WIDA™
UNIVERSITY OF WISCONSIN–MADISON

Technical Report

No. TR-2024-1
April 2024

Validating a new writing scoring scale using multi-faceted Rasch analyses

Lead author: Ping-Lin Chuang. Additional authors: Mark Chapman, Ph.D.; Kyoungwon Bishop, Ph.D.; Ahyoung Alicia Kim, Ed.D.

Wisconsin Center for Education Research
SCHOOL OF EDUCATION
UNIVERSITY OF WISCONSIN–MADISON

Contents

1. Introduction

Rating scales are commonly used in performance-based assessments to ensure that the characteristics of examinee responses (e.g., quality of idea development or language use) can be properly identified. Rating scales are crucial to the quality of assessments; poorly designed scales with unclear descriptors or too few/many criteria can potentially jeopardize test reliability and validity. This means rating scales should be carefully crafted and validated so that scores can be interpreted and used meaningfully. The current study is situated within a larger ongoing project of developing a new scoring scale for the WIDA ACCESS for ELLs[1] Writing test. The project aims to validate the newly developed writing scoring scale by examining its feasibility in differentiating student ability and practical scoring use.

2. Literature review

This section provides the theoretical background of the current study by discussing how rating scales are developed and validated. It also introduces the validation method employed in this study, namely multi-faceted Rasch analysis.

2.1 Scale development and redesign

The development of rating scales can be mainly categorized into two approaches (Turner & Upshur, 2002). The first approach is theory-based and uses "theoretical views about the development of L2 ability" (p. 50) to develop scale descriptors. While these scales have strong theoretical support, they are often criticized as being irrelevant to the test task or unclear due to the use of relative wording. To address these issues, empirically-based scales are developed. These scales require scale developers to identify response characteristics based on sample test responses and create descriptors accordingly. Although empirically-based scales can reflect ability differences effectively, they are also deemed atheoretical and their development can be time-consuming. Despite these drawbacks, empirically-based scales are still favored in performance-based assessments for their content relevance.

The lengthy process of developing a rating scale does not guarantee its permanent use in the same assessment context. Redesigning a rating scale is necessary when the assessment needs to serve different or additional purposes (e.g., Knoch, 2007, 2009) and when there is an update of test content or standards (e.g., Banerjee et al., 2015; WIDA, 2020). For instance, Knoch (2009) compared two scales to determine their applicability in a diagnostic context. One of the scales was originally used in a placement test, with less specific descriptors; the other was empirically developed and had detailed level descriptors. The results showed that when using the empirically-developed scale, raters achieved higher reliability and performed better in terms of distinguishing multiple aspects of writing. This demonstrates the importance of using an appropriate rating scale when administering a test and that there is not a one-size-fits-all scoring rubric that can be used across different tests.

2.2 Scale validation

Once a rating scale has been developed, validation should be performed to ensure its quality and functionality. Scales can be validated quantitatively and qualitatively. Common quantitative validation methods include correlational analyses and multi-faceted Rasch analyses. For a rating scale with multiple criteria, a correlational analysis can be used between scores derived from the scale and external

[1] The WIDA ACCESS for ELLs assessment is taken by K–12 English learners, with the purpose of monitoring their progress toward English language proficiency.

measures to evaluate the constructs different criteria represent (Becker, 2018). A multi-faceted Rasch analysis is commonly performed to examine the psychometric properties of a rating scale. It combines different facets such as examinees, raters, scoring criteria, or test items into one analysis and converts raw scores into a logit interval scale (Linacre, 2004). (See section 2.3.) For qualitative approaches, researchers conduct surveys or interviews with raters or educators (e.g., Becker, 2018; Knoch, 2007). Becker (2018) interviewed teachers in a college-level intensive English program to understand how they felt about the utility of a writing scale. Knoch (2007) surveyed raters by eliciting their opinions about a newly designed scale. Qualitative results can complement quantitative findings by providing insights into the "how" and "why" of rating behaviors. Thus, it is common for validation research to adopt both approaches to gain a complete overview of the quality of a rating scale.

2.3 Conducting multi-faceted Rasch analyses

Rasch modeling is the backbone of a multi-faceted Rasch model, which is an application of item response theory (Bond & Fox, 2015; Linacre, 2004). A Rasch model can be used in dichotomous scoring while a more sophisticated model like the multi-faceted Rasch model is useful in rating constructed responses in performance-based assessment. A multi-faceted Rasch model considers how different characteristics such as examinees, raters, and test items affect test scoring (Myford & Wolfe, 2003). Accordingly, this model can answer different questions regarding rater severity and consistency, item difficulty, and student ability.

Multi-faceted Rasch analyses are frequently conducted in language assessment research. They have been used to evaluate rater behavior (e.g., Eckes, 2005; Goodwin, 2016; Yan, 2014), examine scoring scales (e.g., Becker, 2018; Knoch, 2007; Li, 2022), and perform standard setting (e.g., Hsieh, 2013). For example, Goodwin (2016) compared essay rater behavior in university-level admission and placement tests. She found that a few raters showed statistically significant bias on either test type, thus suggesting additional training for the raters. Li (2022) analyzed the rating behavior of teachers and peer raters, and checked the quality of an EFL (English as a Foreign Language) writing scoring rubric. In terms of rater performance, the teachers were stricter raters than the students. As for the scale quality, his results indicated that the criteria of the scoring rubric were well-designed, but the scoring band could be wider to better differentiate examinees' writing skills. Overall, running multi-faceted Rasch analyses is an effective way to examine rating quality in writing assessments and is a helpful tool in studies involving rater effects or scale validation.

3. Methods

3.1 Context of the study

ACCESS for ELLs is a suite of English language proficiency assessments developed by WIDA and their partner, the Center for Applied Linguistics (CAL). ACCESS is taken by K–12 English learners, and the scores help educators make decisions about students' academic language and determine if they will benefit from English language support services. The Writing domain of ACCESS assesses English writing skills required in school. For each grade-level cluster, there are two test forms, Tier A and Tier B/C. Lower proficiency students take the Tier A test form while higher proficiency students take the Tier B/C test form. Currently, ACCESS written responses are scored by Data Recognition Corporation (DRC) raters (WIDA's scoring vendor for ACCESS) using a six-point holistic scale that also features "plus" score points between the solid scores (i.e., 11 score points in total). However, to reflect the expectations of the WIDA English Language Development Standards Framework, 2020 Edition (i.e., Language for Social and Instructional Purposes, Language for Language Arts, Language for Mathematics, Language for Science,

and Language for Social Studies) and to address the concerns about infrequently used score points of the current scale, a new writing scoring scale was developed in the hope of improving usability for raters.

The new writing scoring scale (see Appendix A) is an eight-point holistic scale, with a raw score range of 0 to 7. WIDA and CAL followed an empirical approach (Turner & Upshur, 2002) to develop the scale by using a corpus of 324 ACCESS responses. These included responses from grades 1 to 12, from both Tier A and Tier B/C, and those targeting multiple WIDA ELD Standards (especially Language for Language Arts and Language for Science responses). The responses were sorted into eight groups and the scoring scale descriptors were developed based on the sorting process and the characteristics of each group. The scale was then reviewed by multiple groups of reviewers: WIDA and CAL staff with expertise in the 2020 Edition, the ACCESS/Screener scoring process, and social justice; a representative group of DRC writing raters; and educators who score WIDA Screener responses.

The current study examines how the newly developed writing scoring scale functions by testing the following four hypotheses:

1. A well-functioning rating scale will result in all score points being used and no single score point being overly used (variation in ratings).
2. A well-functioning rating scale will result in small differences between raters in terms of their leniency and harshness as a group (rater separation).
3. A well-functioning rating scale will result in high rater reliability as indicated by rater point biserial correlations and exact agreement rates (rater reliability).
4. A well-functioning rating scale will result in high candidate discrimination (student discrimination).

Archived ACCESS responses were scored by DRC raters and their rating performances were analyzed psychometrically.

3.2 Data

Responses from 1200 students in grades 1 to 12 (i.e., five grade-level clusters) were selected for the study. There were 12 writing tasks involved, with 100 responses for each. The selection of these responses aimed to replicate the variety of student performances in operational testing. Each grade-level cluster consisted of a unique set of tasks, students, and raters (see Table 1 for the corresponding distribution of rating assignments). The written responses were rated by 51 DRC raters using the newly developed rubric. The goal was to have each rater within a grade-level cluster score all 100 responses to each task, but for operational reasons this was not always possible. There were fewer scores assigned than planned in grade-level clusters 2-3, 4-5, and 6-8 but the final dataset contained 12,040 observations, which is very close to the planned target of 12,000 scores.

Table 1
Data distribution by grade-level cluster, task, and rater

Grade-level cluster	Tasks (Tier A + Tier B/C)	Raters	Number of scores assigned
1	3 (2+1)	11	3,220
2–3	2 (1+1)	10	1,920
4–5	2 (1+1)	10	1,900
6–8	2 (1+1)	9	1,700
9–12	3 (1+2)	11	3,300
Total	12 (6+6)	51	12,040

3.3 Data analysis

The dataset was analyzed using the many-facet Rasch model (MFRM) through the FACETS program version 3.80.0 (Linacre, 2017). We divided the dataset into five subsets to correspond to the five grade-level clusters. To understand how the scale functions in all writing tasks across grade-level clusters, separate analyses were performed using the respective data subsets. For each grade-level cluster, we fitted a rating scale model (RSM) that includes three facets: student, rater, and item (writing task). We examined the logit span of the three facets with the Wright map and collected various statistics related to student ability, rater severity, and scale category. These included student ability estimates and separation indices, allowing us to confirm whether the scale yields high or low student discrimination. Rater severity estimates and separation indices helped us understand how similar raters are in terms of severity. Exact agreements and infit and outfit mean square values provided a glimpse into inter- and intra-rater reliability. Scale category statistics such as score frequency and Rasch-Andrich threshold measures helped determine the psychometric quality of the scale. Then, to understand how individual writing tasks work with the scale, we fitted separate three-facet partial credit models (PCM): student, rater, and item. This allowed us to identify unexpected responses to determine if raters had a problem scoring specific writing tasks or student responses. For both models, group anchoring was performed on students so that the data could be linked.

4. Results

4.1 Descriptive statistics and model fit

Descriptive statistics of the scores assigned by DRC raters are shown in Table 2, Table 3, Figures 1-5, and Appendix B. Tables 2 and 3 present the mean, standard deviation (*SD*), minimum (Min), and maximum (Max) of each writing task and each grade-level cluster. Figures 1-5 show the distribution of score levels for each grade-level cluster while Appendix B contains the detailed raw counts. The lowest and highest scores assigned were 0 and 7 for all grade-level clusters, while the mean raw scores had an ascending trend as the grade-level cluster increased. Some writing tasks had a 0–5 or 0–6 range because they are Tier A tasks for lower proficiency test takers. The highest scores on the rubric are often not expected to be awarded to Tier A responses. For reference, Figures 6-10 and Appendix C show the distribution of score levels with the current scale.

Table 2

Descriptive statistics of assigned scores for each writing task

Grade-level cluster	Writing task	Tier level	Mean (*SD*)	Min - Max
1	Cleaning Up	A	2.16 (1.50)	0 – 6
1	Growing Plants	A	1.84 (1.34)	0 – 6
1	Giant Pandas	B/C	3.36 (1.57)	0 – 7
23	Garden Surprise	A	2.58 (1.62)	0 – 7
23	Changing Water	B/C	3.48 (1.39)	0 – 7
45	Marsh Ecosystem	A	2.10 (1.29)	0 – 5
45	Search for Info	B/C	3.60 (1.59)	0 – 7
68	Illustrator	A	2.97 (1.36)	0 – 6
68	Color and Temperature	B/C	4.09 (1.66)	0 – 7
912	Cherry Trees	B/C	4.54 (1.52)	0 – 7
912	Elasticity Investigation	B/C	4.33 (1.49)	0 – 7
912	Where to Volunteer	A	2.82 (1.34)	0 – 6

Table 3

Descriptive statistics of assigned scores for each grade-level cluster

Grade-level cluster	Mean (*SD*)	Min - Max
1	2.47 (1.61)	0 – 7
23	3.02 (1.58)	0 – 7
45	2.81 (1.62)	0 – 7
68	3.50 (1.61)	0 – 7
912	3.90 (1.64)	0 – 7

Figure 1

Distribution of score levels with the new scale: grade-level cluster 1 frequency graph

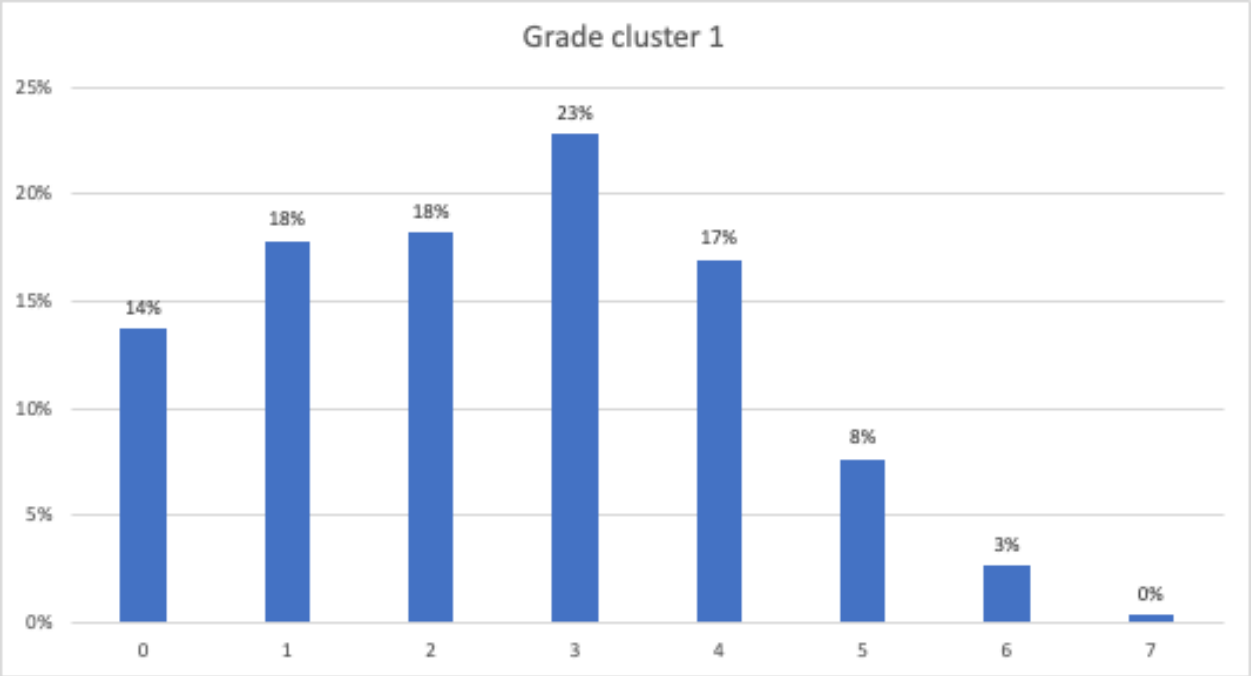


Figure 2

Distribution of score levels with the new scale: grade-level cluster 2–3 frequency graph

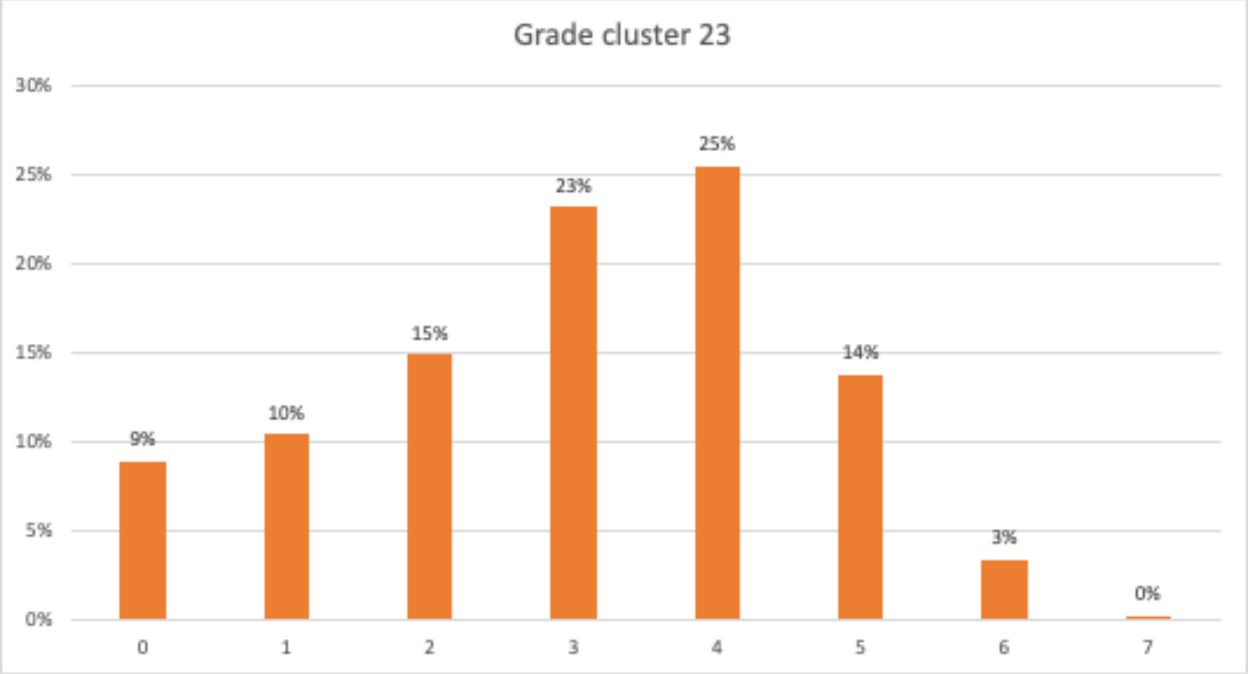


Figure 3

Distribution of score levels with the new scale: grade-level cluster 4–5 frequency graph

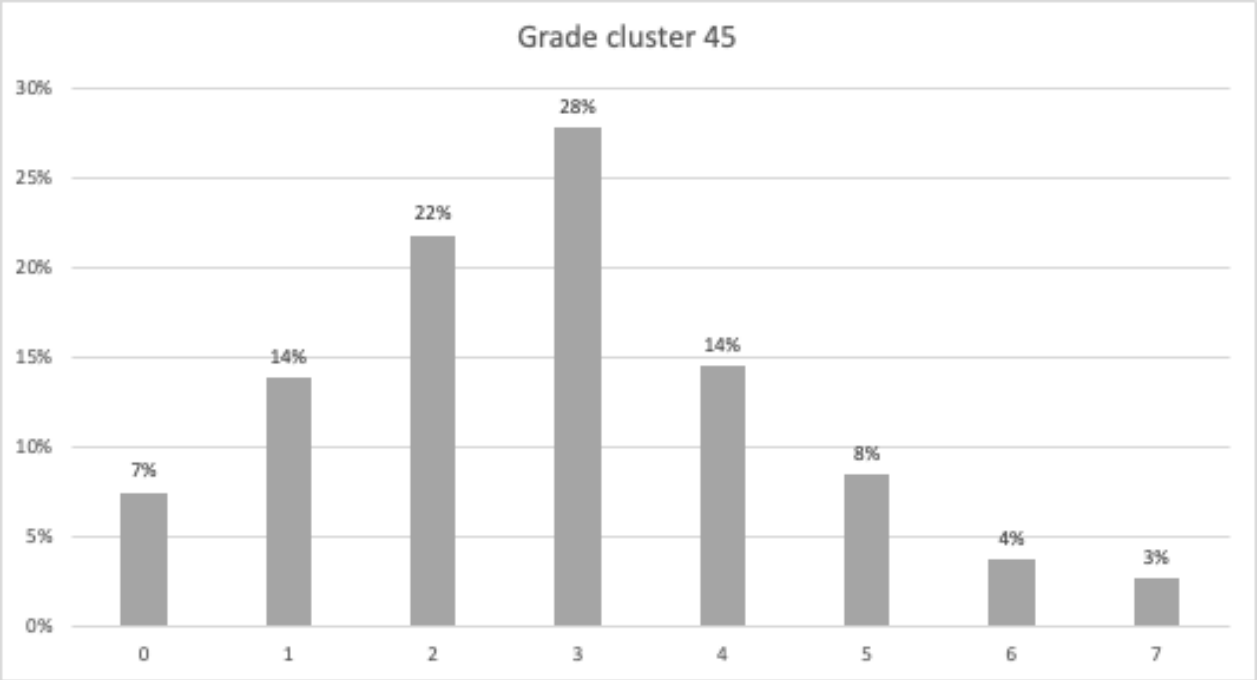


Figure 4

Distribution of score levels with the new scale: grade-level cluster 6–8 frequency graph

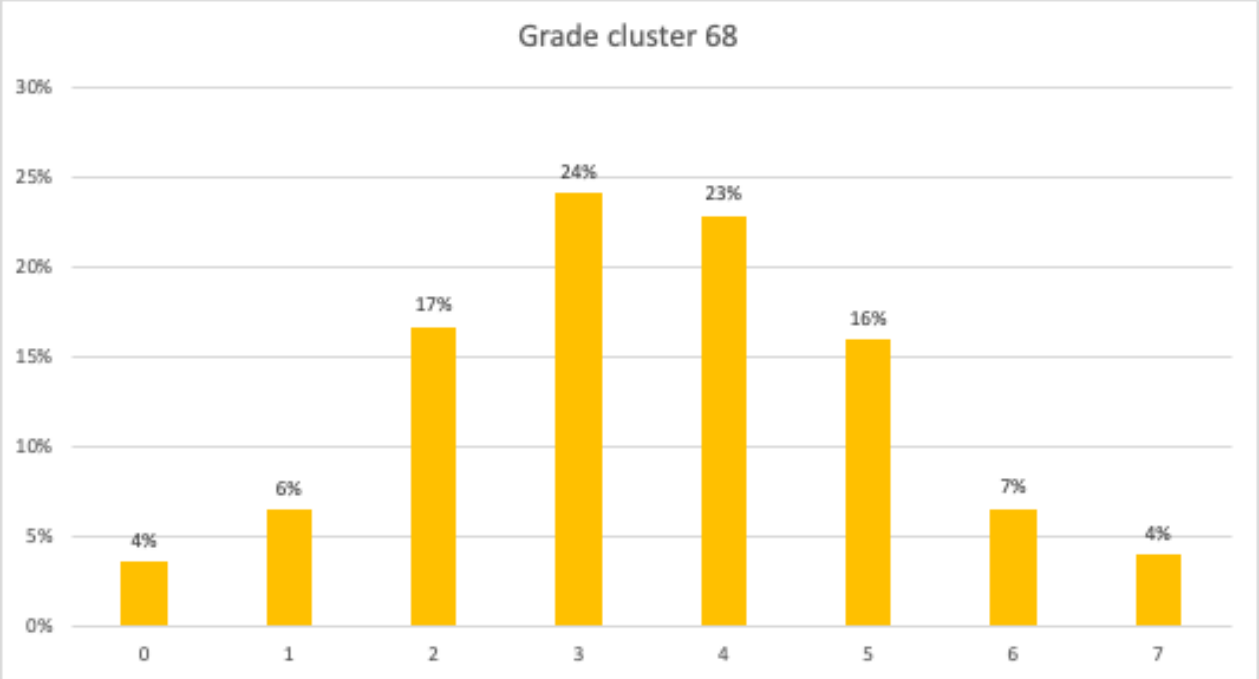


Figure 5

Distribution of score levels with the new scale: grade-level cluster 9–12 frequency graph

Figure 6

Distribution of score levels with the current scale: grade-level cluster 1 frequency graph

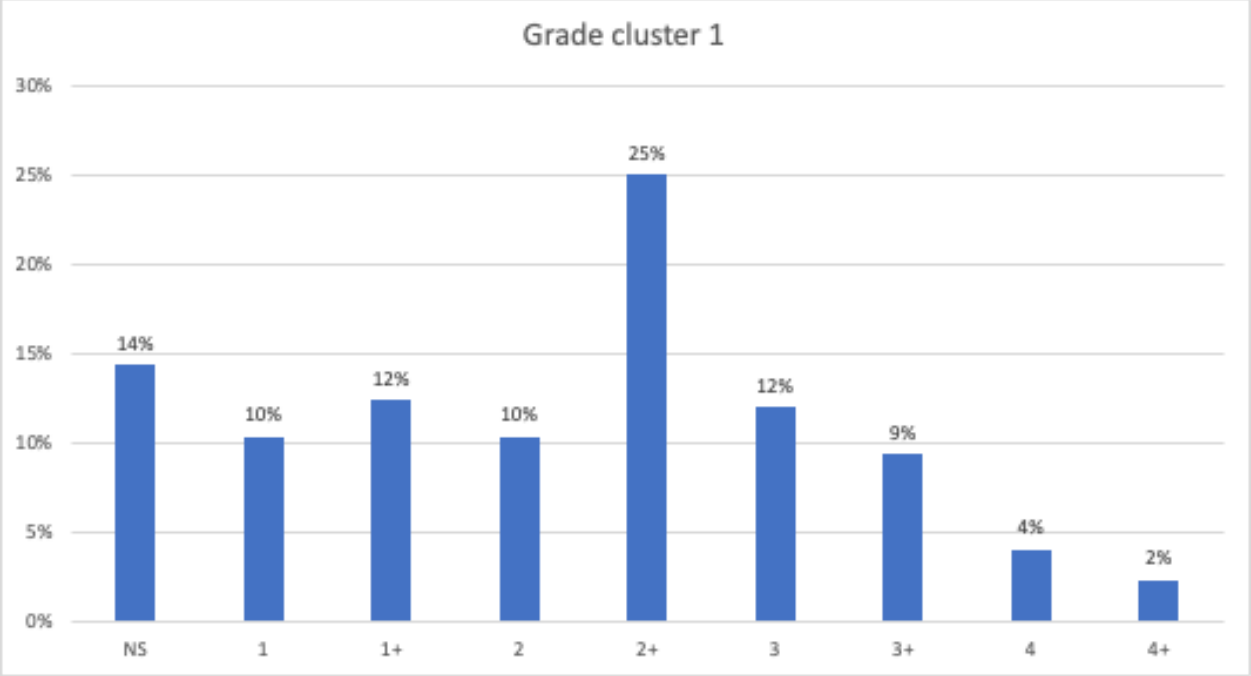


Figure 7

Distribution of score levels with the new scale: grade-level cluster 2–3 frequency graph

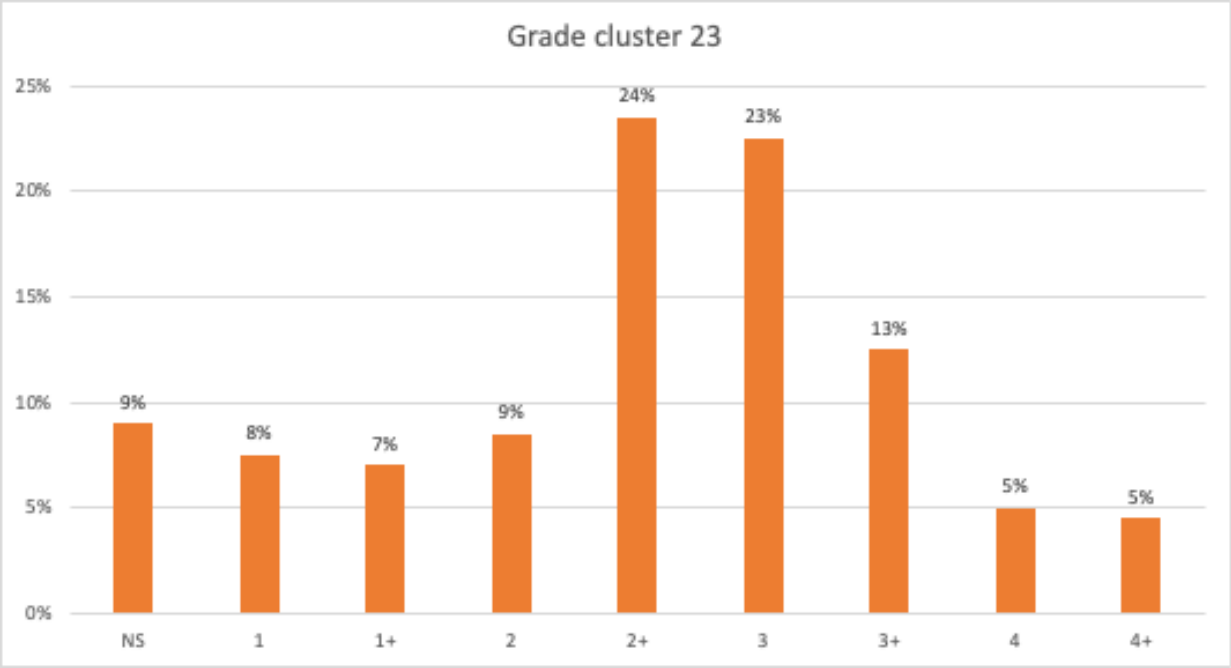


Figure 8

Distribution of score levels with the new scale: grade-level cluster 4–5 frequency graph

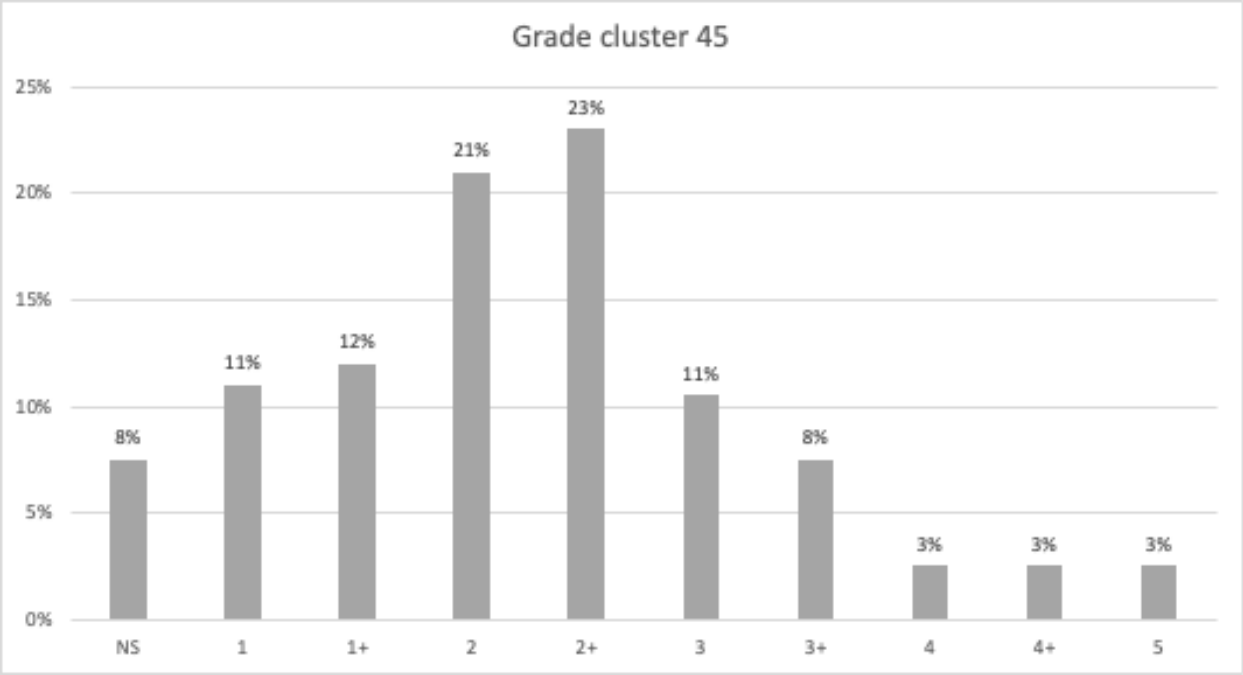


Figure 9

Distribution of score levels with the new scale: grade-level cluster 6–8 frequency graph

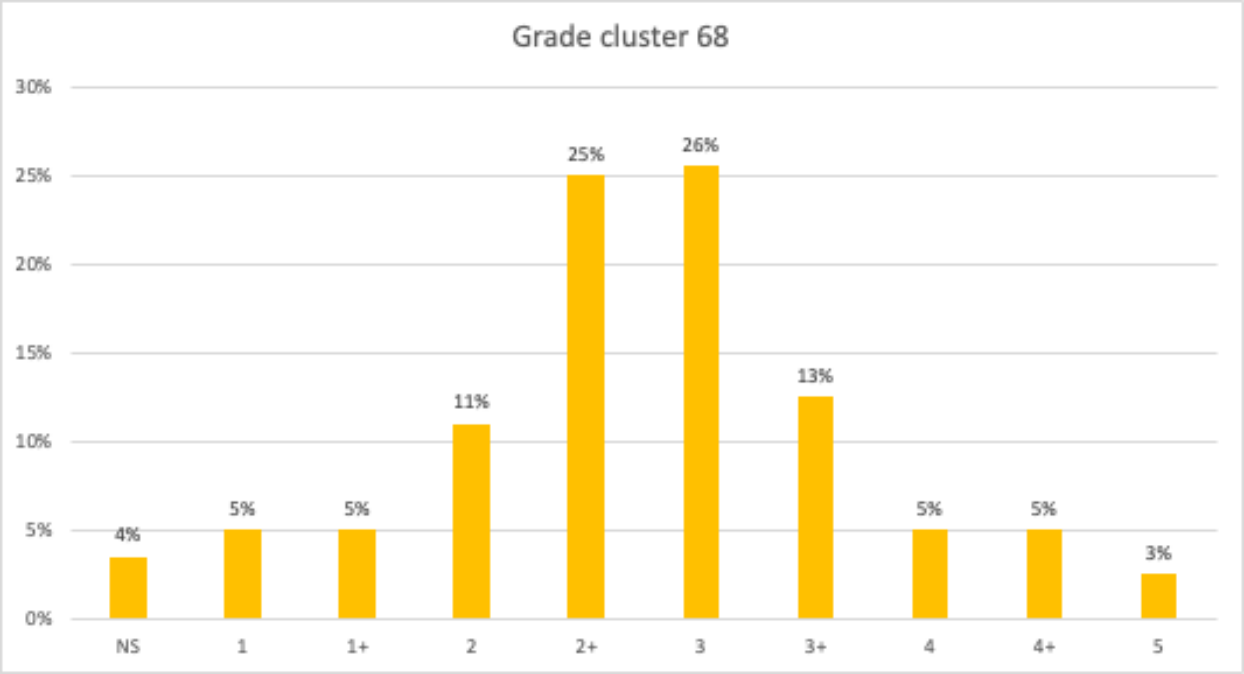
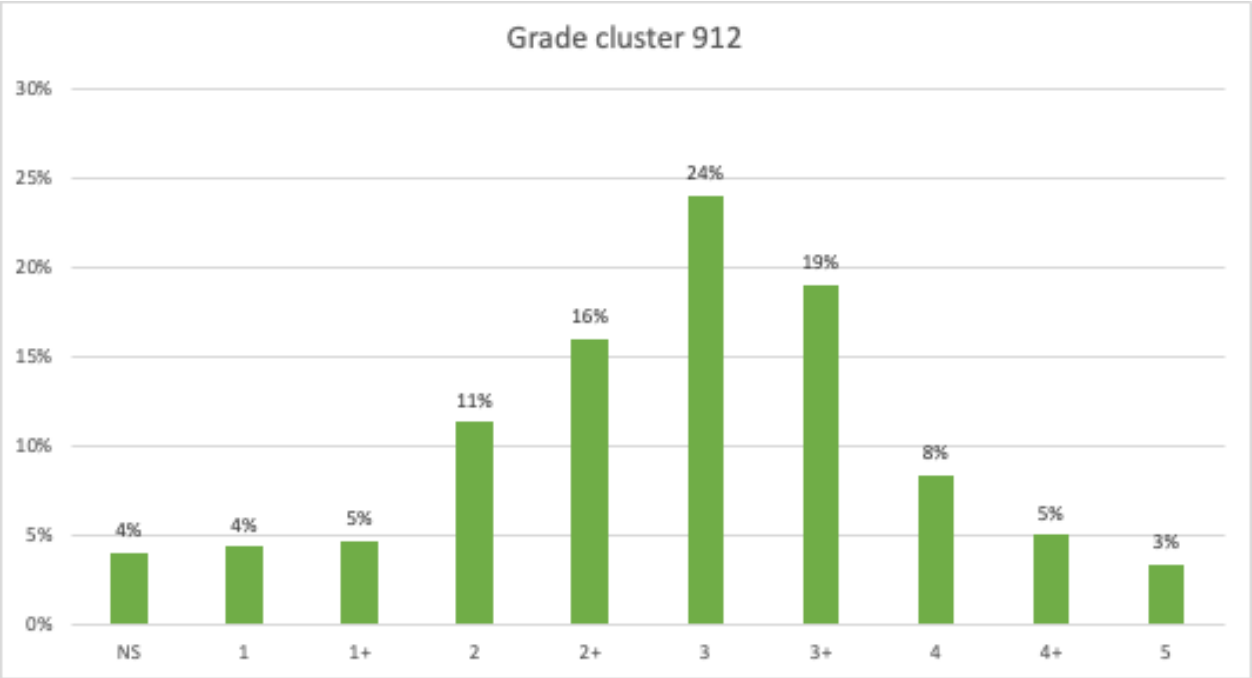


Figure 10

Distribution of score levels with the new scale: grade-level cluster 9–12 frequency graph



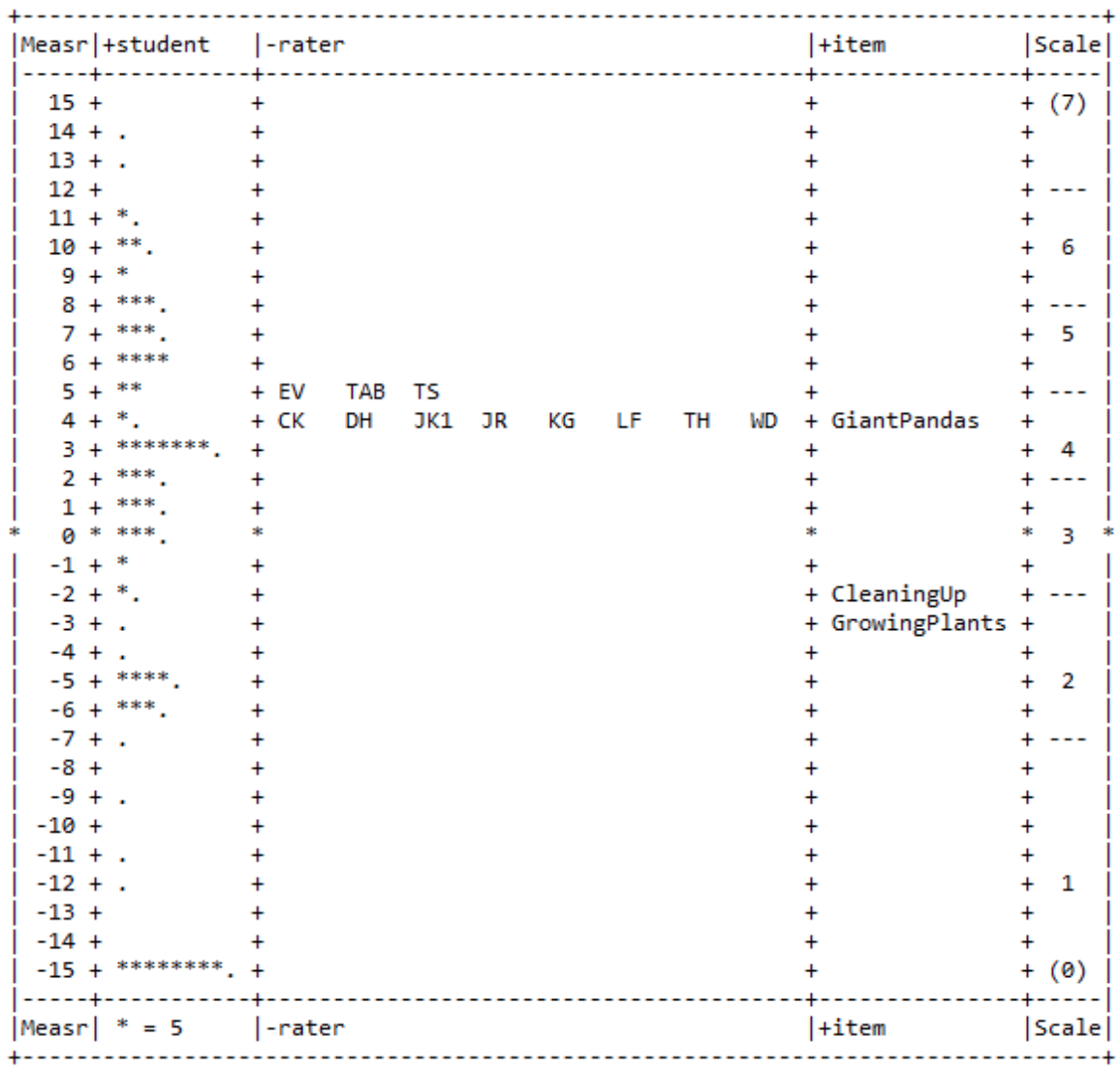
Five rating scale models (i.e., grade-level clusters 1, 23, 45, 68, and 912) with three facets were fitted. For these models, the variance explained by Rasch measures was 88.96%, 88.89%, 91.50%, 89.99%, and 90.84%, respectively. Then, five partial credit models with three facets were fitted. For these models, the variance explained by Rasch measures was 88.98%, 88.59%, 91.09%, 90.00%, and 90.87%, respectively. Having a high data variance (88% to 91%) explained by Rasch measures means that the Rasch model does a good job of accounting for the observed variability in the data using the estimated parameters.

4.2.1 Grade-level cluster 1 (RSM)

Figure 11 shows the Wright map for grade-level cluster 1. The first column (*Measr*) presents the standard value shared by all facets. The second column (*Student*) shows the distribution of all 300 students, with a higher logit value representing higher Writing domain scores. The third column (*Rater*) includes all raters who participated in the scoring and displays their rating severity. Higher values indicate increased rater severity. The fourth column (*Item*) indicates the difficulty level of each writing task, and larger values refer to more difficult tasks. The last column (*Scale*) shows the probabilistic model estimates of the scores, with each horizontal line being the Rasch-Andrich threshold (i.e., the logit interval a student falls when assigned a particular score).

Figure 11

Wright map for grade-level cluster 1 RSM



For this grade-level cluster, student ability measures ranged from –14.96 to 14.28 logits. The student separation ratio was 5.83 and the strata index was 8.10 (with a reliability of .97). This means students' writing ability can be separated into about eight statistically significant levels. A chi-square test also indicates that there were significant differences in students' writing ability (*chi-square* = 17529.0, *df* = 299, $p < .001$).

In terms of rater performance, severity measures ranged from 4.00 to 5.04 logits. The rater separation ratio was 1.74 and the strata index was 2.65 (with a reliability of .75). This suggests that there were about three distinct groups of raters with different degrees of severity. The fixed chi-square value was 43.9 (*df* = 10, $p < .001$), indicating significant differences in rating behaviors. However, these findings should be interpreted with caution. Due to the nature of the dataset (i.e., a large number of ratings provided by each rater), strata indices might be inflated where the standard deviation of the severity measures was much larger than the standard error of each rater's severity estimate. While a strata index of 2.65 seems large, raters were not greatly different in terms of severity as suggested in the Wright map and severity

measures (1-logit difference). As for inter-rater reliability, the exact agreement for this group of raters was 70.9%. The mean point biserial correlation was .73 (*SD* = 0.01; ranging from .70 to .75). Raters were mostly performing consistently as supported by infit (Mean = 0.99, *SD* = 0.27) and outfit (Mean = 0.82, *SD* = 0.22) mean square values. One rater had an infit mean square value larger than 1.5, suggesting a deviation of scoring pattern from what would be expected under the Rasch model.

Additionally, category statistics were reviewed to determine the function of the scale. Table 4 describes the distribution of scores for grade-level cluster 1. The table includes the frequency, the outfit mean square, and the Rasch-Andrich threshold measure of each score level. For evaluation, the following criteria were used: (1) Are there enough data in each score level to provide stable estimates? (2) Do the categories fit the model sufficiently well? (3) Do the thresholds indicate a hierarchical pattern and is the threshold distance enough to distinguish students' abilities? For an ideal model fit, the outfit mean square should be less than 2 and around 1. The threshold distance between score levels should be between 1.4 and 5 logits (Linacre, 2002). As shown in Table 4, the fit statistics were all smaller than 2 and the thresholds increased monotonically. Probability curves (see Figure 12) for the rating scale were also examined. The figure shows that each score level has an outstanding peak. This indicates clear thresholds between score levels, suggesting that the scoring rubric was able to distinguish students' abilities effectively.

Table 4

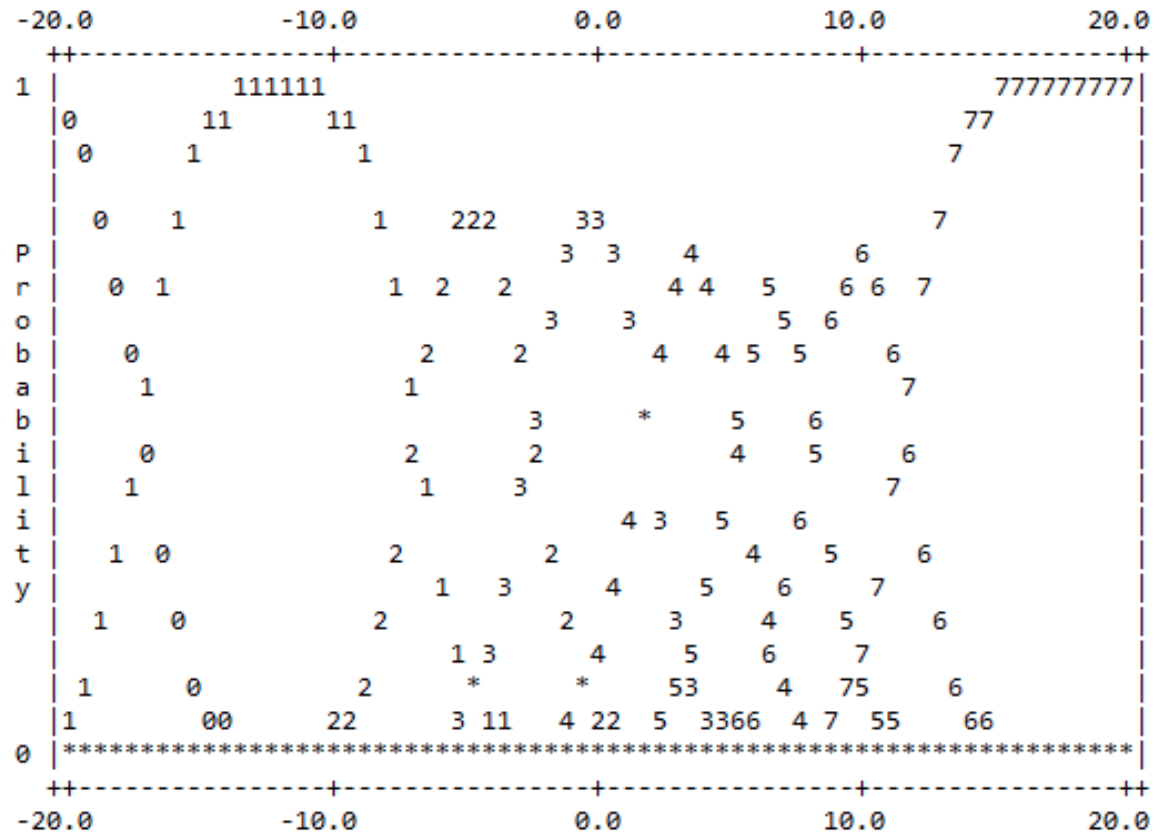
Score distribution for grade-level cluster 1

Score point	Counts (%)	Outfit mean square	Rasch-Andrich threshold measure
0	16 (1%)	0.1	N/A
1	573 (21%)	0.8	-17.30
2	586 (21%)	0.6	-6.85
3	734 (26%)	0.9	-2.47
4	545 (19%)	1.0	1.76
5	245 (9%)	1.1	5.17
6	85 (3%)	1.4	8.12
7	11 (0%)	1.7	11.57

Note: FACETS excluded extreme values during the analysis, so the counts might be slightly different from those in Appendix B. Please refer to Appendix B for the most accurate results.

Figure 12

Probability curves for the rating scale (grade-level cluster 1)

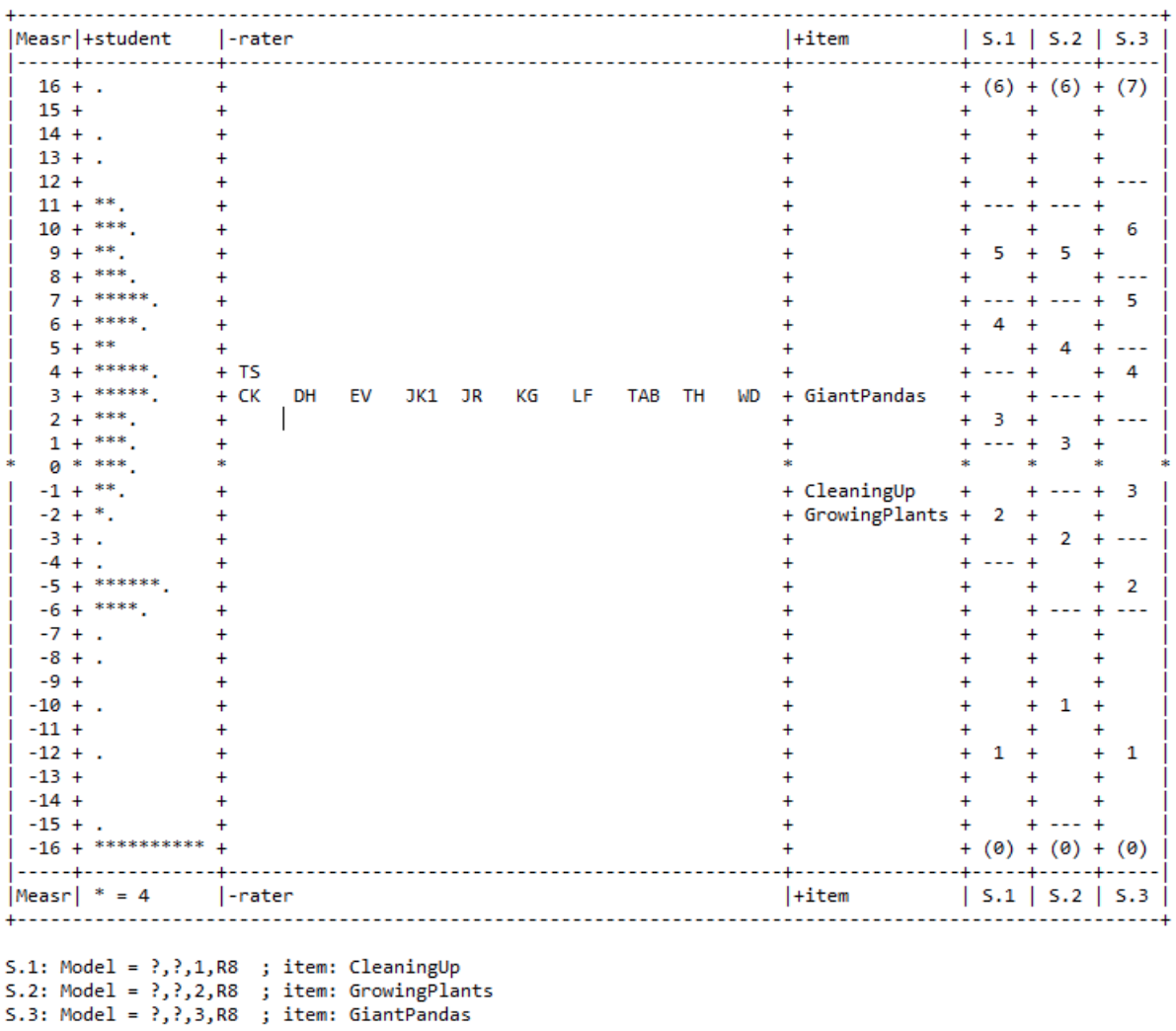


4.2.2 Grade-level cluster 1 (PCM)

Figure 13 shows the Wright map for grade-level cluster 1 with separate probabilistic model estimates of the scores for individual tasks. Unexpected responses for each task were also examined to understand the instances in which individual raters assigned misfitting ratings to misfitting student responses. These misfitting ratings might suggest difficult-to-score responses or inconsistent rating patterns. Such analysis can provide more in-depth insight into rater performance (e.g., flagging raters who assign multiple unexpected scores) and item quality (e.g., identifying writing tasks for which raters have more difficulty providing consistent ratings). For *Cleaning Up*, misfitting ratings involved 8 raters and 24 student responses; for *Growing Plants*, 9 raters and 17 student responses were involved; for *Giant Pandas*, 11 raters and 36 student responses were involved. Tables 5 through 7 highlight cases with high discrepancies between expected and observed scores. Category statistics for individual tasks are in Table 8, and Figures 14 through 16 show their corresponding probability curves. The probability curves for *Giant Pandas* were perhaps the most unclear with unevenly spaced hills and slightly obscure peaks for some score levels.

Figure 13

Wright map for grade-level cluster 1 PCM



S.1: Model = ?,?,1,R8 ; item: CleaningUp
S.2: Model = ?,?,2,R8 ; item: GrowingPlants
S.3: Model = ?,?,3,R8 ; item: GiantPandas

Table 5

Unexpected responses for grade-level cluster 1, Cleaning Up

Student – Rater	Expected rating	Observed rating
25 – DH	1.4	5
41 – DH	4.4	1
4 – LF	1.7	4
76 – JR	2.6	5

Table 6

Unexpected responses for grade-level cluster 1, Growing Plants

Student – Rater	Expected rating	Observed rating
186 – KG	1.6	6
143 – JK1	0.4	3
200 – KG	1.5	4

Table 7

Unexpected responses for grade-level cluster 1, Giant Pandas

Student – Rater	Expected rating	Observed rating
207 – CK	4.1	7
212 – LF	4.0	2
227 – LF	5.2	3

Table 8

Score distribution for individual writing tasks of grade-level cluster 1

Score level	Writing task	Counts (%)	Outfit mean square	Rasch-Andrich threshold measure
0	1	6 (1%)	0.4	N/A
0	2	9 (1%)	0.0	N/A
0	3	1 (0%)	0.7	N/A
1	1	224 (25%)	0.9	-18.81
1	2	279 (32%)	0.6	-14.95
1	3	70 (7%)	0.7	-17.80
2	1	234 (26%)	0.3	-4.42
2	2	243 (28%)	0.5	-5.50
2	3	109 (11%)	1.0	-6.06
3	1	226 (25%)	0.8	0.64
3	2	196 (23%)	0.9	-0.67
3	3	312 (30%)	1.1	-3.24
4	1	149 (16%)	1.0	4.35
4	2	111 (13%)	1.0	3.36
4	3	285 (28%)	1.0	1.69
5	1	62 (7%)	1.7	6.91
5	2	23 (3%)	0.8	7.19
5	3	160 (16%)	0.8	5.55
6	1	9 (1%)	1.0	11.32
6	2	1 (0%)	9.9	10.57
6	3	75 (7%)	0.8	8.25
7	1	N/A	N/A	N/A
7	2	N/A	N/A	N/A
7	3	11 (1%)	2.0	11.62

Note: Task 1 = Cleaning Up; Task 2 = Growing Plants; Task 3 = Giant Pandas

Figure 14

Probability curves for the rating scale (grade-level cluster 1; Cleaning Up)

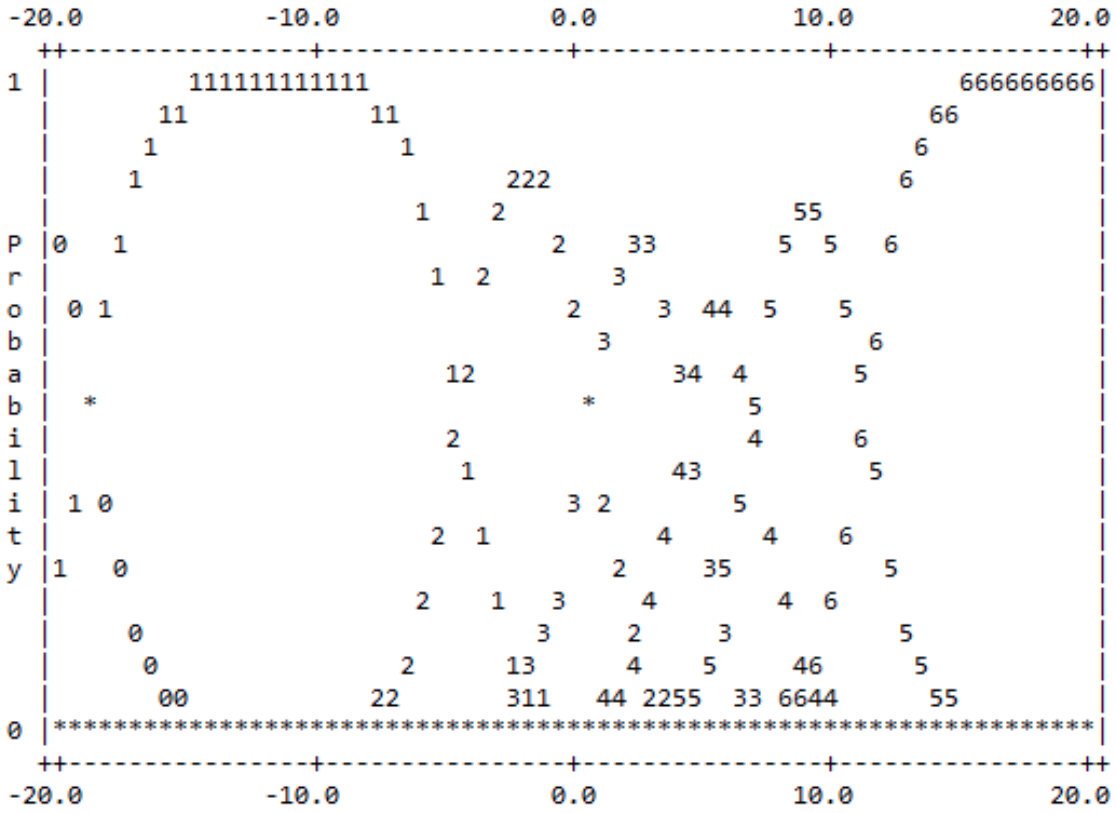


Figure 15

Probability curves for the rating scale (grade-level cluster 1; Growing Plants)

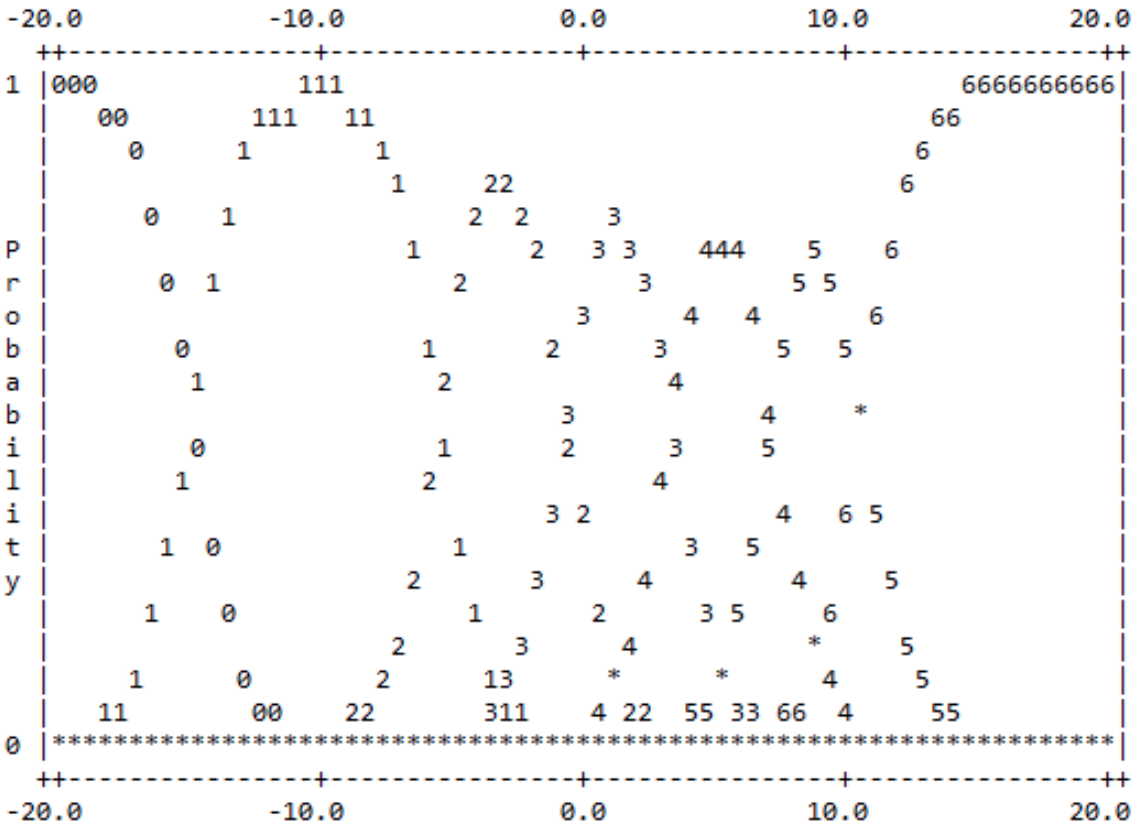
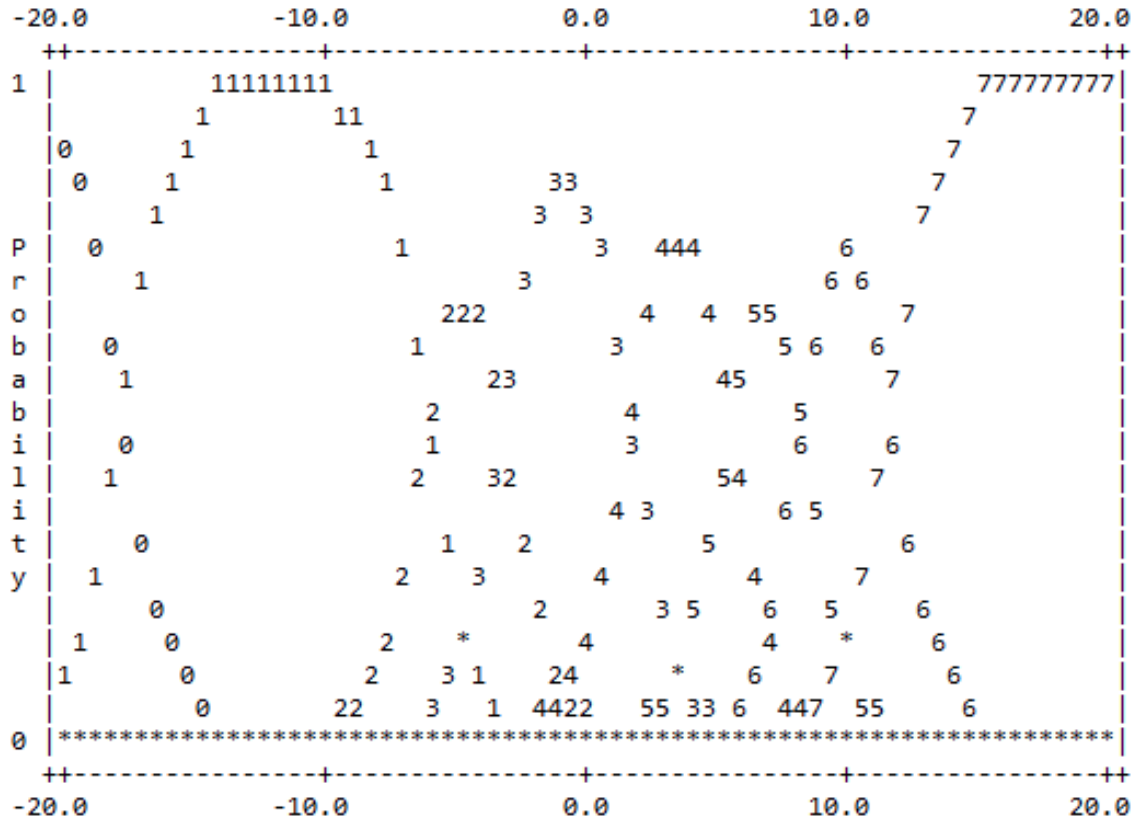


Figure 16

Probability curves for the rating scale (grade-level cluster 1; Giant Pandas)

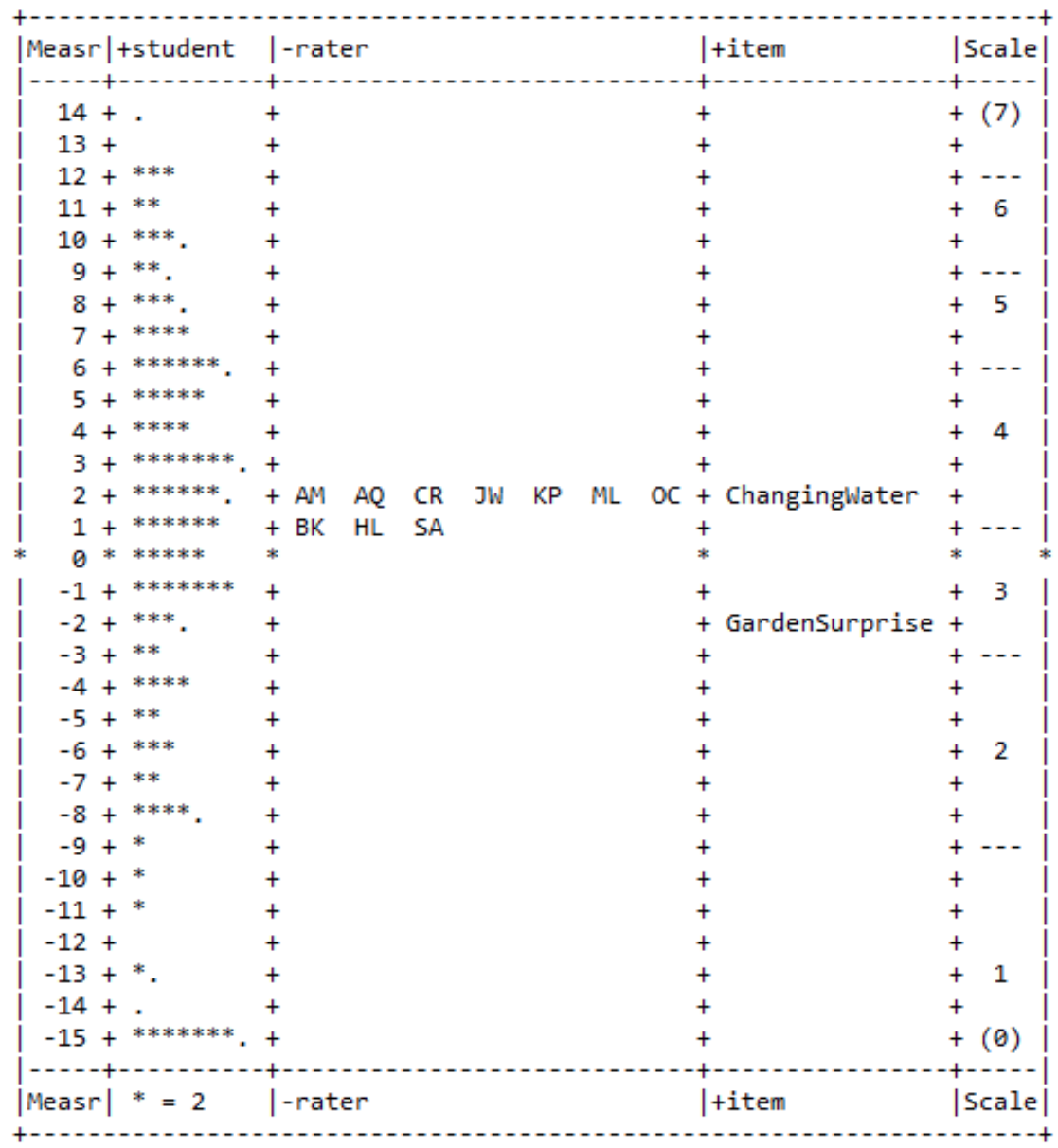


4.3.1 Grade-level cluster 2-3 (RSM)

Figure 17 shows the Wright map for grade-level cluster 2-3. Student ability measures ranged from –14.27 to 13.92 logits. The student separation ratio was 8.44 and the strata index was 11.59 (with a reliability of .99). This means students' writing ability can be separated into about 12 statistically significant levels. A chi-square test also indicates that there were significant differences in students' writing ability (*chi-square* = 14161.3, *df* = 199, $p < .001$).

Figure 17

Wright map for grade-level cluster 2-3 RSM



In terms of rater performance, severity measures ranged from 1.25 to 2.24 logits. The rater separation ratio was 1.65 and the strata index was 2.53 (with a reliability of .73). This suggests that there were about three distinct groups of raters with different degrees of severity. The fixed chi-square value was 37.4 (*df* = 9, $p < .001$), indicating significant differences in rating behaviors. As for inter-rater reliability, the exact agreement for this group of raters was 63.7%. The mean point biserial correlation was .70 (*SD* = 0.01; ranging from .68 to .72). Raters were mostly performing consistently as supported by infit (Mean = 0.98, *SD* = 0.22) and outfit (Mean = 0.93, *SD* = 0.23) mean square values. All raters had desirable infit and outfit mean square values.

Table 9 shows the distribution of scores for grade-level cluster 2-3. According to the table, the fit statistics were all close to 1.0 and the thresholds increased monotonically. Probability curves for the rating scale (see Figure 18) reveal that most score levels have a clear peak.

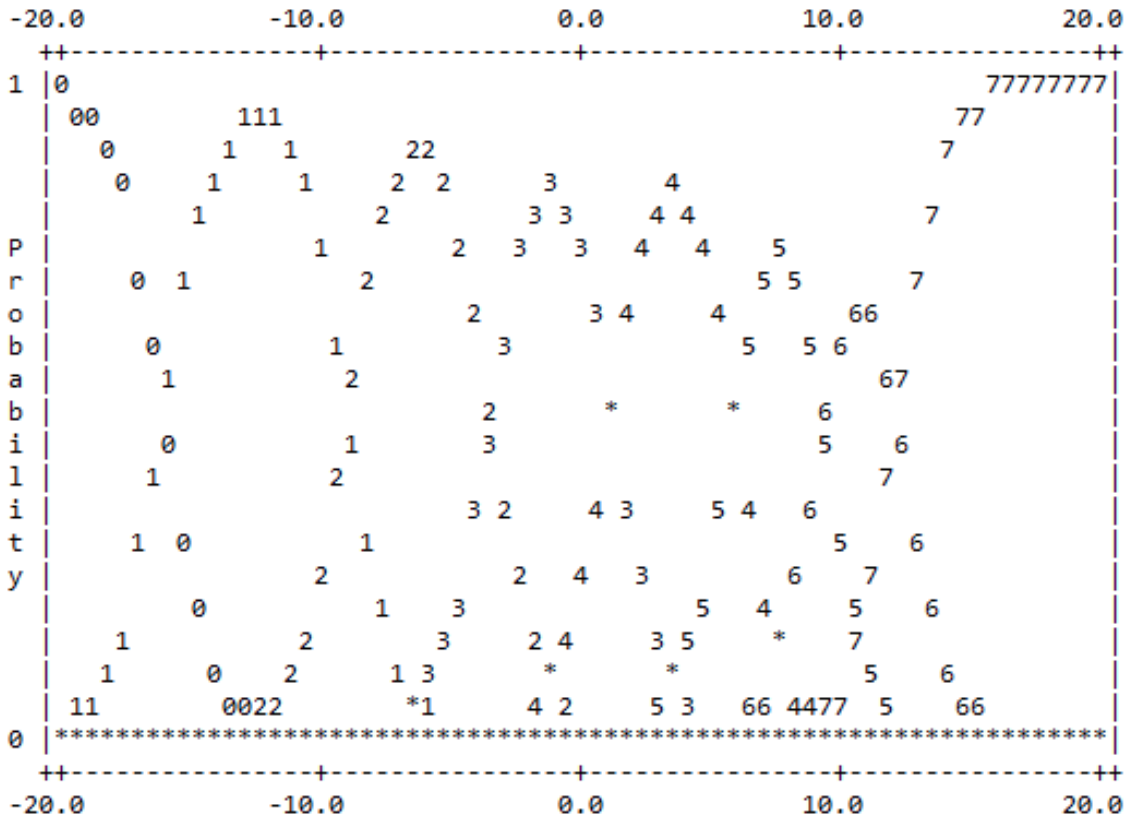
Table 9

Score distribution for grade-level cluster 2-3

Score level	Counts (%)	Outfit mean square	Rasch-Andrich threshold measure
0	22 (1%)	0.6	N/A
1	199 (11%)	0.9	-16.09
2	286 (16%)	0.9	-9.01
3	445 (25%)	1.0	-3.44
4	488 (28%)	1.0	1.16
5	264 (15%)	0.9	5.91
6	64 (4%)	1.2	9.35
7	4 (0%)	0.9	12.11

Figure 18

Probability curves for the rating scale (grade-level cluster 2-3)



4.3.2 Grade-level cluster 2-3 (PCM)

Figure 19 shows the Wright map for grade-level cluster 2-3 with separate probabilistic model estimates of the scores for individual tasks. Unexpected responses for each task were also examined to understand the instances in which individual raters assigned misfitting ratings to misfitting student responses. For *Garden Surprise*, misfitting ratings involved 10 raters and 37 student responses; for *Changing Water*, 10 raters and 33 student responses were involved. Table 10 highlights cases with high discrepancies between expected and observed scores. Category statistics for individual tasks are in Table 11; Figures 20 and 21 show their corresponding probability curves. Both tasks had relatively evenly spaced hills.

Figure 19

Wright map for grade-level cluster 2-3 PCM

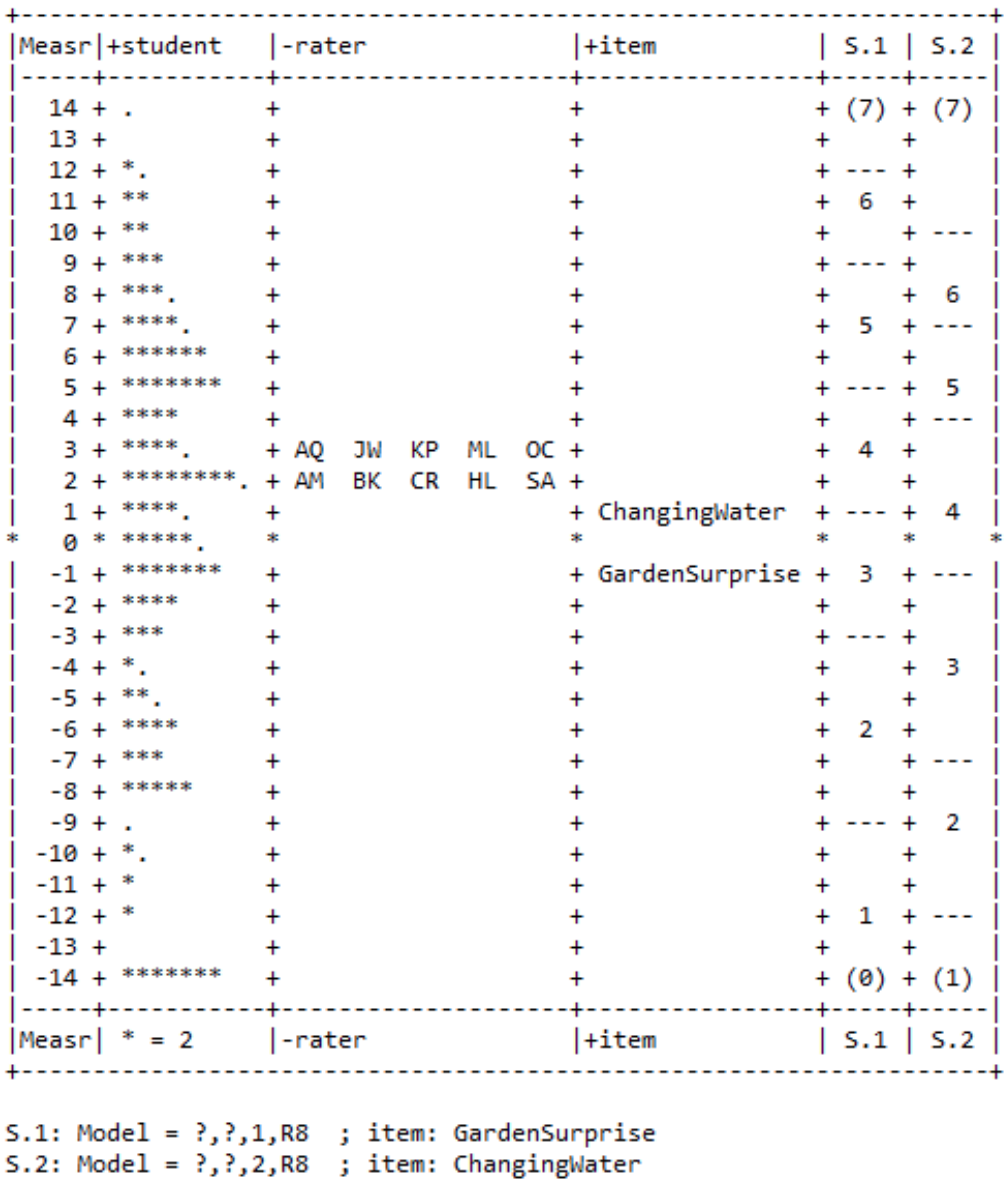


Table 10

Unexpected responses for grade-level cluster 2-3, Changing Water

Student – Rater	Expected rating	Observed rating
112 – AQ	4.2	6
188 – AM	4.7	3

Table 11

Score distribution for individual writing tasks of grade-level cluster 2-3

Score level	Writing task	Counts (%)	Outfit mean square	Rasch-Andrich threshold measure
0	1	22 (2%)	0.7	N/A
0	2	N/A	N/A	N/A
1	1	159 (18%)	0.9	-15.63
1	2	23 (3%)	0.8	N/A
2	1	175 (20%)	0.9	-8.87
2	2	111 (13%)	0.9	-12.17
3	1	212 (24%)	0.9	-3.19
3	2	233 (27%)	1.0	-6.64
4	1	188 (21%)	0.9	1.00
4	2	300 (35%)	1.0	-1.39
5	1	118 (13%)	0.9	5.22
5	2	146 (17%)	0.9	3.81
6	1	15 (2%)	1.1	9.57
6	2	49 (6%)	1.3	6.73
7	1	1 (0%)	1.0	11.90
7	2	3 (0%)	0.9	9.65

Note: Task 1 = Garden Surprise; Task 2 = Changing Water

Figure 20

Probability curves for the rating scale (grade-level cluster 2-3; Garden Surprise)

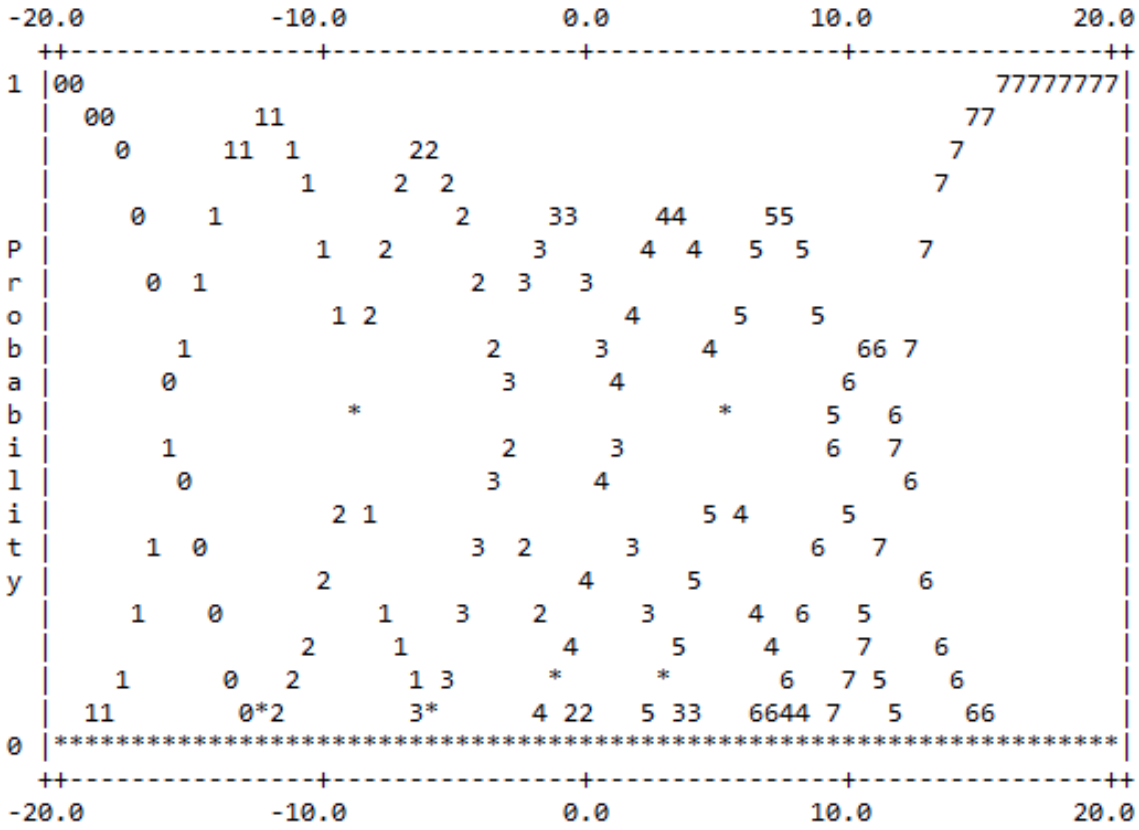
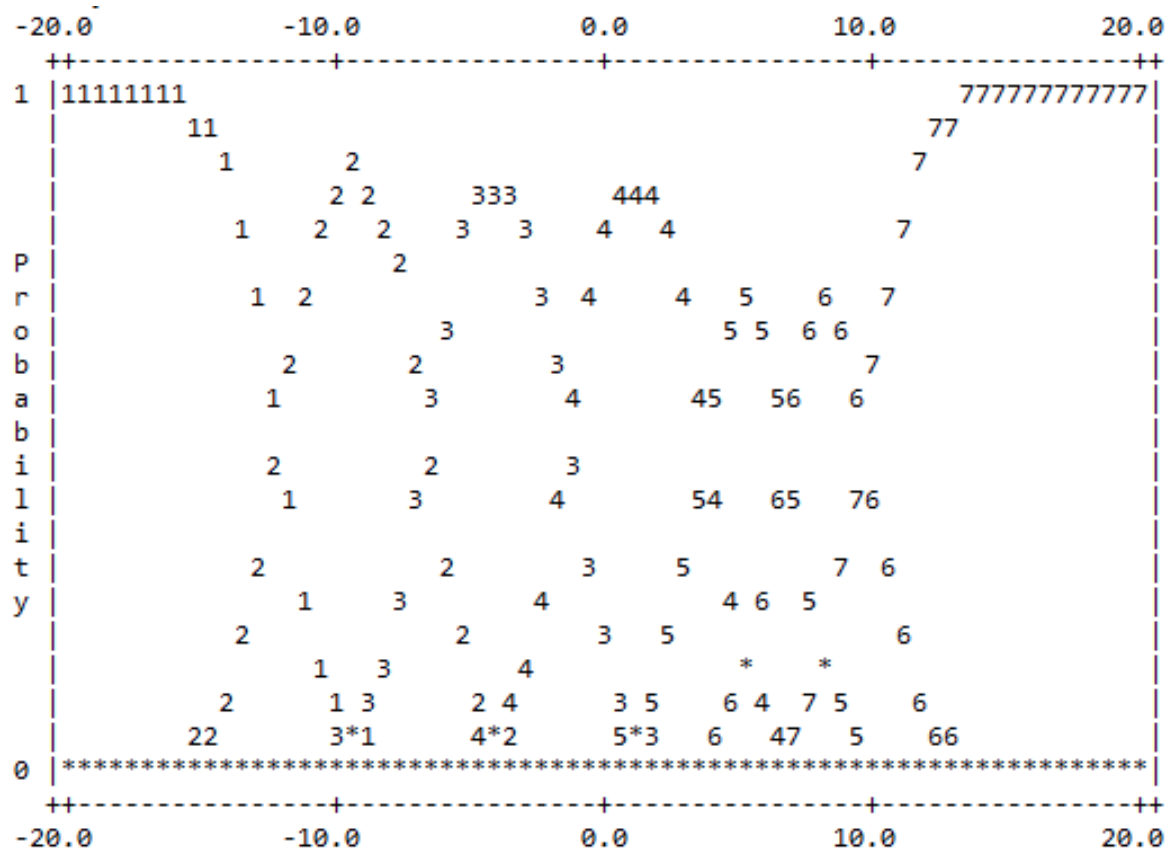


Figure 21

Probability curves for the rating scale (grade-level cluster 2-3; Changing Water)

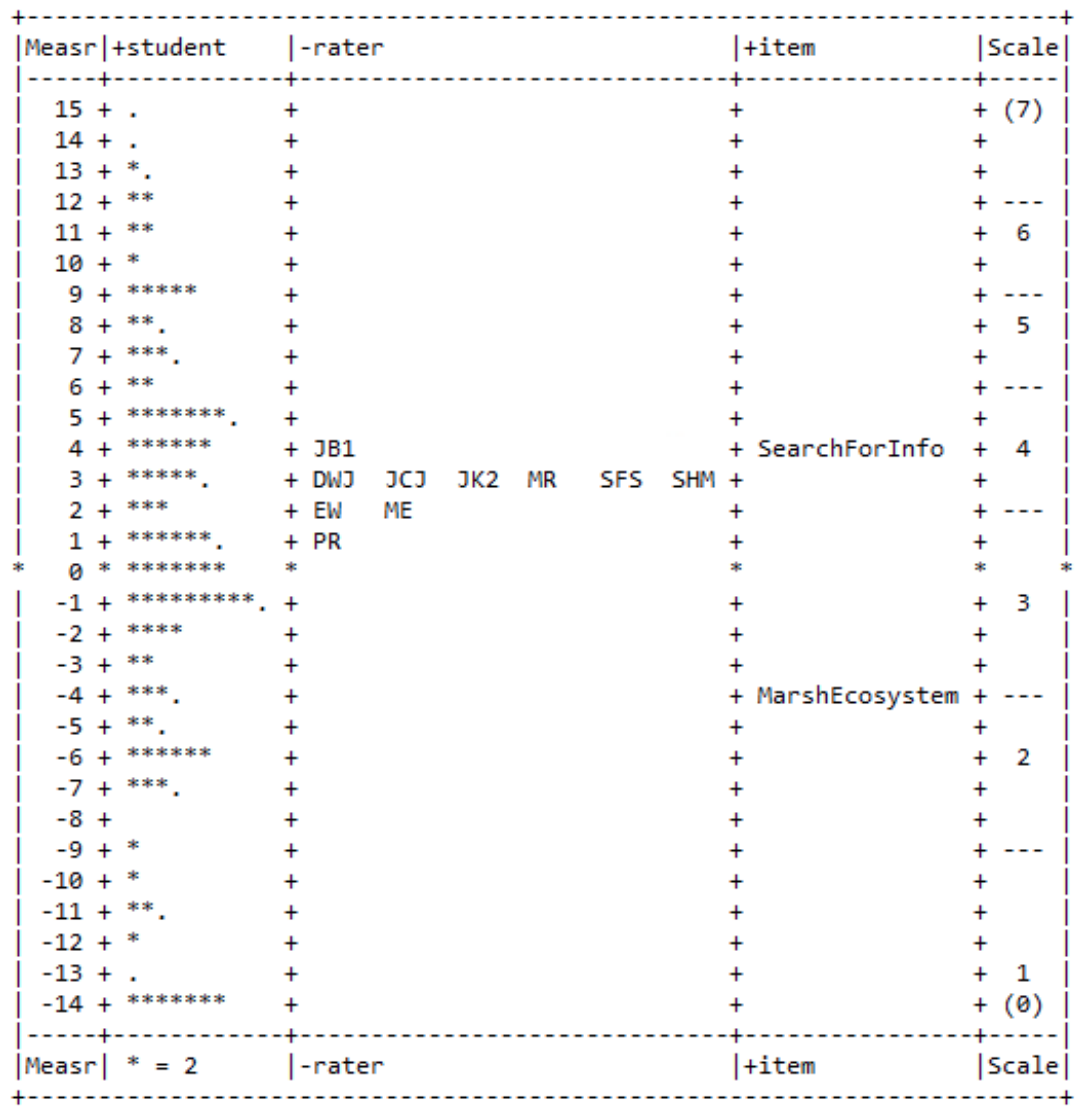


4.4.1 Grade-level cluster 4-5 (RSM)

Figure 22 shows the Wright map for grade-level cluster 4-5. Student ability measures ranged from –13.94 to 14.20 logits. The student separation ratio was 7.73 and the strata index was 10.64 (with a reliability of .98). Students' writing ability can be separated into about 11 statistically significant levels. A chi-square test also indicates that there were significant differences in students' writing ability (*chi-square* = 14072.0, *df* = 199, $p < .001$).

Figure 22

Wright map for grade-level cluster 4-5 RSM



In terms of rater performance, severity measures ranged from 0.99 to 3.54 logits. The rater separation ratio was 3.48 and the strata index was 4.97 (with a reliability of .92). This suggests that there were about five distinct groups of raters with different degrees of severity. The fixed chi-square value was 153.0 (*df* = 9, $p < .001$), indicating raters were behaving significantly differently. It should be noted that the range of rater severity levels for this grade-level cluster was the largest of all, suggesting the need for more training to ensure similar severity across raters. As for inter-rater reliability, the exact agreement for this group of raters was 64.6%. The mean point biserial correlation was .74 (*SD* = 0.03; ranging from .68 to .78). Raters were mostly performing consistently as supported by infit (Mean = 0.97, *SD* = 0.30) and outfit (Mean = 0.89, *SD* = 0.33) mean square values. One rater had large infit (1.74) and outfit (1.68) mean square values.

Table 12 shows the distribution of scores for grade-level cluster 4-5. The fit statistics were mostly around 1.0 and the thresholds increased monotonically. Probability curves (see Figure 23) for the rating scale reveal that most score levels have a clear peak.

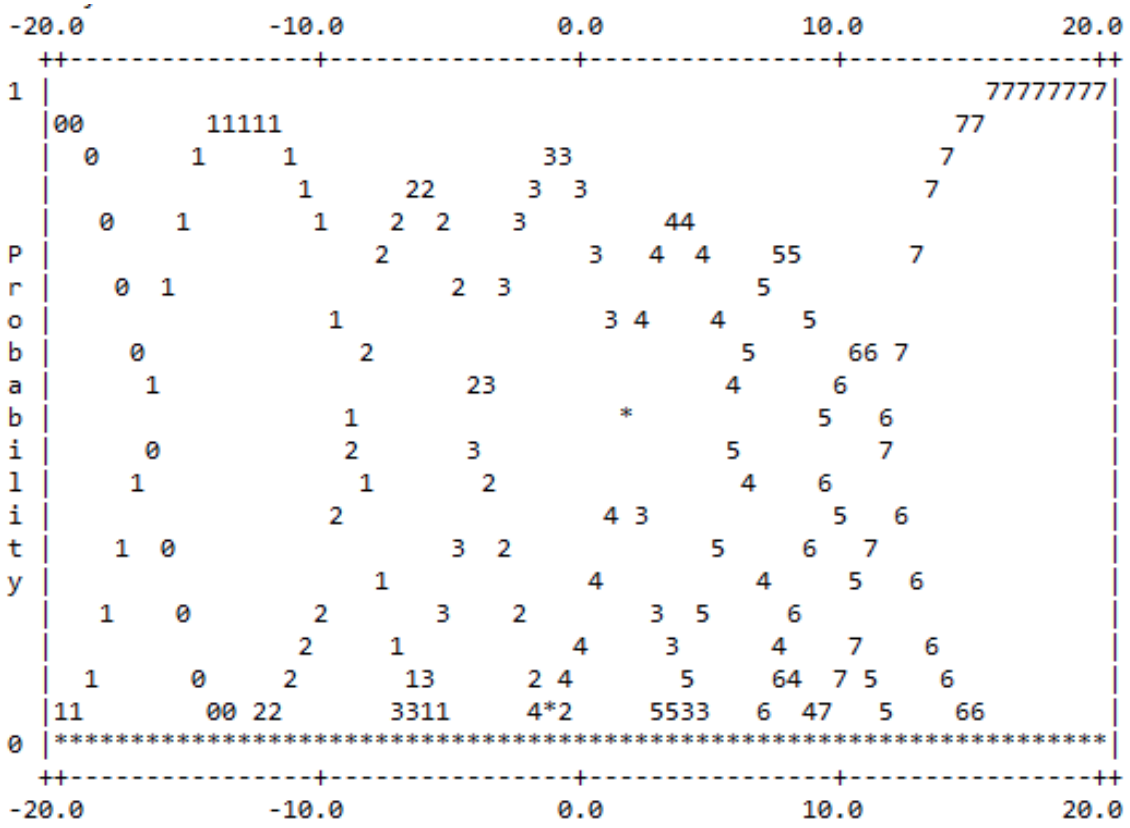
Table 12

Score distribution for grade-level cluster 4-5

Score level	Counts (%)	Outfit mean square	Rasch-Andrich threshold measure
0	52 (3%)	0.3	N/A
1	263 (15%)	1.0	-16.70
2	414 (23%)	0.7	-8.72
3	527 (29%)	0.9	-3.84
4	275 (15%)	0.9	1.73
5	160 (9%)	1.5	6.05
6	70 (4%)	0.7	9.63
7	41 (2%)	1.2	11.86

Figure 23

Probability curves for the rating scale (grade-level cluster 4-5)



4.4.2 Grade-level cluster 4-5 (PCM)

Figure 24 shows the Wright map for grade-level cluster 4-5 with separate probabilistic model estimates of the scores for individual tasks. Unexpected responses for each task were also examined to understand the instances in which individual raters assigned misfitting ratings to misfitting student responses. For *Marsh Ecosystem*, misfitting ratings involved 9 raters and 36 student responses; for *Search for Info*, 9

raters and 26 student responses were involved. Tables 13 and 14 highlight cases with high discrepancies between expected and observed scores. Category statistics for individual tasks are in Table 15; Figures 25 and 26 show their corresponding probability curves.

Figure 24

Wright map for grade-level cluster 4-5 PCM

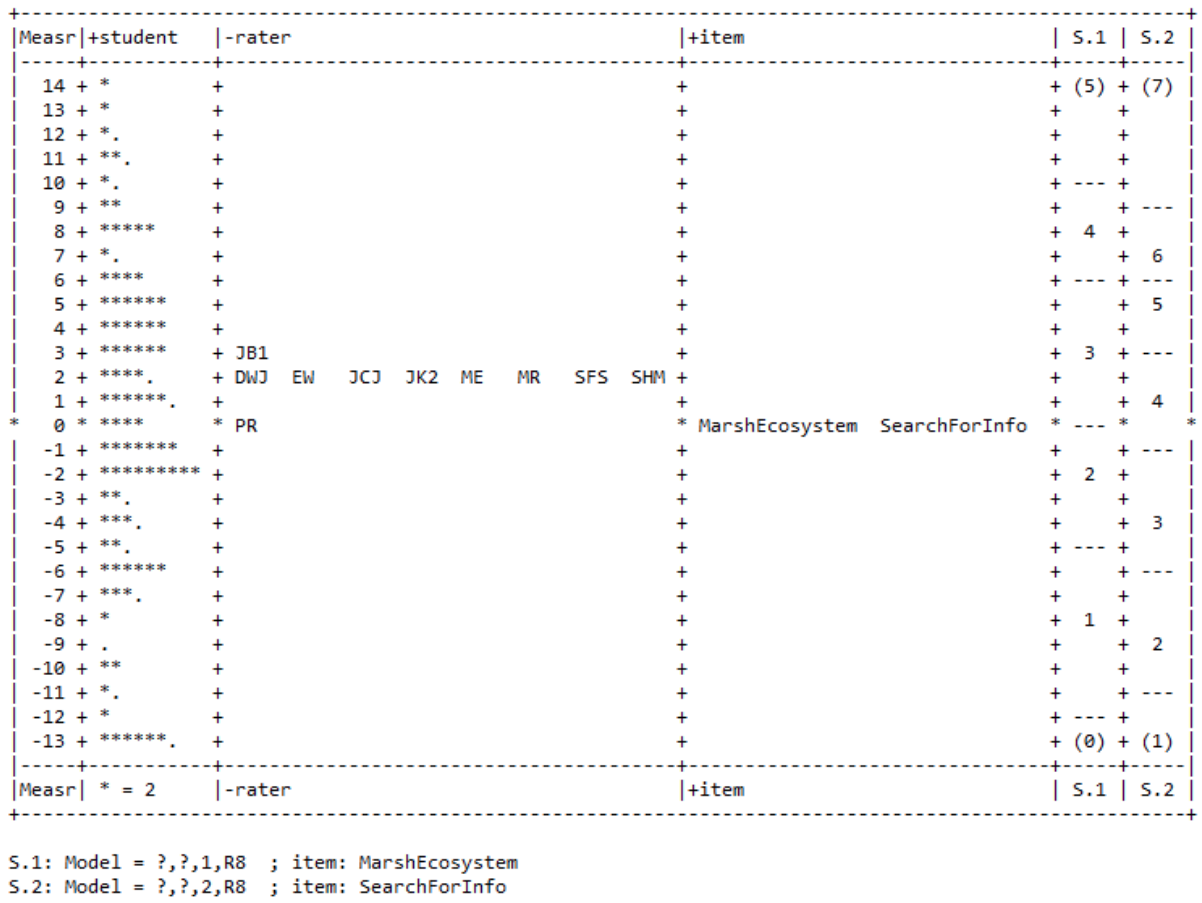


Table 13

Unexpected responses for grade-level cluster 4-5, Marsh Ecosystem

Student – Rater	Expected rating	Observed rating
65 – SHM	2.0	5
44 – ME	0.4	4

Table 14

Unexpected responses for grade-level cluster 4-5, Search for Info

Student – Rater	Expected rating	Observed rating
150 – SHM	3.0	5
168 – ME	5.1	7

Table 15

Score distribution for individual writing tasks of grade-level cluster 4-5

Score level	Writing task	Counts (%)	Outfit mean square	Rasch-Andrich threshold measure
0	1	52 (6%)	0.4	N/A
0	2	N/A	N/A	N/A
1	1	190 (21%)	1.1	-11.84
1	2	28 (3%)	1.0	N/A
2	1	286 (31%)	0.7	-4.71
2	2	128 (15%)	0.9	-10.88
3	1	264 (29%)	0.8	0.33
3	2	263 (31%)	0.8	-6.40
4	1	91 (10%)	1.4	5.86
4	2	184 (22%)	0.7	-0.89
5	1	27 (3%)	2.5	10.36
5	2	123 (15%)	1.1	3.22
6	1	N/A	N/A	N/A
6	2	70 (8%)	0.7	6.35
7	1	N/A	N/A	N/A
7	2	41 (5%)	1.2	8.60

Note: Task 1 = Marsh Ecosystem; Task 2 = Search for Info

Figure 25

Probability curves for the rating scale (grade-level cluster 4-5; Marsh Ecosystem)

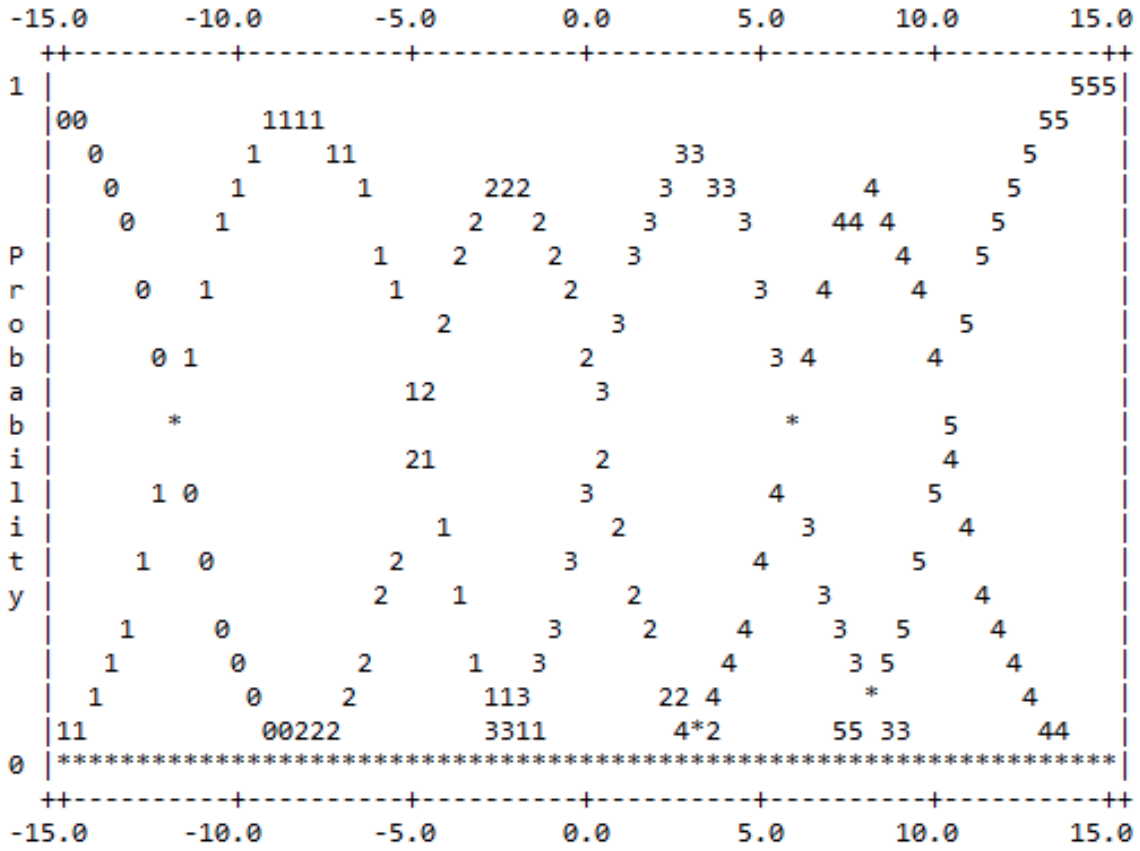
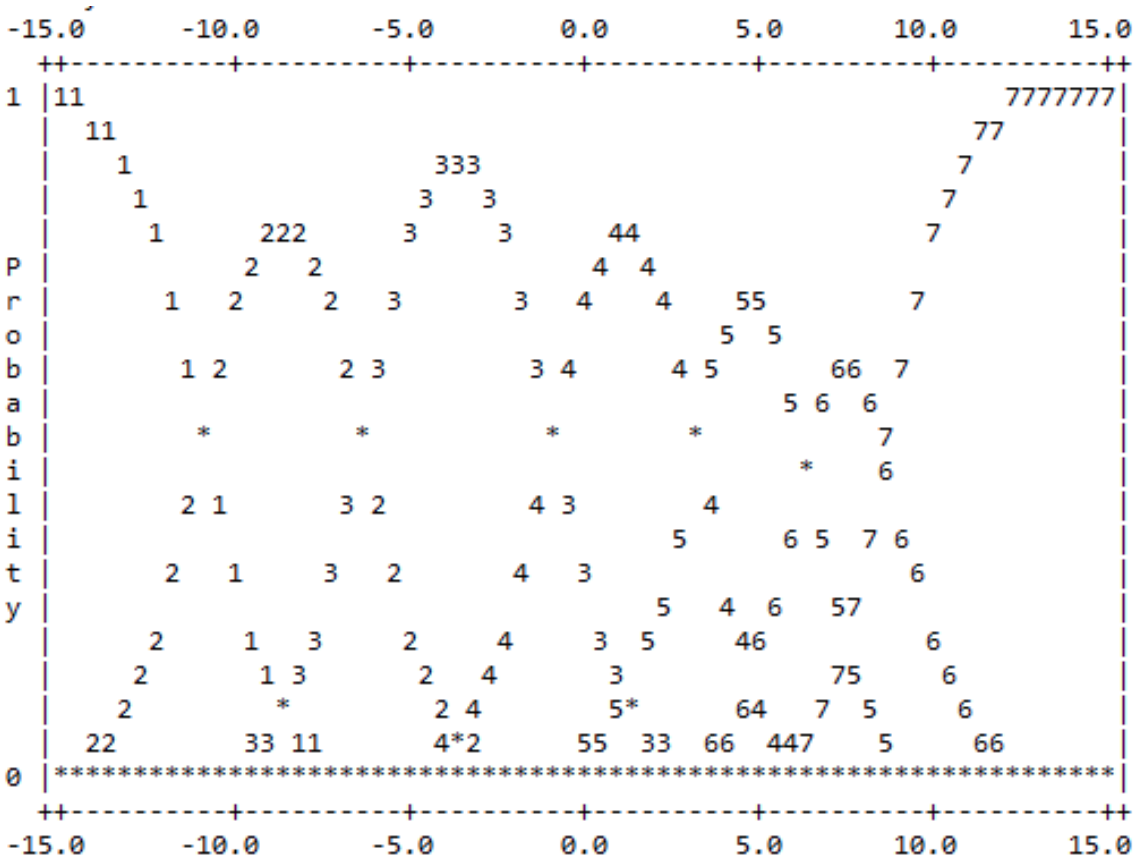


Figure 26

Probability curves for the rating scale (grade-level cluster 4-5; Search For Info)

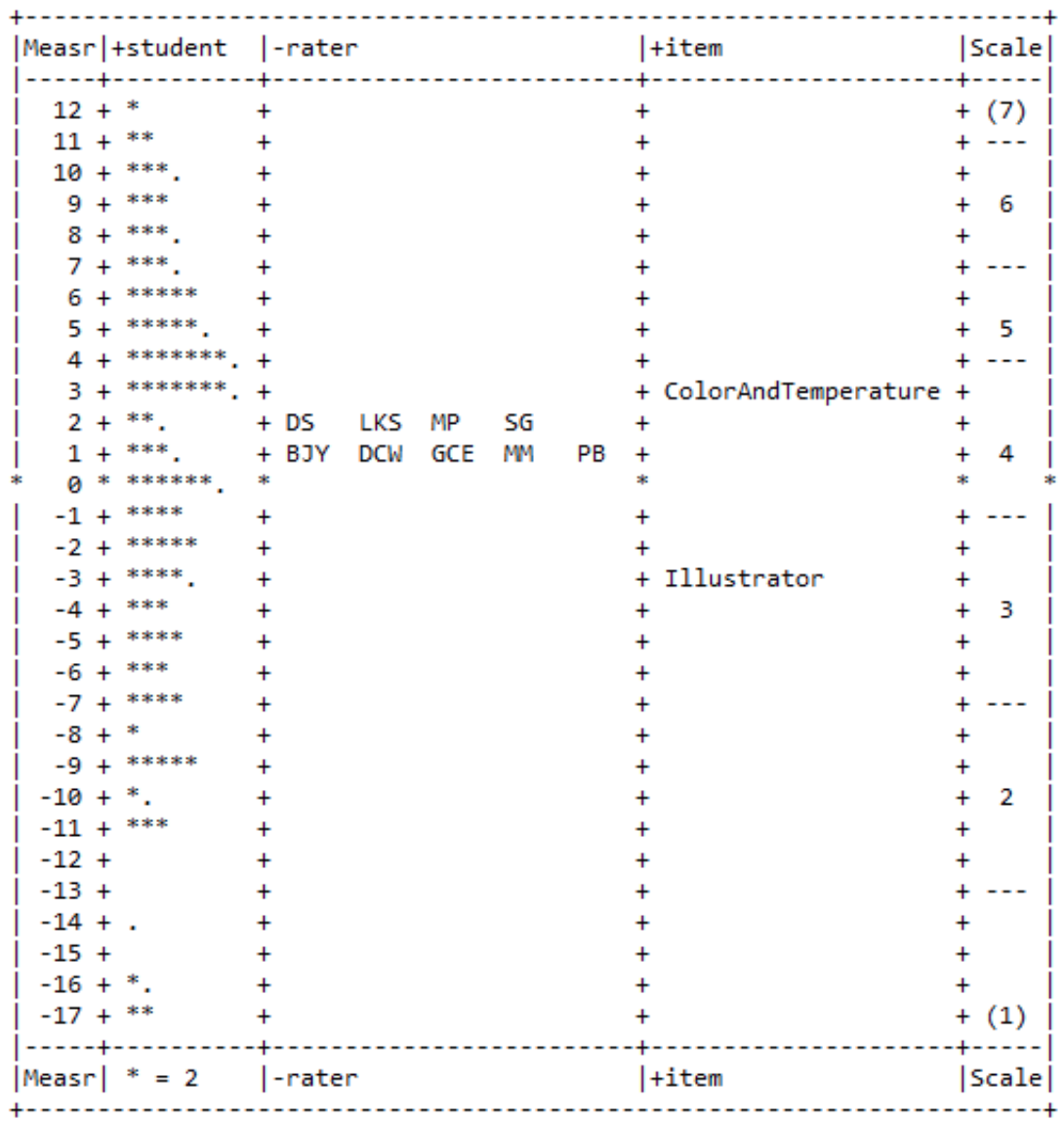


4.5.1 Grade-level cluster 6-8 (RSM)

Figure 27 shows the Wright map for grade-level cluster 6-8. Student ability measures ranged from –16.40 to 11.72 logits. The student separation ratio was 7.85 and the strata index was 10.79 (with a reliability of .98). This means students' writing ability can be separated into about 11 statistically significant levels. A chi-square test also indicates that there were significant differences in students' writing ability (*chi-square* = 12969.4, *df* = 192, $p < .001$).

Figure 27

Wright map for grade-level cluster 6-8 RSM



In terms of rater performance, severity measures ranged from 0.67 to 2.09 logits. The rater separation ratio was 2.47 and the strata index was 3.62 (with a reliability of .86). This suggests that there were about four distinct groups of raters with different degrees of severity. The fixed chi-square value was 71.4 (*df* = 8, $p < .001$), indicating significant differences in rating behaviors. As for inter-rater reliability, exact agreement for this group of raters was 61.3%. The mean point biserial correlation was .71 (*SD* = 0.02; ranging from .68 to .73). Raters were mostly performing consistently as supported by infit (Mean = 0.96, *SD* = 0.35) and outfit (Mean = 0.89, *SD* = 0.35) mean square values. Two raters had out-of-range infit and outfit mean square values, one larger than 1.5 (not fitting the model) and the other smaller than 0.5 (which indicates a level of model fit that is less than ideal).

Table 16 shows the distribution of scores for grade-level cluster 6-8. All the fit statistics were around 1.0. Rasch-Andrich threshold measures increased monotonically. Probability curves for the rating scale (see Figure 28) indicate that most score levels have an outstanding peak and an evenly spaced hill.

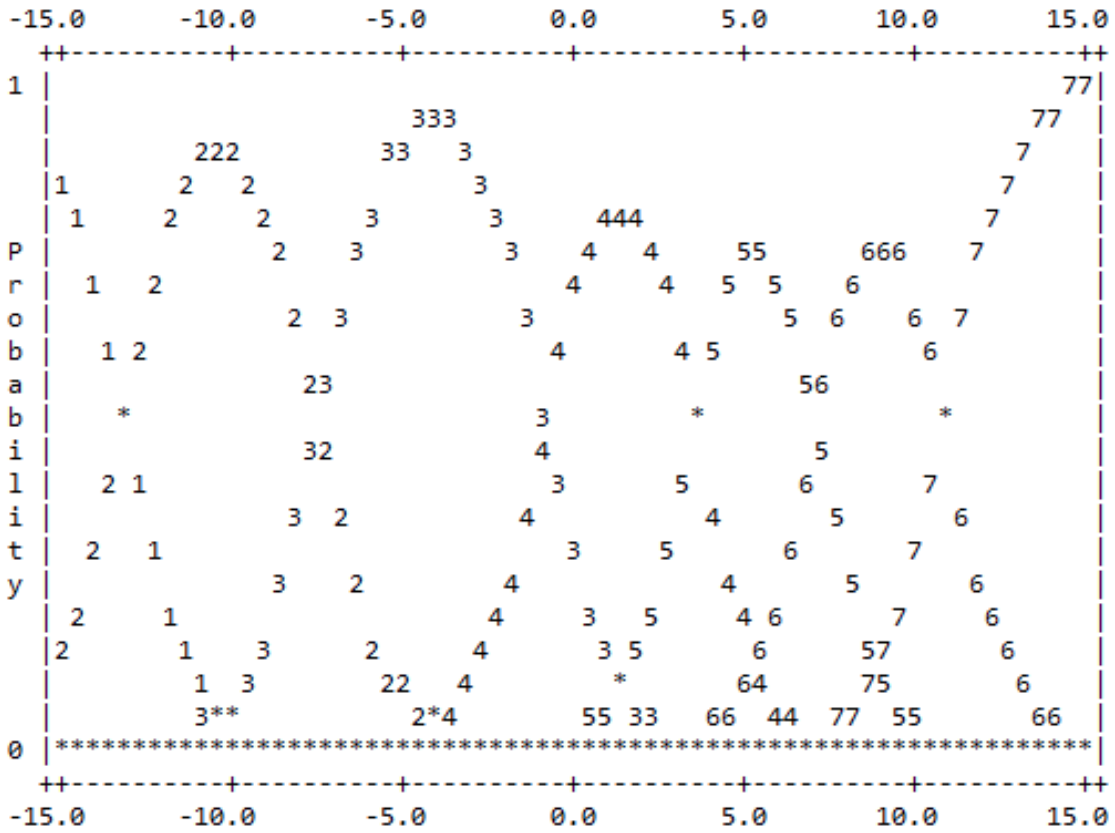
Table 16

Score distribution for grade-level cluster 6-8

Score level	Counts (%)	Outfit mean square	Rasch-Andrich threshold measure
0	N/A	N/A	N/A
1	75 (5%)	0.8	N/A
2	283 (18%)	0.8	-13.22
3	409 (26%)	0.8	-7.48
4	388 (24%)	1.0	-0.81
5	271 (17%)	1.0	3.59
6	111 (7%)	1.2	7.09
7	59 (4%)	1.0	10.83

Figure 28

Probability curves for the rating scale (grade-level cluster 6-8)



4.5.2 Grade-level cluster 6-8 (PCM)

Figure 29 shows the Wright map for grade-level cluster 6-8 with separate probabilistic model estimates of the scores for individual tasks. Unexpected responses for each task were also examined to understand the instances in which individual raters assigned misfitting ratings to misfitting student responses. For *Illustrator*, misfitting ratings involved 8 raters and 31 student responses; for *Color and Temperature*, 8 raters and 32 student responses were involved. Tables 17 and 18 highlight cases with high discrepancies

between expected and observed scores. Category statistics for individual tasks are in Table 19; Figures 30 and 31 show their corresponding probability curves.

Figure 29

Wright map for grade-level cluster 6-8 PCM

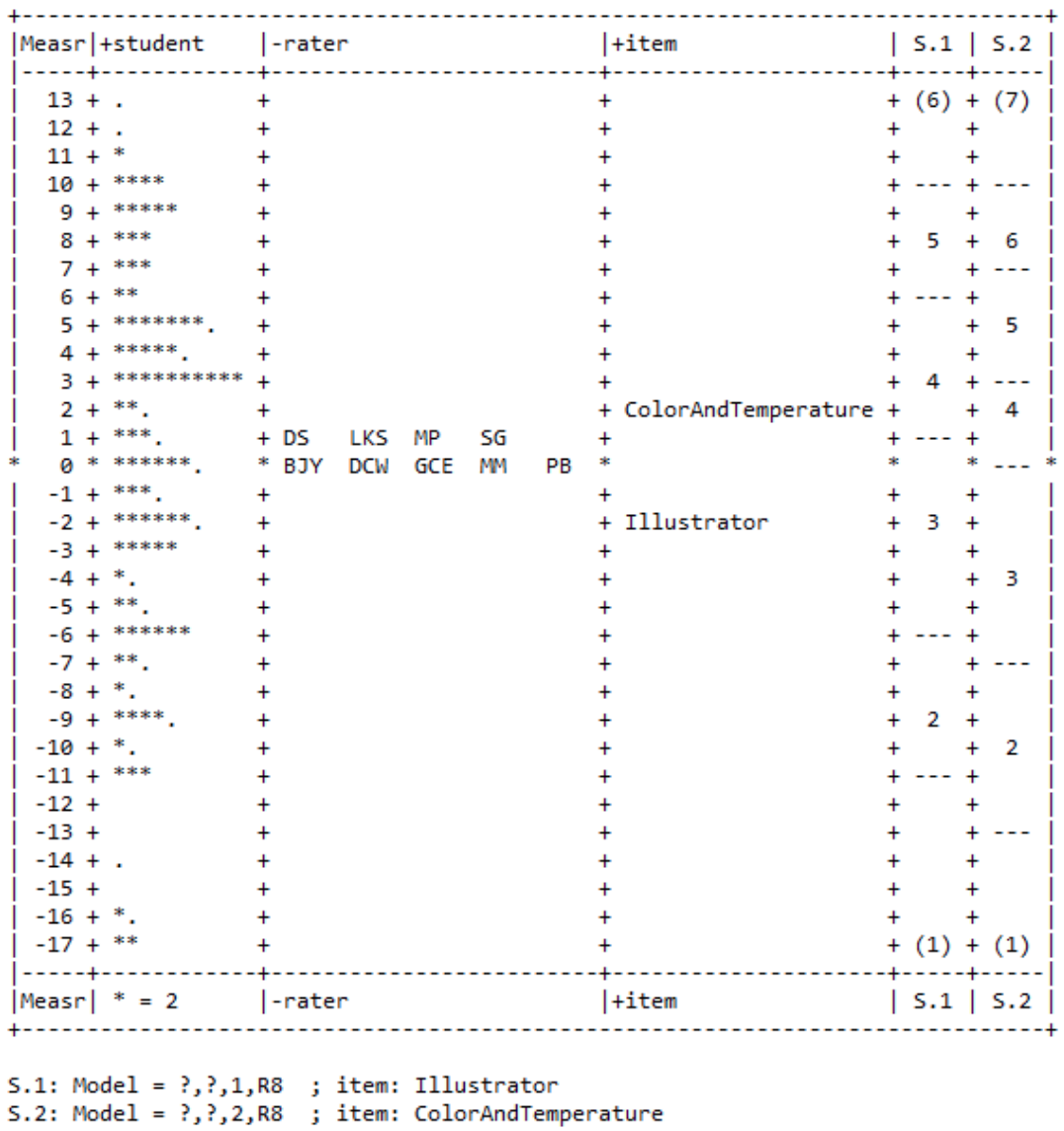


Table 17

Unexpected responses for grade-level cluster 6-8, Illustrator

Student – Rater	Expected rating	Observed rating
67 – PB	1.5	4
66 – PB	2.7	1

Table 18

Unexpected responses for grade-level cluster 6-8, Color and Temperature

Student – Rater	Expected rating	Observed rating
188 – DS	4.1	6
189 – DS	3.3	5
195 – DS	3.3	5
164 – DS	4.3	6
124 – DS	5.7	4

Table 19

Score distribution for individual writing tasks of grade-level cluster 6-8

Score level	Writing task	Counts (%)	Outfit mean square	Rasch-Andrich threshold measure
0	1	N/A	N/A	N/A
0	2	N/A	N/A	N/A
1	1	49 (6%)	0.8	N/A
1	2	26 (3%)	0.9	N/A
2	1	203 (25%)	0.8	-11.29
2	2	80 (10%)	0.8	-12.91
3	1	246 (30%)	0.8	-5.73
3	2	163 (21%)	0.9	-6.86
4	1	207 (25%)	1.0	0.82
4	2	181 (24%)	0.9	-0.38
5	1	113 (14%)	0.9	5.73
5	2	158 (21%)	1.0	3.49
6	1	10 (1%)	1.1	10.47
6	2	101 (13%)	1.2	6.46
7	1	N/A	N/A	N/A
7	2	59 (8%)	1.1	10.20

Note: Task 1 = Illustrator; Task 2 = Color And Temperature

Figure 30

Probability curves for the rating scale (grade-level cluster 6-8; Illustrator)

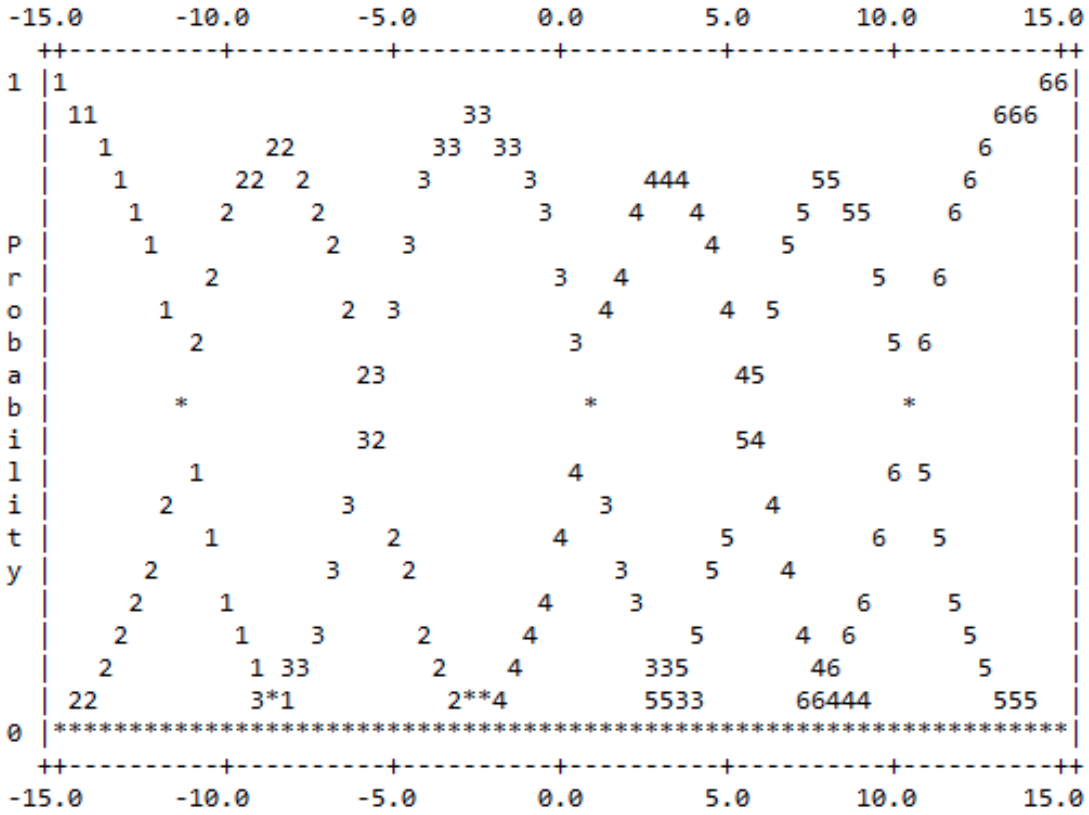
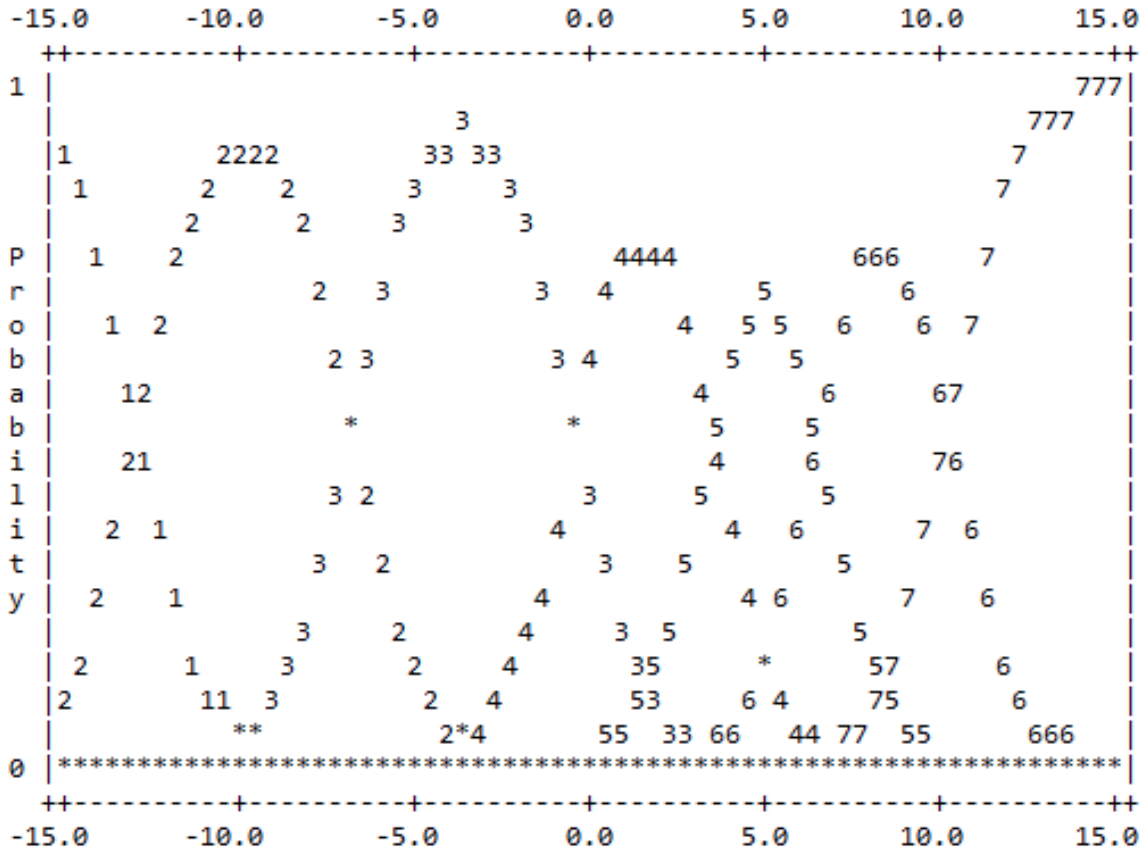


Figure 31

Probability curves for the rating scale (grade-level cluster 6-8; Color And Temperature)

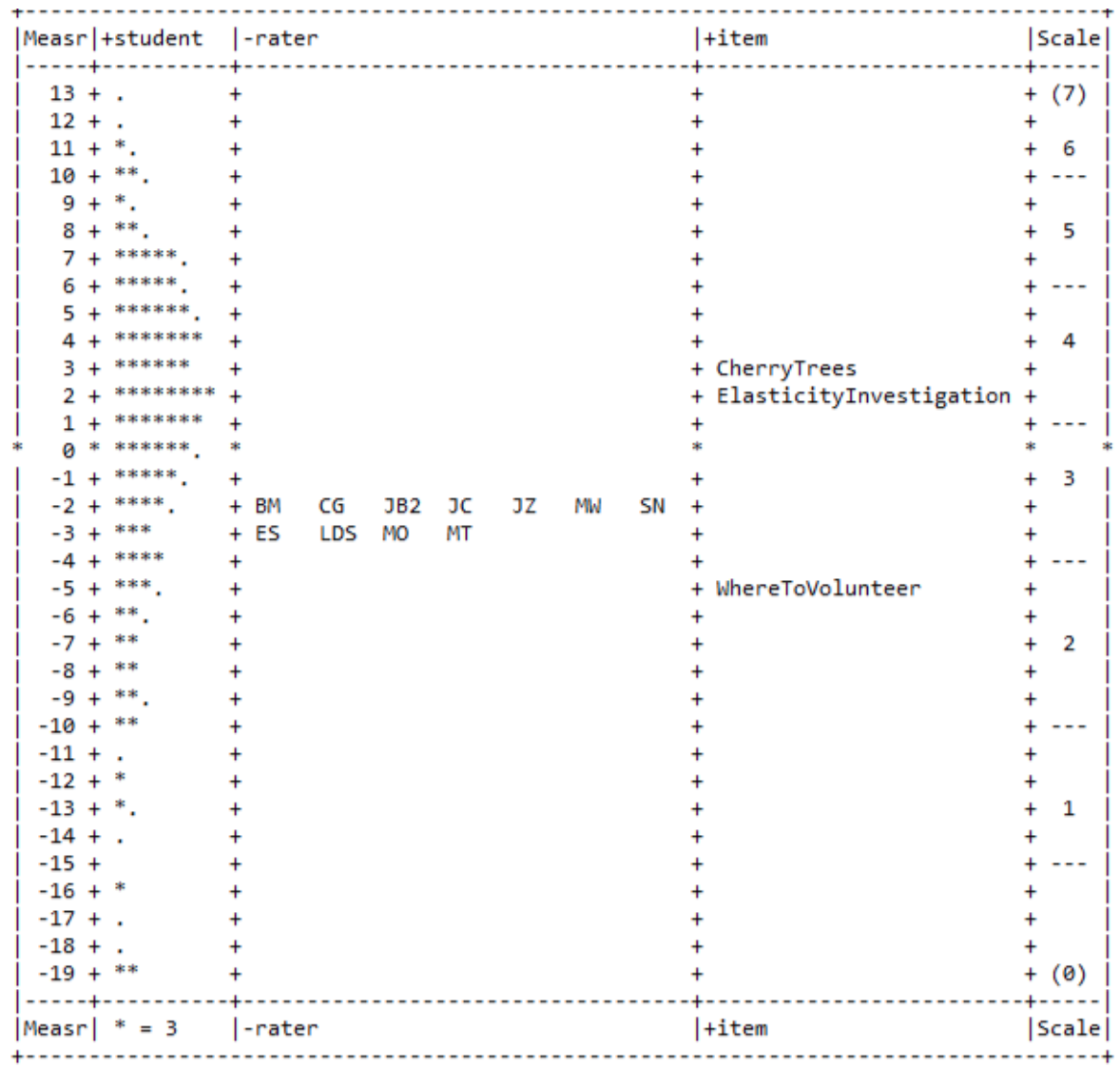


4.6.1 Grade-level cluster 9-12 (RSM)

Figure 32 shows the Wright map for grade-level cluster 9-12. Student ability measures ranged from –18.29 to 12.58 logits. The student separation ratio was 9.18 and the strata index was 12.57 (with a reliability of .99). Students' writing ability can be separated into about 13 statistically significant levels. A chi-square test also indicates that there were significant differences in students' writing ability (*chi-square* = 25203.7, *df* = 299, $p < .001$).

Figure 32

Wright map for grade-level cluster 9-12 RSM



In terms of rater performance, severity measures ranged from −3.01 to −1.80 logits. The rater separation ratio was 3.23 and the strata index was 4.64 (with a reliability of .91). This suggests that there were about five distinct groups of raters with different degrees of severity. The fixed chi-square value was 126.0 (*df* = 10, $p < .001$), indicating raters were behaving significantly differently. As for inter-rater reliability, the exact agreement for this group of raters was 61.1%. It should be noted that the strata index for this grade-level cluster was the second largest among all the grade-level clusters and that the exact agreement rate was the lowest, suggesting the need for more training to ensure similar severity and more consistent ratings across raters. The mean point biserial correlation was .75 (*SD* = 0.02; ranging from .72 to .77). Raters were mostly performing consistently as supported by infit (Mean = 0.99, *SD* = 0.23) and outfit (Mean = 0.95, *SD* = 0.24) mean square values. All raters had desirable infit and outfit mean square values.

Table 20 shows the distribution of scores for grade-level cluster 9-12. The fit statistics were all around 1.0 and the thresholds increased monotonically. Probability curves for the rating scale (see Figure 33) reveal that most curves seem to be evenly spaced hills with an outstanding peak.

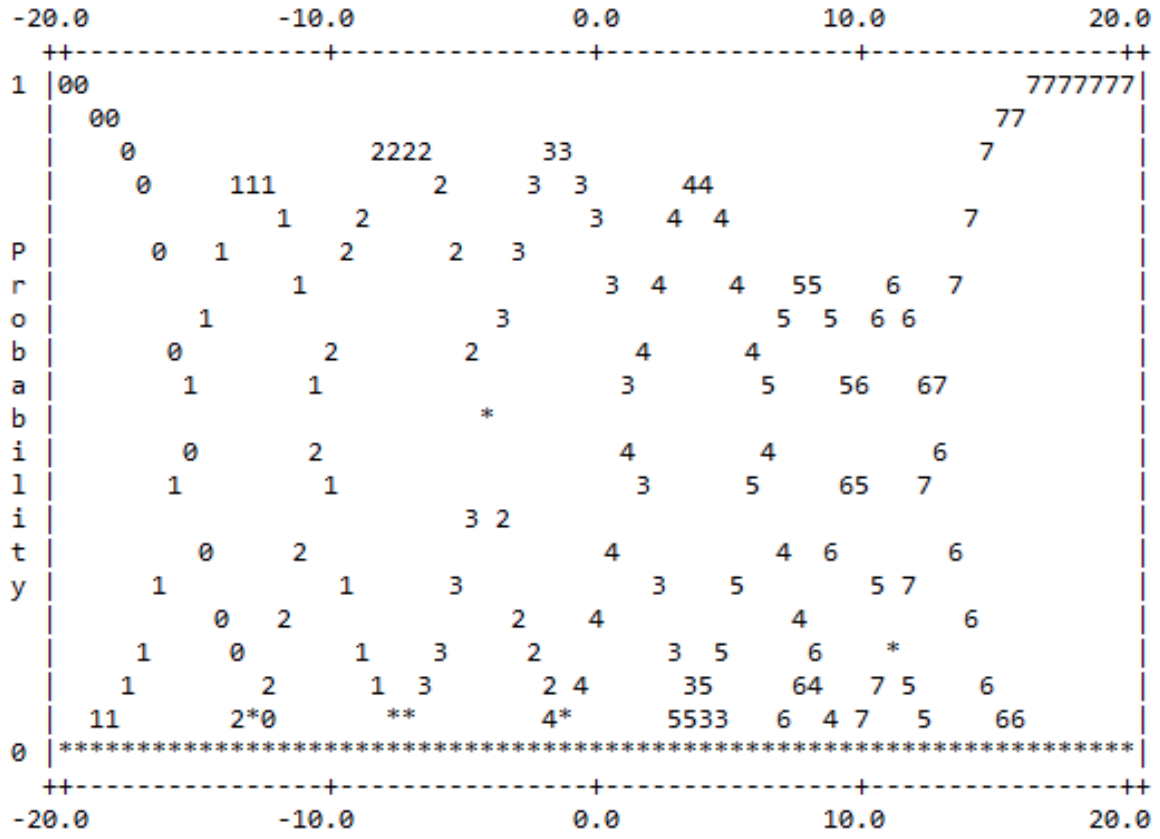
Table 20

Score distribution for grade-level cluster 9-12

Score level	Counts (%)	Outfit mean square	Rasch-Andrich threshold measure
0	13 (0%)	1.2	N/A
1	186 (6%)	0.8	-15.43
2	406 (13%)	0.9	-10.40
3	614 (19%)	0.9	-4.19
4	779 (24%)	1.0	1.39
5	675 (21%)	1.0	6.27
6	396 (12%)	0.9	9.66
7	154 (5%)	1.0	12.71

Figure 33

Probability curves for the rating scale (grade-level cluster 9-12)

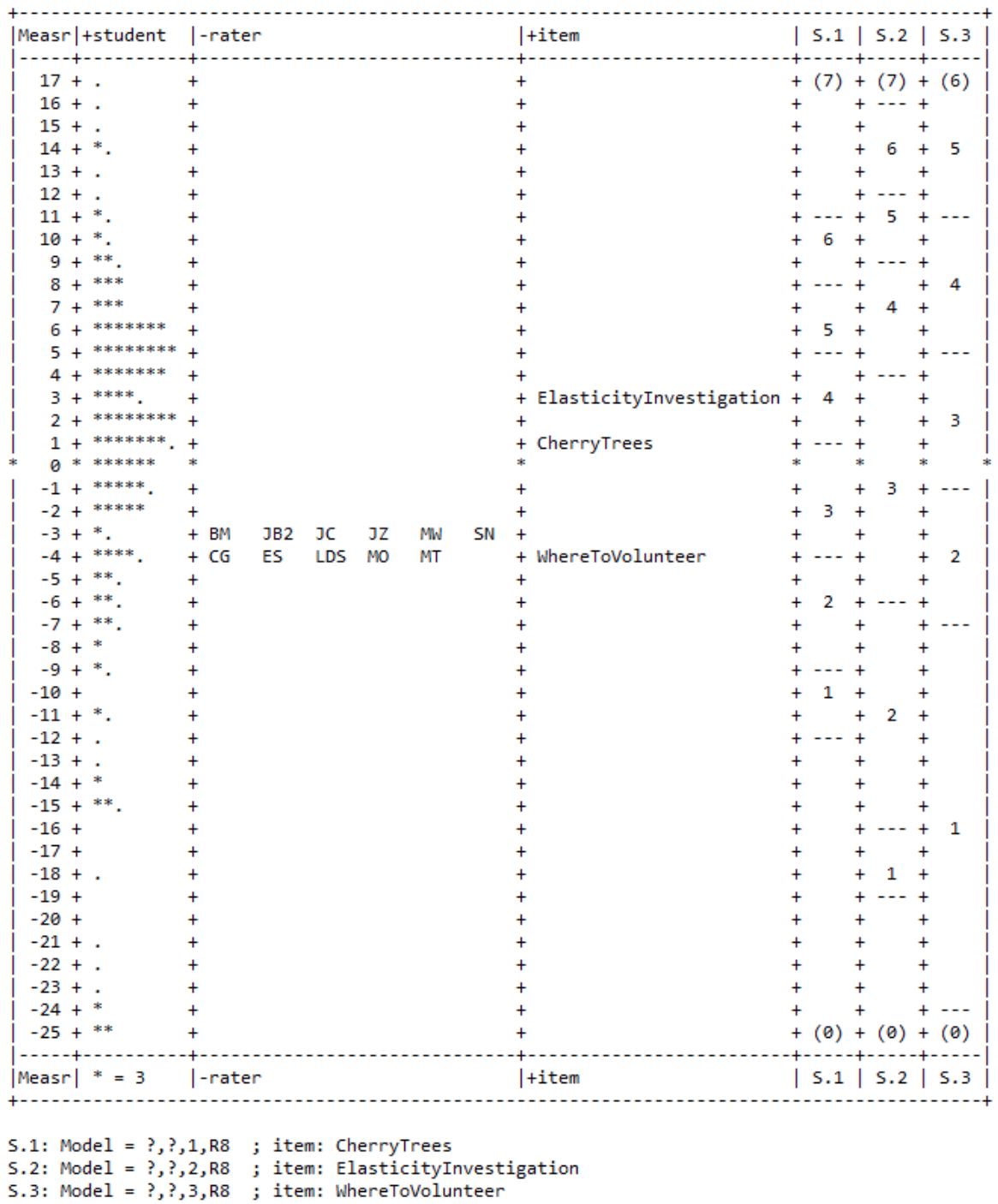


4.6.2 Grade-level cluster 9-12 (PCM)

Figure 34 shows the Wright map for grade-level cluster 9-12 with separate probabilistic model estimates of the scores for individual tasks. Unexpected responses for each task were also examined to understand the instances in which individual raters assigned misfitting ratings to misfitting student responses. For *Cherry Trees*, misfitting ratings involved 10 raters and 17 student responses; for *Elasticity Investigation*, 8 raters and 24 student responses were involved; for *Where to Volunteer*, 11 raters and 36 student responses were involved. Tables 21 through 23 highlight cases with high discrepancies between expected and observed scores. Category statistics for individual items are in Table 24 and Figures 35 to 37 show their corresponding probability curves. Of them all, the probability curves for *Cherry Trees* were perhaps the most unclear with all the hills clustering together. Compared with *Elasticity Investigation*, this task had a relatively high threshold for score levels 0 and 1 and a lower threshold for score levels 6 and 7. This means that the task was not able to distinguish the very proficient (those above a logit value of 11.04) and the least proficient (those below a logit value of -12.29) students.

Figure 34

Wright map for grade-level cluster 9-12 PCM



S.1: Model = ?,?,1,R8 ; item: CherryTrees
S.2: Model = ?,?,2,R8 ; item: ElasticityInvestigation
S.3: Model = ?,?,3,R8 ; item: WhereToVolunteer

Table 21

Unexpected responses for grade-level cluster 9-12, Cherry Trees

Student – Rater	Expected rating	Observed rating
28 – BM	3.9	2
44 – BM	4.9	3
14 – JB2	5.9	4
43 – CG	1.6	0
73 – MW	5.8	4

Table 22

Unexpected responses for grade-level cluster 9-12, Elasticity Investigation

Student – Rater	Expected rating	Observed rating
159 – JB2	5.2	7

Table 23

Unexpected responses for grade-level cluster 9-12, Where to Volunteer

Student – Rater	Expected rating	Observed rating
283 – JB2	2.8	1
289 – JB2	2.7	1

Table 24

Score distribution for individual writing tasks of grade-level cluster 9-12

Score level	Writing task	Counts (%)	Outfit mean square	Rasch-Andrich threshold measure
0	1	3 (0%)	1.8	N/A
0	2	5 (0%)	1.4	N/A
0	3	5 (0%)	0.4	N/A
1	1	30 (3%)	0.2	-12.29
1	2	24 (2%)	0.4	-18.73
1	3	132 (13%)	1.1	-24.39
2	1	53 (5%)	1.1	-8.63
2	2	95 (9%)	0.7	-16.71
2	3	258 (24%)	0.9	-7.05
3	1	151 (14%)	1.0	-3.82
3	2	137 (13%)	0.9	-5.91
3	3	326 (31%)	0.9	-0.91
4	1	266 (25%)	1.0	0.72
4	2	304 (28%)	0.9	4.24
4	3	209 (20%)	1.2	4.83
5	1	280 (26%)	1.0	4.92
5	2	283 (26%)	1.0	9.25
5	3	112 (11%)	1.1	10.79
6	1	204 (19%)	0.9	8.07
6	2	178 (16%)	1.0	12.35
6	3	14 (1%)	0.8	16.72
7	1	91 (8%)	1.0	11.04
7	2	63 (6%)	1.0	15.51
7	3	N/A	N/A	N/A

Note: Task 1 = Cherry Trees; Task 2 = Elasticity Investigation; Task 3 = Where to Volunteer

Figure 35

Probability curves for the rating scale (grade-level cluster 9-12; Cherry Trees)

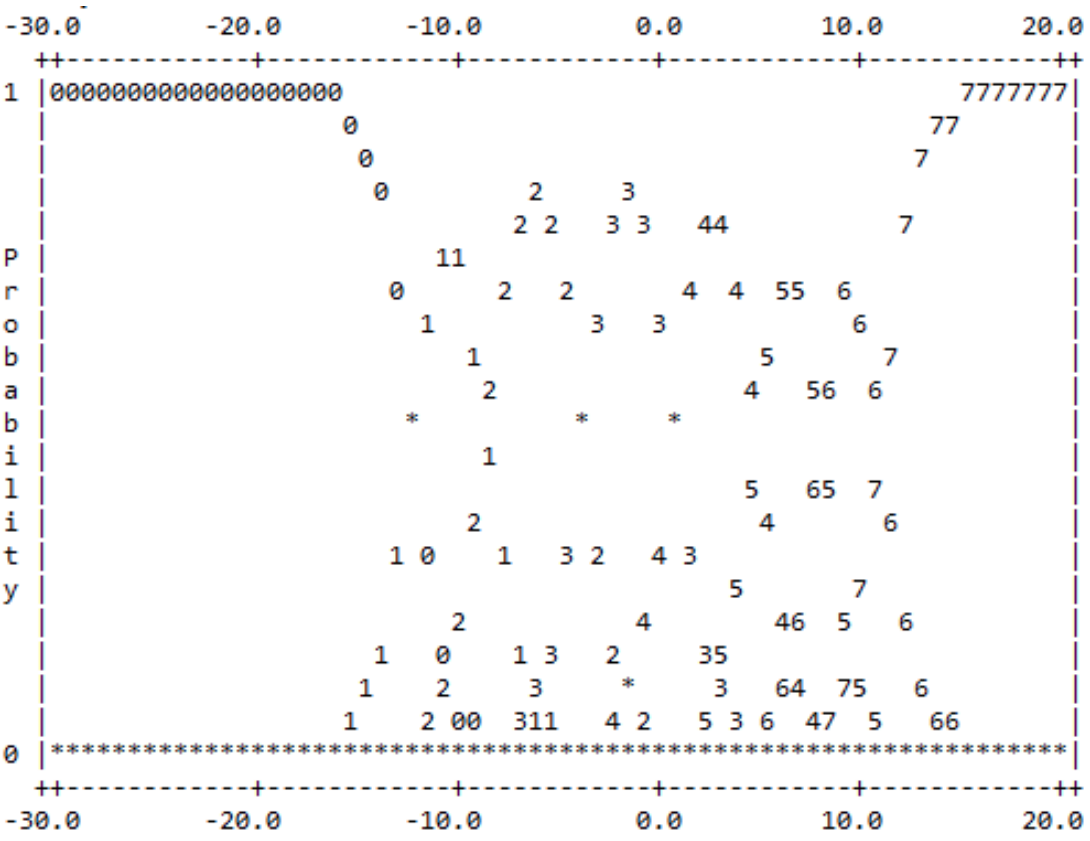


Figure 36

Probability curves for the rating scale (grade-level cluster 9-12; Elasticity Investigation)

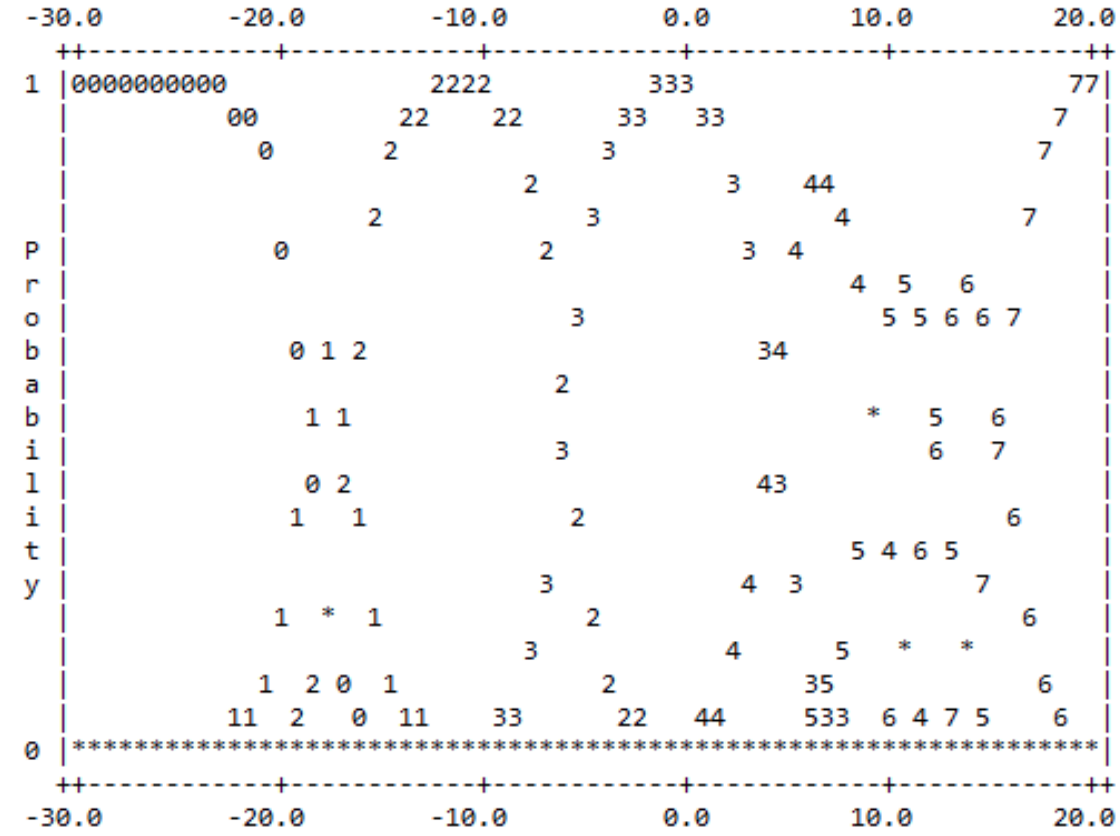
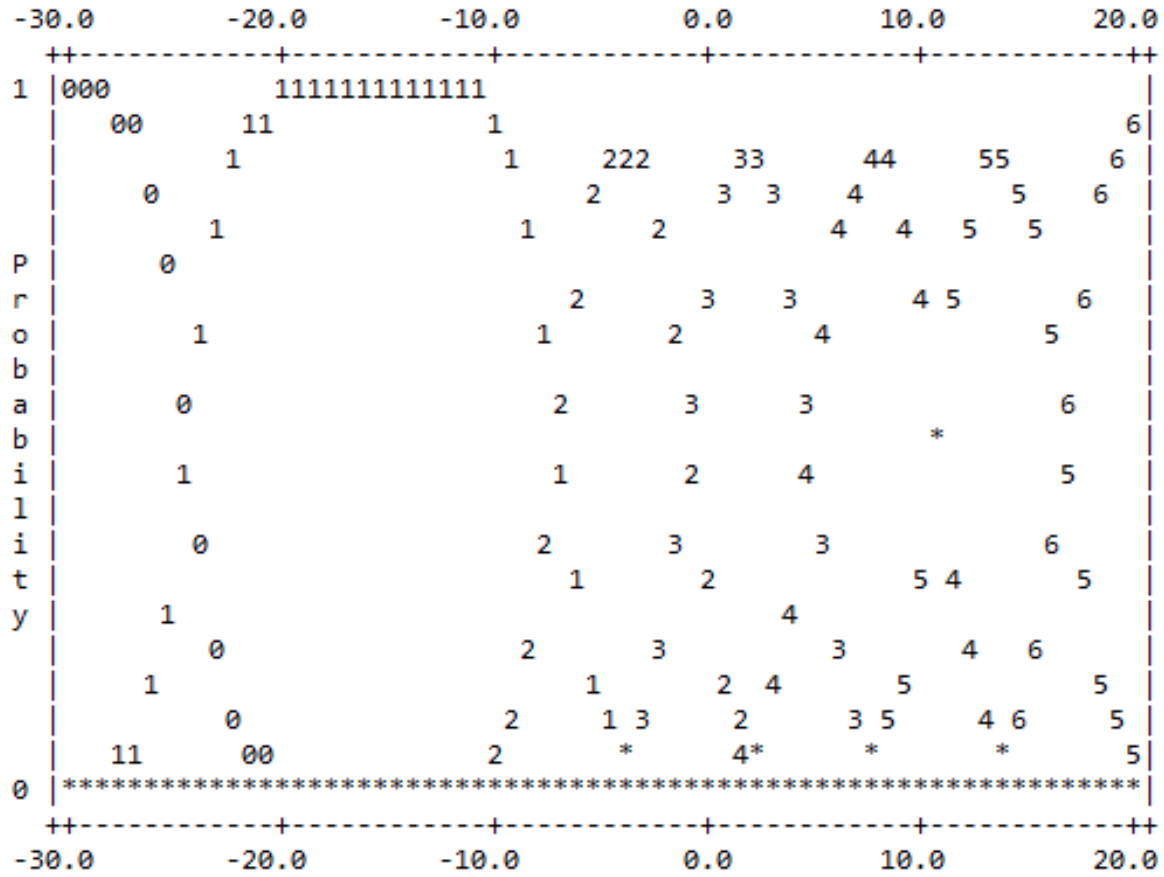


Figure 37

Probability curves for the rating scale (grade-level cluster 9-12; Where to Volunteer)



5. Discussion and conclusions

5.1 The four hypotheses

Based on the results, we reexamined the four proposed hypotheses:

1. A well-functioning rating scale will result in all score points being used and no single score point being overly used (variation in ratings).

This hypothesis was mostly supported by the results. According to the category statistics, there was variation in ratings as all score points were used across all grade-level clusters. The only concern is that score level 7 was not assigned in some tasks and was used very infrequently, with a distribution rate of only 2.5% across the entire dataset. An examination of the frequency graphs and table shows that higher grade-level clusters and Tier B/C tasks elicited more responses that were awarded score level 7, whereas the same score level was rarely used in grade-level clusters 1 and 23 or Tier A tasks. Although the psychometric evidence (e.g., outfit mean square and Rasch-Andrich threshold measure) suggests the feasibility of an eight-point scale, a seven-point scale (i.e., deleting score level 7) can reduce the possibility of leaving score points unused. Whether or not score level 7 should be removed requires careful consideration of various factors, including psychometric quality and practical concerns.

2. A well-functioning rating scale will result in small differences between raters in terms of their leniency and harshness as a group (rater separation).

Some groups of raters were relatively different in severity than others (e.g., grade-level cluster 4-5). It should be noted that rater strata indices across all grade-level clusters were quite large due to the large difference between the standard deviation of the severity measures and the standard error of each rater's severity estimate. This statistical inflation was inevitable because the number of ratings assigned by each rater was large. These findings do not necessarily mean that raters' performances were greatly different. Yet, in the meantime, it should be noted that raters across grade-level clusters were still exhibiting differences in terms of leniency and harshness. More rater training can be conducted to ensure that raters maintain similar severity levels.

3. A well-functioning rating scale will result in high rater reliability as indicated by rater point biserial correlations and exact agreement rates (rater reliability).

Overall, raters were performing consistently individually and as a group. Most raters had infit and outfit mean square values within the range of 0.5 and 1.5, meaning they were maintaining good intra-rater reliability. As a group, they had acceptable inter-rater reliability as indicated by their point biserial correlations (mean range: .70 – .75) and exact agreement rates (mean range: 61.1% – 70.9%). Moving forward, to achieve higher reliability for operational rating, raters might benefit from more thorough training to become more familiar with the scale.

4. A well-functioning rating scale will result in high candidate discrimination (student discrimination).

Candidate discrimination was large for all grade-level clusters. There was at least a 20-logit span in all cases, and the strata indices were around 10. This shows that the scale was able to effectively distinguish students across proficiency levels.

5.2 Implications

This study shows the quality and benefits of empirically developing a writing scoring scale. The validation results suggest the scale's ability to represent test takers of various proficiency levels and its capacity to help raters perform similarly to each other, likely because it captures a range of possible performances based on empirical data. Scale developers can consider adopting this approach to develop task-relevant scales to ensure more accurate scoring. This study also demonstrates the importance of multi-faceted Rasch analysis in validating a scoring scale for an operational writing test. The analysis provided meaningful information including rater severity and student discrimination, allowing for a comprehensive diagnosis of scale functionality. This method is not only applicable to large-scale assessments like WIDA ACCESS, but is also appropriate for smaller-scale local tests or classroom assessments for which sufficient data is collected.

5.3 Limitations and conclusions

The results of this study provide validity evidence for the quality of the new WIDA ACCESS for ELLs writing scoring scale. While the study suggests the scoring scale is well-functioning overall, it has several limitations that should be considered when interpreting the findings. First, due to the characteristics of the dataset, we were not able to investigate all grade-level clusters as a whole but rather had to analyze each grade-level cluster separately. To properly link separate grade-level clusters, common tasks or raters are required. However, neither of these options is feasible due to how DRC rates the test operationally. Raters are not typically trained to score across grade-level clusters; thus, asking DRC raters to do so for this study would have risked having them rate unfamiliar grade-level clusters and writing tasks, which could have introduced more variability into the findings. Additionally, we could not

maintain a balanced design when sampling student responses as there are usually very few highly rated responses. We also only had two to three tasks in each grade-level cluster, limiting the generalizability of our findings. Notwithstanding these limitations, this study still offers important psychometric information regarding the new scale. Future research can supplement quantitative results with qualitative observations such as survey responses or raters' comments on individual writing tasks to gain a complete understanding of the quality of the scoring scale.

References

Banerjee, J., Yan, X., Chapman, M., & Elliott, H. (2015). Keeping up with the times: Revising and refreshing a rating scale. *Assessing Writing, 26*, 5–19. https://doi.org/10.1016/j.asw.2015.07.001

Becker, A. (2018). Not to scale? An argument-based inquiry into the validity of an L2 writing rating scale. *Assessing Writing, 37*, 1–12. https://doi.org./10.1016/j.asw.2018.01.001

Bond, T., & Fox, C. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences.* Routledge.

Eckes, T. (2005). Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis. *Language Assessment Quarterly, 2*(3), 197–221. https://doi.org/10.1207/s15434311laq0203_2

Goodwin, S. (2016). A Many-Facet Rasch analysis comparing essay rater behavior on an academic English reading/writing test used for two purposes. *Assessing Writing, 30,* 21–31. https://doi.org/10.1016/j.asw.2016.07.004

Hsieh, M. (2013). An application of multifaceted Rasch measurement in the Yes/No Angoff standard setting procedure. *Language Testing, 30*(4), 491–512. https://doi.org/10.1177/0265532213476259

Knoch, U. (2007). 'Little coherence, considerable strain for reader': A comparison between two rating scales for the assessment of coherence. *Assessing Writing, 12*, 108–128. https://doi.org/10.1016/j.asw.2007.07.002

Knoch, U. (2009). Diagnostic assessment of writing: A comparison of two rating scales. *Language Testing, 26*(2), 275–304. https://doi.org/10.1177/0265532208101008

Li, W. (2022). Scoring rubric reliability and internal validity in rater-mediated EFL writing assessment: Insights from many-facet Rasch measurement. *Reading and Writing, 35,* 2409–2431. https://doi.org/10.1007/s11145-022-10279-1

Linacre, J. M. (2002). Optimizing rating scale category effectiveness. *Journal of Applied Measurement, 3*(1), 85–106.

Linacre, J. M. (2004). Predicting measures from rating scale or partial credit categories for samples and individuals. *Rasch Measurement Transactions, 18*(1), 972.

Linacre, J. M. (2017). Facets computer program for many-facet Rasch measurement, version 3.80.0. https://www.winsteps.com/facets.htm

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement, 4*(4), 386–422.

Turner, C. E., & Upshur, J. A. (2002). Rating scales derived from student samples: Effects of the scale maker and the student sample on scale content and student scores. *TESOL Quarterly, 36*(1), 49–70. https://doi.org/10.2307/3588360

WIDA. (2020). *WIDA English language development standards framework, 2020 edition: Kindergarten–grade 12.* Board of Regents of the University of Wisconsin System.

Yan, X. (2014). An examination of rater performance on a local oral English proficiency test: A mixed-methods approach. Language Testing, 31(4), 501–527. https://doi.org/10.1177/0265532214536171

Appendix A. The new WIDA ACCESS writing scoring rubric

Score Point 7

- Ideas are coherently organized, using language that connects ideas together smoothly throughout the response achieving a clear flow of text. Response clearly demonstrates features of the intended key language use (Narrate, Inform, Explain, Argue) and content area.
- Response contains a wide variety of sentence structures whose meaning is always clear. Response demonstrates control of complex sentence structures, though sentences may not always be error-free
- Response uses a wide variety of words and phrases appropriately and precisely, with choices that are relevant to the task context.

Score Point 6

- Ideas are well organized and elaborated, using a variety of connectors to create some cohesion. Response demonstrates some features of the intended key language use (Narrate, Inform, Explain, Argue).
- Response contains a variety of sentence structures with consistently clear meaning, though occasional errors may be present.
- Response uses a variety of words and phrases with some precision that are usually appropriate to the task context.

Score Point 5

- Response has a clear organizational pattern with some elaboration. Response uses connectors that join ideas together and these are usually used appropriately.
- Response contains some compound or complex sentence structures with generally clear meaning, though they may include some errors.
- Response uses a range of words and phrases that are generally appropriate and show emerging precision, including some words and phrases related to the task context.

Score Point 4

- Response uses connectors and may have some evidence of an organizational pattern, though longer responses in particular may lack coherence.
- Response contains some compound or complex sentences, though errors may obscure meaning.
- Response uses a range of words and phrases from beyond the stimulus that generally convey the intended meaning.

Score Point 3

- Response shows connected text, which may include some simple connectors (e.g., *and, then, but*), though they may be used repetitively and may not always be used accurately.
- Response contains some complete sentences, though frequent errors may obscure meaning.
- Response uses some original words and phrases, in addition to language drawn from the stimulus

Score Point 2

- Response includes at least one clear, complete sentence, but does not include connected text.
- Response uses a small number of original words and phrases, in addition to language drawn from the stimulus.

Score Point 1

- Response includes at least one recognizable word in English, and may contain attempts at phrases or sentences, but does not include any clear, complete sentences.

Score Point 0: The response

- contains no discernible words in English, though it may contain letters or scribbles [I];
- consists only of text that is completely off task (shows no understanding of or interaction with the prompt) [T];
- is entirely in a language other than English [F];
- consists only of verbatim copied text (with no reformulation or adaptation, though it may contain copying errors) [C];
- is entirely blank [B];
- is partially or entirely plagiarized (i.e., copied or adapted from an external source) [K].

Appendix B. Distribution of score levels with the new scale: Frequency table

Grade-level cluster	Writing task	Counts (%)
1	Cleaning Up	0: 176 (16%) 1: 224 (21%) 2: 234 (22%) 3: 226 (21%) 4: 149 (14%) 5: 62 (6%) 6: 9 (1%)
1	Growing Plants	0: 187 (18%) 1: 279 (27%) 2: 243 (23%) 3: 196 (19%) 4: 111 (11%) 5: 23 (2%) 6: 1 (0%)
1	Giant Pandas	0: 78 (7%) 1: 70 (6%) 2: 109 (10%) 3: 312 (28%) 4: 285 (26%) 5: 160 (15%) 6: 75 (7%) 7: 11 (1%)
1	**All tasks**	**0: 441 (14%) 1: 573 (18%) 2: 586 (18%) 3: 734 (23%)** **4: 545 (17%) 5: 245 (8%) 6: 85 (3%) 7: 11 (0%)**
23	Garden Surprise	0: 132 (13%) 1: 159 (16%) 2: 175 (18%) 3: 212 (21%) 4: 188 (19%) 5: 118 (12%) 6: 15 (2%) 7: 1 (0%)
23	Changing Water	0: 38 (4%) 1: 40 (4%) 2: 111 (12%) 3: 233 (25%) 4: 300 (33%) 5: 146 (16%) 6: 49 (5%) 7: 3 (0%)
23	**All tasks**	**0: 170 (9%) 1: 199 (10%) 2: 286 (15%) 3: 445 (23%)** **4: 488 (25%) 5: 264 (14%) 6: 64 (3%) 7: 4 (0%)**
45	Marsh Ecosystem	0: 132 (13%) 1: 190 (19%) 2: 286 (29%) 3: 264 (26%) 4: 91 (9%) 5: 37 (4%)
45	Search for Info	0: 9 (1%) 1: 73 (8%) 2: 128 (14%) 3: 263 (29%) 4: 184 (20%) 5: 123 (14%) 6: 70 (8%) 7: 50 (6%)
45	**All tasks**	**0: 141 (7%) 1: 263 (14%) 2: 414 (22%) 3: 527 (28%)** **4: 275 (14%) 5: 160 (8%) 6: 70 (4%) 7: 50 (3%)**
68	Illustrator	0: 45 (5%) 1: 76 (8%) 2: 203 (23%) 3: 246 (27%) 4: 207 (23%) 5: 113 (13%) 6: 10 (11%)
68	Color and Temperature	0: 16 (2%) 1: 34 (4%) 2: 80 (10%) 3: 163 (20%) 4: 181 (23%) 5: 158 (20%) 6: 101 (13%) 7: 67 (8%)
68	**All tasks**	**0: 61 (4%) 1: 110 (6%) 2: 283 (17%) 3: 409 (24%)** **4: 388 (23%) 5: 271 (16%) 6: 111 (7%) 7: 67 (4%)**
912	Cherry Trees	0: 14 (1%) 1: 30 (3%) 2: 53 (5%) 3: 151 (14%) 4: 266 (24%) 5: 280 (25%) 6: 204 (19%) 7: 102 (9%)
912	Elasticity Investigation	0: 16 (1%) 1: 24 (2%) 2: 95 (9%) 3: 137 (12%) 4: 304 (28%) 5: 283 (26%) 6: 178 (16%) 7: 63 (6%)
912	Where to Volunteer	0: 49 (4%) 1: 132 (12%) 2: 258 (23%) 3: 326 (30%) 4: 209 (19%) 5: 112 (10%) 6: 14 (1%)
912	**All tasks**	**0: 79 (2%) 1: 186 (6%) 2: 406 (12%) 3: 614 (19%)** **4: 779 (24%) 5: 675 (20%) 6: 396 (12%) 7: 165 (5%)**

Appendix C. Distribution of score levels with the current scale: Frequency table

Grade-level cluster	Writing task	Counts (%)
1	Cleaning Up	NS: 21 (21%) 1: 10 (10%) 1+: 11 (11%) 2: 11 (11%) 2+: 25 (25%) 3: 10 (10%) 3+: 7 (7%) 4: 5 (5%)
1	Growing Plants	NS: 17 (17%) 1: 16 (16%) 1+: 20 (20%) 2: 10 (10%) 2+: 25 (25%) 3: 6 (6%) 3+: 6 (6%)
1	Giant Pandas	NS: 5 (5%) 1: 5 (5%) 1+: 6 (6%) 2: 10 (10%) 2+: 25 (25%) 3: 20 (20%) 3+:15 (15%) 4: 7 (5%) 4+: 7 (7%)
1	**All tasks**	**NS: 43 (14%) 1: 31 (10%) 1+: 37 (12%) 2: 31 (10%) 2+: 75 (25%) 3: 36 (12%) 3+: 28 (9%) 4: 12 (4%) 4+: 7 (2%)**
23	Garden Surprise	NS: 13 (13%) 1: 10 (10%) 1+: 9 (9%) 2: 12 (12%) 2+: 25 (25%) 3: 16 (16%) 3+: 5 (5%) 4: 5 (5%) 4+: 5 (5%)
23	Changing Water	NS: 5 (5%) 1: 5 (5%) 1+: 5 (5%) 2: 5 (5%) 2+: 22 (22%) 3: 29 (29%) 3+: 20 (20%) 4: 5 (5%) 4+: 4 (4%)
23	**All tasks**	**NS: 18 (9%) 1: 15 (8%) 1+: 14 (7%) 2: 17 (9%) 2+: 47 (24%) 3: 45 (23%) 3+: 25 (13%) 4: 10 (5%) 4+: 9 (5%)**
45	Marsh Ecosystem	NS: 13 (13%) 1: 17 (17%) 1+: 16 (16%) 2: 26 (26%) 2+: 18 (18%) 3: 5 (5%) 3+: 5 (5%)
45	Search for Info	NS: 2 (2%) 1: 5 (5%) 1+: 8 (8%) 2: 16 (16%) 2+: 28 (28%) 3: 16 (16%) 3+: 10 (10%) 4: 5 (5%) 4+: 5 (5%) 5: 5 (5%)
45	**All tasks**	**NS: 15 (8%) 1: 22 (11%) 1+: 24 (12%) 2: 42 (21%) 2+: 46 (23%) 3: 21 (11%) 3+: 15 (8%) 4: 5 (3%) 4+: 5 (3%) 5: 5 (3%)**
68	Illustrator	NS: 5 (5%) 1: 5 (5%) 1+: 5 (5%) 2: 15 (15%) 2+: 25 (25%) 3: 25 (25%) 3+: 10 (10%) 4: 5 (5%) 4+: 5 (5%)
68	Color and Temperature	NS: 2 (2%) 1: 5 (5%) 1+: 5 (5%) 2: 7 (7%) 2+: 25 (25%) 3: 26 (26%) 3+: 15 (15%) 4: 5 (5%) 4+: 5 (5%) 5: 5 (5%)
68	**All tasks**	**NS: 7 (4%) 1: 10 (5%) 1+: 10 (5%) 2: 22 (11%) 2+: 50 (25%) 3: 51 (26%) 3+: 25 (13%) 4: 10 (5%) 4+: 10 (5%) 5: 5 (3%)**
912	Cherry Trees	NS: 2 (2%) 1: 3 (3%) 1+: 3 (3%) 2: 5 (5%) 2+: 10 (10%) 3: 25 (25%) 3+: 28 (28%) 4: 14 (14%) 4+: 5 (5%) 5: 5 (5%)
912	Elasticity Investigation	NS: 3 (3%) 1: 3 (3%) 1+: 3 (3%) 2: 5 (5%) 2+: 14 (14%) 3: 32 (32%) 3+: 24 (24%) 4: 6 (6%) 4+: 5 (5%) 5: 5 (5%)
912	Where to Volunteer	NS: 7 (7%) 1: 7 (7%) 1+: 8 (8%) 2: 24 (24%) 2+: 24 (24%) 3: 15 (15%) 3+: 5 (5%) 4: 5 (5%) 4+: 5 (5%)
912	**All tasks**	**NS: 12 (4%) 1: 13 (4%) 1+: 14 (5%) 2: 34 (11%) 2+: 48 (16%) 3: 72 (24%) 3+: 57 (19%) 4: 25 (8%) 4+: 15 (5%) 5: 10 (3%)**

WIDA™
UNIVERSITY OF WISCONSIN–MADISON

Technical Report

No. TR-2024-1
April 2024

Wisconsin Center for Education Research
University of Wisconsin–Madison
1025 West Johnson St., MD #23
Madison, WI 53706

Client Services Center toll free:
(866) 276-7735

help@wida.us

wida.wisc.edu