Examining variation in independent replications of the Bookmark standard setting method on two tests

Michael C. Rodriguez
University of Minnesota

Otto Rego & Fernando Rubio
USAID—Guatemala

April 14, 2009

Paper presented at the annual meeting of the
National Council on Measurement in Education, San Diego, CA.

Examining variation in independent replications of the Bookmark standard setting method on two tests

Background

Accountability in education and the professions has created a very busy industry in the area of performance standard setting. A natural question following the development of a standards-based assessment is: What scores must students or candidates achieve to be classified in one or more performance levels? Several methods for standard setting have been proposed, implemented, modified, and studied. Many of the methods in use have been adequately described elsewhere (e.g., Cizek, Bunch, & Koons, 2004). Perhaps the area of test development that most recently has demanded the greatest efforts in standard setting concerns the establishment of standards for state achievement tests in Mathematics, reading, writing, and science (as required by NCLB).

One of the most popular methods currently in use is the so-called Bookmark standard-setting method (Karantonis & Sireci, 2006). However, little empirical research has been published regarding the strengths and weaknesses of this method, essentially needed for establishing a strong evidence base for its continued use (Karantonis & Sireci). The study reported in this manuscript was designed to provide additional evidence regarding the Bookmark methodology and it's appropriateness in a particular setting, the establishment of performance standards for a test in a developing country with significant opportunity-to-learn limitations.

Evidence suggests that the Bookmark method is gaining attention and becoming more commonly used throughout the country and the world. As argued by Karantonis and Sireci (2006) in their review of the literature on Bookmark standard-setting methods, "it is important that the research base supporting its use continues to grow" (p. 11). The areas of research they identified as needing further attention regarded issues related to validity, the use of the median versus the mean for recommended cut scores, and others.

Kane (2001), among others, provided guidelines for investigating the validity of standard setting methods, consistent with his argument-basis for validation. His internal validity evidence approach includes the examination of the consistency of judges' ratings. These issues are investigated in the current study. Additional evidence with respect to judges' responses to evaluative questions about their participation and the quality of the process and outcomes will contribute to our interpretation of the meaningfulness of their recommended standards.

Background on Current Study

The United States Agency for International Development (USAID) executes its Guatemala projects inside the framework of its Regional Strategy for Central America and Mexico. The work of USAID-Guatemala is within USAID's Strategic Objective 3 (SO3), addressing community health and educational attainment. To achieve the goals of the SO3, USAID-Guatemala reinforces local efforts in the educational arena, supporting the government of Guatemala, through the Ministry of Education (Ministerio de Educacíon, MINEDUC) and civil and social organizations. The goals of these efforts include improving the transparency, efficiency, and effectiveness of the educational system; achieving universal access to primary education; and increasing educational quality.

To support the improvement of the efficiency, equity and quality of the educational system, USAID supported, through a 4 year (2005-2009) grant, the Educational Standards and Research Program. This program, administered by the firm Juárez and Associates, offers technical and financial support to the MINEDUC, utilizing results of educational research and evaluation activities. In addition, the program has developed an active communication and dissemination process that inform the national dialogue on education. An important aspect of this project is the development of the National System of Evaluation.

A first step in this project was the creation of a national set of content standards in several areas of the curriculum, by grade, for K-12 general education public schools. To begin the process of monitoring performance on these new content standards, national standardized assessments have been developed over the past 3 years, in the areas of Mathematics and Language Arts. Following the first operational administration, a low-stakes administration without consequences to students or schools, provisional performance standards were set. Following the first complete administration and analysis of performance levels, additional changes were made to the content standards, requiring subsequent changes to the tests and the need for new performance standards to be set. All of this was by design, as Guatemala has never had national assessments for accountability purposes.

Guatemala has a population of approximately 14 million people and more than 50% speak one of over 20 Mayan languages as their first language. The national education system has worked very hard to provide bilingual (Spanish & Mayan) education for the first three years of school but because some of the over 20 different Mayan languages are spoken by only hundreds of people, the quality of education is quite unequal. Most of the Mayan population live in rural areas where the schools have very few educational resources compared to those in the urban areas thus the learning opportunities for those rural students are very limited. Also, the educational system has serious efficiency problems. About 34% of the students fail their first year and only 42% of the students finish elementary school (essentially grades 1 to 6); less than 10% finish secondary school.

The national content standards recently developed were designed to serve several purposes, including establishing clear content and performance goals for each grade and standardizing the quality of education. Before the year 2006,Guatemalan achievement tests where norm referenced and sample based, designed through university-based evaluation projects. Since then, the national tests were aligned to the new content standards and piloted in 2006 and again in 2008. Beginning 2009, the national assessments are administered by MINEDUC annually near the end of the school year in grades 1, 3, 6, 9 and 12, in Mathematics and Language Arts.

Operational forms have been developed through a strong common-item linking design to facilitate equating across years. The tests, linking, equating, and now standard setting, have all been supported through the use of Rasch scaling. A technical manual was developed during the assessment design and development process and was used as a guide to evaluate the degree to which each step was consistent with the *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 1999).

Because standard setting, as a conceptual framework and a process, was new to Guatemala, the USAID and MINEDUC staff decided to investigate the methodology during the first implementation, with the idea that much could be learned through which a (potentially) more refined methodology could be used for final standard setting once the assessments have been finalized. It was also important to provide validity-related evidence regarding the appropriateness and feasibility of employing any given standard setting method in the Guatemala context. Through extensive review of the literature and guidance from external experts, the teams decided the Bookmark method would be most appropriate method, well suited to setting standards on multiple-choice tests, and one that they could implement with sufficient fidelity. An external consultant with extensive standard setting experience was obtained to help facilitate the standard setting process.

Methods

As part of the investigation of the standard setting methodology, a small study was designed to investigate variability in results from replications of the standard setting process on two tests. The standard setting study included one form of the Language Arts test and one form of the Mathematics test, both at the third grade level. The Language Arts test consisted of 29 items and the Mathematics test consisted of 40 items. All items were multiple-choice items. National data were available following the regular national administration of each test. Each test was developed through standard procedures, including the selection of final items through extensive expert review, piloting, and field-test item analysis.

A note on terminology. The study design, described below, consists of two subject areas, Mathematics and Language Arts; three standard setting panels for each subject area; and three rounds for each panel. When a specific panel is mentioned, for example, Panel M1, the term panel will be capitalized; similarly when a specific round is mentioned, for example Round 1, the term round will be capitalized.

Design

In the Language Arts portion of the study, 47 judges were randomly assigned to 3 independent panels (consisting of 17, 16, and 14 judges). In Mathematics, 51 judges were randomly assigned to three panels (consisting of 18, 17, and 16 judges). Groups were slightly uneven in membership because some judges failed to appear and participate in the process; judges were assigned to their panels prior to their arrival to move along the process. The panels were comprised of Guatemalan education stakeholders, including parents, classroom teachers, and school administrators. The Language Arts Panels were 67% female; the Mathematics Panels were 57% female.

The facilitators were trained together and used the same materials and procedures with each panel. Thus, each panel and their results were completely independent at the person level, including the facilitator and panel members. The process and materials were equivalent in each panel, with the exception of the different subject tests (Mathematics or Language Arts) for each set of three panels.

Judges included grade-specific teachers from various regions of the country that were assigned to the test in their area of primary instructional responsibility (Mathematics or Language Arts). The Guatemala national assessment system includes four performance levels: Unsatisfactory (Insatisfactorio), Should Improve (Debe Mejorar), Satisfactory (Satisfactorio), and Excellent (Excelente). This required panelists to set three cut scores to separate the four performance levels. A standardized (consistent) approach to the Bookmark method was employed by each of the independent panels.

Standard Setting Process

A standard setting report is available that provides detailed explanations and copies of all exhibits, training materials, results from each round, data provided to panelists during the process, and evaluation form. A brief synopsis of the process is provided here to highlight the major steps taken by each panel.

All judges participated in the initial training in a combined session, including an introduction to the tests, the need for content and performance standards, and a review of the Bookmark methodology. Judges then took the test themselves in their content area, proceeded to review the performance level descriptors, and discuss the meaning of each level in terms of students' knowledge, skills, and abilities. Item booklets were ordered based on Rasch measures for each item, employing the RP .50 value. Judges completed the rating tasks (setting bookmarks) in three rounds. Following the first round, panels were provided with feedback including the distribution of ratings from their panel and participated in additional discussions regarding distinguishing characteristics of students at each performance level. Following the second round, panels were provided with feedback including the distribution of ratings from the second round and impact data, with additional discussion regarding student characteristics within each performance level. Following the third round of ratings, panelists completed an evaluation questionnaire. All panelists who began the standard setting process completed the process. Standard setting panelists met over two days in Guatemala City.

Methods of Analysis

Data were collected, as typically done in any standard setting method, for each of the three panels working with the Language Arts test and the three panels working with the Mathematics test. The items identified at each cut in each round comprise the data sources for this study. Three forms of analyses were completed to provide insight into the use of the Bookmark methodology.

To assess agreement across the three panels in each subject area, medians and means were both examined in terms of the item number selected for each of the three cuts. To assess agreement among median cuts and their distributions, the Mean test and the Kruskal-Wallis test were used, capitalizing on the ordinal nature of the item ratings.

Rasch measures also were available for each item, so that cuts assigned as item numbers could be transformed to their Rasch measures, allowing for analysis of variance to study agreement of mean Rasch measures across panels through each round. This allowed us to evaluate shifts across rounds, agreement at each round, and to analyze both mean scores and score variance at each cut in each round for each panel.

In addition, evaluation results including feedback from judges was used to assess the performance of each panel, in terms of their variation from the other two panels within a subject test, shifts in variability from round one to round three, and shifts in each of the three cut scores set at each round (e.g., distances between the three cut scores), given the nature of the feedback and their responses to questions regarding their understanding of the process and confidence in the outcomes.

Results

The results are presented in three layers, including descriptive results reporting on the results from each round for each panel. We analyze agreement in two ways, using raw score (item number) employing ordinal tests and then using Rasch item locations on cut scores employing ANOVA methods. Finally, we evaluate the degree to which variation between panels and within panels can be explained by employing panelist feedback on the evaluation forms.

Complete distribution information for each panel is provided in the Appendix: Tables A to C provide distributions of bookmarks (frequency of selected item numbers) for Mathematics and Tables D to F provide bookmark distributions for Language Arts.

The average selected item numbers at each cut score are summarized in Table 1 for Mathematics and Table 2 for Language Arts. Associated with each Table is a graphical display of the median cuts at each round (see Figures 1 and 2).

Table 1

Mathematics Item Number Cut Score Distribution Summaries by Panel and Round

			Cut Score (29 Items)		
Panel	Round		1st	2nd	3rd
Panel 1	**1**	**Median**	**4.0**	**8.5**	**22.0**
		SD	4.1	5.3	3.2
		Range	15.0	16.0	12.0
	2	**Median**	**4.5**	**8.5**	**22.0**
		SD	3.7	5.1	2.0
		Range	17.0	17.0	8.0
	3	**Median**	**5.5**	**12.0**	**22.0**
		SD	5.0	5.0	2.2
		Range	16.0	16.0	9.0
Panel 2	**1**	**Median**	**5.0**	**11.5**	**23.0**
		SD	2.5	5.2	3.0
		Range	10.0	18.0	10.0
	2	**Median**	**5.0**	**12.0**	**23.0**
		SD	2.2	4.5	3.5
		Range	10.0	15.0	11.0
	3	**Median**	**5.0**	**12.5**	**23.0**
		SD	1.9	3.7	3.0
		Range	8.0	12.0	10.0
Panel 3	**1**	**Median**	**7.0**	**15.0**	**23.0**
		SD	1.9	3.5	2.0
		Range	6.0	13.0	6.0
	2	**Median**	**8.0**	**14.0**	**23.0**
		SD	2.8	3.1	2.2
		Range	11.0	11.0	9.0
	3	**Median**	**8.0**	**14.0**	**23.0**
		SD	2.5	2.5	1.7
		Range	9.0	9.0	7.0

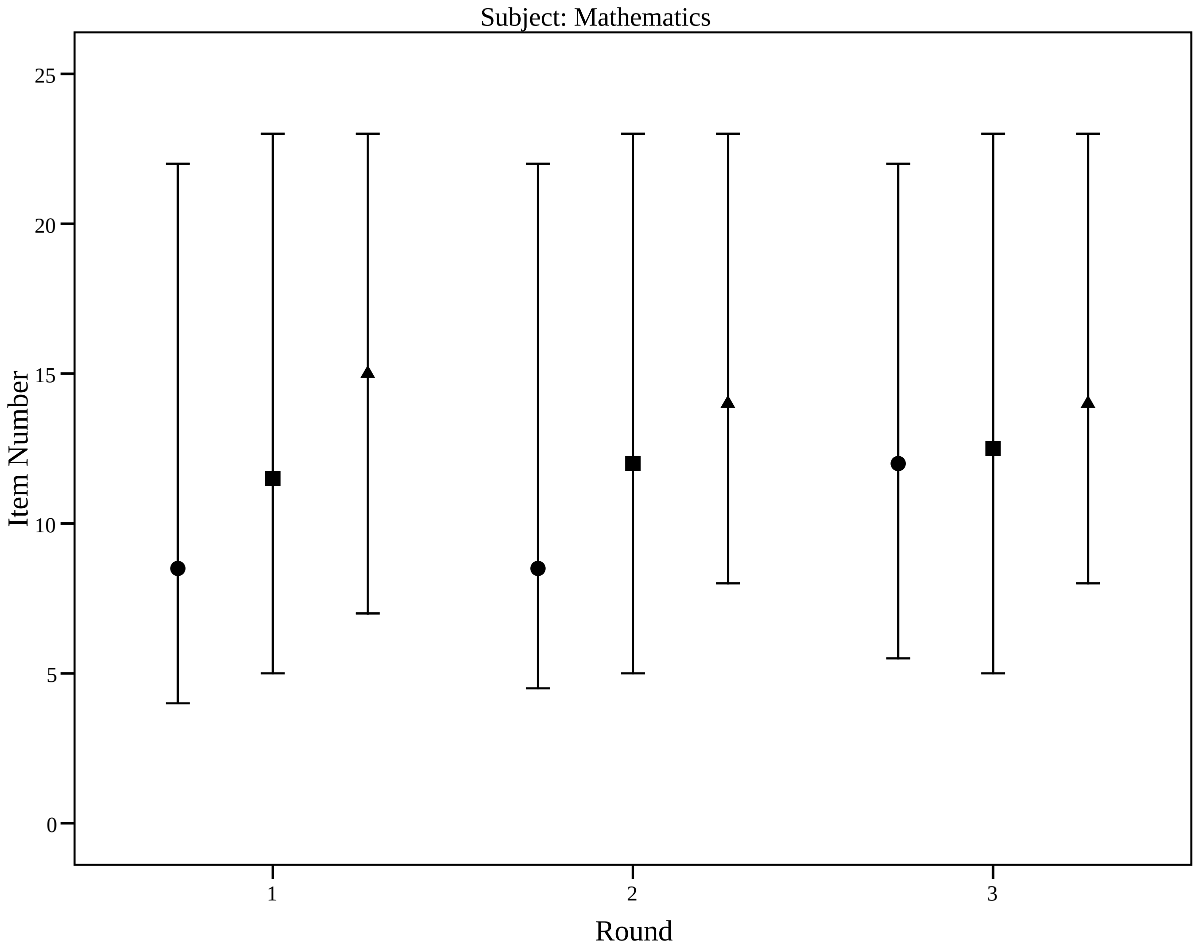


Figure 1. Mathematics cut scores for each panel within each round. The circle represents Panel 1, square represents Panel 2, triangle represents Panel 3.

As can be seen in Figure 1, within Round 1 the three panels varied slightly, more so at the middle cut score (Satisfactory level). At Round 2, variation among panels does not appear to change much. At Round 3, there is less variation at the lowest and middle cut scores.

Table 2

Language Arts Item Number Cut Score Distribution Summaries by Panel and Round

			Cut Score (40 Items)		
Panel	Round		1st	2nd	3rd
Panel 1	**1**	**Median**	**7.5**	**16.5**	**29.5**
		SD	2.5	5.5	6.5
		Range	8.0	19.0	23.0
	2	**Median**	**8.0**	**16.5**	**31.5**
		SD	3.1	4.8	4.5
		Range	11.0	15.0	14.0
	3	**Median**	**9.0**	**17.0**	**32.0**
		SD	4.7	4.8	3.6
		Range	15.0	14.0	11.0
Panel 2	**1**	**Median**	**8.0**	**27.0**	**35.0**
		SD	4.1	5.0	2.0
		Range	15.0	20.0	7.0
	2	**Median**	**8.0**	**27.0**	**36.0**
		SD	4.5	5.1	2.8
		Range	17.0	21.0	10.0
	3	**Median**	**8.0**	**27.0**	**37.0**
		SD	3.3	6.8	2.8
		Range	12.0	26.0	9.0
Panel 3	**1**	**Median**	**6.5**	**16.5**	**29.0**
		SD	3.8	7.7	8.1
		Range	11.0	27.0	32.0
	2	**Median**	**9.5**	**21.0**	**33.5**
		SD	6.2	4.7	5.3
		Range	18.0	17.0	20.0
	3	**Median**	**8.5**	**20.0**	**32.0**
		SD	3.7	6.3	3.9
		Range	14.0	21.0	16.0

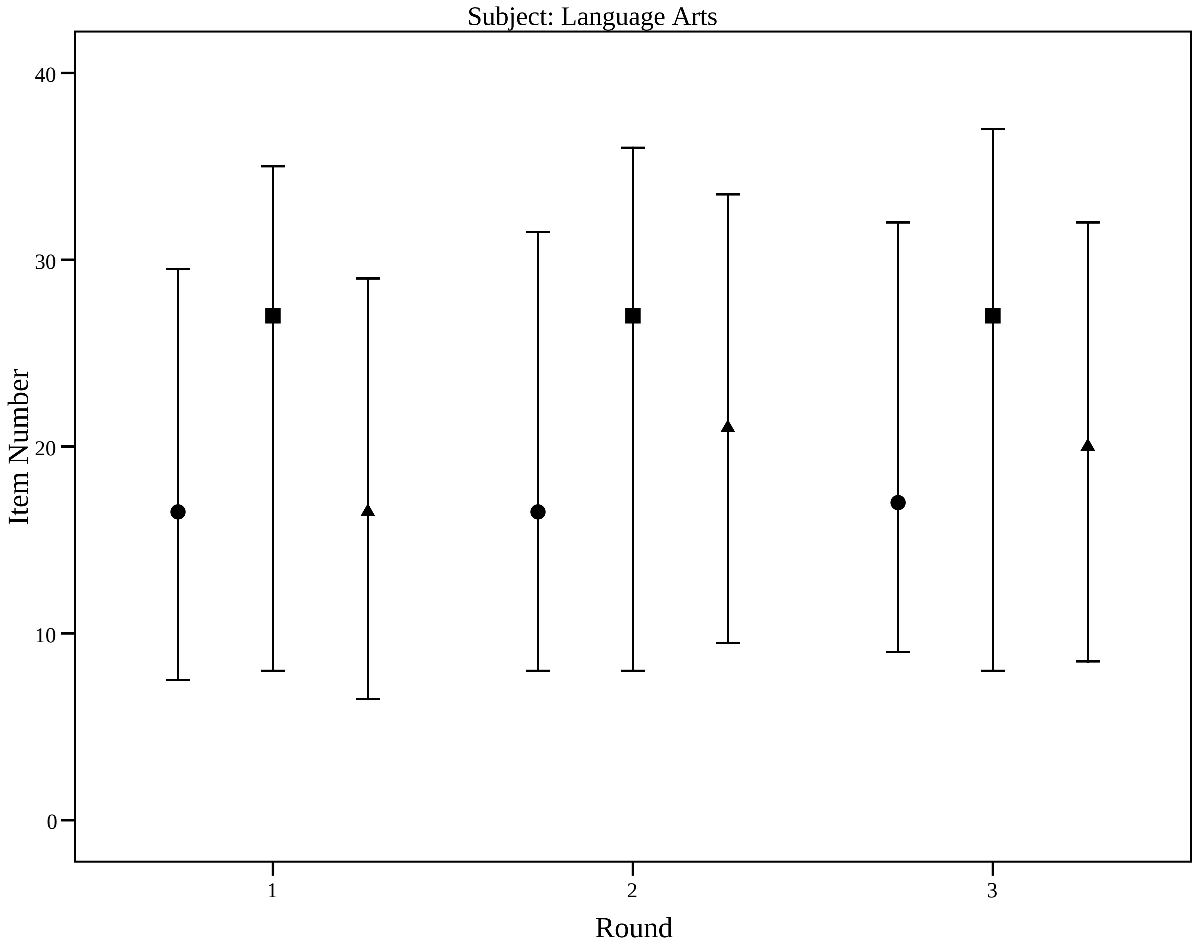


Figure 2. Language Arts cut scores for each panel within each round. The circle represents Panel 1, square represents Panel 2, triangle represents Panel 3.

As can be seen in Figure 2, there was a fair amount of variability between panels within rounds, particularly at the middle cut score (Satisfactory level). Although there were some changes across rounds for each panel, the variability of panels does not appear to change.

Similarly, for Mathematics, we examined summary statistics for each panel at each round given the item Rasch location parameter (*b*-parameter), rather than simple item number. These results are summarized in Tables 3.

Table 3

Mathematics Rasch Item Location Cut Score Distribution Summaries by Panel and Round

			Cut Score		
Panel	Round		1st	2nd	3rd
Panel 1	**1**	**Median**	-1.5	-0.8	0.5
		SD	0.7	0.5	0.8
		Range	2.5	1.6	2.3
	2	**Median**	-1.4	-0.8	0.5
		SD	0.4	0.5	0.5
		Range	1.8	1.8	2.2
	3	**Median**	-1.2	-0.4	0.5
		SD	0.6	0.5	0.7
		Range	1.6	1.6	2.2
Panel 2	**1**	**Median**	-1.3	-0.5	0.9
		SD	0.5	0.5	0.8
		Range	2.0	2.0	2.3
	2	**Median**	-1.3	-0.3	0.9
		SD	0.4	0.5	0.9
		Range	2.0	1.5	2.4
	3	**Median**	-1.3	-0.3	0.9
		SD	0.3	0.4	0.8
		Range	1.0	1.1	2.4
Panel 3	**1**	**Median**	-1.0	-0.1	0.9
		SD	0.2	0.4	0.7
		Range	0.8	1.6	1.7
	2	**Median**	-0.9	-0.2	0.9
		SD	0.3	0.3	0.6
		Range	1.4	1.2	2.2
	3	**Median**	-0.9	-0.2	0.9
		SD	0.3	0.2	0.5
		Range	1.3	0.9	2.1

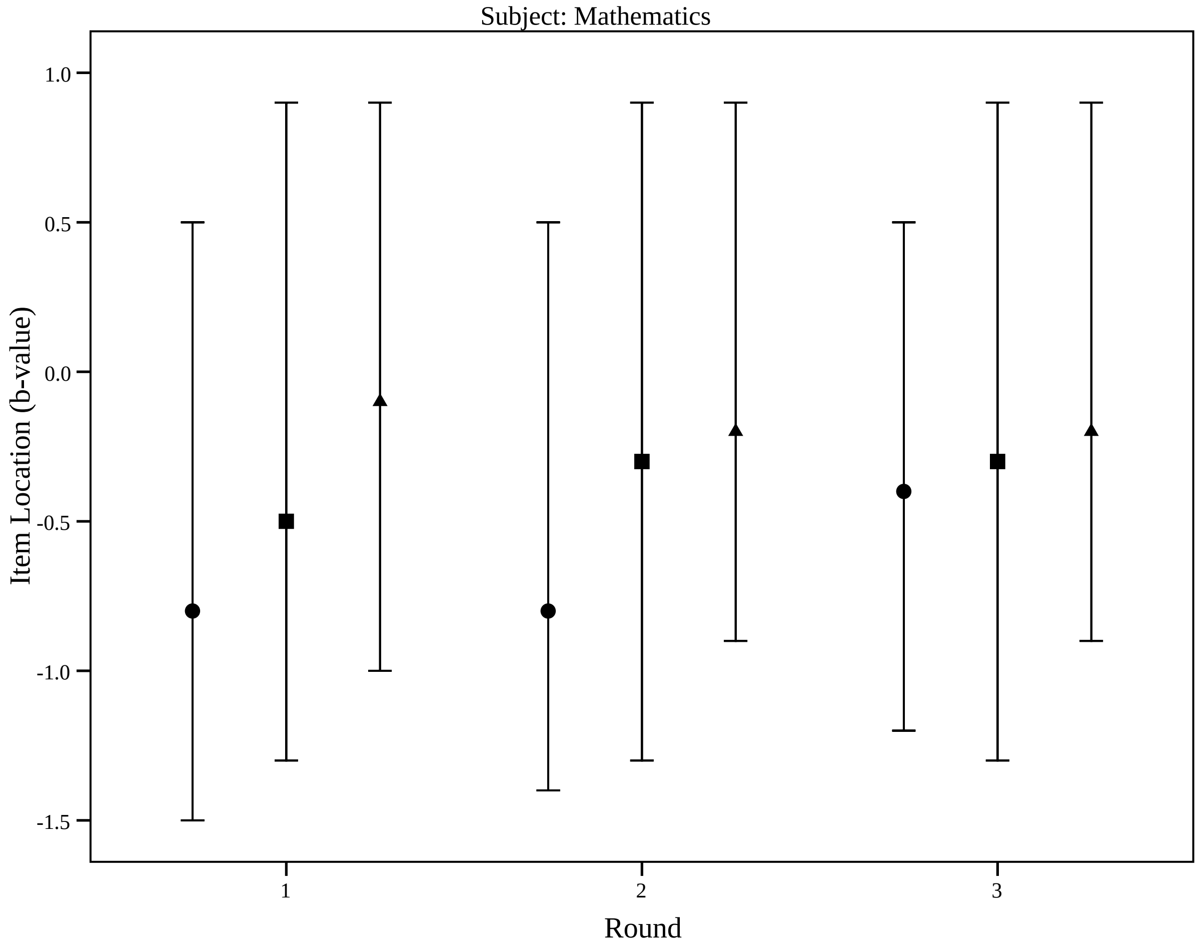


Figure 3. Mathematics cut scores based on item location for each panel within each round. The circle represents Panel 1, square represents Panel 2, triangle represents Panel 3.

The graphical display of cut scores given Rasch item location is similar to that based on item number, but there are noticeable differences. There appears to be more variation in location of the highest cut score (Excellent level); Panel 1 (the first line within each round) is much lower relative to other differences compared to Figure 1 with item number.

Assessing the Difference between Panels

One difference between panels could be defined in terms of the median cut score at each round. The *Median Test* assesses the null-hypothesis that the three panels come from populations with the same median. Based on the results of this test (Table 4), we find that the panels come from populations with different medians for the lowest cut scores at rounds 2 and 3. The evidence suggests that the panels come from populations with the same median cut score at each of the other two positions at each round, at *p*<.01.

Table 4

Median Test for Panel Differences in Mathematics Cut Scores at each Position, each Round

		Cut Score		
Round		1st	2nd	3rd
1	*n*	51	51	51
	Median	5	13	22
	Chi-Square	8.6	5.7	4.5
	df	2	2	2
	p-value	.014	.057	.108
2	*n*	51	51	51
	Median	6	13	23
	Chi-Square	17.9	4.8	1.7
	df	2	2	2
	p-value	**.000**	.089	.421
3	*n*	51	51	51
	Median	6	14	23
	Chi-Square	14.3	0.4	0.4
	df	2	2	2
	p-value	**.001**	.831	.813

Note. H_0: Three panels come from populations with the same median.

For the three Mathematics panels, the Kruskal-Wallis test assessed the hypothesis that the three panels came from the same population regarding their choice of cut score and variation in choices at each performance level. The results (Table 5) indicated no support for this hypothesis for the lowest and middle cut scores at Round 2; the distributions of panel scores differ in Round 2 for the lowest two cut scores.

Table 5

Kruskal-Wallis Test for Panel Differences in Mathematics Cut Scores

		Cut Score		
Round		1st	2nd	3rd
1	Chi-Square	8.0	6.6	3.0
	df	2	2	2
	p-value	.018	.037	.220
2	Chi-Square	16.2	9.6	1.5
	df	2	2	2
	p-value	**.000**	**.008**	.464
3	Chi-Square	7.9	5.3	0.3
	df	2	2	2
	p-value	.019	.072	.856

Note. H_0: Three panels come from the same population.

In Language Arts (Tables 6 and 7), the Median Test indicated significant differences in medians for the middle and top cut scores in Round 1, but not the other rounds, at *p*<.01. These same results were found with the K-W test, suggesting that the distributions of scores across panels were different, as well as the middle cut score at Round 2 and the highest cut score at Round 3.

In summary, in Mathematics, the medians across panels for two cuts across the three rounds were significantly different (lowest cut at Round 2 and 3), whereas two distributions of cut scores were significantly different (both in Round 2). In Language Arts, the medians across panels for two cut scores were significantly different (middle and highest cut in Round 1), whereas four distributions of cut scores were significantly different.

Table 6

Median Test for Panel Differences in Language Arts Cut Scores at each Position, each Round

		Cut Score		
Round		1st	2nd	3rd
1	*n*	47	47	47
	Median	8	20	32
	Chi-Square	2.0	18.1	25.7
	df	2	2	2
	p-value	.369	**.000**	**.000**
2	*n*	47	47	47
	Median	8	22	34
	Chi-Square	0.7	8.1	4.4
	df	2	2	2
	p-value	.694	.017	.112
3	*n*	47	47	47
	Median	8	22	34
	Chi-Square	7.2	6.1	6.3
	df	2	2	2
	p-value	.028	.047	.043

Note. H_0: Three panels come from populations with the same median.

Table 7

Kruskal-Wallis Test for Panel Differences in Language Arts Cut Scores

		Cut Score		
Round		1st	2nd	3rd
1	Chi-Square	8.6	17.7	16.6
	df	2	2	2
	p-value	.014	**.000**	**.000**
2	Chi-Square	0.4	16.4	8.1
	df	2	2	2
	p-value	.816	**.000**	.017
3	Chi-Square	2.9	5.6	11.1
	df	2	2	2
	p-value	.243	.061	**.004**

Note. H_0: Three panels come from the same population.

With the Rasch item location values (*b*-parameters) for the Mathematics panels cuts scores, ANOVA can be used appropriately, given the interval-level nature of the item locations. A MANOVA model was used, with each of the three cut scores (item locations) as the dependent variables and panel and round as fixed factors (Table 8). There were no multivariate significant differences due to the interaction of round and panel (the round result did not depend on the panel) nor on rounds overall (changes in cut scores across rounds were not significant). However, there were significant differences found between panels for all three cut scores. For the lowest two cut scores (reaching Must Improve and reaching Satisfactory), cut scores set by Panel 3 were significantly lower than the other two panels (by a factor of .26 to .44 logits on the Rasch location scale). Table 9 contains the mean Rasch item locations for each cut score in each round by each panel.

Table 8

ANOVA Results for Mathematics Cut Scores based on Rasch Item Location Values

Source	Dependent Variable	Type III Sum of Squares	df	Mean Square	F	Sig.
Panel	**Must Improve**	3.323	2	1.661	8.241	**.000**
	Satisfactory	5.645	2	2.822	14.965	**.000**
	Excellent	.973	2	.487	1.015	.365
Round	Must Improve	1.225	2	.613	3.039	.051
	Satisfactory	.611	2	.305	1.619	.202
	Excellent	.055	2	.027	.057	.945
Panel × Round	Must Improve	.939	4	.235	1.165	.329
	Satisfactory	.301	4	.075	.399	.809
	Excellent	1.298	4	.325	.677	.609

Table 9

Descriptive Statistics of Mathematics Cut Score Rasch Item Locations by Round and Panel

Cut Score	Round	Panel	Mean	SD	*n*
1st	1	M1	-1.41	0.70	18
		M2	-1.31	0.46	16
		M3	-1.07	0.25	17
	2	M1	-1.27	0.41	18
		M2	-1.33	0.43	16
		M3	-0.88	0.34	17
	3	M1	-0.96	0.59	18
		M2	-1.26	0.28	16
		M3	-0.91	0.34	17
2nd	1	M1	-0.59	0.51	18
		M2	-0.50	0.54	16
		M3	-0.08	0.35	17
	2	M1	-0.67	0.52	18
		M2	-0.50	0.48	16
		M3	-0.15	0.27	17
	3	M1	-0.38	0.49	18
		M2	-0.40	0.38	16
		M3	-0.08	0.22	17
3rd	1	M1	0.66	0.76	18
		M2	0.90	0.75	16
		M3	1.04	0.69	17
	2	M1	0.71	0.50	18
		M2	0.90	0.88	16
		M3	0.96	0.57	17
	3	M1	0.92	0.73	18
		M2	1.00	0.81	16
		M3	0.78	0.47	17

As can be seen in Figure 4 regarding Mathematics panels, within panel (e.g., Panel M1), the variation tended to decrease from Round 1 to 3, but this was not uniform. This figure uses the theta values (Rasch item locations) as the outcome (Y-axis). The lowest set of three 95% confidence intervals within each panel constitutes the results for the lowest cut score and so forth. The clearest example of a uniform decrease in variation can be seen in Panel M3 for their highest cuts, where the cut score also decreased slightly from Round 1 to 3. Slight variation in final recommendation from Round 3 of each panel is also observable, with the highest third cut from M2 and the highest second cut from M3; the lowest cut is similar between M1 and M3 with M2 recommending a lower cut. Also note from the figure that the distance between each cut score was much greater for Panel M2 than M3, also, the distance between the lowest 2 cut scores tends to be less – notice the Round 3 shift in the lowest cut for Panel M1.

Figure 5 is essentially the same as Figure 4, with panel and round reversed. Each rectangle contains results from a single round, illustrating how panels varied within each round for each of the three cut scores. This display highlights variation in panels within each round, which is the focus of this analysis, so it will be used for the other subject area.

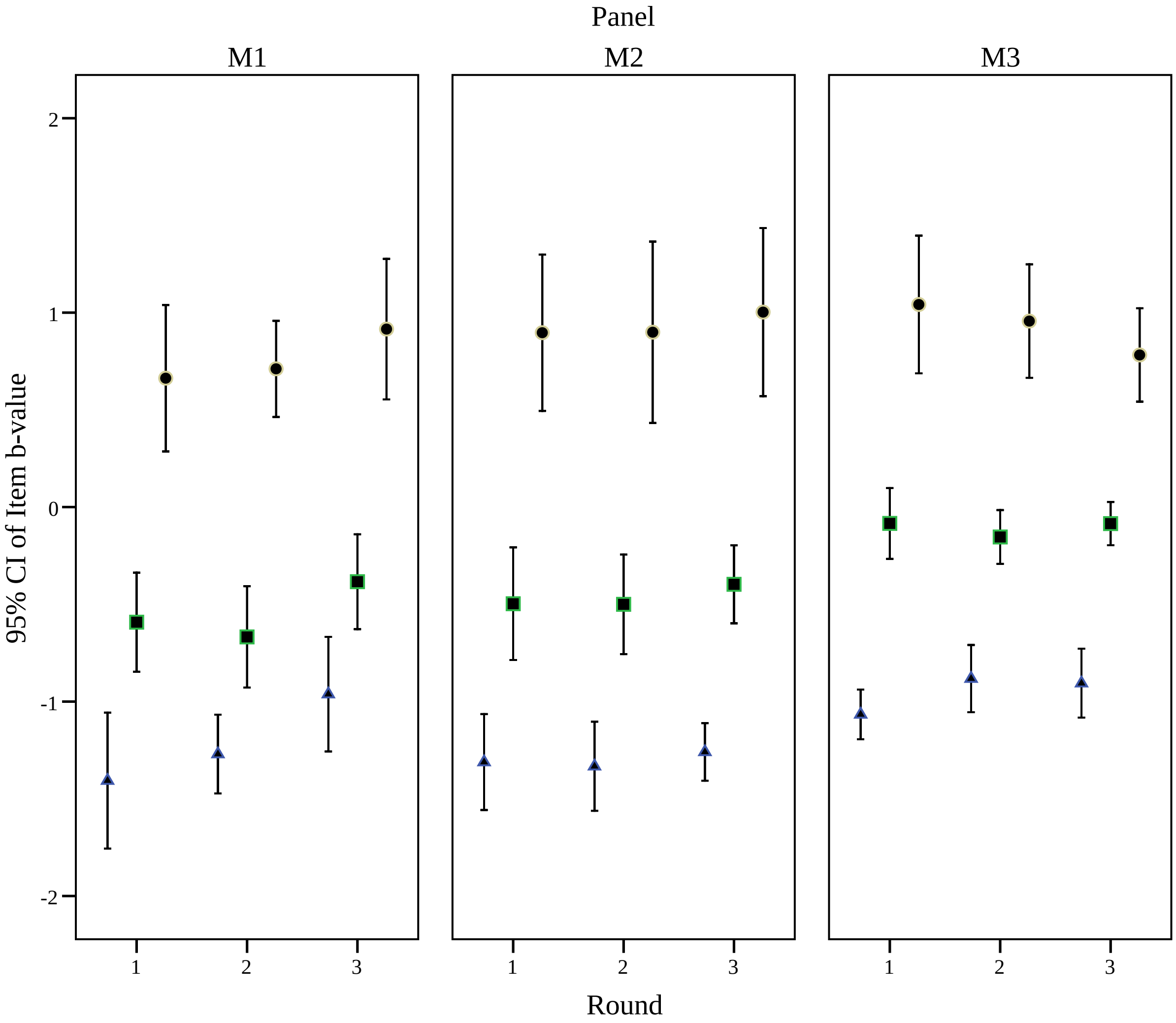


Figure 4. Mathematics cut score 95% confidence intervals given item Rasch locations by panel and round, where the lowest cut scores (must improve, triangles) are uniformly at the bottom of each panel and highest cut scores (excellent, circles) are at the top.

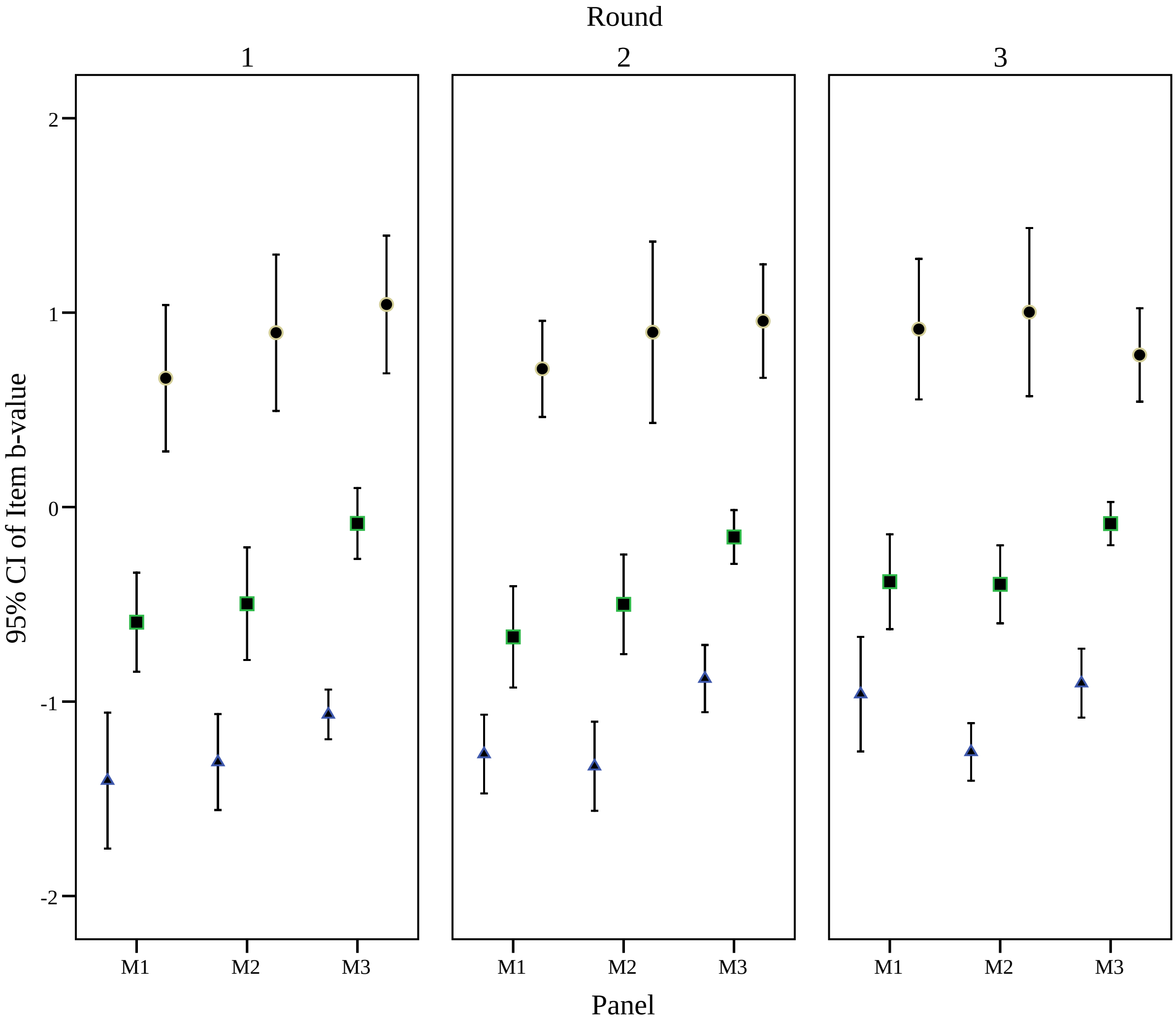


Figure 5. Mathematics cut score 95% confidence intervals given item Rasch locations by round and panel, where the lowest cut scores (must improve, triangles) are uniformly at the bottom of each panel and highest cut scores (excellent, circles) are at the top.

When examining the same display by item number (sequence number) rather than Rasch item location, as in Figure 6, the pattern is similar, but as seen in earlier displays, the analysis of item number creates visually larger differences between cut scores. One cautionary note is that these displays by item number illlustrate mean cut scores with 95% confidence intervals based on the standard error of the mean. Until this point, item numbers (sequence) have been treated as ordinal. In most cases the ordering of the cuts across panels remains the same whether we examine Rasch item locations or raw item numbers, except for the lowest cut score in Round 1 between Panels M1 and M2 (with Rasch item locations, M1 sets a lower cut than M2, but this is not the case with item numbers) and the highest cut score in Round 2 between Panels M1 and M2, where again M1 set a lower cut score than M2 based on Rasch locations, but this reverses slightly with item numbers).

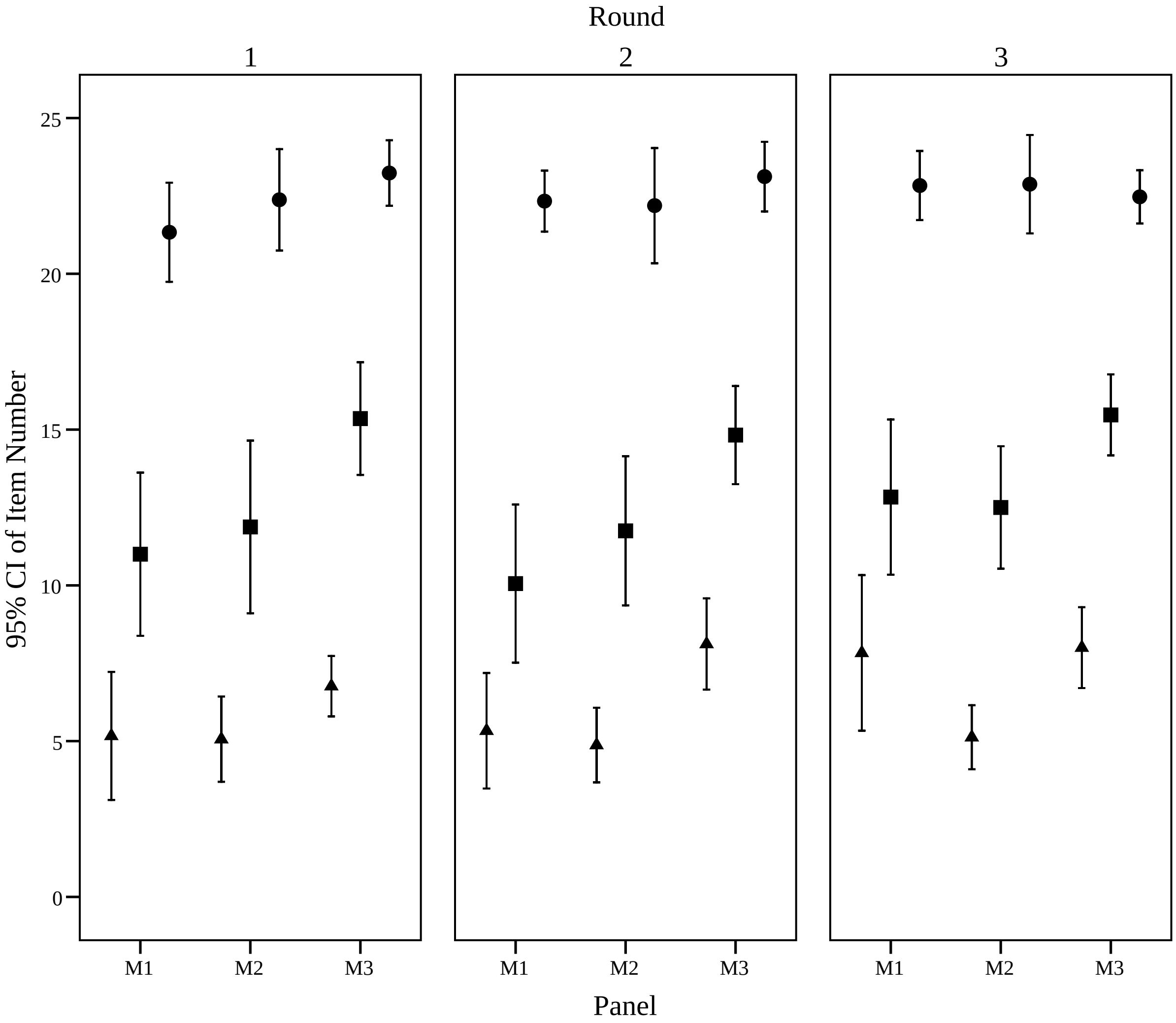


Figure 6. Mathematics cut score 95% confidence intervals given item number by panel and round, where the lowest cut scores are uniformly at the bottom of each panel and highest cut scores are at the top.

In Figure 7, the average item number selected as the cut score at each level in each round is displayed for each panel. One notable feature in this display is the greater variability between panels, particularly Panel L2. This panel tends to set the highest cut scores at each level in each round—notice in Round 1, Panel L2 sets the middle cut score near the top cut scores of the other two panels.

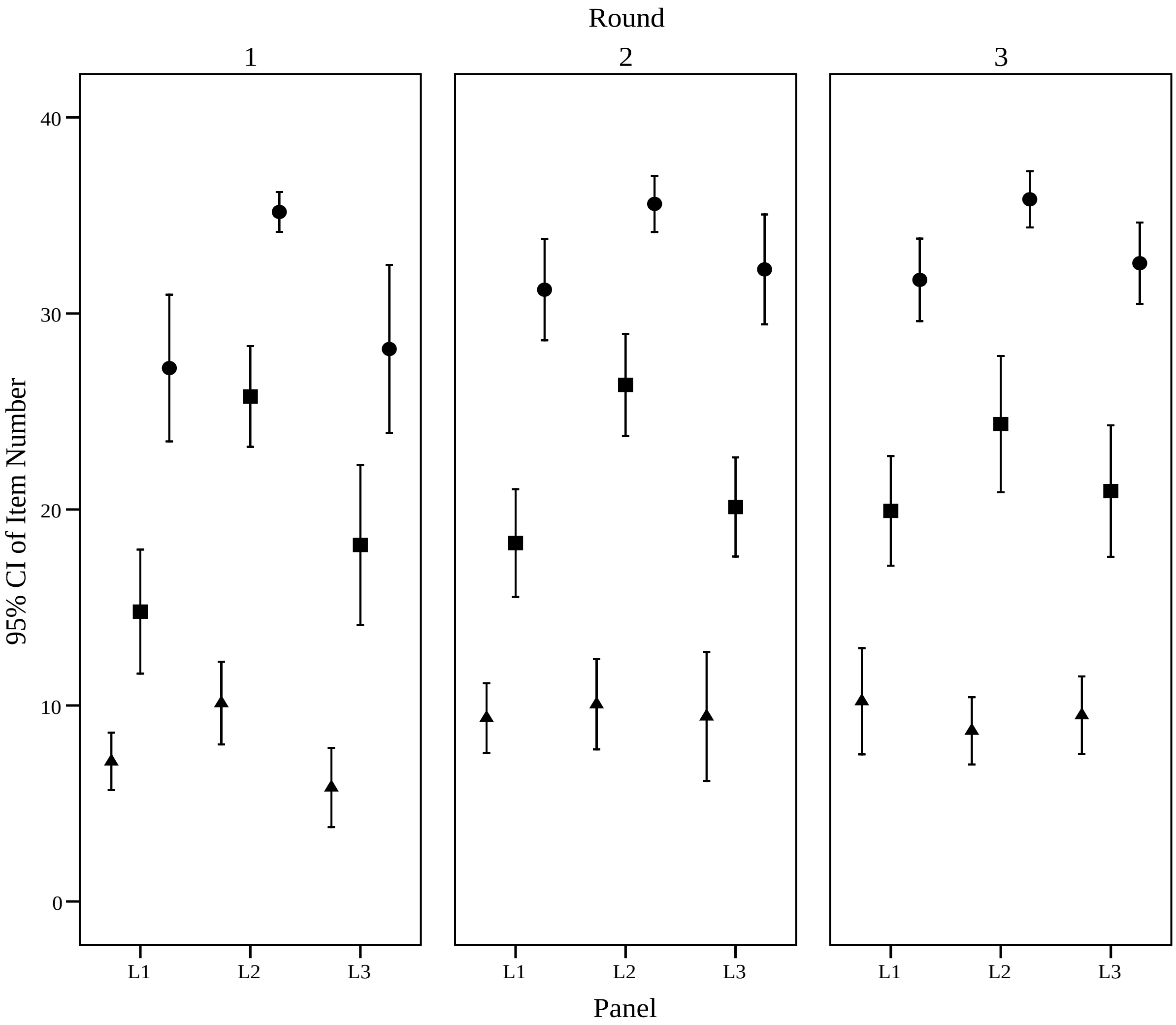
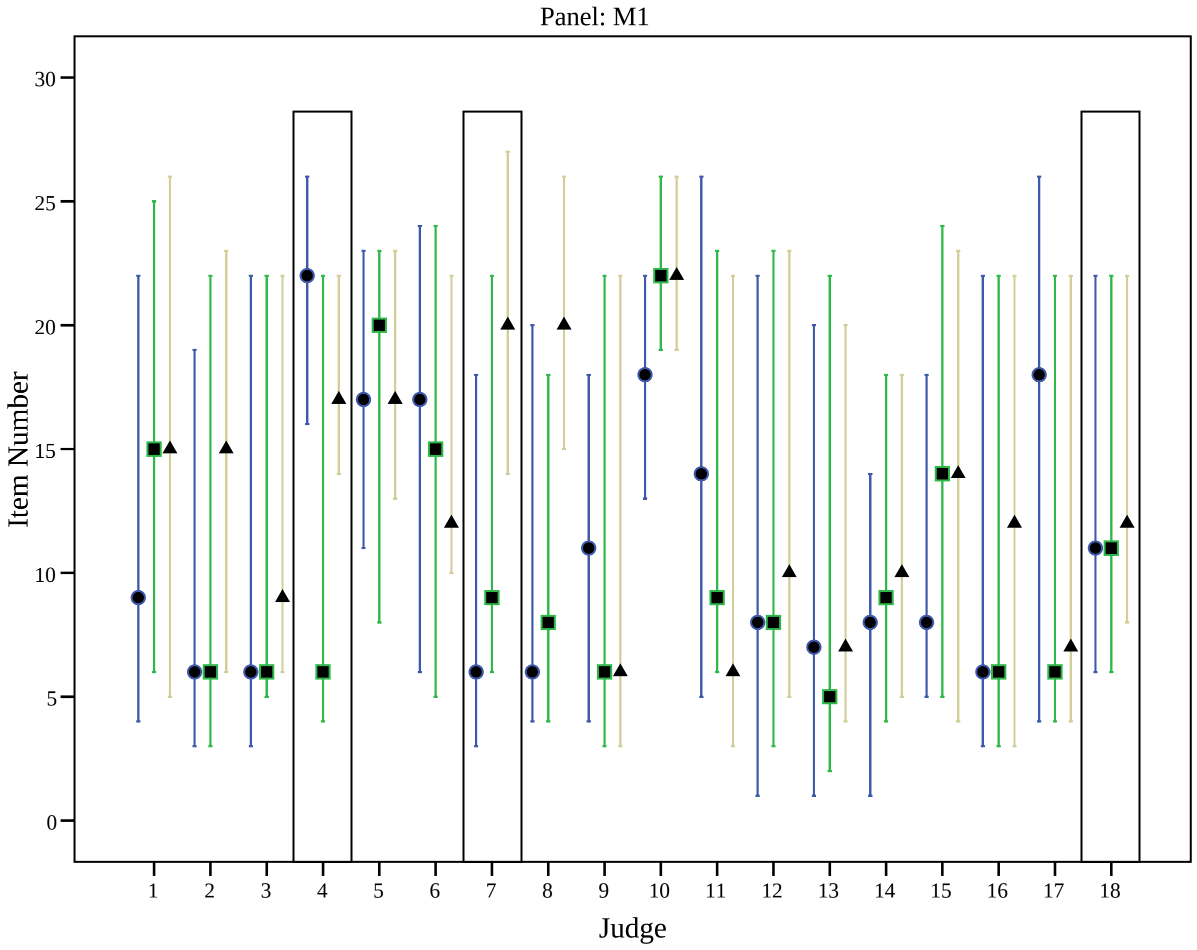


Figure 7. Language Arts cut score 95% confidence intervals given item number by panel and round, where the lowest cut scores are uniformly at the bottom of each panel and highest cut scores are at the top.

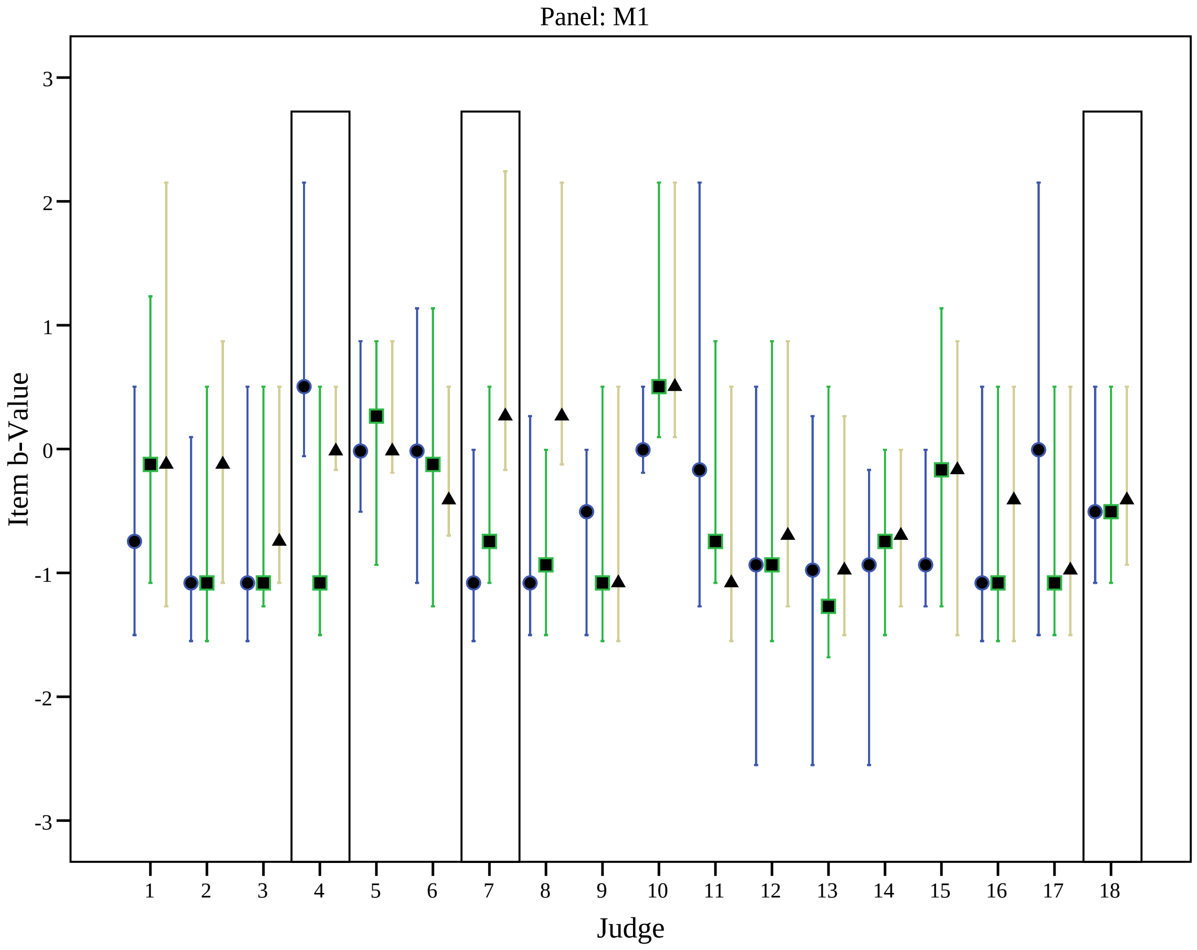
Another graphical display that conveys a great deal of information includes Figures 8 and 9. Here we can see all three cuts for each judge in each round – the lines represent the distance between the lowest cut and the highest cut, with the middle cut represented by the dot in the middle. The rounds follow in sequence from the first line to the third line for each judge. This panel in particular is discussed at length in the manuscript. From this figure we can see that judges were relatively homogenous on their bottom and top cuts, with little movement from Round 1 to 3. One judge in particular (Judge 11) was making a strong statement by severely dropping the middle cut (satisfactory or essentially passing the standard) to Item 6. Information from the evaluation feedback provide insight into scenarios such as these.



Note: For each judge, there are three lines representing the three rounds. Round 1 is a circle, Round 2 a square, Round 3 a triangle. Each line represents the 3 cuts, the lowest point is Must Improve, the middle point is Satisfactory, the highest point is Excellent.

Figure 8. Judge variation across rounds within Panel M1 (Mathematics item numbers).

In these figures, three judges are highlighted (encased in a rectangle). Judge 4 set all three cut scores relatively high in Round 1, significantly dropped each cut score in Round 2, and raised the two lower final cut scores near their original position. Judge 7 shows an increase in all three cut scores across all three rounds. Judge 18 shows very little variation in cut score placement across the three rounds.



Note: For each judge, there are three lines representing the three rounds. Round 1 is a circle, Round 2 a square, Round 3 a triangle. Each line represents the 3 cuts, the lowest point is Must Improve, the middle point is Satisfactory, the highest point is Excellent.

Figure 9. Judge variation across rounds within Panel M1 (Mathematics Rasch item *b*-value).

Figure 9 is essentially the same as the previous Figure 8, except that the scale is now based on Rasch item locations (*b*-values) instead of item number. Similar patterns can be seen, but appear less dramatic, particularly the variation within Judge 4.

Similar figures can be found in the Appendix (Figures A-G) for the remaining panels.

Evaluative Feedback from Judges

Our hope was to use information from the judges' evaluations of the sessions to explore variation in cut scores within and across panels. Unfortunately, this largely resulted in little to no significant relations between the outcomes of standard setting sessions and perceptions of the process. There were a handful of findings that were interesting, yet not unexpected. These are highlighted here. The evaluation questions are provided in the Appendix, translated into English.

To facilitate analysis of Evaluation reports from the judges, item sets were assessed through exploratory factor analysis using Principle Axis Factor extraction. The 5 items regarding the effectiveness of the session (Effective Session) overall were composed of a single factor, with coefficient alpha of .79. The 7 items regarding participants did not result in strong factors and was subsequently divided into 4 parts: a single item about prior knowledge of standard setting (Knew Process), two items regarding participant interest in the process (Participant Interest), a single item stating that the work during the sessions was difficult (Work Was Difficult), and three items regarding the active involvement of the participants (Participant Involvement). The 5 items regarding the quality of the organization and materials of the session (Session Organization) resolved into a single factor, with coefficient alpha of .73. The 7 items regarding the effectiveness of the facilitator (Effective Facilitator) resolved into a single factor, with coefficient alpha of .84.

The first question of the Evaluation Form was: "Do you believe that in the test booklet used to set cut scores, all, some, or none of the items were in order of difficulty, from easiest to most difficult?" Overall, 39% of panelists believed all items were in order of difficulty; 3% believed that none of the items were in order of difficulty (Table 10).

Table 10

Judge Beliefs about Item Order by Panel (Frequencies)

		Items in Order		
Subject Area	Panel	All Items	Some Items	None of the Items
Language Arts	L1	5	9	
	L2	5	12	
	L3	9	6	1
Mathematics	M1	4	13	
	M2	9	6	1
	M3	6	10	1

Based on correlations among all evaluation responses (Table 11), when judges believed the items in the ordered-item-booklet to be in order of difficulty, they tended to report the session to be more effective. Overall, judges reported to be more interested and more actively involved when they viewed the session to be effective, well organized, with an effective facilitator. Participant interest in the activity and their level of active involvement were related (r = .32). Participants who were less interested in the process were more likely to report that the work was very difficult.

Table 11

Spearman Correlations among Judge Evaluation Responses (n = 94 to 98)

		Items in Order	Effective Session	Knew Process	Participant Interest	Work was Difficult	Participant Involvement	Session Organization
Effective Session	*r*	**.230**						
	p	.024						
Knew about Process	*r*	-.148	.136					
	p	.152	.192					
Participant Interest	*r*	.043	**.434**	.111				
	p	.674	.000	.287				
Work was very Difficult	*r*	-.077	-.107	.147	**-.219**			
	p	.456	.301	.160	.033			
Participant Involvement	*r*	.152	**.411**	**.269**	**.319**	-.121		
	p	.138	.000	.009	.002	.242		
Session Organization	*r*	.124	**.562**	.135	**.361**	-.121	**.322**	
	p	.223	.000	.191	.000	.242	.001	
Effective Facilitator	*r*	.114	**.548**	.089	**.453**	-.171	**.326**	**.526**
	p	.265	.000	.393	.000	.096	.001	.000

With only 6 panels, correlations are difficult to assess. However, the purposes here were exploratory and we retained the decision rule of *p*<.05 to identify reliable correlations – requiring correlations to be greater than .7 to be significant. At the panel level (Table 11), when panels viewed the work to be difficult, they tended to report that the session was less effective, that they were less interested in the activity, and the organization of the session was of lower quality. When the session was viewed as more effective, facilitators were seen as more effective and panels were more interested in the process. Difficulty of the work had a much stronger impact at the panel level than at the individual judge level, resulting in three significant correlations (panel level) compared to one at the judge level. In addition, although not significant, if a panel tended to view the work as being very difficult, they were less likely to believe the items in the OIB were in order (*r* = -.61, *p*=.20).

Table 12

Pearson Correlations of Panel Mean Evaluation Responses (n = 6)

		Items in Order	Effective Session	Knew Process	Participant Interest	Work was Difficult	Participant Involvement	Session Organization
Effective Session	*r*	.396						
	p	.437						
Knew about Process	*r*	-.455	-.093					
	p	.365	.860					
Participant Interest	*r*	.393	**.842**	.306				
	p	.441	.036	.555				
Work was very Difficult	*r*	-.607	**-.920**	.080	**-.910**			
	p	.201	.009	.880	.012			
Participant Involvement	*r*	.097	.414	.665	.587	-.423		
	p	.855	.415	.149	.221	.404		
Session Organization	*r*	.563	.641	-.107	.668	**-.835**	.416	
	p	.245	.170	.840	.147	.039	.413	
Effective Facilitator	*r*	.045	**.902**	.051	.690	-.743	.501	.582
	p	.932	.014	.923	.129	.091	.312	.226

None of the panel cut score variance indices were related to aspects of the evaluation of the session, including the cut score levels themselves (Table 13). The only correlation that approached significance was between level of participant involvement and the highest cut score at the Excellent performance level (*r* = .73, *p* = .10).

Table 13

Pearson Correlations of Panel Evaluation Responses with Score Variability Indices (n = 6)

		Within Panels		Between Panels			
		Round 3 Cut Score Variance	Percent Change in Variance	Average Distance in Variance	Should Improve	Satisfactory	Excellent
Items in Order	*r*	-.057	-.370	-.235	-.026	-.071	.007
	p	.915	.470	.654	.962	.894	.990
Effective Session	*r*	.445	-.627	.298	-.087	.237	.375
	p	.376	.183	.566	.870	.651	.464
Knew about Process	*r*	.312	-.259	-.455	.603	.223	.444
	p	.547	.621	.364	.205	.671	.377
Participant Interest	*r*	.280	-.608	-.219	.175	.051	.336
	p	.591	.200	.676	.740	.924	.515
Work was Very Difficult	*r*	-.188	.511	.061	.139	.033	-.175
	p	.721	.300	.909	.792	.951	.741
Participant Involvement	*r*	.682	-.678	-.222	.415	.523	.734
	p	.136	.139	.673	.413	.287	.097
Session Organization	*r*	-.071	-.136	-.279	-.467	-.296	-.116
	p	.894	.797	.592	.351	.568	.827
Effective Facilitator	*r*	.539	-.488	.425	-.213	.308	.404
	p	.269	.327	.401	.686	.553	.427

To describe each index of variability:

1. Within Panel Round 3 Cut Score Variance: This is an index of the variance in the three cut scores set by a panel – the degree to which the three cut scores are spread out.
2. Percent Change in Variance: This is an index of the change in variance in the three cut scores from Round 1 to Round 3 – the degree to which a panel reduces the distance between the three cut scores.
3. Average Distance in Variance: This is an index of the average difference in the variance for a panel compared to the other two panels – the degree to which a panel sets cut scores that vary more than the other panels).
4. Should Improve, Satisfactory, Excellent: These are the actual cut scores set by each panel, based on item sequence number.

When examining indices of variability in cut scores within and between panels, several indices were related, as expected (Table 14). The middle and highest cut scores tended to be higher when there was more variation in the 3 cut scores within panel (clearly a result of the fact that the scores are spread out more). The highest cut score tended to be higher when the percent change in variation in cut scores from Round 1 to Round 3 was lower – when there was less change in panel variance in cut scores, the highest cut score tended to be higher.

Table 14

Pearson Correlations between Panel Level Score Variances and Cut Scores (n = 6)

		Within Panels		Between Panels
		Round 3 Cut Score Variance	Percent Change in Variance from Round 1 to Round 3	Average Distance in Variance from the Other Panels
% Change in Variance	*r*	**-.838**		
	p	.037		
Distance in Variance	*r*	.486	-.188	
	p	.329	.722	
Lowest Cut Score	*r*	.507	-.644	-.188
	p	.305	.167	.722
Middle Cut Score	*r*	**.959**	-.778	.526
	p	.002	.069	.284
Highest Cut Score	*r*	**.966**	**-.899**	.289
	p	.002	.015	.579

None of the correlations among within panel changes in cut score variability from Round 1 to Round 3 and evaluation responses were significant (Table 15).

Table 15

Pearson Correlations among Panel Evaluation Responses and Changes in Within-Panel Cut Score Variability (n = 6)

		Change in Cut Score Variance at Lowest Cut	Change in Cut Score Variance at Middle Cut	Change in Cut Score Variance at Highest Cut
Items in Order	*r*	-.357	-.696	-.479
	p	.488	.125	.336
Effective Session	*r*	-.128	.092	.049
	p	.810	.863	.926
Knew Process	*r*	.757	.128	-.485
	p	.082	.810	.330
Participant Interest	*r*	.353	-.219	-.345
	p	.493	.677	.503
Work was Difficult	*r*	-.033	.298	.205
	p	.951	.567	.697
Participant Involvement	*r*	.226	.041	-.645
	p	.667	.939	.167
Session Organization	*r*	.069	-.497	-.290
	p	.896	.315	.578
Effective Facilitator	*r*	-.087	.368	.205
	p	.870	.473	.697
Change in Variance Lowest Cut	*r*		-.247	-.301
	p		.637	.563
Change in Variance Middle Cut	*r*			.611
	p			.197

Summary & Discussion

This brief discussion is intended to provide a summary of the findings of our examination of variation in the independent replications of the Bookmark procedure in an developing country and to provide some initial guidance for continued investigations. Although most of the results were as excepted, there was significant variation in panel results in some cases, more so in Language Arts than Mathematics, but it is impossible to assess the degree to which this is due to panel variability and real differences in the nature of the assessments in two different areas. The following list summarizes findings:

Variability between Panels:

- In Mathematics, panel median cut scores differed on the lowest cut score at Rounds 2 and 3; cut score distributions differed for both the lowest and middle cut score at Round 2.
- In Mathematics, mean Rasch item location parameters were also examined with ANOVA, where significant differences in mean cut scores were found between panels for all three cut scores (as large as 0.44 logits); Panel 3 tended to set lower cut scores than the other two panels. This result did not significantly differ by round.
- In Language Arts, panel median cut scores and cut score distributions differed on the middle and highest cut scores at Round 1; cut score distributions also differed on the middle cut score at Round 2 and the highest cut score at Round 3.

Changing Variability within Panels:

- Variability of cut scores within panels was examined estimating mean cut scores and their standard errors, resulting in visual inspection of 95% confidence intervals (CI).
- In Mathematics, mean cut scores (based on item sequence number) varied more at the Satisfactory cut score (middle cut score), than the lowest or highest cut scores, at all three rounds.
    - In Round 1, all three panel CIs overlapped within each cut score. In Round 2, two panels differed for the lowest and middle cut scores. In Round 3, two panels differed for the lowest cut score.
    - There was a tendency for the variability within panel to be reduced across rounds, but this effect was not consistent across cut scores or panels.
- In Language Arts, mean cut scores also varied more at the middle cut score than the lowest or highest cut scores.
    - In Round 1, at least 2 panel CIs did not overlap within each cut score. In addition, Panel L2 set their middle cut score in the range of the highest cut score for Panels L1 and L3. In Round 2, the three panels were uniform at the lowest cut score, but two differed for the middle and highest cut scores. In Round 3, 2 panels differed at the highest cut score.
    - There was less of a tendency for within panel variability to be reduced across rounds in Language Arts. We also observe much more variability across panels, except at the lowest cut score.

Variability within Judge across Rounds:

- In examining judge variability, most judges in most panels were reluctant to change their initial cut scores across rounds.
- When examining variability using item sequence number versus Rasch item location parameters, the item number results look much more variable; Rasch item locations tended to reduce the variability between judges and within judges across rounds.
- In a few cases, dramatic score shifts were observed for particular judges across rounds, but these tended to be due to outlier Round 1 cut scores.
- In most panels, judges were less likely to change their lowest cut scores; these were relatively low (between 5 to 8 out of 29 items in Mathematics; about 8 out of 40 items in Language Arts).

Explaining Variation within and between Panels with Evaluative Feedback

- There were interesting relations among different aspects of participant experience and perceptions of the process (described in the manuscript), but the primary interest was examining the ability to explain panel variation.
- With only 6 panels, significance of correlations is difficult to assess. Six of the 48 correlations between feedback and indices of variability were larger than .60.
- The trends we observed were mostly in relations with “Panel % Change in Variance”, which is an index of the degree to which within-panel variance in cut scores changed across rounds.

- Decreases in Panel % Change in Variance tends to be associated with an increased sense of Session Effectiveness, Participant Interest, and Participant Involvement, and a lowered sense that the Work was Very Difficult.
- Participant Involvement also tended to be positively associated with higher variation in cut scores at Round 3 (cut scores differed more from lowest to highest cut scores), and the absolute level of the highest cut scores – panels where participants were more involved set higher Excellent-level cut scores and cut scores that were spread out more (some dependence here).
- Panels reporting higher prior knowledge of the process tended to set higher cut scores at the lowest level (Should Improve), which tended to be relatively low overall.

Do independent standard setting panels vary significantly? We do find that there are significant differences in some cases, including several statistically significant differences, as many as 7 of 18 comparison (employing a conservative significance level of $p< .01$). We also found practically significant differences in cut scores, typically in the range of 0.25 to 0.44 logits on the Rasch item location scale (Theta).

Should we examining panel variability by testing medians, ordinal distributions, or means and variances? We find more, but different differences when examining the entire ordinal distributions of item sequence numbers rather than testing median differences. We also find, in the case of Mathematics, more significant differences among panels using an ANOVA model with Rasch item location parameters rather than ordinal tests of item sequence number.

Are there significant connections between participant perceptions of the process and their involvement (participant evaluative feedback) and cut score variability within and between panels? Yes, to some degree, participant experience and perceptions of the process may help us understand variation within and between panel results. To do a better job of assessing these relations, a larger number of panels is needed, but the trends found here appear to be reasonable and consistent with other findings in the study.

Future Research

Additional work can be done in several areas. It is difficult to separate the results of particular panels and their membership and the particular facilitator. In this study, facilitator and panel membership are confounded. A crossed design would provide stronger information and provide an opportunity to potentially separate facilitator and panel membership effects – requiring multiple panels to be facilitated by the same facilitator.

Additional work is also needed in establishing strong measures of panel experience and perceptions of the process. As we uncover the relations between panel experiences and perceptions with panel outcomes, we can begin to elicit information that may be relevant to the training of panels.

Finally, we found similar results when examining variation in results across panels using item sequence number (the item number judges select for their bookmarks) and the Rasch item location parameters – such that variability between cuts scores at different levels and across judges appears to be less dramatic when examining Rasch item locations. In part, this may be a function of the tendency for item difficulties to not separate items as dramatically as do item sequence numbers, which are uniformly distributed across the location of the first item and the last item. It is important to conduct additional investigations on the use of item sequence number when providing normative feedback to judges prior to Rounds 2 and 3 compared to providing cut score distributions based on the ability needed to respond correctly to the item (Rasch item location). In some of the standard setting literature, there is attention given to the need to focus judges attention on the ability required to get the item right rather than attend to the percent correct metric (e.g., “20 out of 29 items is a lot.”). By providing feedback in terms of relative ability (avoiding the logit metric), rather than number of items (item sequence number), this could help facilitate increased attention on the ability required rather than percent correct thinking.

References

AERA, APA, NCME. (1999). *Standards for educational and psychological testing*. Washington DC, American Educational Research Association.

Cizek, G.J., Bunch, M.B., & Koons, H. (2004). Setting performance standards: Contemporary methods. *Educational Measurement: Issues and Practice, 23*(4), 31-50.

Kane, M.T. (2001). So much remains the same: Conception and status of validation in standard setting. In G. Cizek (Ed.), *Setting performance standards: Concepts, methods and perspectives* (pp. 53-88). Mahwah, NJ: Erlbaum.

Karantonis, A., & Sireci, S.G. (2006). The bookmark standard-setting method: A literature review. *Educational Measurement: Issues and Practice, 25*(1), 4-12.

Appendices

Tables A-F: Frequencies of Item Cut Score Placement across Judges by Panel and Round

Table G: Participant Evaluation of the Standard Setting Panel Process and Outcomes

Figures A-G: Judge variation across rounds within panel

Table A

Frequency of Item Cut Score Placement across Judges, by Panel and Round for Mathematics between Unsatisfactory and Needs Improvement

Unsatisfactory/Needs Improvement									
	Panel 1			Panel 2			Panel 3		
	Round	Round	Round	Round	Round	Round	Round	Round	Round
Item #	1	2	3	1	2	3	1	2	3
1	3		3	1	1				
2		1	3	1	1	2			2
3	4	4	3	3	2	1			1
4	4	4	2	2	2	2	3	1	2
5	2	3		2	4		1	1	
6	2	4	1	4	5	6	4	4	6
7				1		3	2	2	3
8		1	1				5	3	
9				1		1		2	1
10						1	2	1	1
11	1		1	1	1	1			1
12			2					2	
13	1		1						
14									
15								1	
16	1								
17			1						
18									
19		1							

Table B

Frequency of Item Cut Score Placement across Judges, by Panel and Round for Mathematics between Needs Improvement and Satisfactory

	Panel 1			Panel 2			Panel 3		
Item #	Round 1	Round 2	Round 3	Round 1	Round 2	Round 3	Round 1	Round 2	Round 3
3					1				
4				1					
5		1							
6	5	6	2	1		2			
7	1		2	3	3				
8	3	2		1	1	1			
9	1	3	1						
10			2	1	2	1	1	1	
11	2	1		1	1	2	1	1	
12			3	1		2	1	2	2
13				1	1	1	2	2	
14	1	1	1		1	1	3	3	7
15		2	2		1	1	2	3	
16				3	3	4	2		3
17	2		2	1	1		2	1	2
18	2			1	1	1		1	1
19							1	2	
20		1	2						1
21								1	1
22	1	1	1	1			1		
23							1		

Table C

Frequency of Item Cut Score Placement across Judges, by Panel and Round for Mathematics between Satisfactory and Excellent

	Panel 1			Panel 2			Panel 3		
Item #	Round 1	Round 2	Round 3	Round 1	Round 2	Round 3	Round 1	Round 2	Round 3
14	1								
15									
16									
17				1	1			1	
18	3	2	1	2	3	2			
19	1			1	1	1			1
20	2		1						1
21				1	1	1	4	1	2
22	6	9	8	1	1	2	4	4	4
23	1	3	4	5	5	5	3	4	5
24	1	2		1		1	1	2	3
25		1		1			1	3	
26	3	1	3	2	2	2	3	2	1
27			1	1	1	1	1		
28					1	1			

Table D

Frequency of Item Cut Score Placement across Judges, by Panel and Round for Language Arts between Unsatisfactory and Needs Improvement

	Panel 1			Panel 2			Panel 3		
Item #	Round 1	Round 2	Round 3	Round 1	Round 2	Round 3	Round 1	Round 2	Round 3
1							5		
2									
3	2							5	
4			1	1	1	2		1	
5	2	1	1						2
6	1					1	3	1	1
7	2	1		3			3		1
8	3	6	2	6	1	1	2	1	4
9	2	3	7		2	1	1		3
10				2					
11	2							1	1
12							2	1	1
13								2	
14		2		1	1	1		2	2
15				2	1	1			
16		1		1		1			
17			1						
18			2						1
19				1	1				
20								1	
21					1			1	

Table E

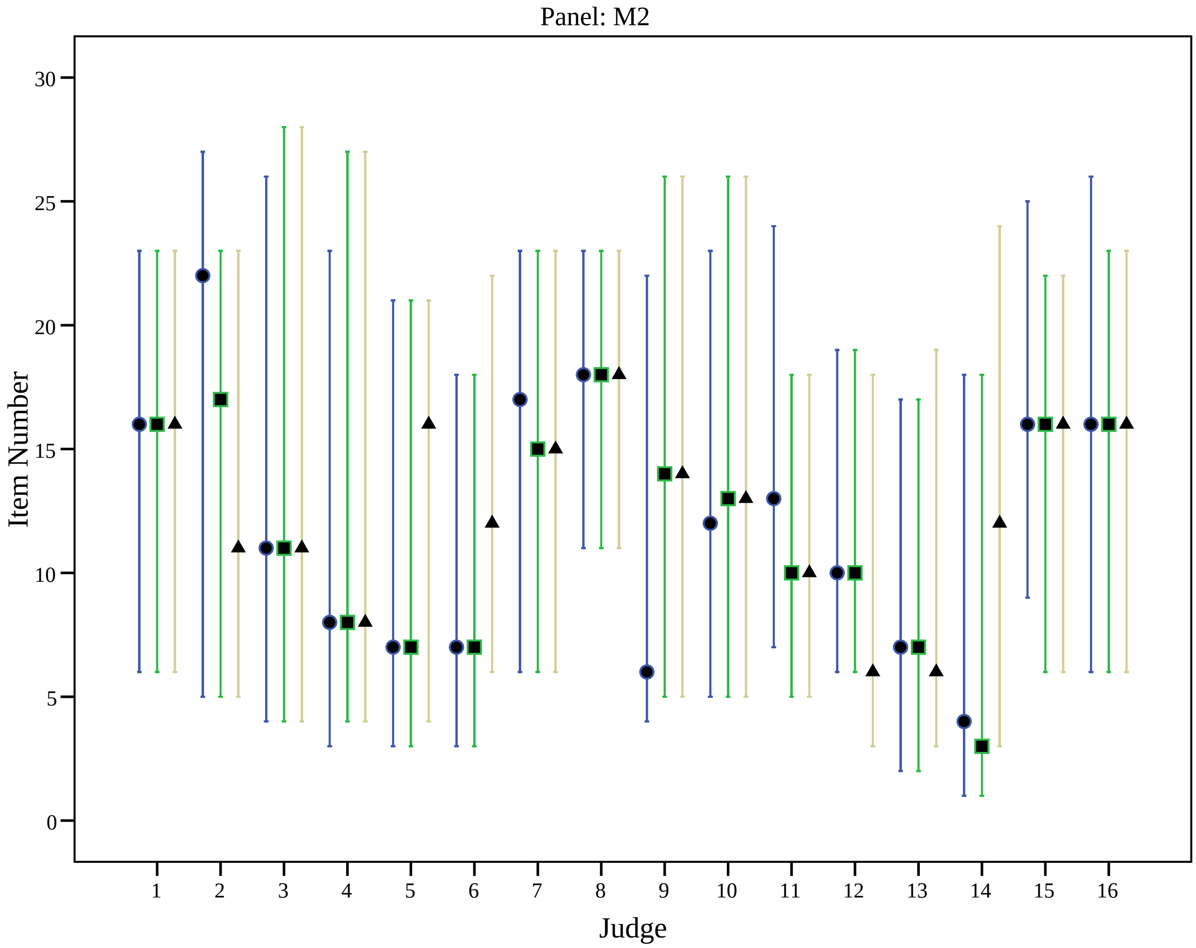
Frequency of Item Cut Score Placement across Judges, by Panel and Round for Language Arts between Needs Improvement and Satisfactory

	Panel 1			Panel 2			Panel 3		
Item #	Round 1	Round 2	Round 3	Round 1	Round 2	Round 3	Round 1	Round 2	Round 3
4	1								
5							1		
6									
7	1								
8									
9	1						1		
10		1		1					
11							1	1	
12	2								
13							2		
14	1	1						1	2
15		2	1		1	1	3		3
16	1	3	2					3	
17	3	2	5						1
18	1					1	2	1	2
19	1		1		1	1		1	
20				1	1			1	
21						1			
22	1		2	1	1	1	1	3	2
23	1	2					1	2	
24		1		1					3
25		2		2	1		1	1	
26				2	2	2			
27			1	1	2	1		1	
28			1	3	3	4	1	1	
29			1	1	1	1	1		2
30				4	2	1			
31						1			
32					1	1	1		
33									
34									
35									1
36					1				

Table F

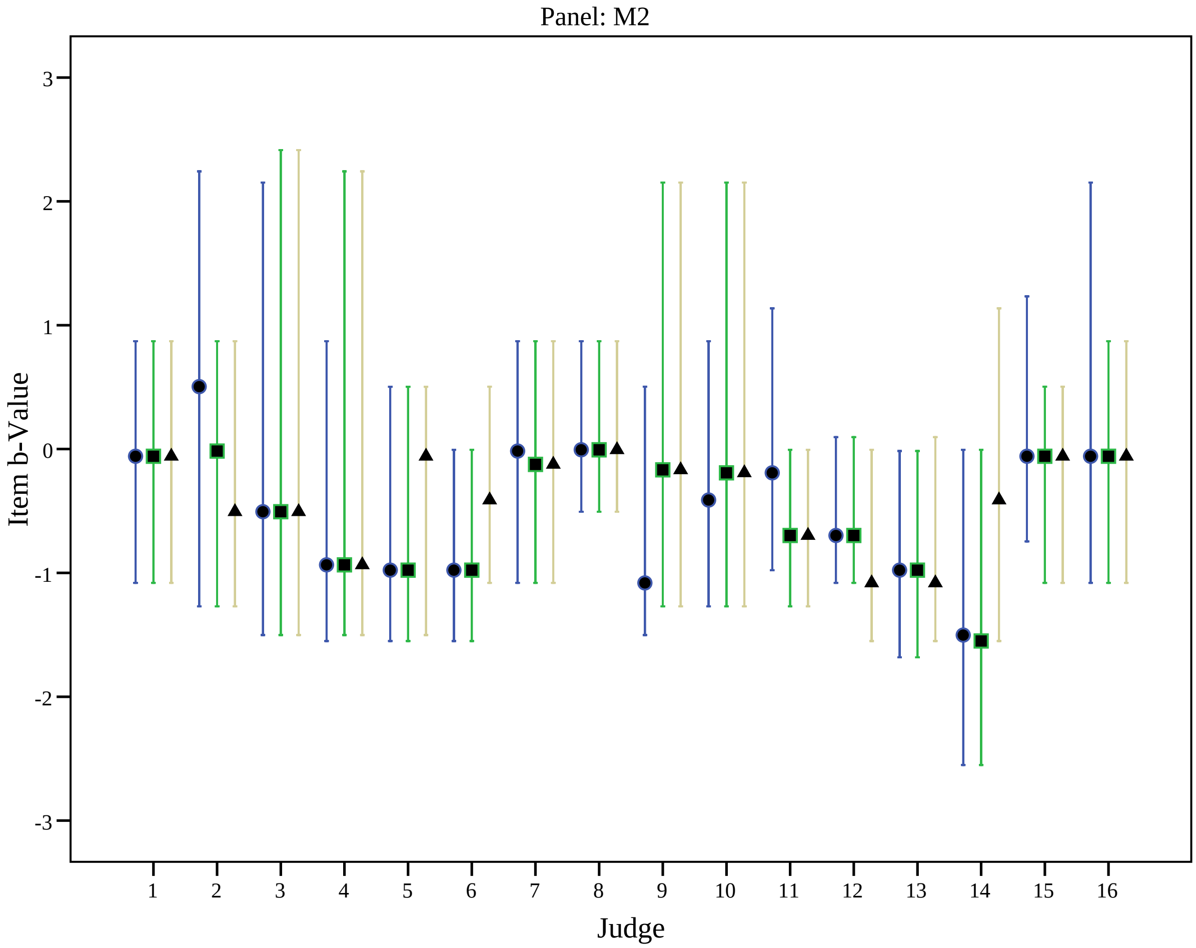
Frequency of Item Cut Score Placement across Judges, by Panel and Round for Language Arts between Satisfactory and Excellent

	Panel 1			Panel 2			Panel 3		
	Round	Round	Round	Round	Round	Round	Round	Round	Round
Item #	1	2	3	1	2	3	1	2	3
8							1		
9-14									
16	2								
17									
18	1								
19								1	
20							2		
21									
22							1		1
23								1	
24		1							
25	1	2	1				1		
26			1				1		
27		1	1						
28	2							1	
29	1						4	1	
30	5	1			1	2			2
31		2	3	1					4
32	1	1	3		2		1	1	2
33				2	1	1	1	3	1
34		1	1	3	2	2		3	
35		4	1	5	2	1		1	2
36			3	1	2	2	2	2	1
37				2		2	1		2
38		1		3	6	6		1	1
39	1					1		1	
40					1		1		



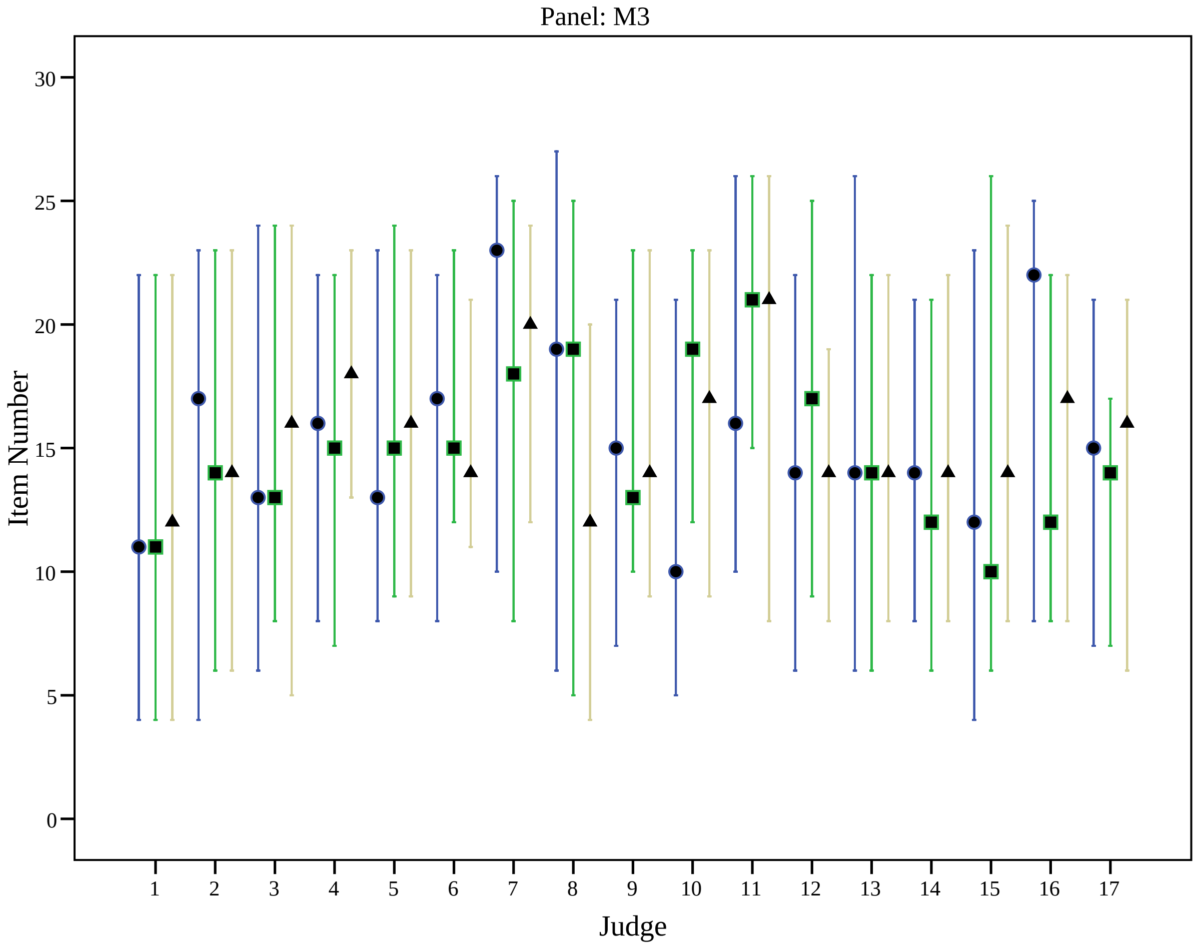
Note: For each judge, there are three lines representing the three rounds. Round 1 is a circle, Round 2 a square, Round 3 a triangle. Each line represents the 3 cuts, the lowest point is Must Improve, the middle point is Satisfactory, the highest point is Excellent.

Figure A. Judge variation across rounds within Panel M2 (Mathematics item numbers).



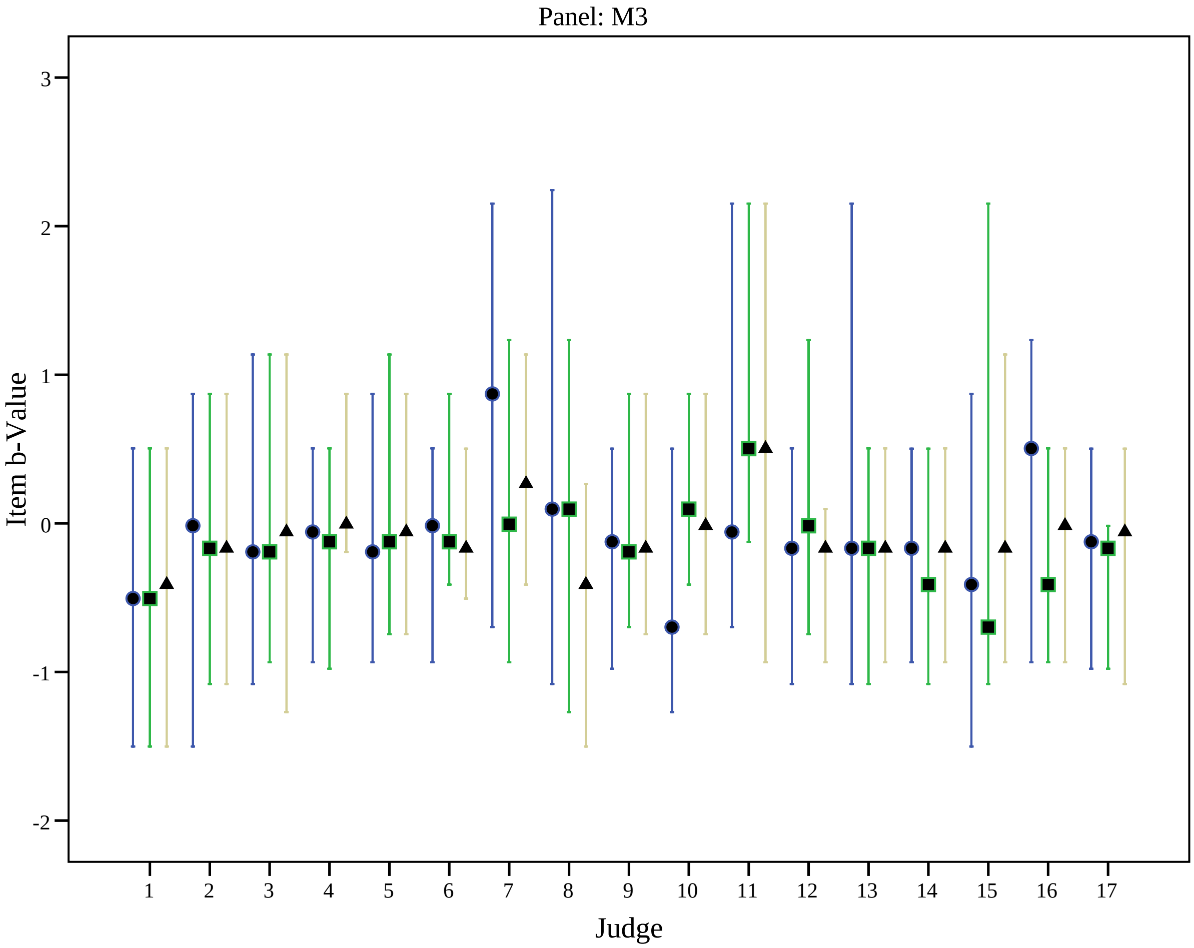
Note: For each judge, there are three lines representing the three rounds. Round 1 is a circle, Round 2 a square, Round 3 a triangle. Each line represents the 3 cuts, the lowest point is Must Improve, the middle point is Satisfactory, the highest point is Excellent.

Figure B. Judge variation across rounds within Panel M2 (Mathematics Rasch item *b*-value).



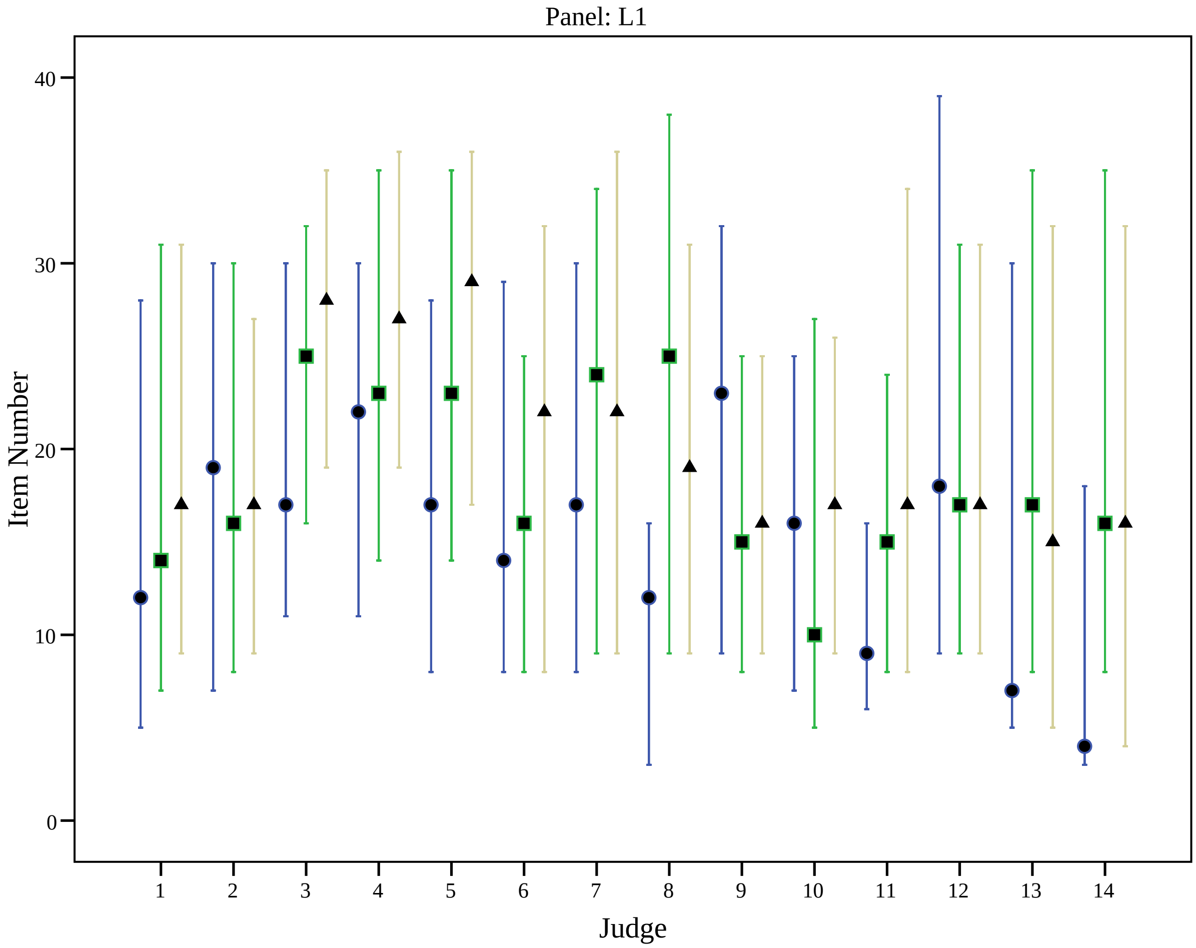
Note: For each judge, there are three lines representing the three rounds. Round 1 is a circle, Round 2 a square, Round 3 a triangle. Each line represents the 3 cuts, the lowest point is Must Improve, the middle point is Satisfactory, the highest point is Excellent.

Figure C. Judge variation across rounds within Panel M3 (Mathematics item numbers).



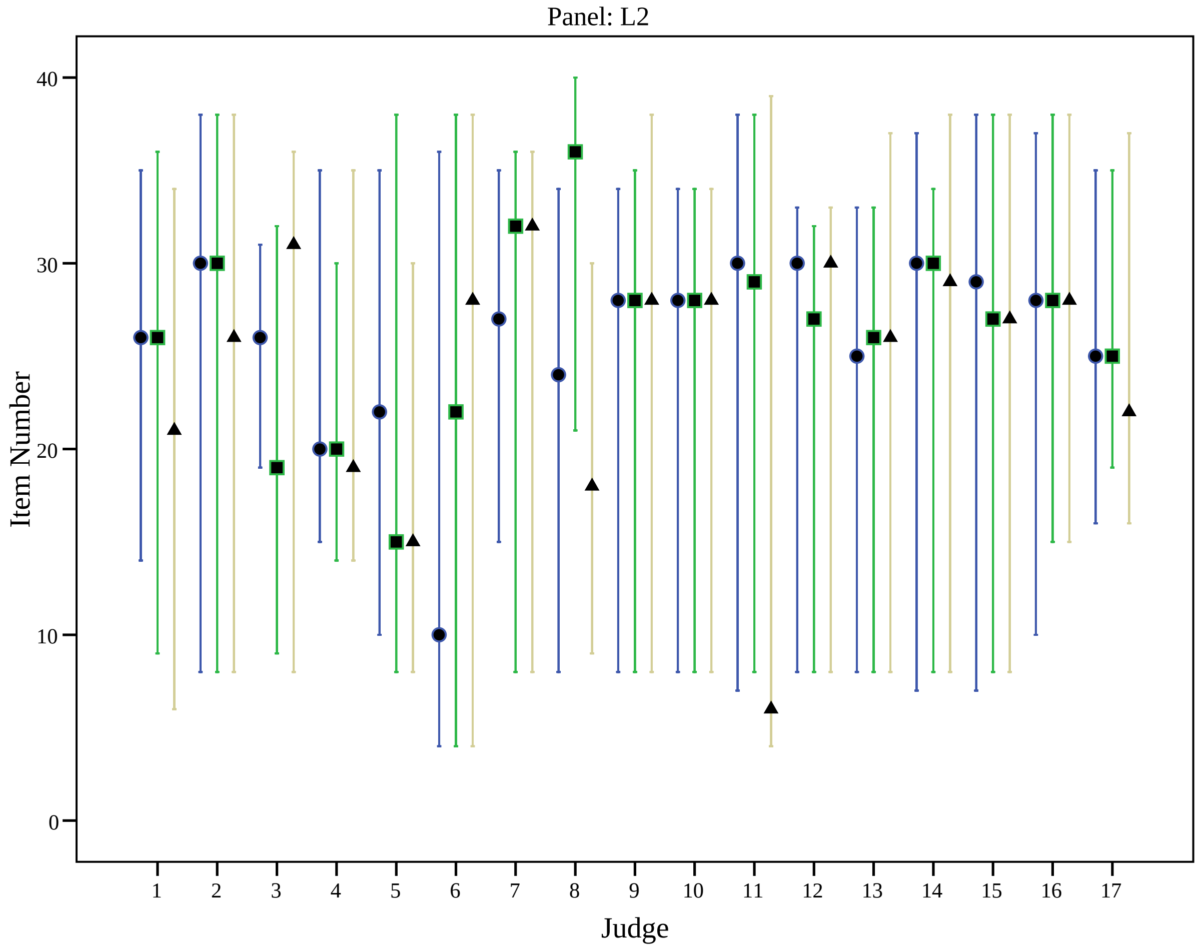
Note: For each judge, there are three lines representing the three rounds. Round 1 is a circle, Round 2 a square, Round 3 a triangle. Each line represents the 3 cuts, the lowest point is Must Improve, the middle point is Satisfactory, the highest point is Excellent.

Figure D. Judge variation across rounds within Panel M3 (Mathematics Rasch item *b*-value).



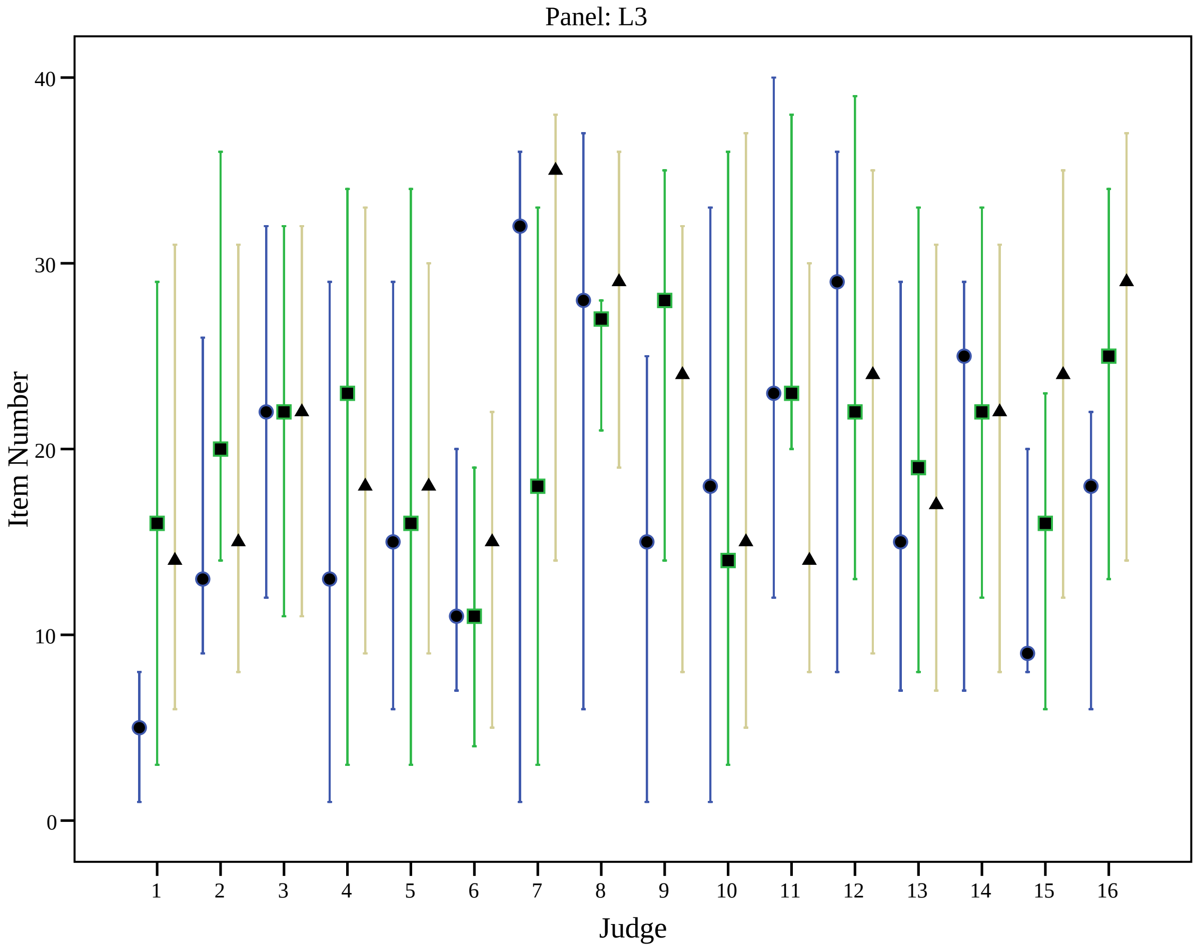
Note: For each judge, there are three lines representing the three rounds. Round 1 is a circle, Round 2 a square, Round 3 a triangle. Each line represents the 3 cuts, the lowest point is Must Improve, the middle point is Satisfactory, the highest point is Excellent.

Figure E. Judge variation across rounds within Panel L1 (Language Arts item numbers).



Note: For each judge, there are three lines representing the three rounds. Round 1 is a circle, Round 2 a square, Round 3 a triangle. Each line represents the 3 cuts, the lowest point is Must Improve, the middle point is Satisfactory, the highest point is Excellent.

Figure F. Judge variation across rounds within Panel L2 (Language Arts item numbers).



Note: For each judge, there are three lines representing the three rounds. Round 1 is a circle, Round 2 a square, Round 3 a triangle. Each line represents the 3 cuts, the lowest point is Must Improve, the middle point is Satisfactory, the highest point is Excellent.

Figure G. Judge variation across rounds within Panel L3 (Language Arts item numbers).

Session Evaluation Questions

A. Session Effectiveness.
The order of the contents of our work seemed adequate.
It seems that the quantity of the work contents is adequate.
The development treated the contents in depth.
The contents have utility for practical application.
The time assigned to work is the adequate.

B. Participant Measures.
1. I knew a lot of the themes that we worked on in the workshop. *

2. The theme that we worked on in the workshop interested me a lot. *i*
3. I learned new know-how with this workshop. *i*

4. The themes that worked on were very difficult for me.*

5. My participation during the two days was active. *p*
6. I called attention to read the material used. *p*
7. I arrived on time to start the work of the workshop. *p*

C. Session Organization.
The content of the support material seemed very good.
The quality of the support material (design and presentation) is very good.
I complied with the schedule assigned to the activities.
The workday was well organized.
The conditions and physical state of the localities are adequate.

D. Facilitator Effectiveness.
I noted that the facilitator prepared the expositions.
The work of the facilitator very was organized.
The facilitator utilized adequately the resources.
The instructions for the development of the activities were clear.
The facilitator showed ability to communicate with the participants.
The facilitator resolved the doubts of the participants.
The facilitator created a climate of participation.

i indicates the items used to measure Participant Interest.

p indicates the items used to measure Participant Involvement.

* indicates single items used as measures.