It was necessary for the moderator to shift the discussion to performance standards. Many in this group thought the test was entirely too easy for a high school student and that successfully answering all items should be required. Some also asked if it would be possible to see a student’s work and give credit based on sound reasoning. When the discussion concluded, the group agreed on the following definition for minimally proficient (note that their definition is more general than the one obtained in the first group):

- Perform simple arithmetic
- Have a basic concept of algebra

- Balance a checkbook
- Comprehend graphs and charts and be able to answer corresponding questions

*Between the two iterations*

The three standard setting procedures were very smoothly implemented in both classes. It is important to note that the time to complete all of the procedures was roughly the same for both classes across all procedures. This could be, perhaps in part, due to the consistency of the facilitators and the panelists surreptitiously observing other groups and wanting to keep pace.

As noted earlier, each procedure included a discussion between iterations. The content of those discussions is briefly outlined below (as obtained directly from field notes):

Class 1 Angoff

- Spoke about personal experiences
- Asked how guessing should be factored in
- Inquired about standards in different countries (some panelists were not from the USA)
- How to distinguish between right and wrong and the idea of a 50% chance of answering an item correctly
- Noted that it was difficult to determine what content should be required of high school students

Class 2 Angoff

- Expressed some sympathy for students taking this test
- Noted that students could be able to answer some items correctly in different contexts
- Asked about the carelessness factor
- Noted that concentration is more important, since answers are already there

#### Class 1 Bookmark

- It was necessary to stop the students from working together initially
- Presented the items around each panelists' bookmark and discussed how they related to the performance standard
- Some students were more outspoken than others

#### Class 2 Bookmark

- Spent time discussing whether or not adding and subtracting fractions without common denominators was necessary for the minimally proficient student; the group decided it was not

#### Class 1 Adaptive

- Everyone said that they would think all examinees should answer all items correctly
- Some spoke about testing in a real life situation
- Noted that problems that included fractions with a common denominator were easier

#### Class 2 Adaptive

- Some disagreement about addition and subtraction of mixed numbers with same denominator
- Questioned how answer options should influence procedure
- Questioned whether or not students should be able to answer practical problems like making change and telling time

#### *Cutscores*

No matter the procedure, each panelist provided a cutscore before and after the group discussion (i.e, after each iteration). For some procedures the cutscore was in raw score units (Angoff) and in other procedures the cutscore was on the theta scale (Bookmark and Adaptive).

Since a Rasch model was used, the raw total score corresponds to exactly one theta value and vice versa. The test characteristic curve, found in Appendix A, was used to map this relationship. A chart with the mapping from raw score scale to theta scale can be found in Appendix B. Aggregated results for the 3 procedural groups are reported on both scales and can be found below:

-----  
Insert Table 1 about here  
-----

-----  
Insert Table 2 about here  
-----

Given the small sample size (there were no more than 34 panelists in a class and no more than 13 panelists assigned to any one procedure), the authors decided it would not be productive

to conduct inferential statistical analyses and draw conclusions from such significance testing. However, some of the descriptive statistics are quite suggestive.

1. Notice that in class one, for the Angoff and adaptive procedures, the standard deviation of cutscores decreased on the second iteration. The increase in standard deviation for the Bookmark procedure was small; it was less than 1 raw score point.

2. For class two, in all cases the standard deviation increased from the first to the second iteration.

3. As far as the resulting cutscores are concerned, there was no consistency across procedures within a given class. Sometimes the average cutscore increased from iteration 1 to iteration 2 and sometimes it decreased.

4. The very large change, compared to the other procedures, from iteration 1 to iteration 2 for the adaptive procedure in class two is somewhat disturbing. Perhaps, the students in that group got lazy and/or bored towards the end of their session.

5. Also, the cutscores spanned the score scale in both classes however, the similarity of the cutscores at iteration 2 for the Angoff and adaptive procedures in class one is encouraging.

The first class seemed to take the activity more seriously and, as noted earlier, was more comfortable with their definition of minimally proficient than class two. This difference between the classes may be related to the decrease/stability of the standard deviation of cutscores from iteration 1 to iteration 2 in class one and the standard deviation increase in class two. Perhaps, this finding suggests that the discussion between iterations was more useful in class one.

### *Survey Results*

A copy of the survey administered at the end of the standard setting procedure can be found in Appendix C of this document and the aggregated results are contained in Appendix D.

Generally, the survey provided a way to obtain the panelists' reaction to the various components of the standard setting process such as: the definition of minimally proficient, training on the standard setting procedure, the group discussion between iterations, and the importance of some of these components in the decision making process. The survey used a 1-5 Likert rating scale.

Again, given the small sample numbers of panelists assigned to each procedure, we decided not to conduct statistical analyses, but the descriptive data are still suggestive. Some of the aggregated results from this survey are highlighted here:

1. While the field notes indicate that class one was more comfortable with their definition of minimally proficient, the survey did not indicate that the participants perceived an appreciable difference between their classes
2. The panelists assigned to the adaptive procedure were less comfortable with their understanding of the procedure compared to the comfort levels assigned to Angoff or Bookmark procedures
3. For all procedures, the definition of minimally proficient and the panelists' perceived difficulty of the items were more important than the group discussion in their decision making process.

The survey also allowed for students to suggest ways to improve the process. The following were the general themes of the responses:

1. Spend more time on the development of the definition of minimally proficient and provide examples.
2. Provide impact data
3. Provide more explanation on how the adaptive procedure works to yield a cutscore

### *General Conclusions*

While the comparison of the results of the three standard setting procedures in each class was generally inconclusive, this study, to a very large extent, served as a proof of concept for the potential usefulness and implementation of the adaptive standard setting procedure. Perhaps, the biggest highlight of the study was the stability/decrease of the standard deviation of cutscores from iteration 1 to iteration 2 and the consistency between the Angoff and adaptive cutscores after iteration 2 in class one. In this class, the adaptive procedure yielded the largest reduction in standard deviation. Note that this class was more serious about the task and agreed on a more focused definition of minimally proficient.

Generally, however, the results of the adaptive procedure were not superior to the other two more traditional approaches. One could reasonably say the results from the adaptive procedure were just as good (or as bad) as the other approaches, though. The implementation of the procedure as outlined in this paper proceeded very smoothly with very few, if any, problems. Of course, the use of computers for administration would allow for a more sophisticated implementation and might result in an even more efficient process than the traditional approaches.

Additionally, it is important to note that the entire standard setting procedure was conducted in less than 3 hours using college students who, despite being capable of answering all the items correctly, were not necessarily prepared nor qualified to make the judgments asked of them. The panelists used in this study represented the best possible convenience sample we could obtain on a college campus, given our absence of funding. Certainly, the quality of the panelists and the time the researchers were able to invest in training them affected the results of

this study. We believe that, particularly important, would have been more time spent on the determination of the definition of the cutscores.

### *Future Research*

The procedure and implementation of the study was reasonably sound, but the quality of the data was poorer than expected. Therefore, the next step for this study would be to make an effort to increase the quality of the data. To do this the following is suggested:

- Use panelists qualified to discuss and make judgments regarding minimal high school mathematics proficiency. This would include teachers and other stakeholders similar to those who are customarily included in such a process for state standard setting.
- Allow a more reasonable amount of time to conduct the study; perhaps, a day or two would be sufficient and would lead to better results.
- If possible, conduct the adaptive procedure on a computer. Panelists would be instructed to complete the test as if they were a minimally proficient high school mathematics student. Essentially, the standard setting process would treat the fixed form exam as if it were a CAT. But, note that the cutscore obtained in this manner would be applied to a fixed form exam and would be comparable to the traditional Angoff and Bookmark procedures that are currently used to set the cutscore on a fixed form exam.

References

- American Educational Research Association, American Psychological Association, & National Council of Measurement in Education (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.
- Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2 ed., pp. 508-600). Washington, DC: American Council on Education.
- Ebel, R. L. (1965). *Measuring educational achievement*. Englewood Cliffs, NJ: Prentice-Hall.
- Huynh, H. (2006). A clarification on the response probability criterion RP67 for standard setting Based on bookmark and item mapping. *Educational measurement: Issues and practices*, 25(2), 19-20.
- Impara, J. C., & Plake, B. S. (1997). Standard setting: An alternative approach. *Journal of Educational Measurement*, 34 (4), 353-366.
- Lewis, D. M., Mitzel, H. C., & Green, D. R. (1996, June). Standard setting: A Bookmark approach. In D. R. Green (Chair), *IRT-based standard setting procedures utilizing behavioral anchoring*. Symposium conducted at the Council of Chief State School Officers National Conference on Large-Scale Assessment, Phoenix AZ.
- Maryland State Department of Education. (1990). Maryland functional math test II: Technical report.
- Sireci, S. G. & Clauser B. E. (2001). Practical issues in setting standards on computerized adaptive tests. In G. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 355-369). Mahwah, NJ: Erlbaum.

- Sireci, S. G., Patelis, T., Rizavi, S., Dillingham, A., & Rodriguez, G. (2000). Setting Performance standards on the ACCUPLACER elementary algebra test (Laboratory of psychometric and evaluative research, rep. No. 368). School of Education, University of Massachusetts, Amherst, MA.
- Wang, N. (2003). Use of Rasch IRT model in standard setting: An item mapping method. *Journal of Educational Measurement, 40*, 231-253.
- Zieky, M. J. (2001). So much remains has changed: How the setting of cutscores has evolved since the 1980s. In G. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 53-88). Mahwah, NJ: Erlbaum.

Table 1

Total cutscore and theta cutscore for Class 1

total score						
class 1	Angoff iter. 1	Angoff iter. 2	Bookmark iter. 1	Bookmark iter. 2	Adaptive iter. 1	Adaptive iter. 2
median	46.00	45.00	35.00	27.00	38.00	45.50
average	47.46	45.62	32.45	30.09	42.30	47.30
standard deviation	7.13	5.35	5.65	6.52	9.27	6.73
average change (first - second)	1.85		2.36		-5.00	
N	13		11		10	

theta						
class 1	Angoff iter. 1	Angoff iter. 2	Bookmark iter. 1	Bookmark iter. 2	Adaptive iter. 1	Adaptive iter. 2
median	0.83	0.75	0.11	-0.51	0.33	0.85
average	0.99	0.82	-0.11	-0.27	0.65	1.02
standard deviation	0.63	0.41	0.41	0.46	0.75	0.58
average change (first - second)	0.18		0.16		-0.37	
N	13		11		10	

Table 2

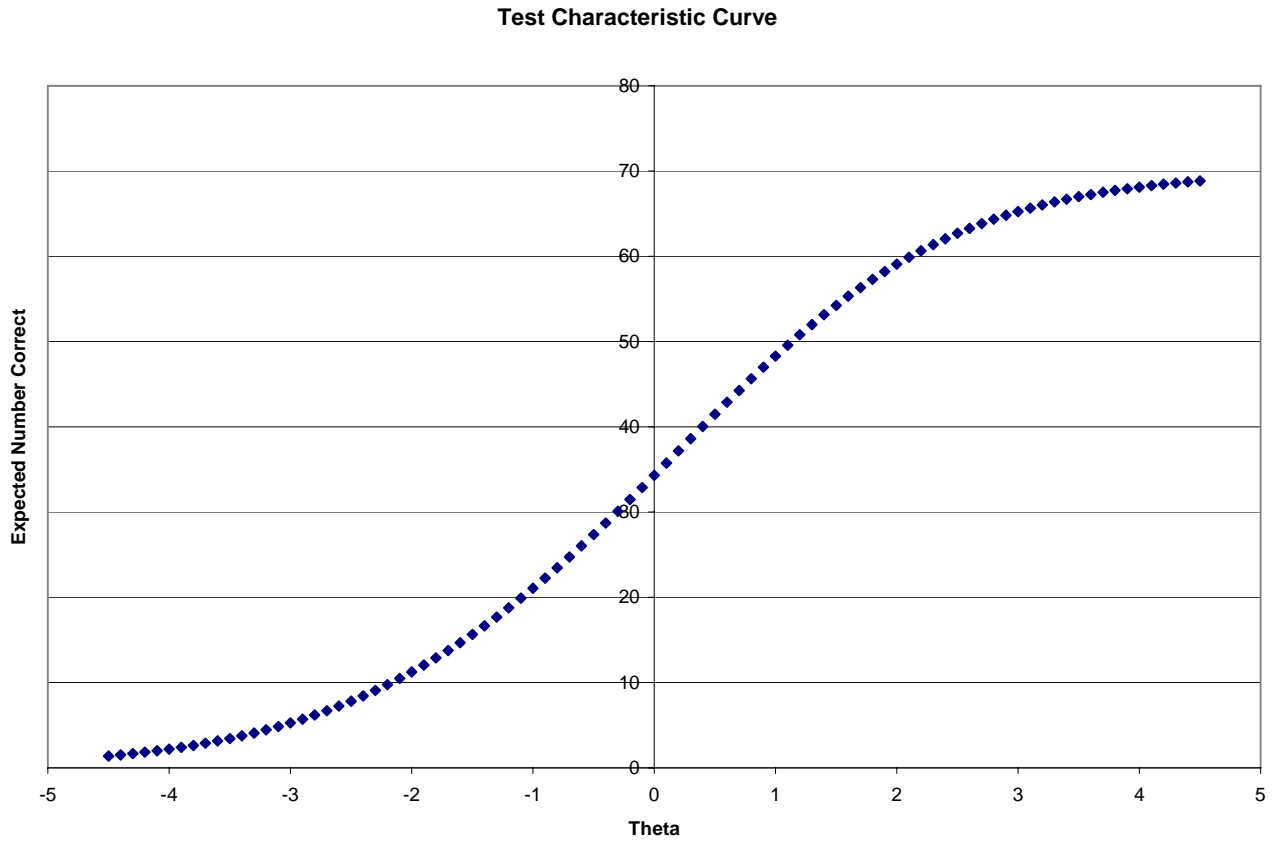
Total cutscore and theta cutscore for Class 2

total score						
class 2	Angoff itr. 1	Angoff itr. 2	Bookmark itr. 1	Bookmark itr. 2	Adaptive itr. 1	Adaptive itr. 2
median	59.00	53.00	37.00	38.00	39.00	49.00
average	57.44	55.89	37.78	41.00	41.56	52.22
standard deviation	5.32	6.57	4.71	6.56	5.39	5.59
average change (first - second)	1.56		-3.22		-10.67	
N	9		9		9	

theta						
class 2	Angoff itr 1	Angoff itr 2	Bookmark itr 1	Bookmark itr 2	Adaptive itr 1	Adaptive itr 2
median	1.99	1.39	0.24	0.29	0.35	1.08
average	1.90	1.84	0.27	0.54	0.55	1.44
standard deviation	0.60	0.95	0.32	0.50	0.39	0.53
average change (first - second)	0.06		-0.27		-0.88	
N	9		9		9	

Appendix A

Test Characteristic Curve:



## Appendix B

Mapping from Raw Score Scale to Theta Scale:

total correct	corresponding theta
0	-6.87
1	-4.84
2	-4.1
3	-3.65
4	-3.33
5	-3.06
6	-2.84
7	-2.64
8	-2.47
9	-2.31
10	-2.17
11	-2.03
12	-1.91
13	-1.79
14	-1.67
15	-1.57
16	-1.46
17	-1.37
18	-1.27
19	-1.18
20	-1.09
21	-1
22	-0.92
23	-0.84
24	-0.76
25	-0.68
26	-0.6
27	-0.53
28	-0.45
29	-0.38
30	-0.31
31	-0.23
32	-0.16
33	-0.09
34	-0.02
35	0.05

total correct	corresponding theta
36	0.12
37	0.19
38	0.26
39	0.33
40	0.4
41	0.47
42	0.54
43	0.61
44	0.68
45	0.75
46	0.83
47	0.9
48	0.98
49	1.06
50	1.14
51	1.22
52	1.3
53	1.39
54	1.48
55	1.57
56	1.67
57	1.77
58	1.88
59	1.99
60	2.12
61	2.25
62	2.39
63	2.55
64	2.73
65	2.94
66	3.19
67	3.5
68	3.94
69	4.66
70	5.88





Appendix D

Results of Panelist Survey by Class and Standard Setting Procedure:

Question
<i>Q1: What did you think of the adequacy of our definition of <b>minimally proficient</b>?</i>
<i>Q2: What did you think of the adequacy of the <b>training</b> on your procedure?</i>
<i>Q3: Did you feel that you really understood the <b>process for setting the cut score</b> in your group?</i>
<i>Q4: What did you think of the usefulness of the <b>group discussion</b> with those assigned to your procedure?</i>
<i>Q5: How confident were you with <b>performing the standard setting task</b>?</i>
<i>Q6: How important in making your decision was the Description of <b>Minimally Proficient</b>?</i>
<i>Q7: How important in making your decision was Your <b>Perceived Difficulty</b> of the Items?</i>
<i>Q8: How important in making your decision was The <b>Group Discussion</b>?</i>

	Class 1 Angoff	Class 2 Angoff	Class 1 Bookmark	Class 2 Bookmark	Class 1 Adaptive	Class 2 Adaptive
<i>Q1</i>	3.31	3.22	2.64	3.22	3.80	3.22
<i>Q2</i>	3.58	3.44	3.45	3.22	3.50	2.89
<i>Q3</i>	3.23	3.89	3.18	3.33	2.60	2.56
<i>Q4</i>	3.77	3.44	3.64	3.67	2.80	3.22
<i>Q5</i>	3.38	3.78	3.00	3.00	3.40	3.56
<i>Q6</i>	4.58	4.11	4.27	3.89	3.80	4.89
<i>Q7</i>	4.42	4.22	3.55	4.11	4.20	4.67
<i>Q8</i>	3.42	3.67	3.64	3.78	3.30	3.22