

Running Head: PANELIST EVALUATIONS AND CUT SCORES IN STANDARD SETTING

Examining panelist evaluations and cut scores in a standard setting study: Bilingual study

Elaine M. Rodeck

Tzu-Yun Chin

Susan L. Davis

Barbara S. Plake

Buros Center for Testing
University of Nebraska-Lincoln

Poster to be presented at the 2007 NCME Graduate Student Poster Session

Questions regarding this paper should be directed to:

Elaine M. Rodeck

Buros Center for Testing

University of Nebraska-Lincoln

ERodeck1@unl.edu

Examining panelist evaluations and cut scores in a standard setting study: Bilingual study

Introduction

The objective of this study is to examine potential relationships between panelist evaluations and changes in ratings between different rounds of a standard setting in a bilingual assessment program. In particular, this study examines panelist evaluations to determine if panelists' perceptions of the standard setting are related to adjustments they make in their recommended cut scores across rounds of the process as a function of their language specific group. The study also examines whether panelist evaluations change across rounds of the standard setting.

This study is based on a standard setting conducted for a high school mathematics test composed of multiple-choice and constructed response items. The test was designed for a population of students who speak and receive primary instruction in either English or French (both language versions of the test exist). Although multiple cut scores were recommended during this process, this particular research study focused on panelist evaluations for one cut score across three rounds of ratings. The cut score chosen for this study is the middle cut score, and represents the cut point between performance categories two and three among four possible performance categories. While this particular cut score is most useful for policy decisions, the full standard setting activity involved the establishment of cut scores to inform broad policy decisions that consider the historical performance of students on similar assessments, changes in student population, and changes in curriculum.

Existing literature in the field of standard setting focuses primarily on the methodological aspects of standard setting (e.g., Hambleton, 2001; Livingston & Zieky, 1982; Cizek, Bunch & Koons, 2004). Specifically, the focus of the literature is on selecting and training panelists, selecting methodologies for standard setting, addressing judgement in standard setting, evaluating the validity evidence of a standard setting, and examining the impact of passing scores (e.g., Cizek, 2001; Kane, 1994).

Concerning panelists, the existing literature focuses on the selection and training of panelists, and denotes the importance of collecting panelists' perceptions of the standard setting through an evaluation (e.g., Hambleton, 2001; Kane, 1994). Although evaluation data is often collected during a standard setting, it is usually only examined for any negative comments that would threaten the validity of the standard setting process. This study examines potential relationships between panelist evaluations and changes in ratings between rounds of a standard setting. The goal of this study is to better understand how panelists' perceptions of the standard setting process are related to their subsequent ratings.

Method

Background

The standard setting was conducted over two days and included two groups, panelists who primarily speak English or French, using an English language or French language assessment designed to measure student achievement in 9th grade mathematics. The two forms of the assessment (English and French) were built independently using the same table of specifications. There were no common items across the two forms. The 39 panelists selected (20

English-speaking and 19 French-speaking) were educators familiar with either language, and knowledgeable about the high school mathematics curriculum and students who participate in the testing program. Geographic representation of panelists was sought, with a goal of representing the student population across the region covered by the standard setting. Students in the testing program for which the standard setting was conducted speak and receive primary instruction in either English or French.

The purpose of the panelist evaluation completed at the conclusion of the standard setting is to collect panelist reactions and perceptions for various components of the standard setting study. Panelists completed ratings for several dimensions on the evaluation. The dimensions analyzed in this study include: (i) confidence with item performance predictions for each round (confidence), (ii) comfort in making item performance predictions for each round (comfort), and (iii) time allotted to complete item performance predictions for each round (time).

Training

To ensure all panelists had an opportunity to receive the intended standard setting training experience, training for the standard setting was provided with the assistance of translators who speak both English and French fluently. The training took place in one room with translators at the back of the room providing simultaneous translation. To accommodate both English and French panelists, panelists were provided with headsets with one channel dedicated to the English language and a second channel dedicated to the French language. Training was conducted in both English and French by a team of two facilitators, with one facilitator speaking English and a second facilitator speaking French. Two presentation screens were used throughout the presentation, one screen displayed material in English and a second screen displayed material in French. The French-speaking facilitator was bilingual in both French and English. The training session material was presented once with English and French facilitators alternately presenting the training material. With simultaneous translation provided throughout the training session, panelists were able to experience the entire training activity in either English or French. All training documents, including slides and practice items, were provided to panelists in their respective language. This was made possible by adapting all training materials from one language to another (i.e., including the same material) to strive for consistency in the training experience for both English and French panelists.

The training session included informing panelists about the purpose of the standard setting and tasks they would use to complete their ratings, providing panelists with the opportunity to experience the test in quasi-operational conditions, and allowing panelists to discuss the skills and competencies of students who were just at the boundary for the performance categories. To assist panelists in gaining a perception of skills and competencies of boundary level students across the performance categories, panelists reviewed student test booklets from a prior test administration. Examples of student work for the training session was obtained from a prior administration of an earlier version of the test. For each of the three cut points, a small group of panelists in language-specific groups reviewed the work of students who performed just above the cut point. The full English and French language panelist group then reconvened. Each language-specific small group presented their perceptions of the skills and competencies of the student work that was just at the boundary for the cut point they examined. This presentation process was done sequentially for the English and French groups, with simultaneous translation provided throughout the presentation. At this point panelists were divided into language-specific groups for the duration of the standard setting. Panelists then

participated in either an English or French language-specific practice session. In this session panelists experienced the rating procedures to be used for the operational test items, and learned about the feedback they would receive between rounds of ratings.

Operational

The English and French tests were created using the same test framework and specifications. However, since items on the English and French tests were different, alignment studies were carried out to ensure that the assessments matched the test framework and test specifications. Both English and French tests consisted of multiple-choice and constructed response questions. Multiple-choice items are scored 0 for an incorrect response and 1 for a correct response. Short constructed response items are scored 0 for an incorrect response and 1 for a correct response. Other constructed response items were scored on a four-point rubric with score point values identified in item-specific rubrics.

A variation of the Angoff (1971) standard setting method was used and is described below. For multiple-choice and short constructed response items, panelists were asked to estimate the proportion of students who possess skills just at the boundary for each performance category and who will answer the items correctly. Round one ratings of the multiple-choice items were completed using multiples of 5's with second and third round ratings, completed using the full 0 – 100 point proportion scale. For other constructed response items, panelists were asked to estimate the mean score of students who are just at the boundary for each of performance categories.

Panelists completed three rounds of ratings, with feedback provided to panelists between rounds. Feedback provided between rounds varied across the rounds and sequentially included individual cut point values, and summary statistics for the full language panel's three cut points (mean, median, standard deviation, 25th percentile, 75th percentile, minimum and maximum values). In addition, panelists were given current test performance data for students based on data from a test administration conducted the prior year: for multiple choice items, the percentage of students who answered the items correctly; for constructed response items the mean, standard deviation, and distribution of scores across the score values.

After round one, in small groups, panelists discussed skills consistent with student work at each of the cut points. A group spokesperson was identified for each small group. This spokesperson was asked to summarize the discussion for the full language-specific group. Following the large group discussion, panelists were asked to make their second round of ratings. Following the second round of ratings, panelists were also informed, based on their round two results, of the proportion of students who would be classified into performance categories 1, 2, 3, and 4. Panelists were also told the proportion of students who historically have been classified into these four performance categories. After receiving this feedback information, panelists made their third and final ratings of the estimated percentage of students at each performance category who will answer the item correctly for multiple-choice and short constructed response items or the mean score for each performance category for the other constructed response items.

At the conclusion of the standard setting, panelists were asked to anonymously complete an evaluation that sought to measure a) panelist satisfaction with the training activities, b) how well panelists understood the tasks they were to complete, and the information they received between rounds, and c) whether panelists felt they had sufficient time to learn the tasks they used and to implement these tasks in making their ratings of test questions (time). Panelists were also asked about their comfort in making their item ratings (comfort) and how confident they were

that the procedures used in the standard setting would yield appropriate cut scores for the four performance categories (confidence). The three factors of confidence, comfort and time were chosen for this study because each factor was assessed over the three rounds of ratings. The evaluation forms for the English and French panelists were adapted from one another (i.e., included the same questions) to facilitate data analysis.

To ensure confidentiality of panelist ratings, panelists selected an identification number they used for all ratings and reporting. This identification number was used when the panelists made their item ratings across the rounds and also when they completed their evaluations. Therefore, even though individual panelists' identify was protected, we were able to connect individual panelist's evaluations with their item ratings across rounds.

Results

Cut scores

The average cut scores for the English and the French groups obtained from the three rounds of the standard setting are reported in Table 1. To examine the change in the recommended cut scores across three rounds, repeated measures ANOVA's were conducted separately for the English and for the French versions.

Mauchly's test indicated that the sphericity assumption might not hold for the English version ($W=.63, \chi^2=8.23, df=2, p=.016$) or the French version ($W=.47, \chi^2=12.69, df=2, p=.002$). Therefore, Huynh-Feldt adjustment for the degrees of freedom was applied to the repeated measures ANOVA's. The results suggested that there were no statistically significant differences in the cut scores across rounds for the English version ($F_{(1.6,29.5)}=1.47, p=.25$) and for the French version ($F_{(1.4,24.7)}=.16, p=.77$).

Table 1. Average cut scores for the English ($n=20$) and the French ($n=19$) groups.

	English	French
Round 1	33.32 (3.57)	33.48 (2.91)
Round 2	32.03 (1.88)	33.27 (1.85)
Round 3	32.74 (1.92)	33.26 (1.85)

Note. Standard deviations (*SD*) are reported in parentheses.

Evaluation Ratings

The average ratings of confidence, comfort, and time allocation obtained from the English and the French groups for the three rounds are reported in Table 2 and Figure 1. Panelists provided their ratings on four-point rating scales. Responses coded with high scores represent a higher degree of confidence, a greater degree of comfort, and more than sufficient time allocated for a round.

Table 2. Average evaluation ratings for the English ($n=20$) and the French ($n=19$) group.

	Confidence		Comfort		Sufficient time	
	English	French	English	French	English	French
Round 1	2.75 (.55)	2.74 (.45)	2.95 (.51)	2.42 (.51)	3.45 (.51)	3.05 (.62)
Round 2	3.10 (.64)	3.47 (.51)	3.15 (.67)	3.32 (.48)	3.30 (.57)	3.26 (.45)
Round 3	3.65 (.49)	3.74 (.45)	3.70 (.47)	3.74 (.45)	3.50 (.51)	3.58 (.51)

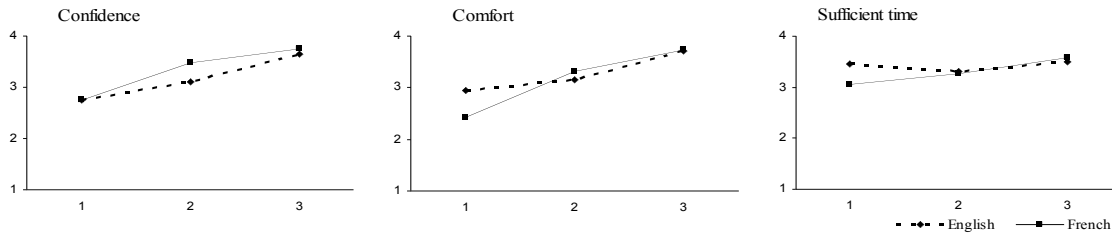


Figure 1. Average evaluation ratings

Repeated measures ANOVA’s with one between (language groups with 2 levels) and one within factors (round, with 3 levels) were conducted to examine (1) change in the evaluation ratings across three rounds, (2) difference between the evaluation ratings from the English and the French groups, and (3) interaction effects of the language group and standard setting round on the evaluation ratings. Confidence, comfort and time allocation ratings were analyzed separately. Mauchley’s tests suggested that sphericity might be assumed for the comfort ratings ($W=.97, \chi^2=1.25, df=2, p=.54$) and time allocation ratings ($W=.99, \chi^2=.32, df=2, p=.85$). However, the sphericity assumption might not be sustained for the confidence ratings ($W=.80, \chi^2=8.04, df=2, p=.02$). Therefore, repeated measures ANOVA’s were carried out with sphericity assumption except for analyzing the confidence ratings where Huynh-Feldt adjustment was applied. The results are presented in Table 3.

Table 3. Repeated Measures ANOVA

Source		<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>
Confidence ratings (with Huynh-Feldt Adjustment)					
Within	Round	17.71	1.78	9.93	37.32***
	Round x Language	.79	1.78	.44	1.66
	Error(Round)	17.56	65.96	.27	
Between	Language	.65	1	.65	1.90
	Error	12.65	37	.34	
Comfort ratings					
Within	Round	20.82	2	10.41	54.46***
	Round x Language	2.66	2	1.33	6.96**
	Error(Round)	14.14	74	.19	
Between	Language	.35	1	.35	.80
	Error	15.98	37	.43	
Time allocation ratings					
Within	Round	1.96	2	.98	5.61**
	Round x Language	1.20	2	.60	3.45*
	Error(Round)	12.90	74	.17	
Between	Language	.41	1	.41	.82
	Error	18.51	37	.50	

* $p < .05$; ** $p < .01$; *** $p < .001$

The results presented in Table 3 show that the panelists' confidence ratings did not differ across language groups. However, there was a statistically significant round effect on confidence ratings ($F_{(1,78,65.96)}=37.32, p<.001$). Post hoc comparisons indicate that panelists' confidence increased between rounds ($p<.001$).

There was an interaction effect of round and language group on the panelists' comfort ratings ($F_{(2,74)}=6.96, p=.002$). Post hoc comparisons between language groups reflect that, during the first round, the English group were more comfortable than the French group ($\bar{X}_{\text{English, Round 1}}=2.95, \bar{X}_{\text{French Round 1}}=2.42, p=.002$), whereas the second and third round comfort ratings did not differ between language groups. An alternative post hoc approach to follow up on the significant interaction effect was to condition on the language group and compare across rounds (i.e., simple main effect of round). These comparisons show that (1) for the English group, the comfort ratings did not differ between Round 1 ($\bar{X}_{\text{English Round 1}}=2.95$) and Round 2 ($\bar{X}_{\text{English Round 2}}=3.15$) whereas there was a significant increase ($p<.001$) in comfort ratings from Round 2 to Round 3 ($\bar{X}_{\text{English Round 3}}=3.70$), and (2) for the French group, there was a significant increase in comfort ratings ($p<.001$) from Round 1 ($\bar{X}_{\text{French Round 1}}=2.42$) to Round 2 ($\bar{X}_{\text{French Round 2}}=3.32$) and a significant increase ($p=.002$) from Round 2 to Round 3 ($\bar{X}_{\text{French Round 3}}=3.74$).

There was also an interaction effect of round and language group on the panelists' time allocation ratings ($F_{(2,74)}=3.45, p=.037$). Post hoc comparisons reflect that, during the first round, the English group perceived more than sufficient time was allocated to the task when compared to the French group ($\bar{X}_{\text{English, Round 1}}=3.45, \bar{X}_{\text{French, Round 1}}=3.05, p=.035$), whereas the second and third round time allocation ratings did not differ between language groups. Alternative post hoc comparisons across rounds show that (1) for the English group, there were no statistically significant differences in time allocation ratings across rounds, and (2) for the French group, panelists reported more than sufficient time to work on the task ($p=.001$) in Round 3 ($\bar{X}_{\text{French, Round 3}}=3.58$) than in Round 2 ($\bar{X}_{\text{French, Round 2}}=3.26$) although the difference in time allocation rating between Round 1 ($\bar{X}_{\text{French, Round 1}}=3.05$) and Round 2 was not statistically significant ($p=.111$).

Absolute Value Change in Cut Scores

Multiple regressions were used to explore the relationships between the evaluation ratings and the cut score changes. In addition, differential impacts of the evaluation ratings on the cut score changes between the language groups were also examined. Two absolute value cut score changes (i.e., Round 1 to Round 2 absolute value changes and Round 2 to Round 3 absolute value changes) were used as dependent variables and were analyzed separately. Each set of absolute value changes were regressed on the language group, the three evaluation ratings (i.e., confidence, comfort, and time allocation) of the early round, and the cross-products of language and evaluation ratings. Language groups were dummy coded and English group was set to be the reference group (i.e., English=0 and French=1). Evaluation ratings were grand mean centered in the regression equations to reduce possible multicollinearity introduced by including the cross-product terms.

Regression coefficients for predicting Round 1-Round 2 absolute value changes are reported in Table 4. Forty percent of the Round 1-Round 2 absolute value change score variances were accounted for by the language, evaluation ratings and their interactions ($R^2=.40, F_{(7,31)}=3.00, p=.016$). Table 4 shows that there was a statistically significant language group

effect on the Round 1-Round 2 absolute value changes ($b=-2.38$, $t=-3.48$, $p=.002$), where the French group had smaller magnitude of cut score modification from Round 1 to Round 2 when compared to the English group. There was also a statistically significant regression coefficient for the comfort ratings ($b=-2.70$, $t=-2.68$, $p=.012$) and for the comfort by language interaction ($b=3.71$, $t=2.71$, $p=.011$). These results indicate that, while the comfort level was negatively associated with the magnitude of cut score modification for the English group, the opposite relationship was suggested for the French group. For the French group, the higher the comfort level, the greater magnitude the cut scores were modified from Round 1 to Round 2.

Table 4. Regression coefficients for predicting Round 1-Round 2 absolute value changes ($N=39$)

	<i>b</i>	<i>SE(b)</i>	<i>t</i>	<i>p</i>
Intercept	4.02***	.48	8.34	<.001
Language	-2.38**	.68	-3.48	.002
Confidence	1.63	.89	1.83	.077
Comfort	-2.70*	1.00	-2.68	.012
Time	1.00	.86	1.16	.257
Confidence x Language	-2.15	1.35	-1.59	.122
Comfort x Language	3.71*	1.37	2.71	.011
Time x Language	-1.89	1.14	-1.66	.107

* $p<.05$; ** $p<.01$; *** $p<.001$

Regression coefficients for predicting Round 2-Round 3 absolute value changes are reported in Table 5. Only eight percent of the Round 2-Round 3 absolute value change score variances were accounted for by the language, evaluation ratings and their interactions and this regression model was not statistically significant ($R^2=.08$, $F_{(7,31)}=.38$, $p=.908$). Table 5 shows that none of the variables, including interactions, were statistically significant predictors for the Round 2-Round 3 absolute value changes.

Table 5. Regression coefficients for predicting Round 2-Round 3 absolute value changes ($N=39$)

	<i>b</i>	<i>SE(b)</i>	<i>t</i>	<i>p</i>
Intercept	1.32**	.36	3.66	.001
Language	-.66	.52	-1.27	.214
Confidence	-.09	.97	-.09	.926
Comfort	-.05	.91	-.05	.959
Time	.25	.63	.40	.694
Confidence x Language	.05	1.27	.04	.971
Comfort x Language	.70	1.44	.49	.627
Time x Language	-1.11	1.21	-.91	.368

* $p<.05$; ** $p<.01$; *** $p<.001$

Discussion

A review of the panelists' evaluations showed panelists' understanding of the standard setting process and panelists' confidence in the results and procedures. Thus, all of the panelists' data has been used in this study. There was no missing data in the study. No significant differences in average cut score across three rounds, between the English and French language groups, or interactions were found. This finding provides good procedural support for the study

(Kane, 1994). While panelists' confidence ratings did not differ across language groups, panelists' confidence did increase in later rounds. This finding should not be surprising since the standard setting process is new for most panelists and panelists gain confidence in the process in later rounds.

The interaction effect of round and language group on panelists' comfort ratings showed that the English group was more comfortable than the French group during the first round, while the second and third round comfort ratings did not differ between language groups. A plausible explanation for this finding may be that English is the first language in the region in which the standard setting was conducted. While every effort was made to ensure equivalence in the standard setting experience, there may be a natural tendency on the part French group to be more cautious and less comfortable when initially placed in a new environment (e.g., standard setting). As the French group becomes more comfortable, fewer differences would be expected and this was the finding in this study for rounds two and three. An alternate post hoc approach was conducted by conditioning on the language group and comparing across rounds. These analyses showed that while comfort ratings for the English group did not differ between Rounds 1 and Round 2, there was a significant increase in comfort ratings between Round 2 and round 3. The English group was initially more comfortable with the process and had to make a larger change for any change to be significant. Consistent with the earlier suggestion that the French group will become more comfortable over the later rounds, it is reasonable that the analyses show a significant increase in the comfort ratings for each of Round 1 to Round 2, and Round 2 to Round 3.

An analysis of the interaction effect of round and language group on panelists' time allocation ratings showed that the English group perceived more time was allocated to the task when compared to the French group for the first round with no difference between groups noted in Round 2 or Round 3. It is plausible that this finding is consistent with the earlier suggestion that the French group was less comfortable initially, but became more comfortable in later rounds. Alternate post hoc comparisons of the interaction effect of round and language show that there were no statistically significant differences in time allocation ratings for the English group. This is not unexpected since the English group reported that sufficient time was available for all three rounds. Analyses for the French group are not surprising with the French group reporting significantly more time to work on the task in Round 3, with no significant difference in time allocated for Round 1 and Round 2. Even though there was no significant difference for the French group in time allocated for Round 1 and Round 2, it is worth noting that the average rating for time allocated increased for the French group from Round 1 to Round 2.

Multiple regressions were used to explore relationships between the evaluation ratings and absolute value cut score changes, showed different results for Round 1-Round 2 and Round 2-Round 3. While forty percent of the change in cut scores between Rounds 1 and 2 was accounted for by the language, evaluation ratings and their interactions, only eight percent of the change in cut scores between Rounds 2 and 3 were accounted for by the same predictor variables. These results show that key predictor variables are omitted in the multiple regression equation. This is especially the case for Round 2-Round 3. The French group showed a significantly smaller magnitude of change in cut score from Round 1 to Round 2. The English group showed fewer changes in cut score when the panelist was more comfortable. However, the findings indicate that the French group showed greater changes in cut score as the panelist comfort level increased. The relationship between comfort level and cut score modification is unexpected. A plausible explanation may be that panelists in the French group who were more

cautious in initially making changes to the cut score became more comfortable over subsequent rounds.

Educational Contribution

This study was designed to inform the educational community by examining potential relationships between panelist evaluations and changes in the ratings between rounds of a standard setting. The perception of the standard setting process by the panelist is important as the final recommended cut score is based on the panelists' professional judgments. While existing literature examines psychometric and statistical methodologies of standard setting, it does not extend to the examination of the panelist and the associated nuances in evaluation and cut scores. Ultimately, the role of the panelist is critical in the establishment of cut scores used to inform policy decisions that impact candidates, schools, and agencies.

This study also makes contributions to the field of standard setting in bilingual settings. Often, test developers are asked to create translated versions of the assessment for use in bilingual settings. In this study, two versions of the examination were created independently, with the assessments designed to measure the same table of specifications. This is a fairly unique situation in bilingual assessment programs. Another unique feature of the study is the use of two parallel, yet simultaneous, standard setting procedures rather than setting the cutscores on the base examination (often English) and using an equating strategy to obtain equivalent cutscores on the other language assessment. This study provided promising evidence that a parallel, simultaneous standard setting can provide comparable results across the two assessments designed to measure the same performance categories. Future research will include further analysis of change in panelist cut score at multiple cut points.

References

- Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational Measurement*, (2nd ed., pp. 508-600), Washington, DC: American Council on Education.
- Cizek, G.J., Bunch, M.B., Koons, H. (2004). An NCME instructional module on setting performance standards: Contemporary methods. *Educational Measurement: Issues and Practice*, 23(4), pp. 31-50.
- Cizek, G.J., (Ed.). (2001). *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.
- Kane, M. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research*, 64(3), 425 – 461.
- 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.