

When adaptation is not an option: An application of bilingual standard setting

Susan L. Davis

Chad W. Buckendahl

Barbara S. Plake

Buros Center for Testing  
University of Nebraska – Lincoln

Paper presented at the International Test Commission Conference in Brussels, Belgium  
(July, 2006).

Correspondence concerning this paper should be directed to:  
Susan L. Davis  
Buros Center for Testing  
sdavis7@unl.edu

## Abstract

As an alternative to test adaptation, tests can also be developed simultaneously in multiple languages. These tests may be distinctly different, even though they are used for the same intended purpose. In such a situation, one challenge is setting passing standards for each exam that results in the same meaning of the scores. This paper describes an application of a standard setting process used for a bilingual high school literacy assessment that was constructed under these conditions. This methodology was designed to address the specific challenges presented by this testing program including maintaining equivalent expectations for performance across different student populations. The validity evidence collected for this method will be discussed along with recommendations for future practice.

## When adaptation is not an option: An application of bilingual standard setting

### Introduction

Tailored suits are made to fit a person precisely; therefore, no two are exactly alike. Similarly, testing programs are unique in that they are designed to fit the specific characteristics of a population in a way that measures the intended constructs as precisely as possible. This customization process includes accommodating variations in the language characteristics of the population by making the test applicable to the intended population of interest. Many times, an existing test is translated and/or adapted from an original language to a second or third language to accommodate different sub-populations within a population. However, this often occurs as an afterthought and the adaptation process may result in significant and substantial changes to the intended properties of test items (e.g., content, difficulty, meaning) due to nuances of the second language or characteristics of the second population.

An alternative to the adaptation process is to simultaneously develop tests for multiple populations or sub-populations. This approach presents its own unique challenges including that of setting performance standards. There exists substantial research on various methods for setting performance standards (e.g., Angoff, 1971; Cizek, 2001; Impara & Plake, 1997; Jaeger, 1989; Livingston & Zieky, 1982; Plake & Hambleton, 2000). This literature introduces new methods, variants on existing methods, and ways to evaluate standard setting methods (e.g., Hambleton, 2001; Kane, 1994, 2001). Despite the wealth of literature on standard setting in general, very little research exists on the comparability of passing standards on similar tests in different languages that are intended for the same purpose. The purpose of this paper is to discuss the issues involved in setting performance standards for a bilingual testing program that chose not to translate or adapt the base form of the test and an application of one method for doing so.

The *Test Adaptation Guidelines* (ITC, 2000) suggest “The test developer should provide specific information on the ways in which the socio-cultural and ecological context of the populations might affect performance on the test, and should suggest procedures to account for these effects in the interpretation of results” (guideline I.4). Interpreting scores for tests that are available in multiple languages may be based on one standard for performance or separate performance standards for each version of the test. This likely depends on the intended use of the scores.

This paper describes a study conducted for a testing program that required students within the same geographic region to demonstrate their literacy skills (Reading and Writing) as a high school graduation requirement. Two sub-populations of students are the focus of the study. One subpopulation receives instruction and speaks English as their primary language. The second sub-population receives instruction and speaks French as their primary language. The literacy expectations as defined by the policy are the same for each group of students (English and French); however, the two versions of the test are not translations, but rather separate tests. Both versions were created from the same test framework and test specifications but the test items are different. Alignment studies were conducted to ensure that each assessment matched the intended framework and specifications.

The unique characteristics of this testing program require special consideration when attempting to set passing standards for each version of the assessment. Given that neither version is a translation/adaptation of the another, one cannot assume that a single passing standard would have the same meaning between the tests. However, because the policy-defined expectations for passing are the same for each sub-population of students, one would want the standard setting process to result in recommended passing standards that uphold these expectations across the two assessments. Strategies considered in the design of this standard setting study considered guidelines from current literature on

evaluation frameworks for standard setting methods (e.g., Hambleton, 2001; Kane, 1994, 2001).

Our goal in this process was to design a standard setting process that would produce recommended passing standards - unique to each sub-population - that held students to the same expectations regardless of the respective language sub-population. Therefore, we utilized research on evaluation of standard setting methods to identify the critical elements that are relevant to the validity of a standard setting, specifically focused on the process used to estimate the standard. Suggestions were taken from Kane (1994), Hambleton (2001), and the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, and the National Council on Measurement in Education, 1999).

Kane (1994) suggests three elements to consider when assessing the validity evidence of passing standards. The first component of this evaluation is procedural validity evidence, which pertains to the appropriateness of the standard setting method and how it was applied to the testing program. The other two components of Kane's framework include internal validity evidence (reliability of the results) and external validity evidence (classifications consistent with other information about student ability). Although these latter two elements are important to assessing the validity of a set of passing standards, this paper will focus on the primary source of evidence available in this study: procedural evidence of validity.

Hambleton (2001) presents a list of 20 questions designed to evaluate the quality of a standard setting process that cover topics including selection of the method, selection of panelists, and implementation of the method. The *Standards* (1999) provide suggestions for sound processes when setting performance standards. The method described in this paper was designed to keep as many of these elements as possible the same in the dual processes used to determine the individual passing standards.

### *Standard Setting*

The first step in a standard setting process is to select a method. No single method is appropriate for all situations; rather, one must weigh the advantages and disadvantages of various methods with respect to a particular assessment (Kane, 1994). Hambleton (2001) notes the importance of choosing a method that is appropriate to the assessment and one that has been field tested before operational use and the *Standards* (AERA, APA, NCME, 1999) require adequate documentation of the process used.

The second step is choosing the panel of judges that will provide information used to estimate the recommended passing standard. It is critical that a panel is qualified to make judgments about the expectations of student performance, familiar with the target population (Kane, 1994), represents various constituencies (Hambleton, 2001), and is of adequate size to produce acceptable standard errors (around the cut scores; Kane, 1994).

An important component of any standard setting process is the training of the panelists to prepare them for making estimates about student performance. Some elements of training include informing panelists of the purpose of the assessment, the intended use of the test scores, providing panelists with descriptions of the performance categories, allowing panelists to experience the assessment to become familiar with it, and training with the method that includes a practice activity (Hambleton, 2001; Kane, 1994).

With respect to the operational standard setting, panelists should be instructed to engage in a systematic activity where they can apply what they have learned in training, as well as their own expertise, and provide independent judgments about expectations for student performance. Kane (1994) and Hambleton (2001) both discuss the use of an iterative process where panelists are given more than one opportunity to provide ratings. Often, after the first round of independent ratings, panelists are provided feedback about their estimates as well as the estimates of the entire panel. Kane (1994), Hambleton (2001), and the *Standards* (AERA, APA, & NCME, 1999) note the importance of

supplementing this feedback with data on student performance (e.g., p-values for items) and impact/consequential data that shows the hypothetical classification of students into performance categories if the first round cut score estimates were used. As a final step in the operational standard setting process, Hambleton (2001) and Kane (1994) suggest collecting feedback or evaluation data from panelists to assess their perception of the process and confidence in their ratings.

The critical elements of a standard setting process highlighted by Kane (1994), Hambleton (2001), and the *Standards* (1999) were considered when the standard setting process and method were designed for the bilingual testing program described in this paper. In the subsequent sections, we describe the specific process used and evaluate the parallel processes against the criteria described above. In the final section, we provide recommendations for future practice.

## Method

### *Process*

As described above, this test was designed to document the literacy skills of high school students in a bilingual educational setting. The literacy test consists of Reading and Writing components and students must meet a passing standard that represents a compensatory score for the combination of these two components. The Reading portion includes selected response items, short constructed response items, and extended constructed response items. The Writing portion consists of four extended constructed response items. In this educational setting, some schools are English speaking whereas the others are French speaking. The literacy test was created in both an English version and a French version from the same content framework and test specifications; however, the items are different (i.e., items were not translated from one language to another). The goal of the standard setting workshop was to recommend a passing standard for each version (e.g., pass/fail cut point). These recommendations were then to be presented to a policy board that would determine the exact passing score for each version of the assessment.

Panels were recruited from both English and French speaking schools. Each panel was comprised of 20 experts with experience in their respective language. The panels were selected in a way to have coverage of the geographic region (including rural and urban schools) and the different types of schools. The panelists were teachers and administrators who had experience with the subject matter and familiarity with the target population. The English panelists had an average of 17 years of experience and the French panelists had an average of 16 years of experience.

The standard setting process took two days. This standard setting was designed in a way to maximize the similarity between the processes and reduce any facilitator effect, and therefore, many of the components of the process were conducted with panelists in one large group (both English and French). However, some components did have to be conducted in language-specific groups. The process is described below and detailed in Figure 1. As shown in the Figure, some components of the standard setting process were conducted with one large group (N=40) whereas others were conducted in smaller, language-specific groups (N=20).

The first day began with an orientation conducted in both English and French. Panelists were welcomed and informed of the purpose of the standard setting process and the schedule for the meeting. A team of two facilitators led this meeting and facilitated the subsequent training session; one facilitator spoke English, the second facilitator was responsible for conducting the orientation and training in French but also spoke English fluently. Panelists were gathered in one large room for the orientation and training and each was given a packet of materials (e.g., copy of the content standards, orientation and training PowerPoint) in their native language. The two facilitators stood in the front of the room with dual presentation screens behind them: one screen displaying content in English, the other in French.

Translators were seated in a booth in the back of the room and their translations of the orientation and training were made available to the panelists via individual headphones. The orientation was presented in sections, with each section being presented by the English facilitator first (simultaneous French translation available via the headphones) followed by the same section presented by the French facilitator (simultaneous English translation available via the headphones). Therefore, panelists had the opportunity to hear the orientation presentation twice: once from their language facilitator, and a second time as a translation of the other language facilitator. In addition, as the panelists asked questions or held discussions with the facilitator in front of the large group, translations were provided via the headphones.

The second part of the orientation was a discussion among panel members about the skills and competencies of the “target student”. In this particular standard setting process, the target student was one whose work was deemed to be just barely acceptable in literacy (e.g., student who would be considered just barely passing). For this activity, panelists divided into small groups by domain (Reading and Writing) and by language (English and French). It was necessary at this point to separate the groups by language so the discussion could flow smoothly without the interruption of translation. Following the small group discussions, the panelists reconvened in one large group (both Reading and Writing, English and French) and each small group presented their panel’s ideas of the skills and competencies necessary to be considered passing for Reading and Writing. The two facilitators lead the discussion (which was also translated via the headphones) and the final list of skills and competencies were transcribed and distributed to the panelists for their reference during the operational standard setting.

At this point in the standard setting workshop, panelists were separated into their language-specific groups where they would remain until the end of the workshop. Each facilitator was responsible for leading the activities in their respective groups. Panelists were given the opportunity to practice using each standard setting method. Each method (described below) used an iterative (2-round) process and panelists were provided feedback between the rounds.

### *Standard Setting Methods*

Passing scores were derived separately for each of component of the assessment. For the Reading component, two approaches were used, a modified Angoff (1971) and an Extended Angoff (Hambleton & Plake, 1995). The modified Angoff method was used for the selected response items. The specific modification used for this application was the Yes/No method (Impara & Plake, 1997). In this method, panelists are asked to envision a typical high school student who just has the needed qualifications to pass the literacy requirement, based on the curriculum across subjects (target student). With this student in mind, the panelists independently decided whether this student would be able to answer the items correctly (YES, 1 point) or not (NO, 0 points). Panelists were asked to make this decision for each of the selected response items that comprise the Reading test. For the short constructed response and the extended constructed response items, panelists were again asked to consider a typical student who has just barely mastered the necessary curriculum to pass the literacy requirement. The panelists are asked to estimate how many of the available score points the student would likely earn on each question.

The total number of points assigned by a panelist across the items that comprise the Reading component was summed to compute their “Round 1” performance estimates. These estimates represented, for each panelist, the Reading passing score they would recommend based on the score estimates they provided during the first round. These values were averaged across panelists to determine the Round 1 estimate for Reading. The median, minimum, maximum, standard deviation, 75<sup>th</sup> percentile, 25<sup>th</sup> percentile, and semi-interquartile range statistics were also determined. Panelists were provided with a report that showed their item judgments for each of the Reading test items, summary statistics from the entire panel’s Round 1 ratings, performance information from the most

recent administration of the test (p-values), and impact data (percent of students who would pass the literacy test if the Round 1 estimate was used as the cut score). Using this information, panelists had the opportunity to revise their Round 1 ratings. These revised ratings, called Round 2 ratings, were used in the manner identified above to calculate the Round 2 results.

For the Writing component, panelists used the Analytical Judgment Method (Plake & Hambleton, 2000). For this method, the four tasks (extended constructed response items) were considered sequentially. Thirty student papers had been pre-selected to represent student performance across the score distribution and the scores were not revealed. For each task, panelists were instructed to sort the student papers into three piles: Exemplary work, Acceptable work, and Unacceptable work. The pile of papers representing Exemplary work was then set aside. From the Acceptable pile, the panelists were asked to identify the three that they felt were the weakest. From the Unacceptable pile, the panelists were instructed to select the three strongest papers. Therefore, for each task, each panelist identified six student papers, three that represented work that was just barely acceptable and three that represented work that was just barely unacceptable. Mean scores were computed by panelist, by task. A panelist's mean score represents the average score earned by the six papers they selected. Following the completion of the panelists' first round of estimates across the four tasks, summary statistics were reported.

Each panelist was given four feedback reports, one for each of the four tasks, showing their individual average score from the six papers they selected, the overall panel mean, performance information for students from the most recent administration of the test (mean score earned by all students who took the item), and impact data based on the most recent administration. Based on this information, panelists were asked to reconsider their first round ratings and select, if they desired, different student papers for some or all of the six papers that represented work of students who performed either slightly below acceptable or slightly above acceptable on that task. The data from the panelists' second ratings were also summarized consistent with the analyses from the first round.

After the Round 2 results had been entered for each of the panelists, they were asked to complete an evaluation of the standard setting workshop. This evaluation sought information from the panelists regarding their satisfaction with several of the workshop activities including the orientation, practice, and operational rating components. Panelists were asked if sufficient time was allotted to the tasks they were asked to complete and how comfortable and confident they felt with their ratings. Panelists were also given an opportunity to provide confidential comments.

## Results

From the perspective of the facilitators, the process seemed to flow well and the language differences did not seem to produce any issues during the large group orientation or training. This may be a function of the advance preparation and the aid of the translators. For the Reading component<sup>1</sup>, the median recommended cut scores for the two panels differed by six points on a 200-point scale. This difference appears very small and the similarity is underscored by the impact data; the percent of students who would pass the exam with these recommended cut scores differed only by less than one percent. For the Writing component, the scores differed by only one point on a 16-point scale; however, the impact data suggested that six percent more of the French students would pass the exam as compared to the English. These results (both the similarity in the

---

<sup>1</sup> The Reading Round 2 results for the French panel indicated substantial differences between Round 1 and Round 2. Close inspection of the item-level responses suggested that panelists may have misunderstood the p-values. Therefore, Round 1 results are presented here (and were also recommended for use in setting the final passing standard).

Reading recommended cut scores and difference in the Writing recommended cut scores) do not necessarily provide evidence that the panels were or were not holding the same expectations for students. This could only be determined by considering these results in relation to the existence (or lack thereof) of differences in the ability of students across language groups.

Through the evaluations, both panels provided similar perceptions of the process and the time allocated for each component of the training and the operational standard setting (see Table 1). Notably, many panelists provided comments indicating the dual presentation and translation was redundant and was not an efficient use of time. Both panels reported similar levels of comfort and confidence in the ratings they provided for each round using the various methods. Although both groups indicated strong levels of comfort with and confidence in their ratings overall, the English panel reported higher levels of comfort than the French panel in their Round 2 Angoff judgments ( $q.8, t(38)=0.97, p<.05$ ). The French panel rated the organization of the workshop significantly higher than the English panel ( $q.17, t(37^2)=3.25, p<.01$ ). Given this finding, it was interesting to note that two French panelists provided comments indicating they felt the large group orientation/training was inefficient and the entire process should have been conducted in language specific groups.

### Summary

Through this process, we attempted to facilitate a standard setting method that would result in separate recommended cut scores for two different language versions of a literacy assessment. Referring back to the critical elements of a standard setting as defined by Kane (1994), Hambleton (2001), and the *Standards* (AERA, APA, & NCME, 1999), we attempted to keep the process the same between the two language panels by overlapping as much of the process as possible. A method was designed that provided a systematic process that could be conducted in parallel across the two language groups. The standard setting processes were well documented and studied in the literature and had been used successfully with these types of assessments.

For panel selection, we attempted to ensure the panels were as similar as possible in terms of size, experience with their respective target populations, and representation of the different geographical areas. The training process was conducted primarily in one large group to establish one set of characteristics of the target student and reduce any potential facilitator effect. Panelists worked in their language-specific groups while defining the “target student” because we did not want to hinder this process with translations in the small group. As an important feature, the small groups reconvened into a larger group to present their work and the combined target student descriptions were made available to all panelists during the operational activity. In addition, panelists were separated into their language-specific groups for the practice activity as this required panelists to look at items from their language-specific tests and receive feedback data that was specific to their language group. For the operational standard setting, the two panels followed the same systematic process and were given the same types of feedback data (although the student performance and impact data was based on two different sub-populations). Finally, the panelists responded to the same evaluation and this information was used to assess the comparability of the experience between the two panels.

However, there were limitations in the study. Except for the procedural validity evidence suggested by the study design and the evaluation data collected at the conclusion of the study, there was limited evidence that the meaning of the passing scores across the two groups were equivalent (as this could not be determined by comparison of the cut scores or impact data). Additional studies that examine this important validity

---

<sup>2</sup> These analyses were conducted with independent samples t-tests. The degrees of freedom are different for the two reported statistics because one English panelist did not complete the final item (q.17).

question should be conducted as situations like the one described here may become more prevalent with greater variation in language testing.

### Recommendations

Based on consideration of the key elements of a standard setting and the experience and evaluation of this standard setting process, we offer three recommendations for practitioners when conducting a standard setting process with a bilingual test. First, it is important that as much of the process as possible is the same across language groups. Whenever feasible, it would be important to conduct the process together for all language groups to minimize facilitator effect and demonstrate to panelists the similarity of the procedure between versions of the test. By making the processes as similar as possible, one will have a better opportunity to achieve the goal of having panelists conceptualize a common standard (and performance that is on either side of the standard) and judge students' performance, regardless of language group, using the same expectations.

Our second recommendation pertains to the specific process described in this paper. Although we found the process to work quite well, the panelists' evaluations suggest that the dual presentations and translations may not be needed, as panelists did not feel they benefited from hearing the training material twice. Although the panelists may not completely understand the purpose of conducting the orientation and training as one large group, it remains critical to the validity of the results of the workshop and we want to minimize and hindrance that the language barrier imposes. Uses of this process in the future may include a different strategy. For example, facilitators may be able to alternate their training presentations so each panelist could hear the entire training in their own language; half from a facilitator in one language and the other half as a translation of the other facilitator in the second language.

Finally, as mentioned above, additional validity evidence could be collected to determine whether panelists were holding students to the same standards for performance. The recommended cut scores derived from this process suggested the impact would be very similar for one subject area and only slightly different in the other subject area. However, given that the exact abilities of each sub-population are unknown, one cannot conclude that students were being held to the same expectations. Rather, empirical evidence is needed to support the idea that both panels were using one common understanding of the abilities of the minimally competent student and serve as validity evidence for this process. One strategy that might accomplish this goal would be to have panelists rate the expected performance of the target student on a common set of items. Within this specific situation, this could be accomplished by translating ~10 items from one language to another to achieve a common (anchor) set that each panel would rate. The common items would be presented in the same fashion as the remaining test items to ensure that the panelists would rate them in the same way they did the other items. If this strategy were applied in the illustration described in this paper, we would have recommended that half of the items on the common-item set be English items translated to French, and the other half would be French items translated to English. Given the different characteristics of the languages, items would have to be selected based on the ability for them to be directly translated without changing meaning. If the two panels were interpreting the characteristics of the target students similarly, the recommended cut scores should converge on the common item set across the panels.

### Conclusions

Setting passing scores on bilingual testing programs that do not rely on translation/adaptation to create their respective versions creates a challenge to the common interpretation of the scores. The standard setting application described in this paper was an example of one strategy for how to respond to this challenge. The evidence of comparability between the standard setting panels rested primarily on procedural

validity evidence to support the common meaning of the passing scores. As testing programs become more complex, new methods and strategies for processes such as standard setting will be developed and continually tested and improved. More important, studies that collect evidence to evaluate the valid interpretations of these scores given their common use will continue to be needed.

Figure 1

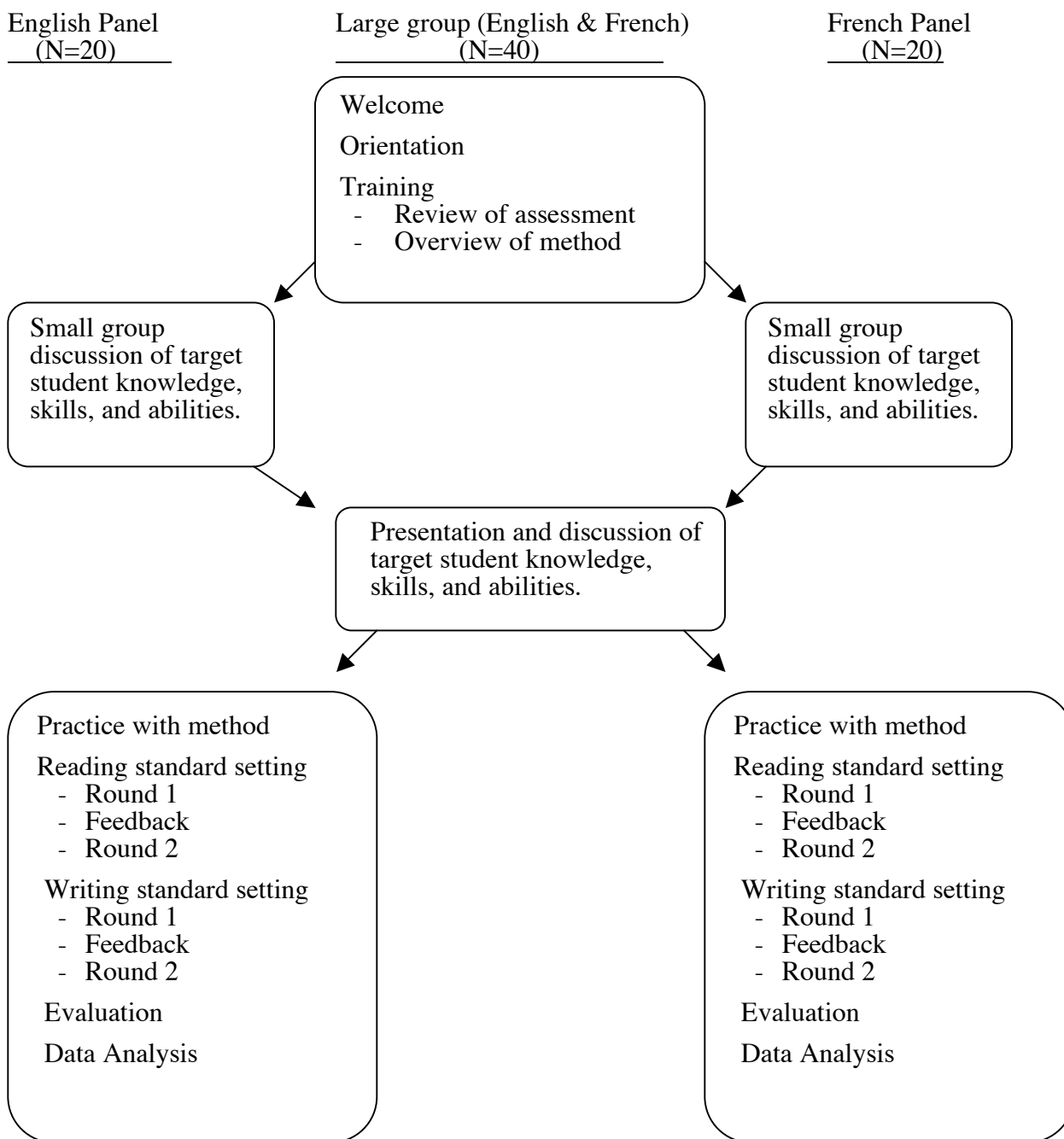
*Process used for Standard Setting Workshop*

Table 1

*Panelist Evaluation Data*

Question	English Average	French Average	Total Panel Average
<b>Part 1: Training</b>			
1. Success ( <i>6 = very successful to 1 = very unsuccessful</i> )			
a. Orientation	4.45	5.05	4.75
b. Training	4.80	4.95	4.88
c. Practice	5.25	5.05	5.15
d. Overall	4.90	4.65	4.78
2. Time Allocated ( <i>6 = Totally Adequate to 1 = Totally Inadequate</i> )			
a. Orientation	4.45	5.50	4.98
b. Training	5.10	5.35	5.22
c. Practice	4.90	5.30	5.10
d. Overall	4.80	5.30	5.05
3. How would you rate the amount of time allocated to training? <i>1 = too much time to 3 = too little time</i>	1.60	1.80	1.70
<b>Part 2: Round 1 Standard Alignment Activity</b>			
4. How confident did you feel with your Round 1 item performance predictions? <i>4 = Confident to 1 = Not Confident</i>	3.45	3.40	3.42
5. How comfortable did you feel about your Round 1 item performance predications? <i>4 = Comfortable to 1 = Not Comfortable</i>	3.60	3.30	3.45
6. How did you feel about the time allotted for Round 1? <i>4 = More than enough time to 1 = More time needed</i>	3.10	3.45	3.28
<b>Part 3: Round 2 Standard Alignment Activity</b>			
7. How confident did you feel with your Round 2 item performance predictions? <i>4 = Confident to 1 = Not Confident</i>	3.80	3.55	3.68
8. How comfortable did you feel about your Round 2 item performance predications? <i>4 = Comfortable to 1 = Not Comfortable</i>	3.85	3.45	3.65
9. How did you feel about the time allotted for Round 2? <i>4 = More than enough time to 1 = More time needed</i>	3.15	3.55	3.35
10. How confident are you that the Round 2 cut score will provide an appropriate performance level for OSSLT? <i>4 = Confident to 1 = Not Confident</i>	3.70	3.55	3.62
<b>Part 4: Round 1 Analytical Judgment Method</b>			
11. How confident did you feel with Analytical Judgments on the responses? <i>4 = Confident to 1 = Not Confident</i>	3.45	3.30	3.38
12. How comfortable did you feel with Analytical Judgments on the responses? <i>4 = Comfortable to 1 = Not Comfortable</i>	3.40	3.35	3.38

*Continued on next page*

Question	English Average	French Average	Total Panel Average
<b>Part 5: Round 2 Analytical Judgment Method</b>			
13. How confident did you feel with your specific Analytical Judgments? 4 = <i>Confident</i> to 1 = <i>Not Confident</i>	3.35	3.35	3.35
14. How comfortable did you feel with you specific Analytical Judgments? 4 = <i>Comfortable</i> to 1 = <i>Not Comfortable</i>	3.35	3.25	3.30
15. How did you feel about the time? 4 = <i>More than enough</i> to 1 = <i>More time needed</i>	3.05	3.30	3.18
<b>Part 6: Overall Evaluation</b>			
16. Overall, how would you rate the success of the Standard Setting Activity? 4 = <i>Totally Successful</i> to 1 = <i>Totally Unsuccessful</i>	3.05	3.40	3.22
17. How would you rate the organization of the Standard Setting Activity? 4 = <i>Totally Successful</i> to 1 = <i>Totally Unsuccessful</i>	2.95	3.55	3.25

## References

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.
- 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.
- Cizek, G.J. (Ed.), *Setting performance standards: Concepts, methods, and perspectives*, Mahwah, NJ: Lawrence Erlbaum Associates.
- Hambleton, R.K. (2001). Setting performance standards on educational assessments and criteria for evaluating the process. In G.J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives*, pp. 89-116. Mahwah, NJ: Lawrence Erlbaum Associates.
- Hambleton, R.K., and Plake, B.S. (1995). Extended Angoff procedures to set standards on complex performance assessments. *Applied Measurement in Education*, 8, 41-56.
- Impara, J. C. & Plake, B. S. (1997). An alternative approach to standard setting. *Journal of Educational Measurement*, 34(4), 355-368.
- International Test Commission (2000). *ITC Test Adaptation Guidelines*. Author. Retrieved March 17, 2006 from <http://www.intestcom.org>.
- Jaeger, R.M. (1989). Certification of student competence. In R.L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 485-514). New York: MacMillan Publishing Co.
- Kane, M. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research*, 64(3), 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: Lawrence Erlbaum Associates, Inc.
- Livingston, S. A. and Zieky, M.J. (1982). *Passing scores: A manual for setting standards of performance on educational and occupational tests*. Princeton, N.J.: Educational Testing Service.
- Plake, B. S., & Hambleton, R. K. (2000). A standard-setting method designed for complex performance assessments: Categorical assignments of student work. *Educational Assessment*, 6(3), 197-215. Washington, DC: American Council on Education.