WIDA™
UNIVERSITY OF WISCONSIN–MADISON

Technical Report

No. TR-2024-2
August 2024

Development of a New WIDA Writing Scoring Rubric for Grades 1–12

Authors: Mark Chapman, PhD; Ping-Lin Chuang, PhD; Tanya Bitterman; Heather Elliott

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

Contents

Project aim and research hypotheses

The main aim of this project was to develop a new scoring rubric grounded in the *WIDA English Language Development Standards Framework, 2020 Edition: Kindergarten–Grade 12* (hereafter, *WIDA ELD Standards Framework, 2020 Edition* or *2020 Edition*). This rubric will be used for scoring responses to the writing tasks on ACCESS for ELLs Online, ACCESS for ELLs Paper, WIDA Screener Online, and WIDA Screener Paper. The need for this project was established during a series of meetings between WIDA and The Center for Applied Linguistics (CAL) which focused on a thorough review of the WIDA ELD Standards Framework, 2020 Edition. During these meetings, WIDA and CAL staff identified components of WIDA assessments that would require revision to align with the 2020 Edition.

Two features of the WIDA ELD Standards Framework, 2020 Edition that differed from previous editions prompted the need for a new writing rubric. The first was the shift to grade-level cluster-specific proficiency level descriptors. Whereas previous editions had featured a single set of descriptors for grades 1–12, the 2020 Edition includes different sets of descriptors for grade-level clusters 1, 2–3, 4–5, 6–8, and 9–12. The new writing scoring rubric needed to incorporate the proficiency ranges described across all the five grade-level clusters. The second was the greater emphasis on the discourse dimension of language in the 2020 Edition. The WIDA ELD Standards Framework has consistently described three dimensions of language: discourse, sentence, and word/phrase. In the 2020 Edition, the discourse dimension was expanded into three different criteria: *organization of language, cohesion of language,* and *density of language* (see Table 1). In previous editions of the ELD Standards Framework, there had been only a single criterion under the discourse dimension, so the new writing scoring rubric needed to encode this greater emphasis on discourse to reflect the 2020 Edition.

Table 1

Excerpt from WIDA ELD Standards Framework, 2020 Edition

Dimension	Criteria	Focus on …	Sample Language Features
Discourse	Organization of language	How ideas are coherently organized to meet a purpose through organizational patterns characteristic of the genre	Whole text organizational patterns, such as introduction, body, conclusion; claim, evidence, reasoning
Discourse	Cohesion of language	How language connects ideas within and across sentences and discourse using a range of cohesive devices	Cohesive devices, such as repeated words, synonyms, pronoun substitution, connectors
Discourse	Density of language	How information in noun groups is expanded or consolidated	Noun groups expanded with resources, such as adjectives or other modifiers added before nouns, prepositional phrases following nouns, nominalization
Sentence	Grammatical complexity of language	How relationships are expressed with clauses through simple, compound, and complex sentences	Simple, compound, complex sentences; coordinating, subordinating conjunctions; dependent and independent clauses
Word/Phrase	Precision of language	How everyday, cross-disciplinary, and technical language more effectively conveys precise meaning	A variety of words and phrases such as, adverbials of time, manner, and place; verb types; abstract nouns

Terminology

This report makes repeated reference to two different instruments, the *writing scoring* ***scale*** and the *writing scoring* ***rubric***. The *writing scoring scale* refers to the WIDA Writing Scoring Scale Grades 1-12, which was introduced in 2015 and is the instrument that is being replaced by the newly developed WIDA Writing Scoring Rubric Grades 1-12 (*writing scoring rubric*). This report describes the development steps taken to create the new *writing scoring rubric*. The new instrument is called the *writing scoring rubric* to bring consistency to different instruments used across multiple assessments that serve similar purposes.

Additional justification for a new writing rubric

While the need to design a new scoring rubric that aligned with the 2020 Edition was the main impetus for this project, several features of the writing scoring scale used from academic years 2015-16 to 2024-25 to score writing responses on ACCESS Online, ACCESS Paper, WIDA Screener Online, and WIDA Screener Paper, provided motivation to correct issues that have existed with this instrument. This scale was grounded in the 2012 Amplifications of the WIDA

English Language Development Standards and was developed using a theoretical approach to scale design, which leaned heavily on the performance definitions described in that publication. The resulting scale had a number of issues when used operationally and the opportunity to resolve these issues provided further motivation to the team developing the new writing scoring rubric.

The primary issue with the writing scoring scale was that the higher score points on the scale were rarely awarded. The top three score points (5, 5+, and 6) were so rarely awarded that they had to be consolidated into a single raw score point for psychometric purposes. This issue is described in more detail in the ACCESS for ELLs Annual Technical Report (see, for example the *Annual Technical Report for ACCESS for ELLs Online English Language Proficiency Test Series 601, 2022-2023 Administration*, p. 77-78). The treatment of raw scores is shown in Table 2 (reproduced from the *Annual Technical Report for ACCESS for ELLs Online English Language Proficiency Test Series 503, 2021–2022 Administration*).

Table 2
Truncation of writing scoring scale raw score range

Rating to raw score conversion (Writing)

Rating	Raw Score
Nonscorable	0
1	1
1+	2
2	3
2+	4
3	5
3+	6
4	7
4+	8
5	9
5+	9
6	9

Tables 3 and 4 demonstrate the low rates at which the higher score points on the writing scoring scale are awarded. Table 3 shows raw scores awarded to responses to Tier A writing tasks on 2021–22 ACCESS Online test. Table 4 shows raw scores awarded to responses to Tier B/C writing tasks on the 2021–22 ACCESS Online test.

Table 3

Raw score distributions on Tier A tasks

Raw Score	Tier A Task 1	Tier A Task 2
0	18.45%	21.79%
1	11.91%	10.90%
2	11.43%	12.04%
3	20.91%	20.25%
4	27.10%	25.11%
5	8.09%	8.51%
6	1.97%	1.36%
7	0.13%	0.04%
8	0.01%	0.00%
9	0.00%	0.00%

Table 4

Raw score distributions on Tier B/C tasks

Raw Score	Tier B/C Task 1	Tier B/C Task 2
0	1.04%	2.05%
1	1.77%	1.09%
2	7.70%	6.36%
3	18.11%	15.44%
4	41.56%	40.34%
5	20.25%	25.98%
6	7.55%	7.72%
7	1.69%	0.92%
8	0.30%	0.09%
9	0.03%	0.01%

These raw score point distributions highlight the issue of the disuse of the three highest score points. The distributions also indicate another issue with the writing scoring scale: the frequent use of single score points on the scale. For example, responses to the Tier B/C tasks are often awarded score point 4 (2+ on the scoring scale) and score point 5 (3 on the scoring scale). In 2021–22 over 60% of responses were awarded these two raw score points. These score point distribution data indicate that the writing scoring scale may not discriminate well between student writing performances in the middle of the range of observed scores. These score point distributions also demonstrate that the plus score points (converted raw scores of 2, 4, 6 in Tables 3 and 4) were frequently awarded by raters using the writing scoring scale.

An additional issue with the writing scoring scale has been the reporting of rater reliability data. As the scoring scale has a relatively large number of raw score points (12 score points for raters and 10 score points psychometrically), raw scores that were both exact and adjacent were considered in agreement when calculating rater agreement data. When interpreting rater agreement data calculated this way, the reported agreement rates may appear inflated. Clarifying the calculation and reporting of rater reliability data was another stated aim of the new writing scoring rubric project.

Finally, the writing scoring scale was 0–6 for raters, which was the source of some score interpretation confusion, particularly in the context of WIDA Screener. WIDA Screener writing responses are scored by local test administrators. Test scores are reported in terms of WIDA proficiency levels, also on a 0–6 scale. WIDA Screener test administrators commonly believed that the raw score awarded on the writing test would directly equate to the reported proficiency level. However, that is not how the score calculation and reporting works. The reported proficiency level scores are typically lower than the awarded raw score. A final stated aim of the project to develop the writing scoring rubric is to mitigate the confusion resulting from the raw score range and reported score range being the same.

These issues with the writing scoring scale drove the project team to set goals for the development of the new writing scoring rubric. These goals were framed as hypotheses that serve to define the qualities of a well-functioning new rubric and relate directly to the issues with the writing scoring scale. The hypotheses are:

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

These qualities serve as the hypotheses that the newly developed writing scoring rubric would need to meet in order to be considered a success.

Relevant literature

The qualities of a well-functioning rating scale, as described in the previous section, are detailed in numerous publications (e.g. Becker, 2018; Hamp-Lyons, 1991; Knoch 2007, 2009; Weigle, 2002). In particular, the work of Knoch (2009) and Weigle (2002) was influential in determining the technical qualities (variation in ratings; rater separation; rater reliability; student discrimination) of a writing scoring rubric that the project would aim to meet. These qualities also addressed concerns with the writing scoring scale described previously.

The process that the project team followed in developing the new writing scoring rubric followed that of Turner and Upshur, which has been published in several papers (Turner, 2000; Turner & Upshur, 2002; Upshur & Turner, 1995). Turner and Upshur promote a data-informed approach to rubric development whereby the number of score points on the rubric and the rubric descriptors are derived from a systematic review of authentic student performances. That is, the rubric emerges from the language that is elicited by the test tasks.

Turner & Upshur (2002: 52) summarize this approach as:

> *A group of scale constructors, generally L2 teachers, is given a sample of writings or recorded oral performances. Working without a rating scale, the raters first arrive at a consensus on assignment of the sample performances into a specified number of levels and then identify and describe salient features that distinguish performances at adjacent levels. In this way, scale descriptors emerge from holistically scaled samples.*

Methods

As described above, the project team followed the approach of Turner & Upshur and utilized a data-informed approach to the development of the new writing scoring rubric. Following the development of the first draft of the new rubric, an extensive series of reviews was conducted by both internal and external reviewers. Finally, the new rubric was used by a team of trained raters at Data Recognition Corporation (DRC), WIDA's scoring and test delivery vendor, to try out when scoring student responses. The resulting data were analyzed using multi-faceted Rasch analyses (MFRA) to investigate the technical qualities of the new rubric. In this section, we describe the methods employed in each phase of the project.

Phase 1: Rubric development

Writing assessment and psychometrics specialists from WIDA and CAL worked together to establish a representative corpus of student writing performances that would serve as the starting point for the data-informed approach to rubric development detailed above. This corpus of student responses to writing tasks from ACCESS for ELLs needed to cover grades 1–12, include responses to both Tier A and B/C tasks, sample from different WIDA Standards and Key Language Uses, and represent the full range of operational scores observed on the assessment. Content specialists and psychometricians (lead psychometricians from both WIDA and CAL) jointly constructed a corpus of student responses that met these criteria. The responses were drawn primarily from the 2021–22 administration of ACCESS, though responses to one task were collected from the 2020–21 administration. The composition of the corpus (n=324) is shown in Table 5.

Table 5

Corpus of student responses: Number of responses in corpus by score point

Grade-level cluster	Tier	Standard	Key language use	1	1+	2	2+	3	3+	4	4+	5,5+,6	TOTAL
1	A	SC	Inform	4	4	4	5	5	5	0	0	0	**27**
1	A	LA	Inform	4	4	4	5	5	5	0	0	0	**27**
1	BC	SC	Explain	0	0	0	5	5	5	4	4	4	**27**
2-3	A	LA	Narrate	4	4	4	5	5	5	0	0	0	**27**
2-3	BC	SC	Argue	0	0	0	5	5	5	4	4	4	**27**
4-5	A	SC	Explain	4	4	4	5	5	5	0	0	0	**27**
4-5	BC	LA	Argue	0	0	0	5	5	5	4	4	4	**27**
6-8	A	LA	Narrate	4	4	4	5	5	5	0	0	0	**27**
6-8	BC	SC	Explain	0	0	0	5	5	5	4	4	4	**27**
9-12	A	LA	Argue	4	4	4	5	5	5	0	0	0	**27**
9-12	BC	LA	Inform	0	0	0	5	5	5	4	4	4	**27**
9-12	BC	SC	Explain	0	0	0	5	5	5	4	4	4	**27**
N/A	N/A	N/A	**TOTAL by score point**	**24**	**24**	**24**	**60**	**60**	**60**	**24**	**24**	**24**	**324**

The score points shown in Table 5 are those of the writing scoring scale that was being replaced. The project team decided to include three writing tasks at grade-level clusters 1 (two tasks from Tier A and one from Tier B/C) and 9–12 (one task from Tier A and two tasks from Tier B/C) to help ensure that the corpus contained sufficient responses representing both the floor and ceiling of performances elicited on the test. At the other grade-level clusters (2–3, 4–5, 6–8) there is one task for each grade-level cluster and tier combination. In terms of score point representation, we slightly overrepresented score points in the middle of the scale as these are the ones that are used most commonly during operational scoring and where the scale may fail to discriminate between performances.

Utilizing these responses to ACCESS Writing tasks, a team of WIDA and CAL writing specialists read and reviewed the 324 writing samples. This team consisted of two writing specialists from WIDA and three from CAL. The members of this team had extensive experience rating ACCESS Writing responses, creating rater training materials for ACCESS Writing operational scoring, and designing rubrics. Several members of the team also had extensive experience teaching writing skills to K–12 students, both in the United States and internationally.

Prior to reviewing the student responses, team members read the guidance of Turner and Upshur (2002). They then read the corpus of student responses in full, and divided it into two groups based on performance level: high and low. This reading and ranking was conducted independently and informed by a directive to read for the three dimensions of language encoded in the WIDA Standards: discourse, sentence, and word/phrase. Each team member recorded their decisions in a spreadsheet and then the group met to review these decisions. The meetings were conducted via Zoom, recorded, and transcribed. During the review meetings, the team members discussed the responses that had split judgements. Typically, responses that four team members agreed upon (high or low group judgment) were assigned to the consensus decision without further discussion. Responses with a 3:2 split were discussed by the team. In some cases, a consensus decision was possible, but in other cases it was not. When a consensus decision could not be reached the response was removed from the corpus for future rounds of reading. After each meeting, the process was repeated, with each set of responses being further divided into high and low groups of responses. For example, after the first round the agreed upon "high" group of responses was read by the team and further divided into two subgroups. The same approach was taken with the "low" group form the first round. As Turner and Upshur recommend, the team members met after each round and the process continued until they all agreed that there were no further meaningful divisions of performances to be made.

After determining the number of score points to be described in the new scoring rubric, the project team created the first draft of the new rubric. The authoring process was directly informed by transcripts and video recordings from the corpus review meetings and by the textual characteristics of the responses representing each score point. After reviewing the recordings of the corpus review meetings, the first draft author then closely read student responses that all corpus readers had agreed upon, indicating that these responses were strongly representative of the assigned score point. Score points 1 and 7 were drafted first, to create the floor and ceiling of the new rubric. The other score points were then drafted iteratively, again drawing heavily on responses at each score point that all corpus readers had agreed upon. The first draft of the new rubric was initially authored by one member of the project team and then reviewed extensively by the other four readers of the corpus. Draft 1 of the new writing scoring rubric is shared in Appendix A.

Phase 2: Internal reviews of the new rubric

Following the creation of the first draft of the new writing scoring rubric, it underwent a series of reviews by WIDA and CAL staff, which are described in Table 6. Each of these review phases was conducted by an independent group of reviewers. In addition to the relevant professional assessment experience of these reviewers, many of them also have experience as educators of multilingual learners in grades 1-12.

Table 6

Internal review of new scoring rubric

Activity	Parties involved	Review aims	Reviewers
Corpus reader reviews	WIDA and CAL	Check the score point descriptors are consistent with decisions made during corpus review; consider "+" score points	Reviewers all participated in the series of meetings to establish the number of score points
Standards reviews	WIDA and CAL	Check the new scale is not in conflict with the WIDA ELD Standards Framework, 2020 Edition, and Proficiency Level Descriptors	Reviewers all had deep familiarity with the WIDA ELD Standards Framework, 2020 Edition, and how assessments operationalize the standards
Content reviews	WIDA and CAL	Check the progression of the scale score points and descriptors; apply the new scale to a set of representative responses	Reviewers all had deep familiarity with creating rater training materials and developing writing tasks
SJCT reviews	WIDA	Check the descriptors are consistent with the WIDA Can Do Philosophy and social justice values	Reviewers were all members of the WIDA social justice change team that advises WIDA on social justice issues

Phase 3: Educator reviews

After the new writing scoring rubric was revised in line with feedback provided by the reviewers described above, the new rubric was provided to educators for their review. Educators are a critical group of stakeholders for the new writing scoring rubric as it will be used to score responses to the WIDA Screener Writing test. Unlike ACCESS for ELLs, which is scored by trained raters at DRC, WIDA Screener Writing responses are locally scored by test administrators, most of whom are educators. WIDA recruited a group of educators who provided self-reported data on years of experience scoring the WIDA Screener Writing test, and the number of students tested annually. These educators were recruited from across different geographical regions of the consortium and had experience testing students and scoring writing responses from grades 1–12. For their review, educators (n=23) were provided with the following materials:

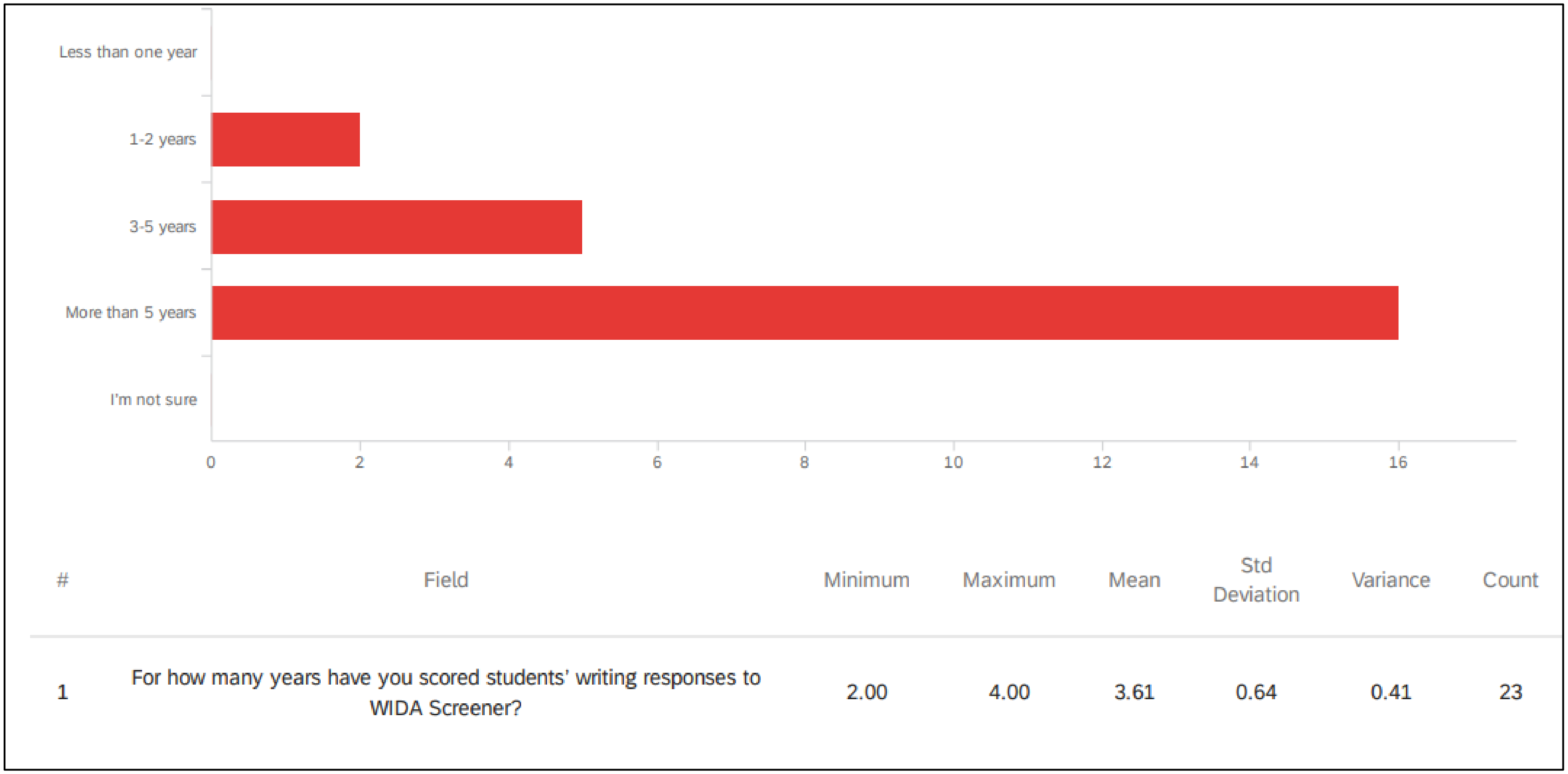
- An instructional video to introduce the new writing scoring rubric
- Two sets of anchor responses to review

    - One Tier A Set and one Tier B/C set representing two different grade-level clusters
- A set of student responses (five responses to each of two tasks) to practice scoring using the new writing scoring rubric

Educators completed a Qualtrics survey to provide their feedback on the new rubric. The following graphs show key findings from the survey questions, along with some relevant and supporting quotes from the open-ended survey items. One point to consider with regard to educator feedback is that WIDA Screener is administered to incoming students, most commonly at the start of the school year. Among students who take WIDA Screener are many beginners and newcomers who have not previously studied English. As such, educators who score WIDA Screener Writing responses may award a higher proportion of low score points than is the case with the ACCESS test.

Figure 1

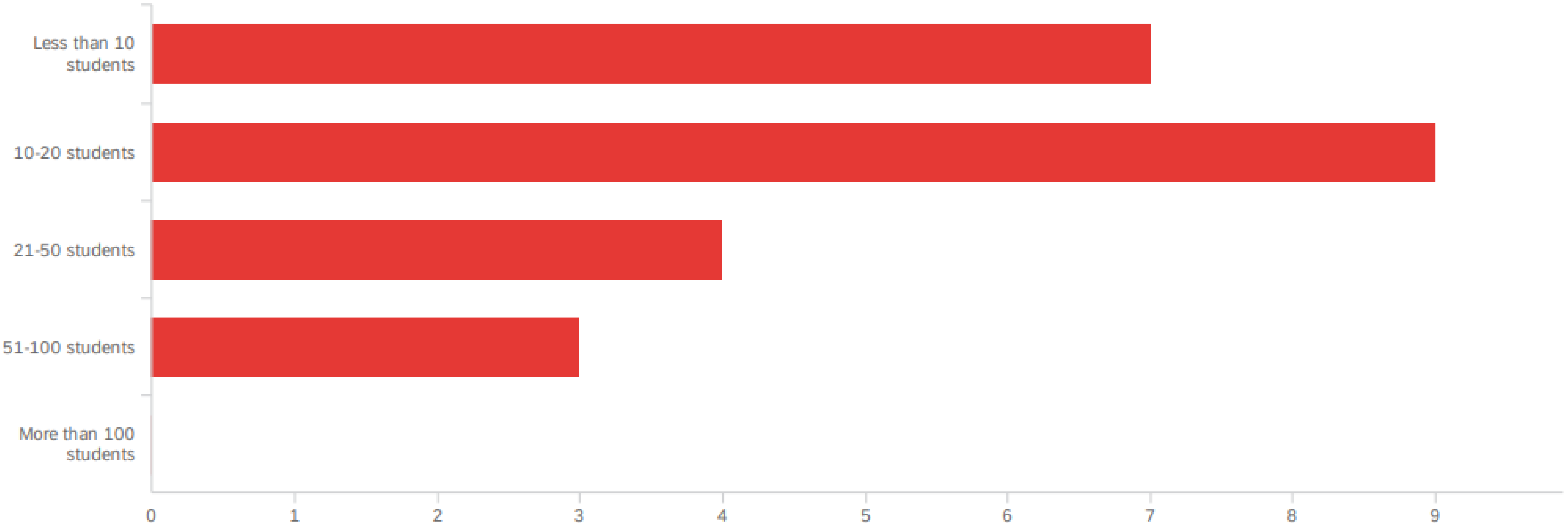
Respondents' experience scoring WIDA Screener Writing test



#	Field	Minimum	Maximum	Mean	Std Deviation	Variance	Count
1	For how many years have you scored students' writing responses to WIDA Screener?	2.00	4.00	3.61	0.64	0.41	23

Figure 2

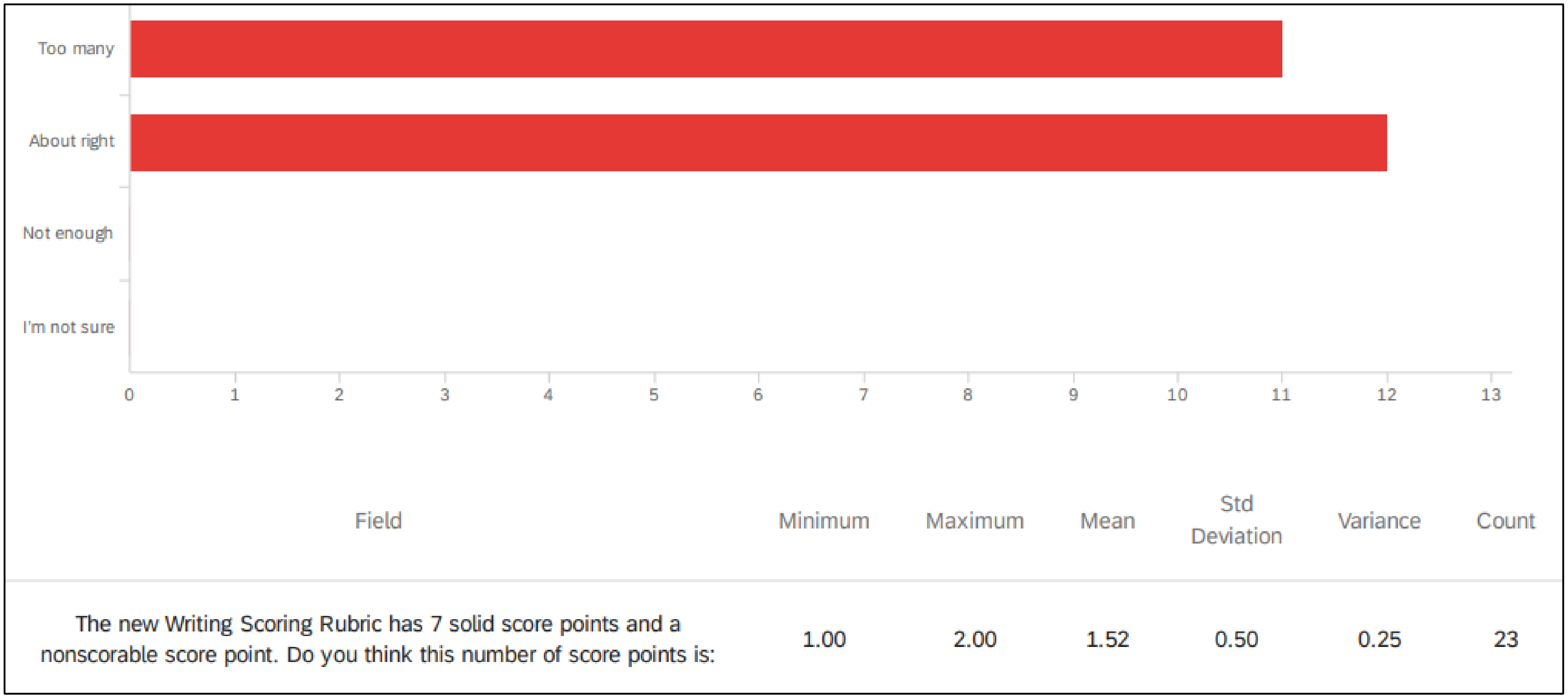
Respondents' annual volume of WIDA Screener Writing scoring



Field	Minimum	Maximum	Mean	Std Deviation	Variance	Count
Approximately how many students' writing responses on WIDA Screener do you score in one year?	1.00	4.00	2.13	0.99	0.98	23

Figure 3

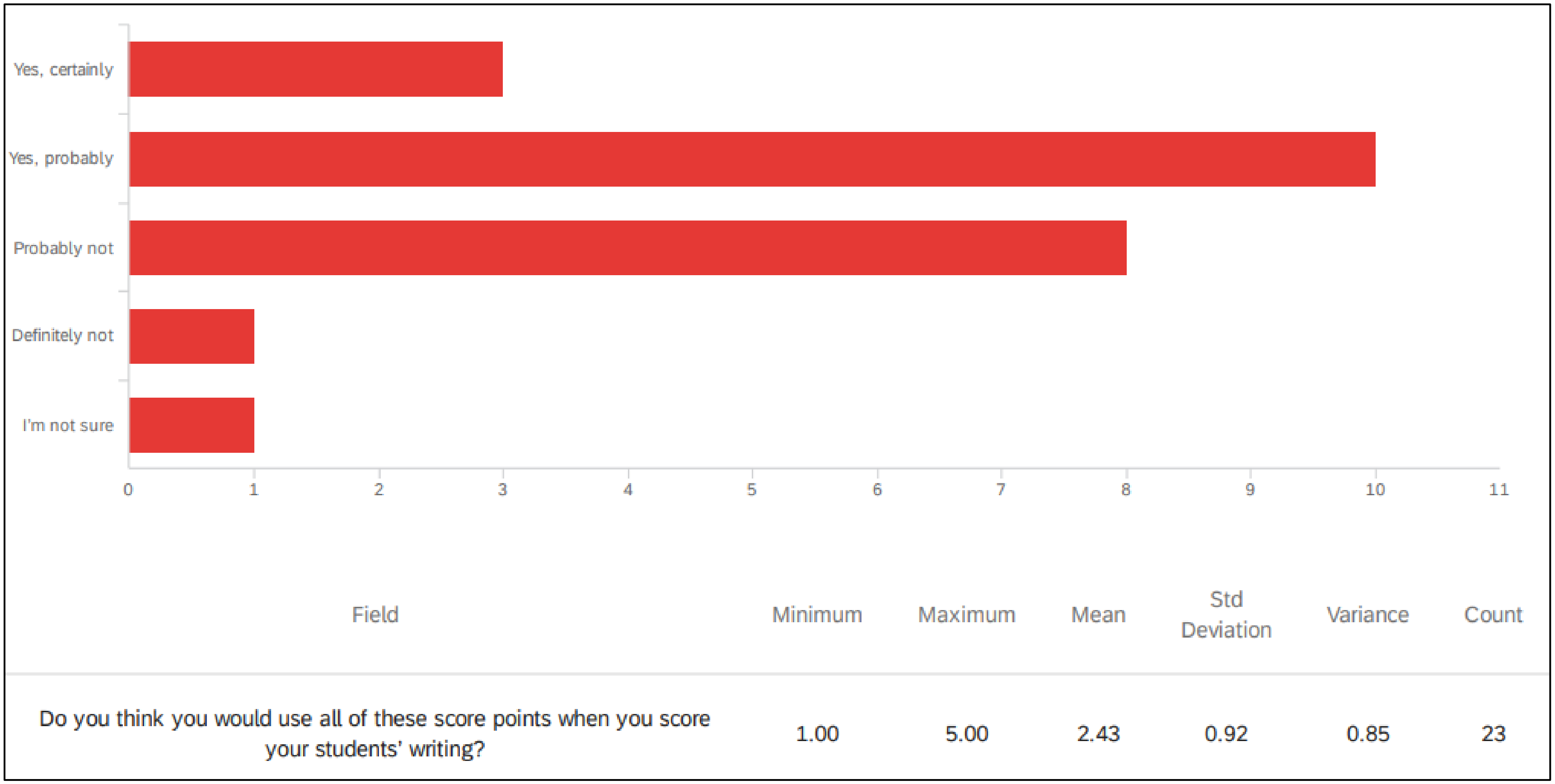
Opinion of number of new score points



Field	Minimum	Maximum	Mean	Std Deviation	Variance	Count
The new Writing Scoring Rubric has 7 solid score points and a nonscorable score point. Do you think this number of score points is:	1.00	2.00	1.52	0.50	0.25	23

Figure 4

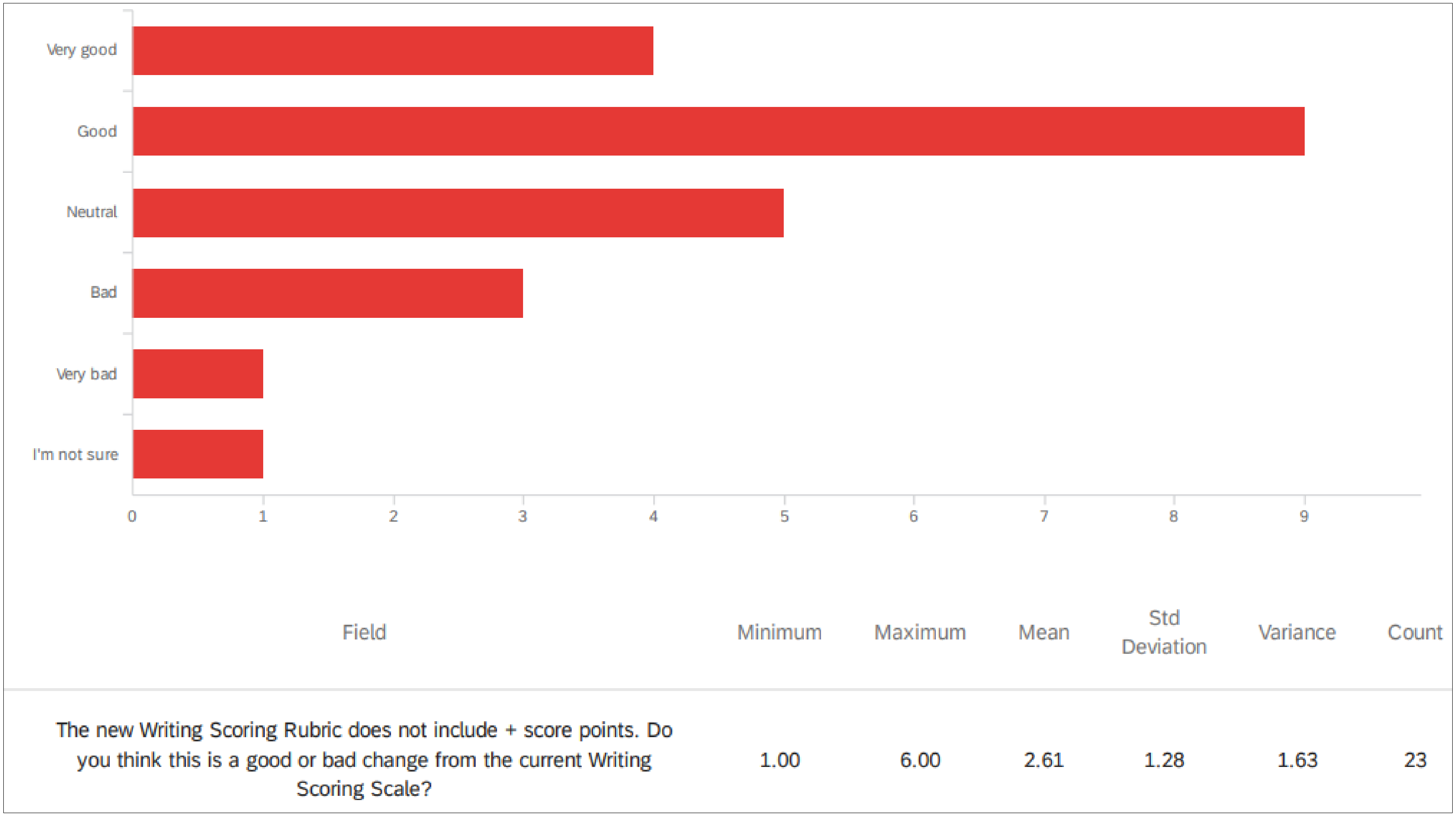
Opinion of likely use of new score points



Field	Minimum	Maximum	Mean	Std Deviation	Variance	Count
Do you think you would use all of these score points when you score your students' writing?	1.00	5.00	2.43	0.92	0.85	23

Figure 5

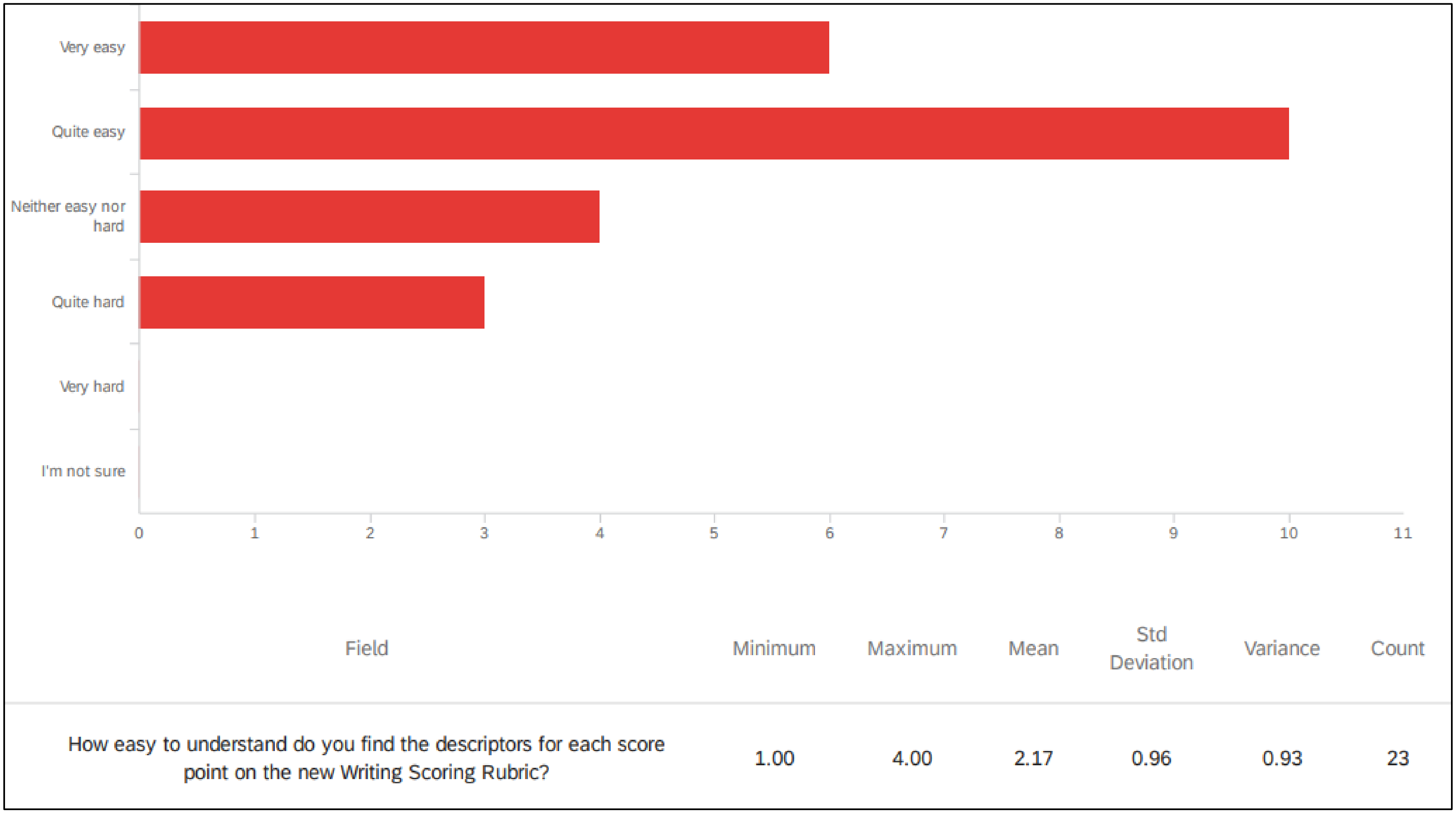
Opinion of removal of plus score points



Field	Minimum	Maximum	Mean	Std Deviation	Variance	Count
The new Writing Scoring Rubric does not include + score points. Do you think this is a good or bad change from the current Writing Scoring Scale?	1.00	6.00	2.61	1.28	1.63	23

Figure 6

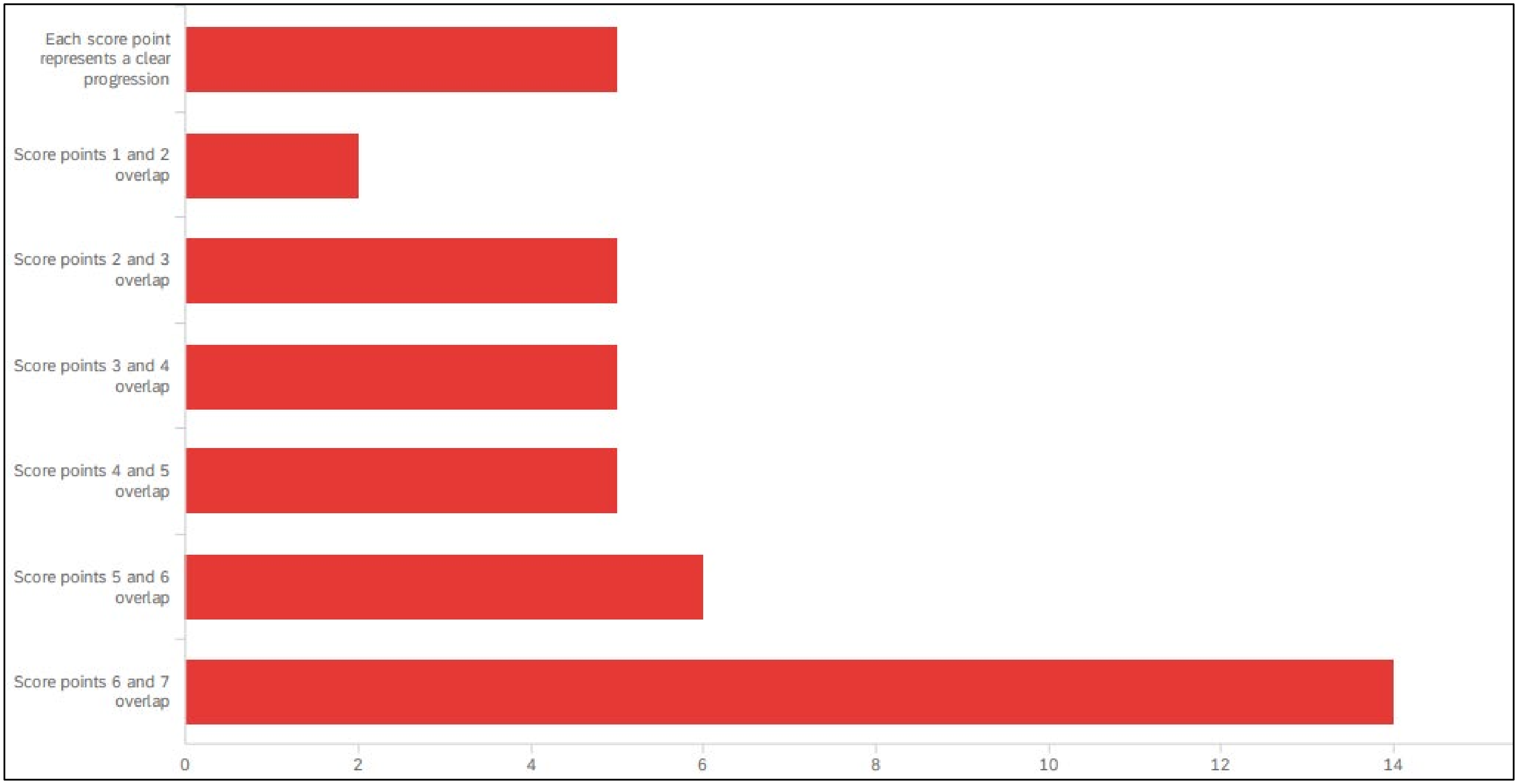
Ease of understanding of new rubric



Field	Minimum	Maximum	Mean	Std Deviation	Variance	Count
How easy to understand do you find the descriptors for each score point on the new Writing Scoring Rubric?	1.00	4.00	2.17	0.96	0.93	23

Figure 7

New score points that are challenging to distinguish

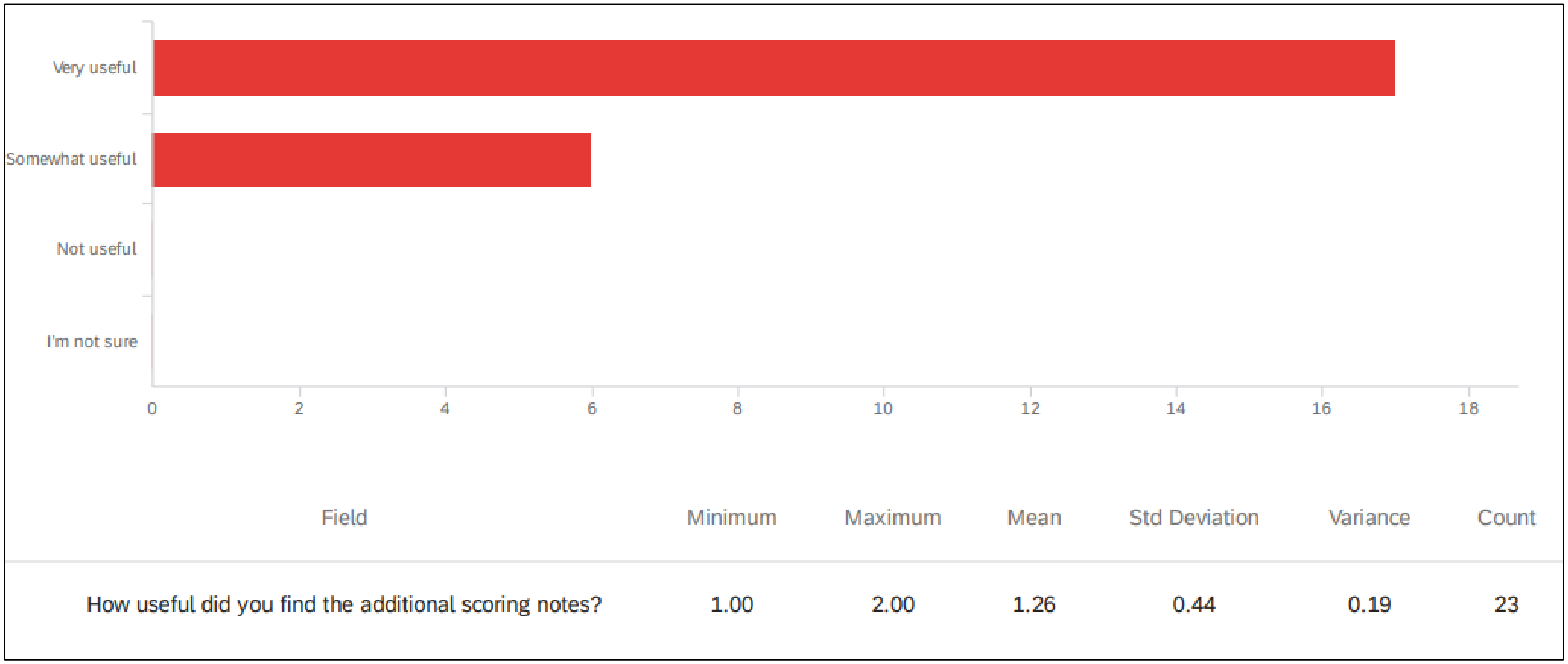


Survey respondents were then asked to elaborate on reasons why they had difficulty distinguishing between specific score points, with selected responses being shared here.

> *I indicated the scoring elements I struggled with. As expected, they were in the middle. I think one of the most difficult was when there was a single compound sentence and calling that a 4, especially because mechanics are not a factor in the scoring.*
>
> *I feel like students hit these two score points midway, so it's hard to know what exactly is the bump that gets a student from a 6 to a 7.*
>
> *On some I had a hard time determining between a score of 2 or 3 and then 6 or 7. I could see points being made for either score.*
>
> *The difference between a 6 and 7 does not seem to be significant enough or evident enough to warrant having 7 score points in addition to the nonscorable.*
>
> *No difficulty understanding but the scoring notes, glossary and connector descriptions were very helpful.*

Figure 8

Usefulness of scoring notes



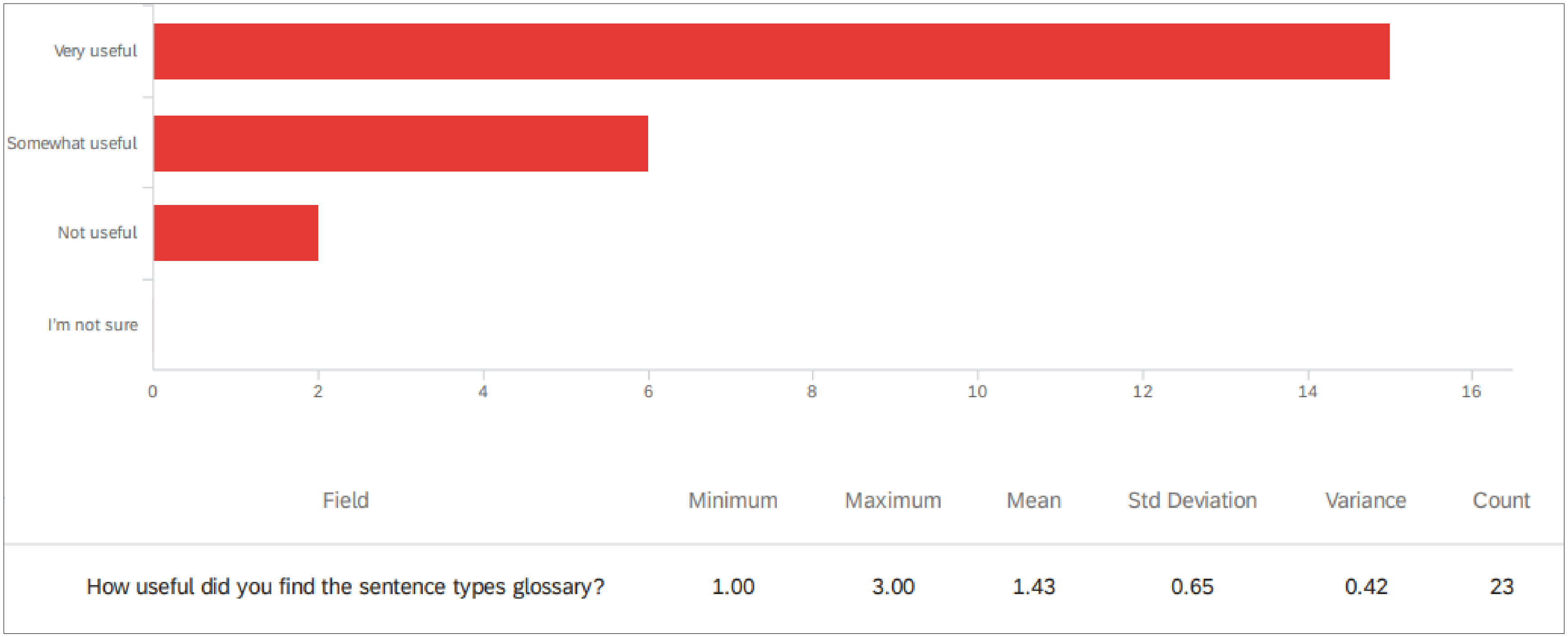
Field	Minimum	Maximum	Mean	Std Deviation	Variance	Count
How useful did you find the additional scoring notes?	1.00	2.00	1.26	0.44	0.19	23

Survey respondents were then asked to suggest any other scoring notes that should be provided, with selected responses being shared here.

> *I think that the rubric is very complex and a content area teacher who is not an English teacher will struggle with it. They are not used to identifying different sentence types, dependent and independent clauses, etc. It may not be content they are comfortable with, so they may not be comfortable evaluating it. An example of this would be my colleague who teaches the EL science courses.*
>
> *Perhaps some clarification on the distinction between language drawn from the stimulus vs language beyond the stimulus. Clarifying this could include examples, does the stimulus include audio or just written language, etc. I also thought the "Languages other than English" section is very much needed and appreciated its nuance.*
>
> *Yes, I think in addition to connectors, there needs to be some statements that indicate how clear the progression of ideas are. Just because the student uses more sophisticated connectors doesn't mean that their ideas flow smoothly. Does that make sense?*
>
> *Length of the response and how accurately a student answers the question. Sometimes students use the correct language and their response is related to the prompt but they didn't actually answer correctly.*

Figure 9

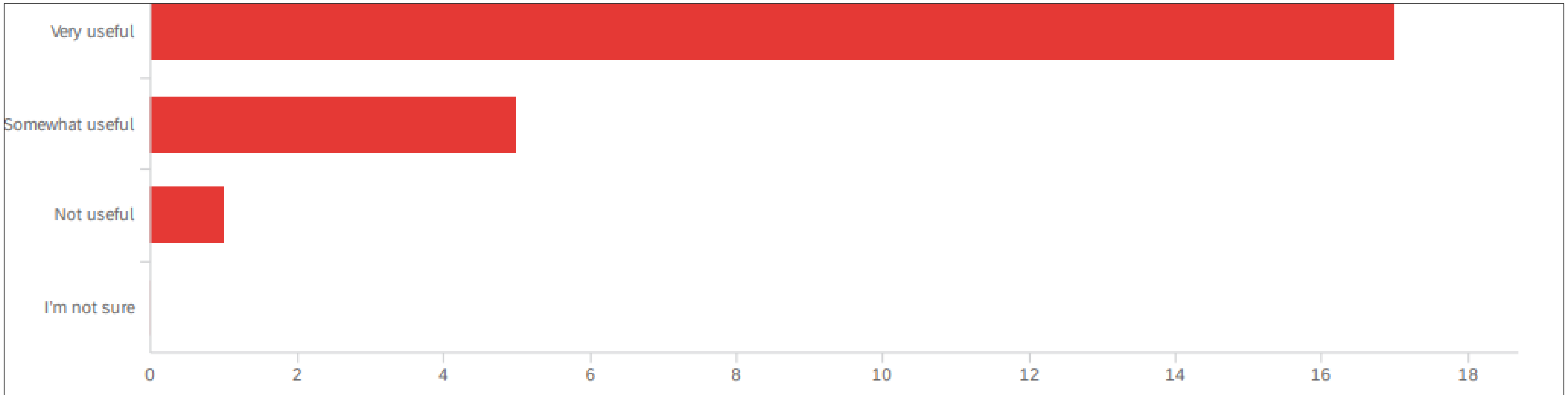
Usefulness of sentence types glossary



Field	Minimum	Maximum	Mean	Std Deviation	Variance	Count
How useful did you find the sentence types glossary?	1.00	3.00	1.43	0.65	0.42	23

Figure 10

Usefulness of connectors glossary



Field	Minimum	Maximum	Mean	Std Deviation	Variance	Count
How useful did you find the connectors glossary?	1.00	3.00	1.30	0.55	0.30	23

Survey respondents were then asked whether they had any other feedback on the new rubric that had not yet been captured, with selected responses being shared here.

> *I think it is a really good idea to do away with the "+." I think it created both confusion and more inaccuracies in scoring since it allowed for looser interpretations.*
>
> *I think this is a much stronger rubric that will help educators discern between genuine language proficiency areas of need and those who are simply just struggling writers, but who have a generally strong control of the English language itself.*
>
> *I like this rubric a lot better! And it helped that you provided multiple anchors with rationale, plus the glossary. Thank you!*
>
> *It looks more specific for discourse and sentence levels, but too generic for the word level criteria. Since we are in academic setting, use of academic vocabulary should be acknowledged and scored at a higher level.*
>
> *It took some time to get used to the 7 point scale (without plus and minus), but I do like it. I would like more clarity about the qualifications of each level, especially for levels 6 and 7.*

In general, responses to the educator survey about the new writing scoring rubric were positive. The feedback from these educators offered support for the removal of the "+" score points and indicated that the descriptor wording in the new rubric is understandable for educators. The additions to the expanded glossary were warmly received by educators.

After analyzing the educator feedback, the project team determined that descriptors for score points 6 and 7 needed to be revised. The revisions focused on more clearly delineating between the two score points without raising the bar for responses to be awarded a score of 7.

Phase 4: Rubric validation with DRC raters

After the new writing scoring rubric was revised in line with feedback provided by the reviewers described above, the new rubric was provided to DRC, whose trained raters score student responses to the ACCESS for ELLs Writing test. DRC raters were asked to score sets of responses from two or three tasks, depending on the grade-level cluster. Within each grade-level cluster all DRC raters scored all responses.

The following requests were made to DRC:

- Number of tasks = 12
- Number of responses = 100 per task

- These responses should span the operational (OP) score distribution, per tier and grade-level cluster
- Number of raters per task = 10
- Within a grade-level cluster all raters score all responses

These requests were made to DRC so the resulting dataset of scores awarded by DRC raters could be used for multi-faceted Rasch analyses to investigate the technical qualities of the new scoring rubric. To mirror the approach taken during operational scoring, DRC raters scored only responses from a single grade-level cluster. Raters were not asked to score responses from grade-level clusters with which they were unfamiliar.

Table 7 shows the distribution of writing tasks provided to DRC raters by grade-level cluster and tier.

Table 7
Task distribution by grade-level cluster and tier

Grade-level cluster	Tier A	Tier B/C	Total
1	2	1	3
2–3	1	1	2
4–5	1	1	2
6–8	1	1	2
9–12	1	2	3

DRC raters were provided with sets of training materials specific to each grade-level cluster. These materials consisted of anchor responses and 10 training samples for each task, which provided an opportunity for DRC raters to practice applying the new rubric. DRC raters then completed the scoring of the responses to tasks within their assigned grade-level cluster. Following the completion of scoring, WIDA and CAL project members held a debrief call with DRC senior raters to gather feedback on DRC raters' experiences of scoring using the new rubric.

The dataset of DRC rater scores was provided to WIDA and analyzed using multi-faceted Rasch analyses (mfra). These analyses allowed the project team to confirm whether the hypotheses related to the qualities of a well-functioning rubric were met or not. DRC raters also provided qualitative feedback to WIDA on the use of the new writing scoring rubric for operational scoring.

Features of the new writing scoring rubric

The writing scoring rubric underwent multiple rounds of review and revisions via the processes described in the previous sections. Some of the major decisions made based on the input from these reviews were:

- The new writing scoring rubric features eight score points (0-7). A majority of reviewers offered support for the 0–7 raw score range, though some reviewers reported that score points 6 and 7 were difficult to distinguish and should be consolidated. Descriptors for these score points were revised to make them more distinguishable.
    - For example, greater emphasis was placed on describing the extent to which responses demonstrated features of the intended key language uses (KLUs) and relevant content area.
- The plus score points (e.g., 4+) that were a feature of the WIDA Writing Scoring Scale Grades 1-12 are not included in the new WIDA Writing Scoring Rubric Grades 1-12. Reviewers, including internal WIDA reviewers, educators, and DRC reviewers, unanimously supported the removal of the plus score points in the new rubric. Reviewers commented that the shift away from using "+" in the score points would help make scoring more straightforward and may contribute to increased rater reliability.
- Score points 3 through 7 include three descriptors, one for each dimension of language encoded in the WIDA Standards. Score points 1 and 2 include one and two descriptors respectively, reflecting the observation that student responses at these score points tend largely to feature writing at the word/phrase (SP1) and sentence (SP2) dimensions. Discourse descriptors are typically not relevant to these responses.
- Educators requested that the new writing scoring rubric add more detail to the scoring notes and glossary sections. Guidance is now included on how to rate responses that include languages other than English in the rubric scoring notes for the first time.
- Reviewers consistently commented that the new scoring rubric is an improvement on the writing scoring scale, which will be easier to use operationally for both DRC raters and educators.

Rasch analyses of the DRC rater dataset

The aim of conducting Rasch analyses on the dataset of DRC rater scores was to investigate the technical quality of the new rubric, specifically the variation in ratings, rater separation, rater reliability, and student discrimination, as articulated in the research hypotheses for this project.

We selected 1,200 student responses from grades 1 to 12 (i.e., five grade-level clusters) for this component of the study. There were 12 writing tasks involved, each comprising 100 responses. The selection of these responses aimed to replicate the variety of student performances in operational testing. This meant sampling student responses from across all of the raw score points. The project team decided to oversample at the higher score points because of the project aim to create score descriptors that reflected the ceiling of student performances on the test, but not to include descriptors that described aspirational performance expectations, as seen on the writing scoring scale. Each grade-level cluster consisted of a unique set of tasks, students, and raters (see Table 8 for the data distribution). Each of the student responses within each grade-level cluster was written by a unique student. The responses were rated by 51 DRC raters using the newly developed rubric. The original plan was to have all raters within a grade-level cluster score all responses. Although this goal was not completely attained, all responses within a grade-level cluster were scored by most of the DRC raters. The final dataset contained 12,040 observations.

Table 8

Data distribution by grade-level cluster, task, and tier

Grade-level cluster	Tasks (Tier A + Tier B/C)	Responses	Raters	Number of scores assigned
1	3 (2+1)	300	11	3,220
2–3	2 (1+1)	200	10	1,920
4–5	2 (1+1)	200	10	1,900
6–8	2 (1+1)	200	9	1,700
9–12	3 (1+2)	300	11	3,300
Total	12 (6 + 6)	1,200	51	12,040

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 because there were five grade-level clusters. To understand how the scoring rubric 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 rubric category. These included student ability estimates and separation indices, allowing us to confirm if the scoring rubric 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 scoring rubric. Then, to understand how individual writing tasks work with the rubric, we fitted separate three-facet partial credit models (PCM): student, rater, and item. This allowed us to retrieve unexpected responses to check 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.

Results

Overview of descriptive statistics and model fit: Multi-facet Rasch analyses (MFRA) findings

Descriptive statistics of all scores assigned by DRC raters are shown in Table 9 and Figures 11–20. Table 9 presents the mean, standard deviation (*SD*), minimum (Min), and maximum (Max) for each writing task in each grade-level cluster. Figures 11–15 show the distribution of score points for each grade-level cluster. 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 16–20 show the distribution of score points with the current writing scoring scale for the set of responses scored by DRC in this study.

Table 9

Descriptive statistics of assigned scores

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
1	All tasks	A + B/C	**2.47 (1.61)**	**0 – 7**
2–3	Garden Surprise	A	2.58 (1.62)	0 – 7
2–3	Changing Water	B/C	3.48 (1.39)	0 – 7
2–3	Both tasks	A + B/C	**3.02 (1.58)**	**0 – 7**
4–5	Marsh Ecosystem	A	2.10 (1.29)	0 – 5
4–5	Search for Info	B/C	3.60 (1.59)	0 – 7
4–5	Both tasks	A + B/C	**2.81 (1.62)**	**0 – 7**
6–8	Illustrator	A	2.97 (1.36)	0 – 6
6–8	Color and Temperature	B/C	4.09 (1.66)	0 – 7
6–8	Both tasks	A + B/C	**3.50 (1.61)**	**0 – 7**
9–12	Where to Volunteer	A	2.82 (1.34)	0 – 6
9–12	Cherry Trees	B/C	4.54 (1.52)	0 – 7
9–12	Elasticity Investigation	B/C	4.33 (1.49)	0 –7
9–12	All tasks	A + B/C	**3.90 (1.64)**	**0 – 7**

Figure 11

Distribution of score points with the new scoring rubric for grade 1

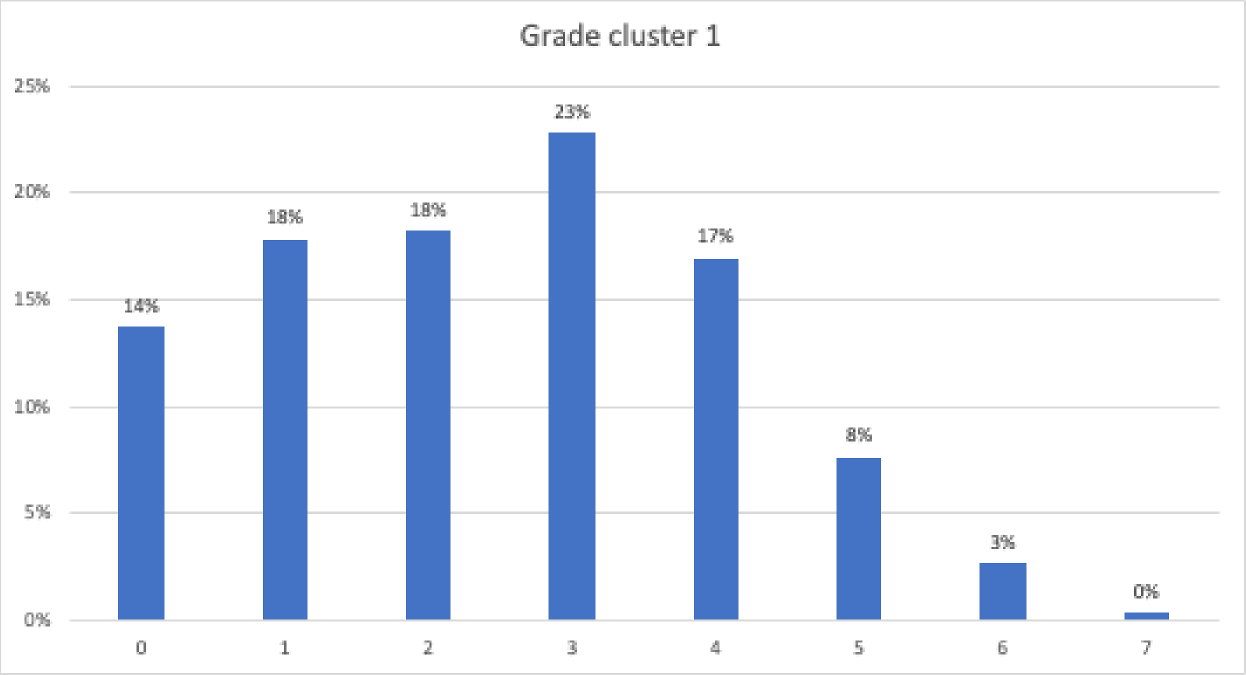


Figure 12

Distribution of score points with the new scoring rubric for grades 2–3

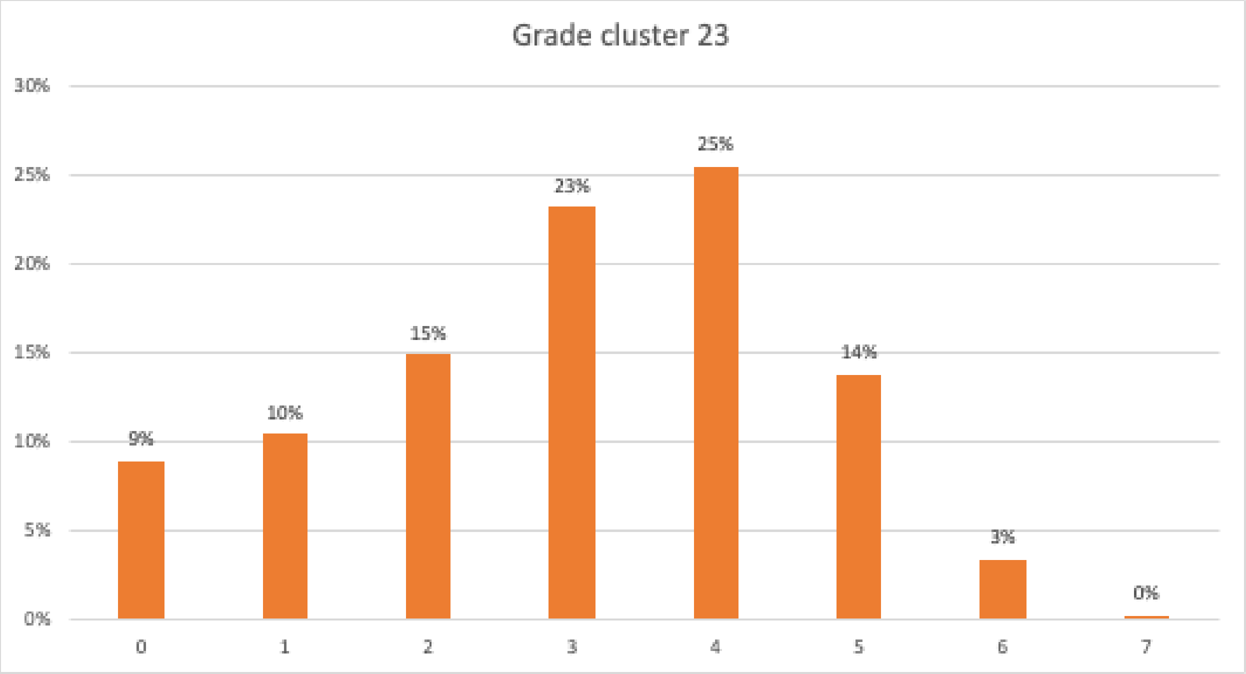


Figure 13

Distribution of score points with the new scoring rubric for grades 4–5

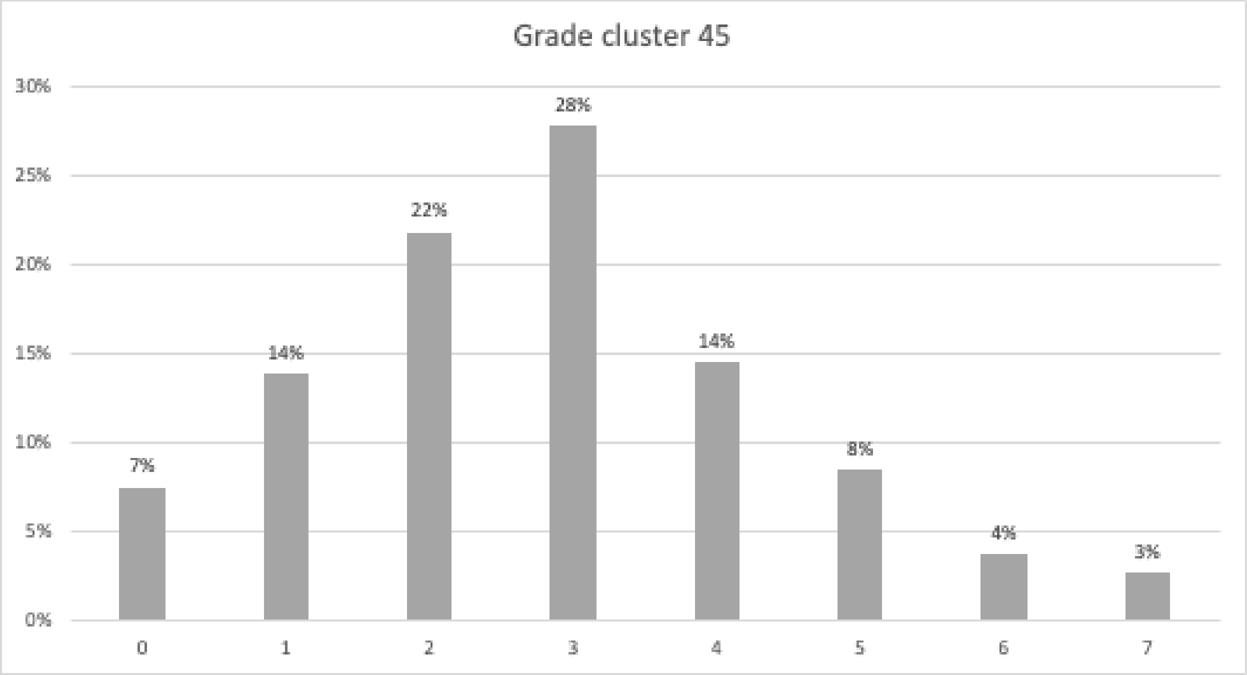


Figure 14

Distribution of score points with the new scoring rubric for grades 6–8

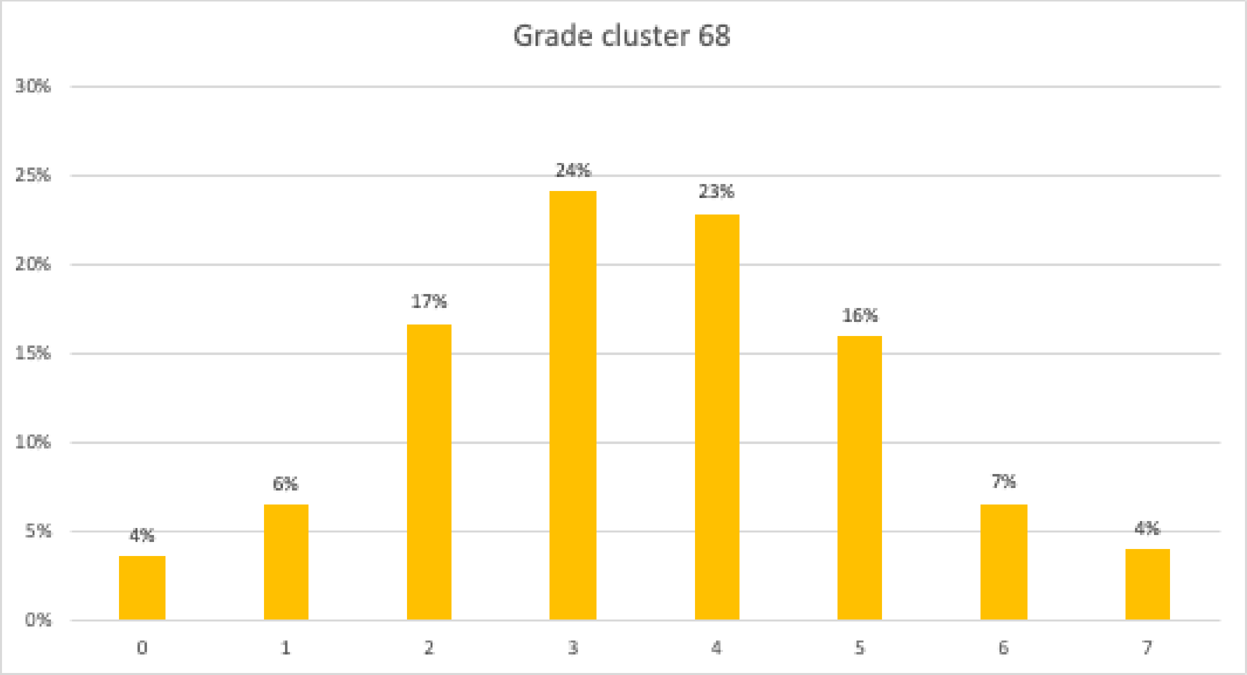


Figure 15

Distribution of score points with the new scoring rubric for grades 9–12

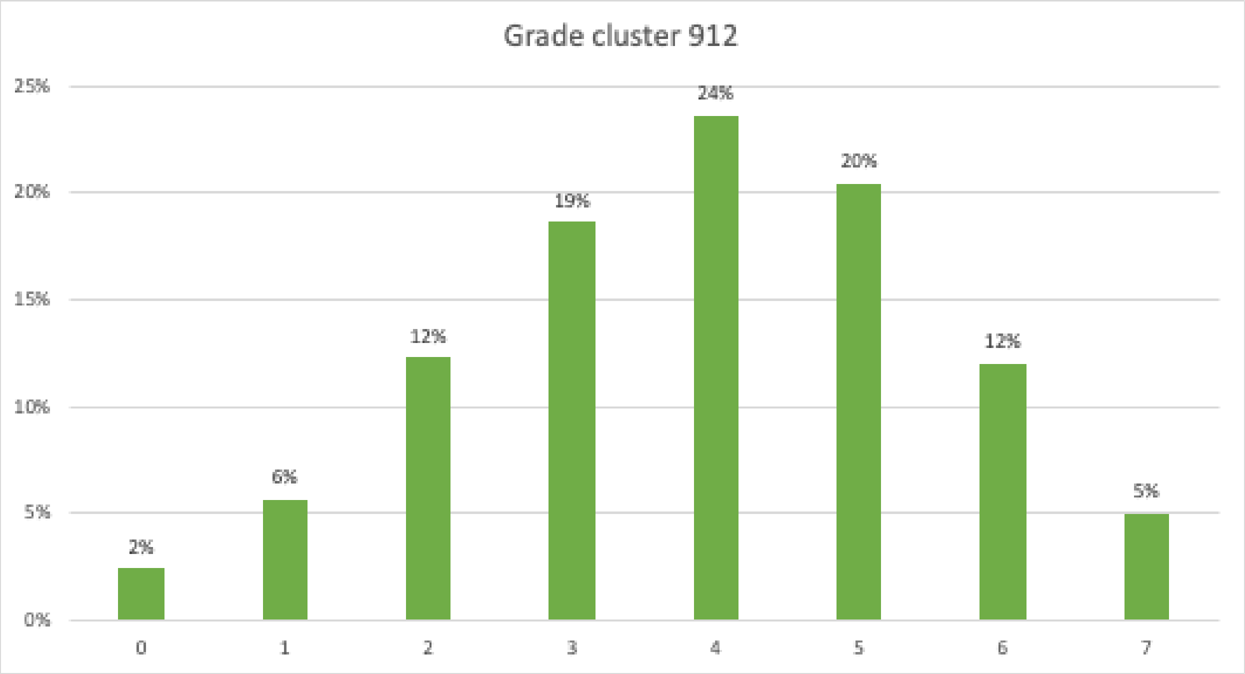


Figure 16

Distribution of score points with the current scoring scale for grade 1

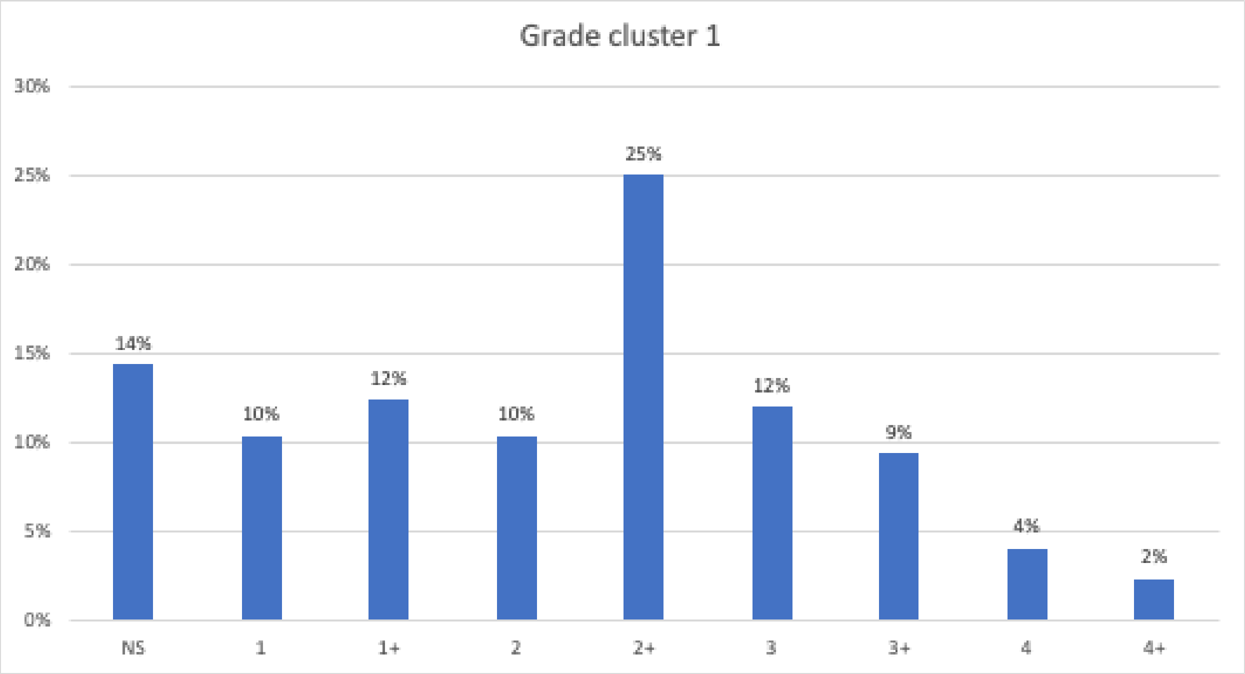


Figure 17

Distribution of score points with the current scoring scale for grades 2–3

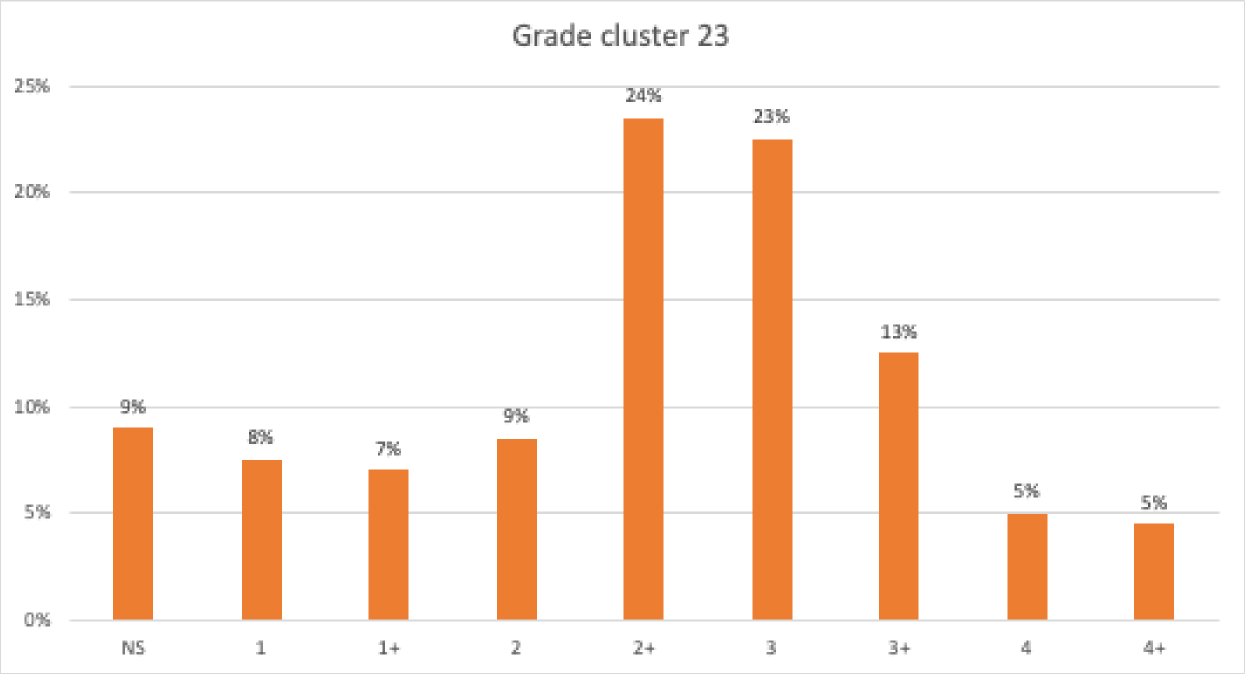


Figure 18

Distribution of score points with the current scoring scale for grades 4–5

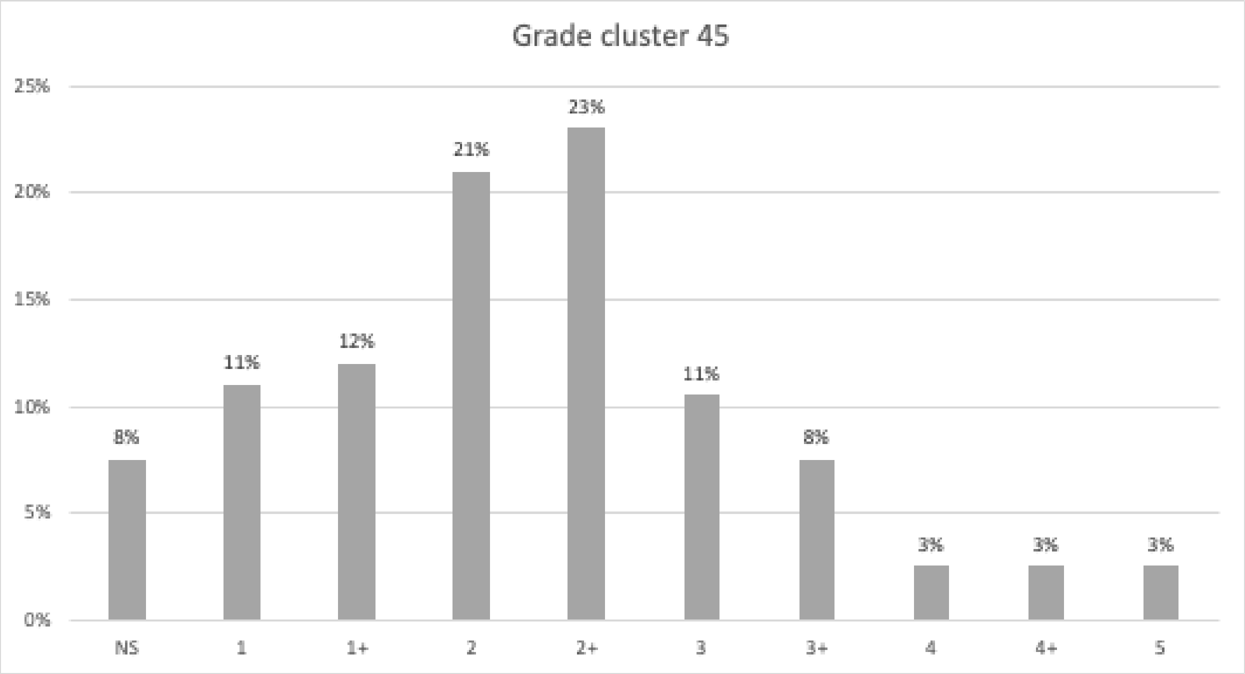


Figure 19

Distribution of score points with the current scoring scale for grades 6–8

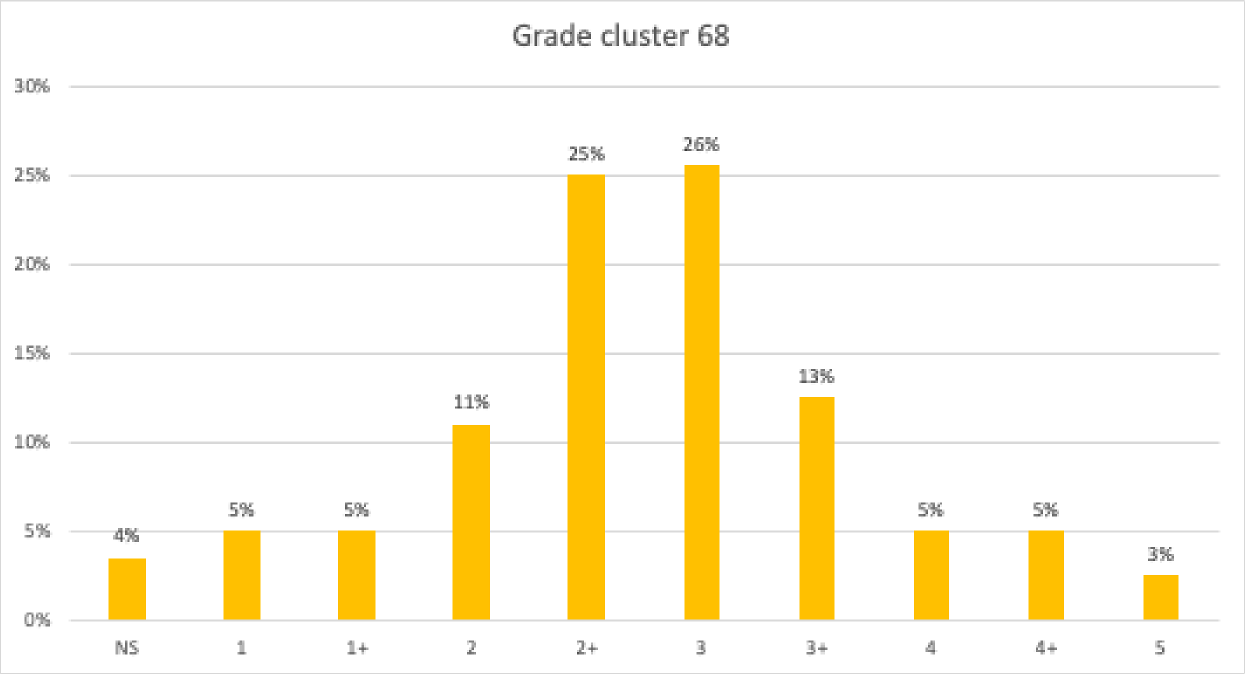
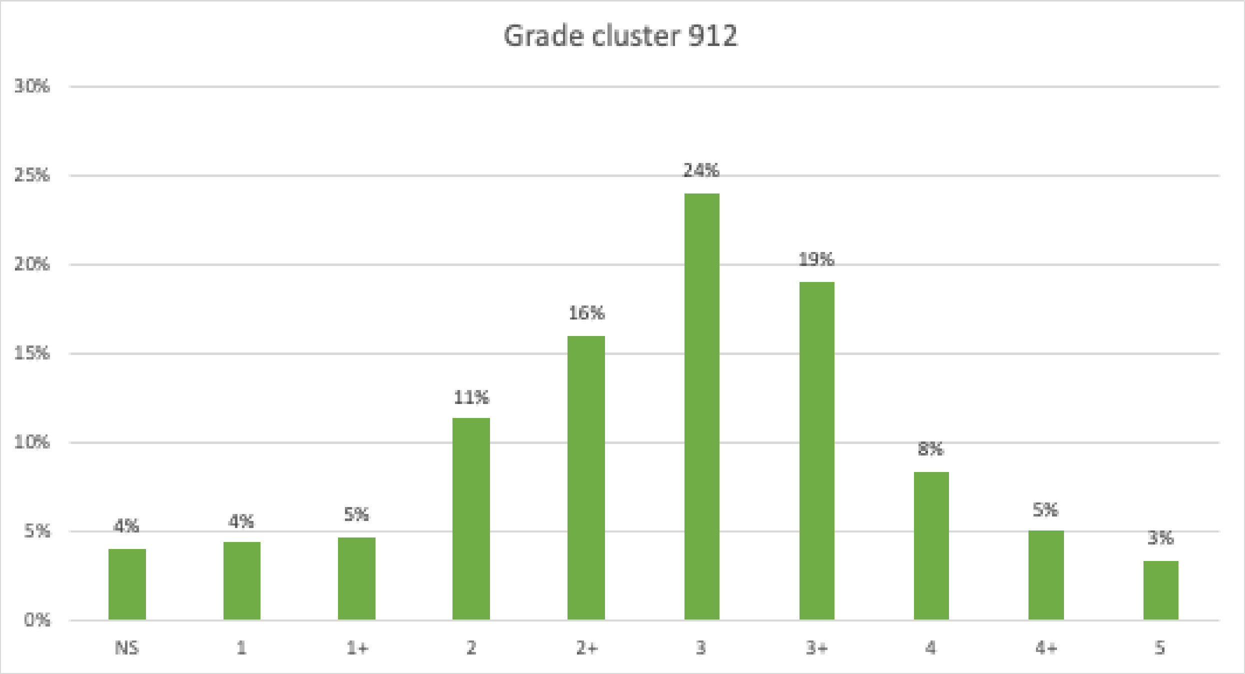


Figure 20

Distribution of score points with the current scoring scale for grades 9–12



Five rating scale models (i.e., grade-level clusters 1, 2–3, 4–5, 6–8, and 9–12) 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.

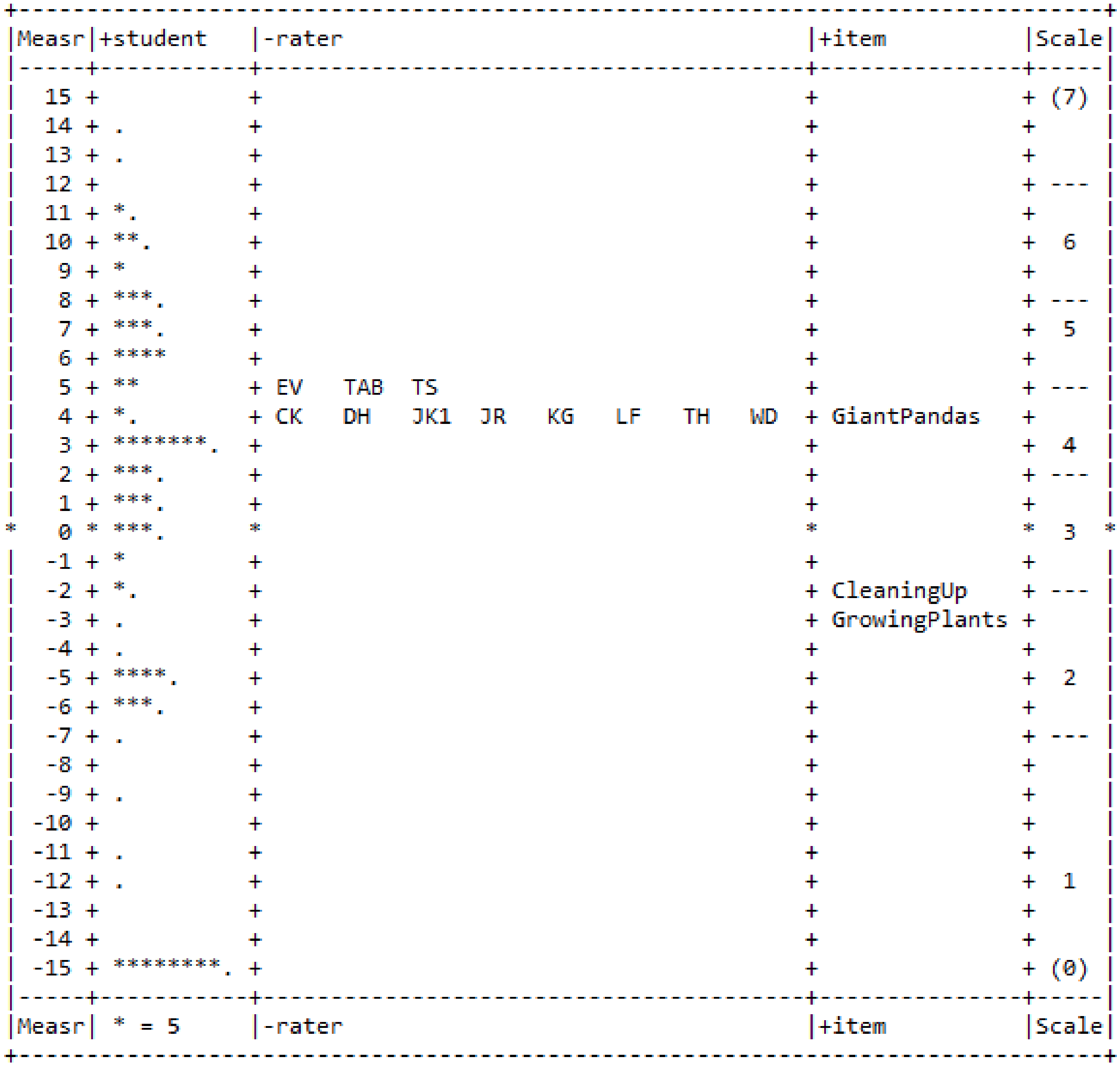
In this report, we present only a subset of the Rasch findings. A full report and discussion of the Rasch findings may be found in a separate report. Here, in the spirit of readability, we present the full Rasch findings for grade 1. Then, for the other grade-level clusters we present a shortened version of the Rasch findings that are most relevant to the research hypotheses listed earlier in this report.

Grade-level cluster 1 (RSM)

Figure 21 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 English language writing skills. The third column (*Rater*) includes all raters who participated in the scoring and displays their rating severity. The higher the measure is, the harsher the rating is. The fourth column (*Item*) indicates the difficulty level of each writing task, where 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 21

Wright map for grade-level cluster 1 RSM



```
+--------------------------------------------------------------------------------+
|Measr|+student   |-rater                                  |+item          |Scale|
|-----+-----------+----------------------------------------+---------------+-----|
|  15 +           +                                        +               + (7) |
|  14 + .         +                                        +               +     |
|  13 + .         +                                        +               +     |
|  12 +           +                                        +               + --- |
|  11 + *.        +                                        +               +     |
|  10 + **.       +                                        +               +  6  |
|   9 + *         +                                        +               +     |
|   8 + ***.      +                                        +               + --- |
|   7 + ***.      +                                        +               +  5  |
|   6 + ****      +                                        +               +     |
|   5 + **        + EV   TAB  TS                           +               + --- |
|   4 + *.        + CK   DH   JK1  JR   KG   LF   TH   WD  + GiantPandas   +     |
|   3 + *******.  +                                        +               +  4  |
|   2 + ***.      +                                        +               + --- |
|   1 + ***.      +                                        +               +     |
*   0 * ***.      *                                        *               *  3  *
|  -1 + *         +                                        +               +     |
|  -2 + *.        +                                        + CleaningUp    + --- |
|  -3 + .         +                                        + GrowingPlants +     |
|  -4 + .         +                                        +               +     |
|  -5 + ****.     +                                        +               +  2  |
|  -6 + ***.      +                                        +               +     |
|  -7 + .         +                                        +               + --- |
|  -8 +           +                                        +               +     |
|  -9 + .         +                                        +               +     |
| -10 +           +                                        +               +     |
| -11 + .         +                                        +               +     |
| -12 + .         +                                        +               +  1  |
| -13 +           +                                        +               +     |
| -14 +           +                                        +               +     |
| -15 + ********. +                                        +               + (0) |
|-----+-----------+----------------------------------------+---------------+-----|
|Measr| * = 5     |-rater                                  |+item          |Scale|
+--------------------------------------------------------------------------------+
```

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 10 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 10, the fit statistics were all smaller than 2 and the thresholds increased monotonically.

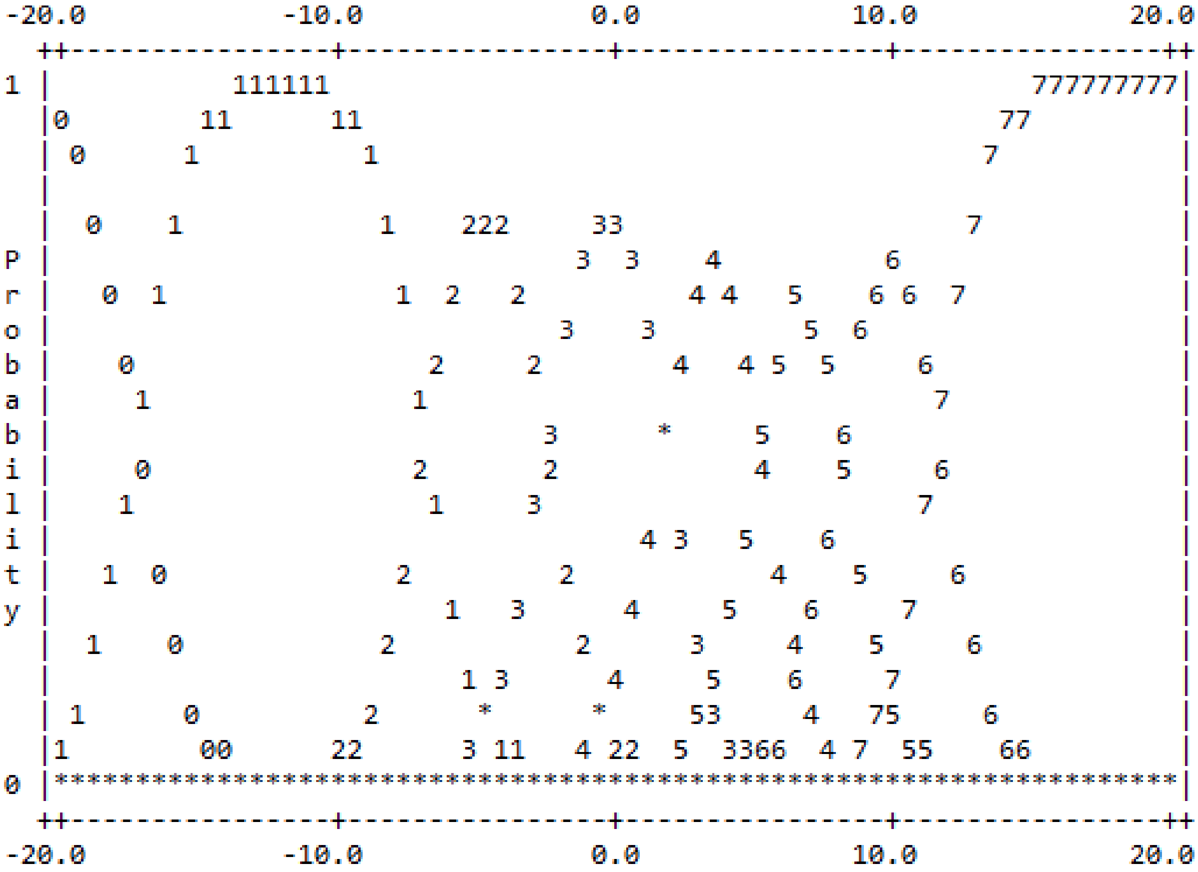
Table 10
Score distribution for grade-level cluster 1

Score level	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

Probability curves for the rating scale (see Figure 22) were also examined to assist the evaluation. 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.

Figure 22

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



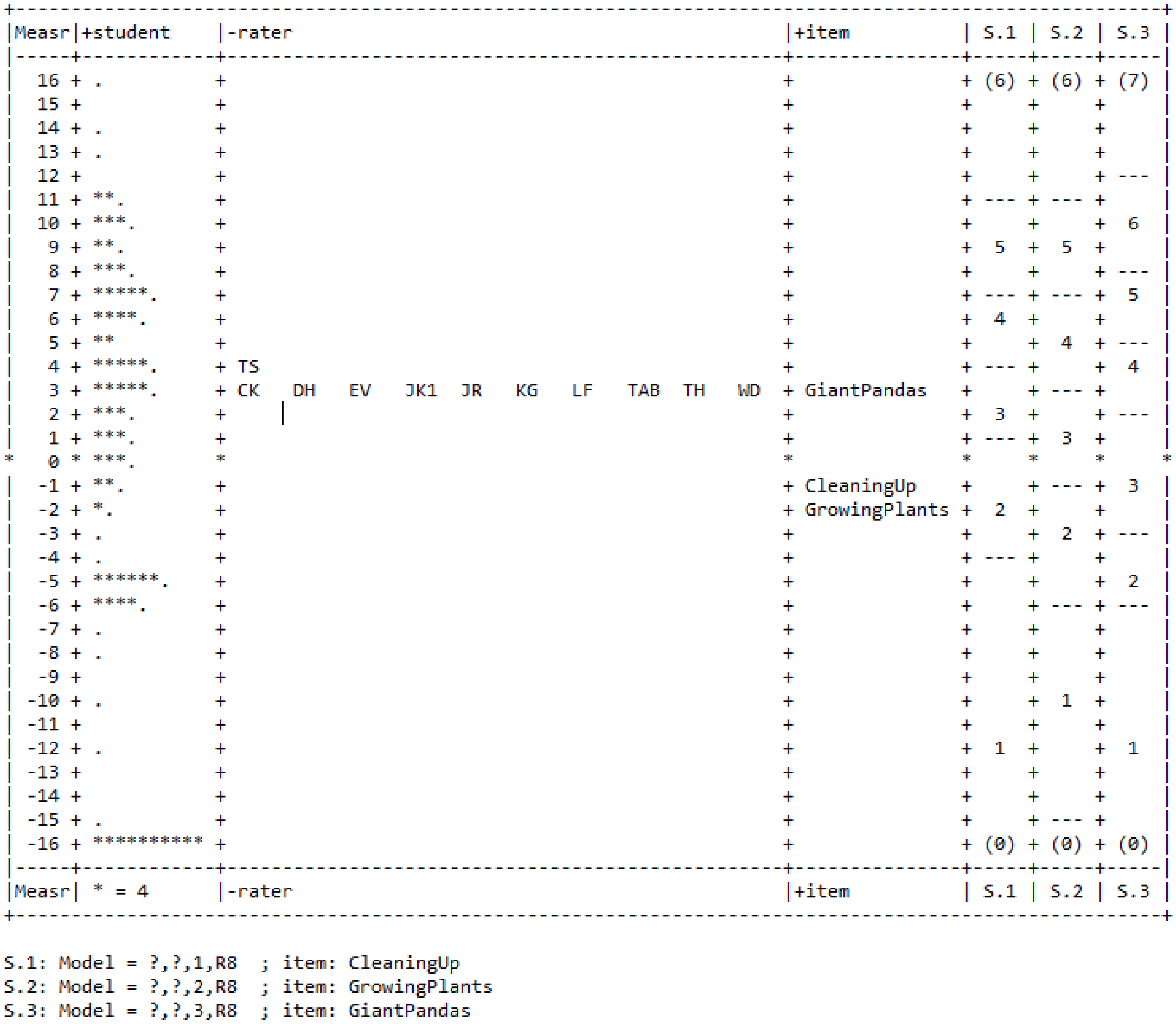
```
-20.0            -10.0              0.0              10.0              20.0
  ++----------------+----------------+----------------+----------------++
1 |           111111                                           777777777|
  |0        11      11                                       77         |
  | 0      1          1                                     7           |
  |                                                                     |
  |  0    1            1    222     33                     7            |
P |                                3  3    4          6                 |
r |   0  1              1  2   2          4 4   5    6 6  7             |
o |                               3    3         5  6                   |
b |    0                  2     2        4   4 5  5     6               |
a |     1                1                               7              |
b |                              3      *     5    6                    |
i |     0                2       2            4    5     6              |
l |    1                  1     3                       7               |
i |                                    4 3   5    6                     |
t |   1  0              2         2            4    5     6             |
y |                        1   3      4     5    6     7                |
  |  1    0            2           2      3     4    5     6            |
  |                         1 3      4     5    6     7                 |
  | 1      0          2      *      *     53     4   75     6           |
  |1        00      22      3 11   4 22  5  3366  4 7  55    66         |
0 |*********************************************************************|
  ++----------------+----------------+----------------+----------------++
-20.0            -10.0              0.0              10.0              20.0
```

Grade-level cluster 1 (PCM)

Figure 23 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 that 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.

Figure 23

Wright map for grade-level cluster 1 PCM



```
+-------------------------------------------------------------------------------------------------------+
|Measr|+student     |-rater                                            |+item          | S.1 | S.2 | S.3 |
|-----+------------+--------------------------------------------------+---------------+-----+-----+-----|
|  16 + .          +                                                  +               + (6) + (6) + (7) |
|  15 +            +                                                  +               +     +     +     |
|  14 + .          +                                                  +               +     +     +     |
|  13 + .          +                                                  +               +     +     +     |
|  12 +            +                                                  +               +     +     + --- |
|  11 + **.        +                                                  +               + --- + --- +     |
|  10 + ***.       +                                                  +               +     +     +  6  |
|   9 + **.        +                                                  +               +  5  +  5  +     |
|   8 + ***.       +                                                  +               +     +     + --- |
|   7 + *****.     +                                                  +               + --- + --- +  5  |
|   6 + ****.      +                                                  +               +  4  +     +     |
|   5 + **         +                                                  +               +     +  4  + --- |
|   4 + *****.     + TS                                               +               + --- +     +  4  |
|   3 + *****.     + CK   DH   EV   JK1  JR   KG   LF   TAB  TH   WD  + GiantPandas   +     + --- +     |
|   2 + ***.       +     |                                            +               +  3  +     + --- |
|   1 + ***.       +                                                  +               + --- +  3  +     |
*    0 * ***.       *                                                  *               *     *     *     *
|  -1 + **.        +                                                  + CleaningUp    +     + --- +  3  |
|  -2 + *.         +                                                  + GrowingPlants +  2  +     +     |
|  -3 + .          +                                                  +               +     +  2  + --- |
|  -4 + .          +                                                  +               + --- +     +     |
|  -5 + ******.    +                                                  +               +     +     +  2  |
|  -6 + ****.      +                                                  +               +     + --- + --- |
|  -7 + .          +                                                  +               +     +     +     |
|  -8 + .          +                                                  +               +     +     +     |
|  -9 +            +                                                  +               +     +     +     |
| -10 + .          +                                                  +               +     +  1  +     |
| -11 +            +                                                  +               +     +     +     |
| -12 + .          +                                                  +               +  1  +     +  1  |
| -13 +            +                                                  +               +     +     +     |
| -14 +            +                                                  +               +     +     +     |
| -15 + .          +                                                  +               +     + --- +     |
| -16 + ********** +                                                  +               + (0) + (0) + (0) |
|-----+------------+--------------------------------------------------+---------------+-----+-----+-----|
|Measr| * = 4      |-rater                                            |+item          | S.1 | S.2 | S.3 |
+-------------------------------------------------------------------------------------------------------+

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

Table 11 highlights cases with high discrepancies between expected and observed scores. Category statistics for individual tasks are in Table 12 and Figures 24 to 26 show their corresponding probability curves. Among all, the probability curves for *Giant Pandas* were perhaps the most unclear with unevenly spaced hills and slightly obscure peaks for some score levels.

Table 11

Unexpected responses for grade-level cluster 1

Writing task	Student – Rater	Expected rating	Observed rating
Cleaning Up	25 – DH	1.4	5
Cleaning Up	41 – DH	4.4	1
Cleaning Up	4 – LF	1.7	4
Cleaning Up	76 – JR	2.6	5
Growing Plants	186 – KG	1.6	6
Growing Plants	143 – JK1	0.4	3
Growing Plants	200 – KG	1.5	4
Giant Pandas	207 – CK	4.1	7
Giant Pandas	212 – LF	4.0	2
Giant Pandas	227 – LF	5.2	3

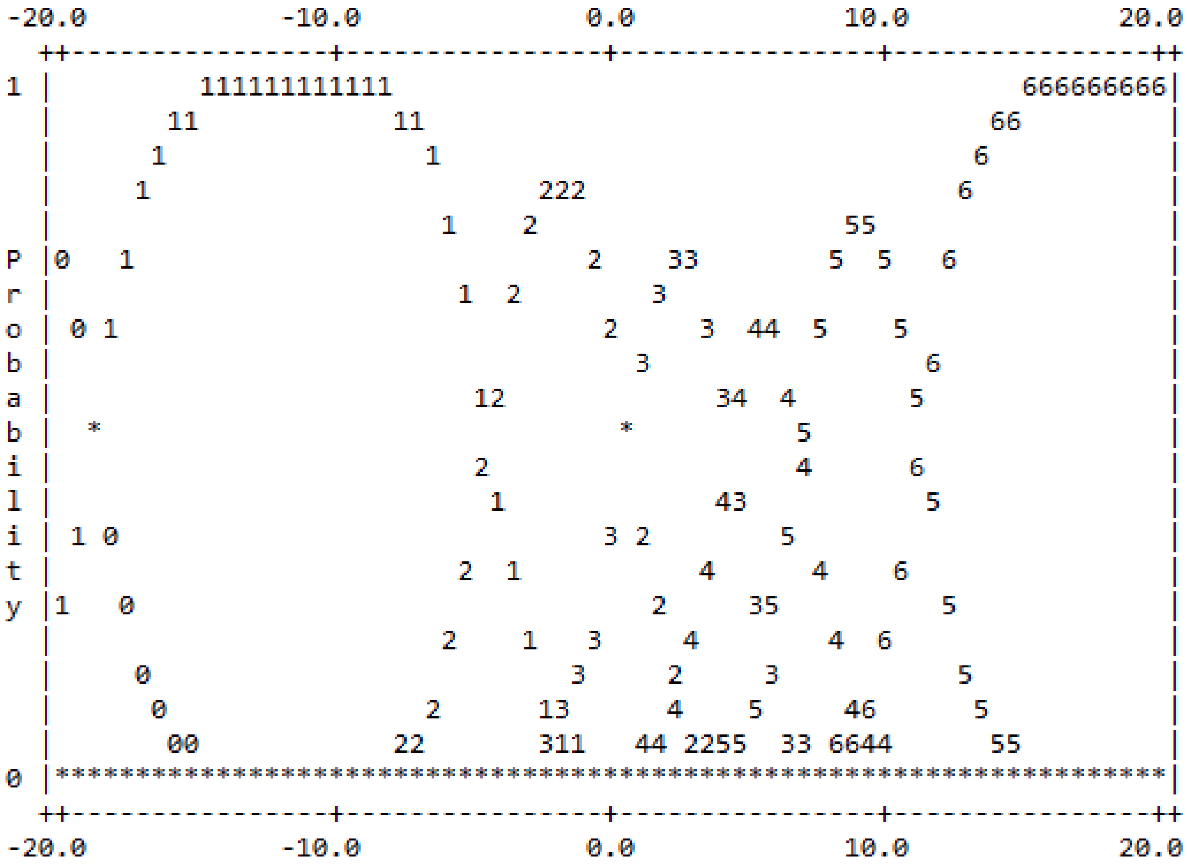
Table 12
Score distribution for individual writing tasks of grade-level cluster 1

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

Note: *Cleaning Up* and *Growing Plants* are Tier A; *Giant Pandas* is Tier B/C

Figure 24

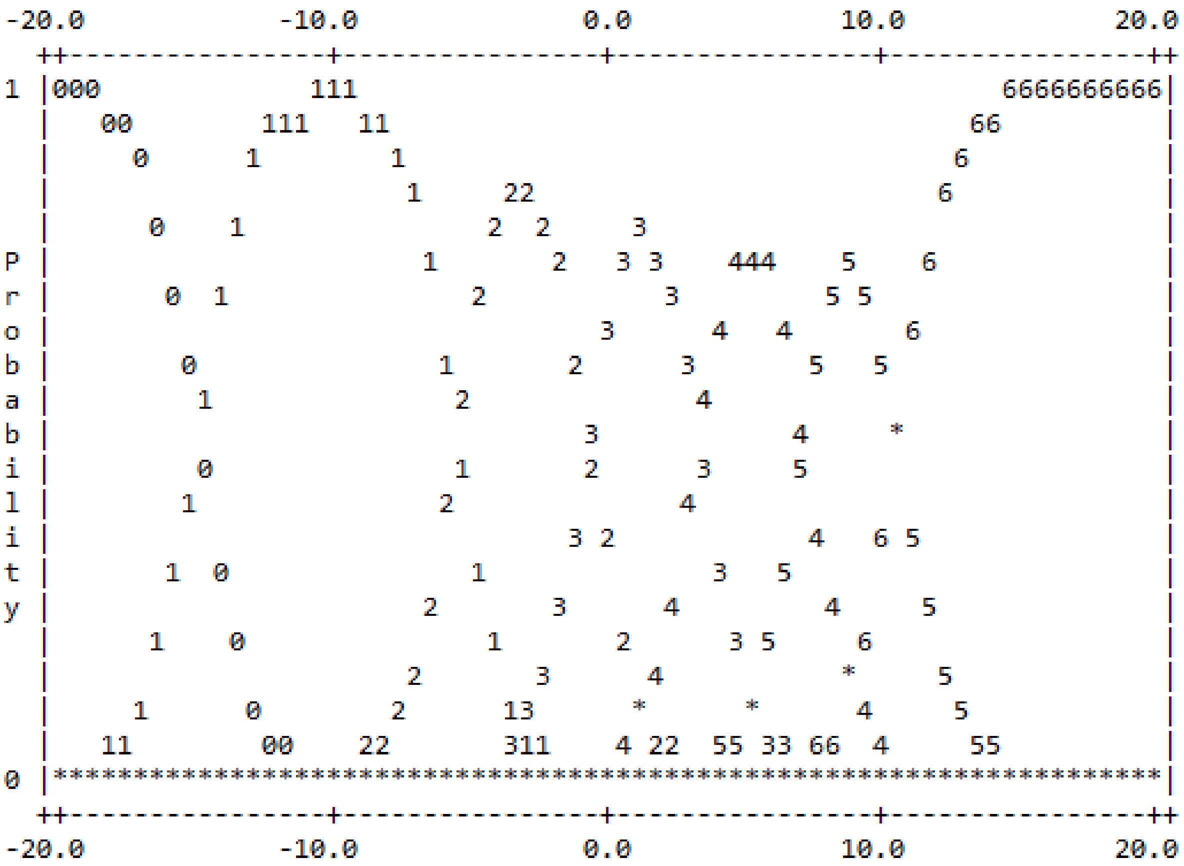
Probability curves for the rating scale (grade-level cluster 1; Tier A Cleaning Up)



```
-20.0            -10.0              0.0             10.0             20.0
  ++----------------+----------------+----------------+----------------++
1 |         111111111111                                       666666666|
  |       11            11                                   66         |
  |      1                1                                 6           |
  |     1                        222                       6            |
  |                        1    2                   55                  |
P |0   1                            2    33        5  5   6             |
r |                         1  2        3                               |
o | 0 1                              2     3  44  5    5                |
b |                                    3                 6              |
a |                          12             34  4       5               |
b |  *                                *          5                      |
i |                          2                   4      6               |
l |                           1             43           5              |
i | 1 0                              3 2        5                       |
t |                         2  1           4      4    6                |
y |1   0                                2     35          5             |
  |                        2    1   3     4        4  6                 |
  |     0                          3     2     3           5            |
  |      0                2      13      4    5     46      5           |
  |       00            22       311   44 2255  33 6644      55         |
0 |*********************************************************************|
  ++----------------+----------------+----------------+----------------++
-20.0            -10.0              0.0             10.0             20.0
```

Figure 25

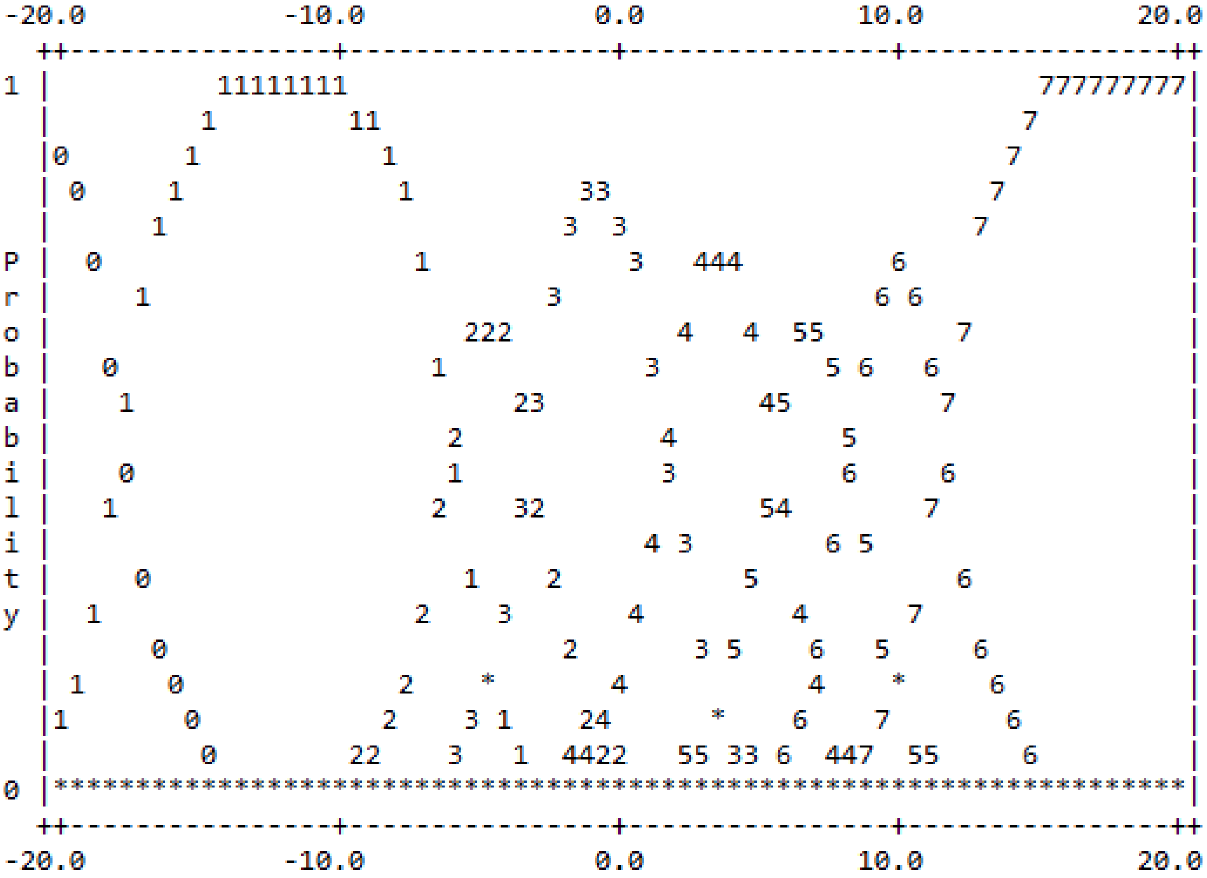
Probability curves for the rating scale (grade-level cluster 1; Tier A Growing Plants)



```
-20.0            -10.0             0.0             10.0             20.0
  ++----------------+---------------+----------------+----------------++
1 |000             111                                       6666666666|
  |   00        111   11                                   66          |
  |     0      1        1                                 6            |
  |                      1    22                         6             |
  |      0    1              2  2     3                                |
P |                       1      2   3 3    444    5    6              |
r |       0  1              2           3         5 5                  |
o |                                 3      4   4       6               |
b |        0               1      2      3       5   5                 |
a |         1              2              4                            |
b |                                3            4     *                |
i |         0               1      2      3     5                      |
l |        1              2              4                             |
i |                               3 2            4   6 5               |
t |       1  0               1             3   5                       |
y |                      2       3      4         4     5              |
  |      1    0               1      2      3 5     6                  |
  |                     2       3      4           *     5             |
  |     1      0       2      13      *      *      4     5            |
  |   11        00   22       311    4 22  55 33 66  4     55          |
0 |********************************************************************|
  ++----------------+---------------+----------------+----------------++
-20.0            -10.0             0.0             10.0             20.0
```

Figure 26

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



```
-20.0            -10.0              0.0             10.0             20.0
  ++----------------+----------------+----------------+----------------++
1 |          11111111                                          777777777|
  |         1        11                                       7         |
  |0       1           1                                     7          |
  | 0     1             1          33                       7           |
  |      1                        3  3                     7            |
P |  0                   1            3   444         6                 |
r |     1                        3                   6 6                |
o |                         222          4   4  55        7             |
b |   0                   1            3          5 6   6               |
a |    1                       23             45         7              |
b |                        2            4          5                    |
i |    0                   1            3          6     6              |
l |   1                   2    32             54        7               |
i |                                    4 3        6 5                   |
t |     0                   1    2           5            6             |
y |  1                   2    3       4         4      7                |
  |      0                        2       3 5    6   5     6            |
  | 1     0             2    *       4           4    *     6           |
  |1       0           2    3 1    24      *    6    7       6          |
  |         0        22    3   1  4422   55 33 6  447  55     6         |
0 |*********************************************************************|
  ++----------------+----------------+----------------+----------------++
-20.0            -10.0              0.0             10.0             20.0
```

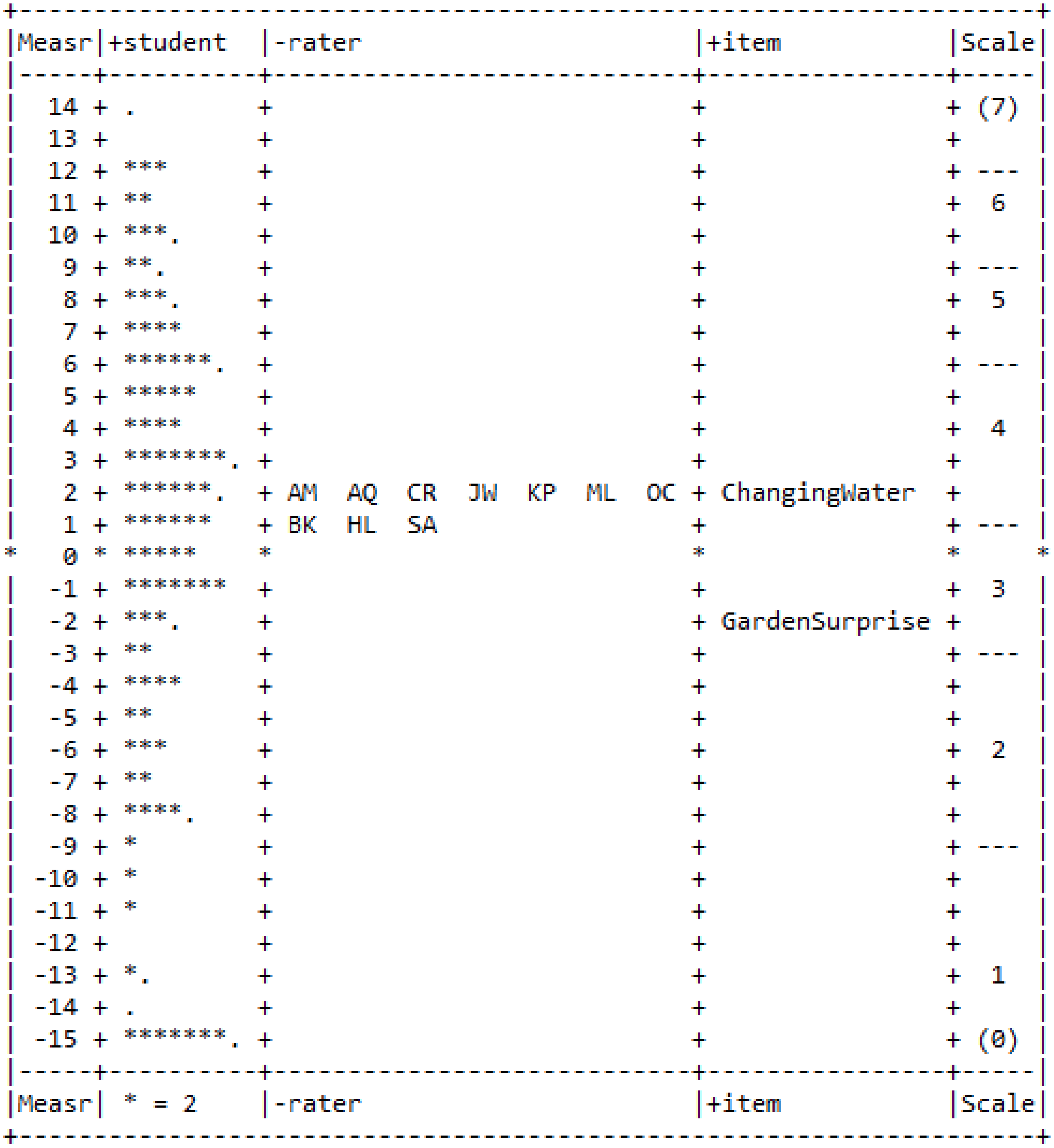
Grade-level cluster 2–3 (RSM)

Please note that for this and subsequent grade-level clusters, this report presents only a summary of the Rasch findings. The full Rasch findings are detailed in a separate report.

Figure 27 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). Therefore, 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 27

Wright map for grade-level cluster 2–3 RSM

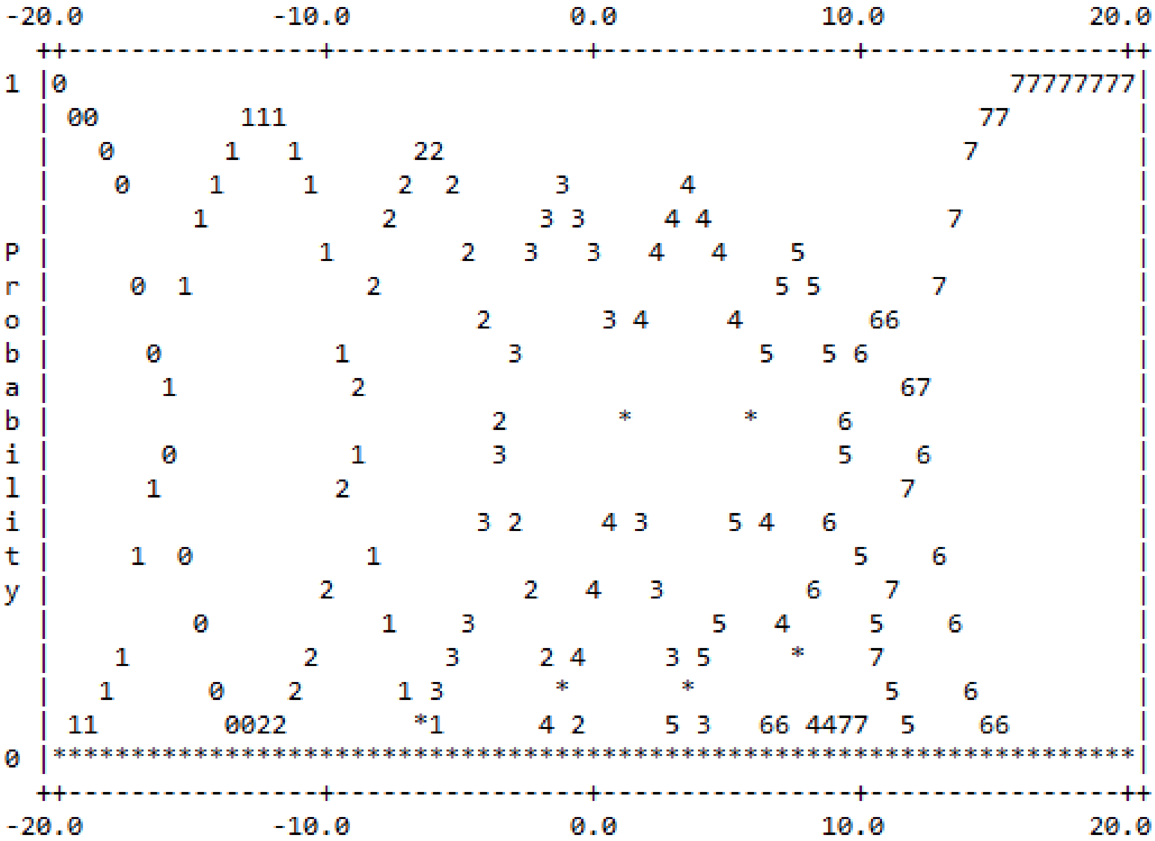


```
+-----------------------------------------------------------------------+
|Measr|+student   |-rater                        |+item            |Scale|
|-----+-----------+------------------------------+-----------------+-----|
|  14 + .         +                              +                 + (7) |
|  13 +           +                              +                 +     |
|  12 + ***       +                              +                 + --- |
|  11 + **        +                              +                 +  6  |
|  10 + ***.      +                              +                 +     |
|   9 + **.       +                              +                 + --- |
|   8 + ***.      +                              +                 +  5  |
|   7 + ****      +                              +                 +     |
|   6 + ******.   +                              +                 + --- |
|   5 + *****     +                              +                 +     |
|   4 + ****      +                              +                 +  4  |
|   3 + *******.  +                              +                 +     |
|   2 + ******.   + AM  AQ  CR  JW  KP  ML  OC   + ChangingWater   +     |
|   1 + ******    + BK  HL  SA                   +                 + --- |
*   0 * *****     *                              *                 *     *
|  -1 + *******   +                              +                 +  3  |
|  -2 + ***.      +                              + GardenSurprise  +     |
|  -3 + **        +                              +                 + --- |
|  -4 + ****      +                              +                 +     |
|  -5 + **        +                              +                 +     |
|  -6 + ***       +                              +                 +  2  |
|  -7 + **        +                              +                 +     |
|  -8 + ****.     +                              +                 +     |
|  -9 + *         +                              +                 + --- |
| -10 + *         +                              +                 +     |
| -11 + *         +                              +                 +     |
| -12 +           +                              +                 +     |
| -13 + *.        +                              +                 +  1  |
| -14 + .         +                              +                 +     |
| -15 + *******.  +                              +                 + (0) |
|-----+-----------+------------------------------+-----------------+-----|
|Measr| * = 2     |-rater                        |+item            |Scale|
+-----------------------------------------------------------------------+
```

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.

Figure 28

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



```
-20.0            -10.0              0.0             10.0             20.0
  ++----------------+----------------+----------------+----------------++
1 |0                                                            77777777|
  | 00         111                                            77        |
  |   0       1   1       22                                 7          |
  |    0     1     1     2  2      3       4                            |
  |         1           2         3 3     4 4               7           |
P |                 1        2   3   3   4   4    5                     |
r |     0  1           2                         5 5       7            |
o |                           2       3 4     4        66               |
b |      0           1          3               5   5 6                 |
a |       1           2                                  67             |
b |                            2       *       *     6                  |
i |       0           1        3                     5    6             |
l |      1           2                                   7              |
i |                           3 2     4 3     5 4   6                   |
t |     1  0           1                              5    6            |
y |                 2            2   4   3         6    7               |
  |         0           1    3               5   4     5    6           |
  |    1           2        3     2 4     3 5     *    7                |
  |   1      0    2      1 3       *       *            5    6          |
  | 11        0022        *1      4 2     5 3   66 4477  5    66        |
0 |*********************************************************************|
  ++----------------+----------------+----------------+----------------++
-20.0            -10.0              0.0             10.0             20.0
```

Table 13 shows the distribution of scores for grade-level cluster 2–3 by task. As shown in the table, the fit statistics were all around 1 and the thresholds increased monotonically. The probability curves for the scoring rubric (see Figure 28) reveal that most score points have a clear peak.

Table 13
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	Garden Surprise	22 (2%)	0.7	N/A
0	Changing Water	NA	NA	N/A
1	Garden Surprise	159 (18%)	0.9	-15.63
1	Changing Water	23 (3%)	0.8	NA
2	Garden Surprise	175 (20%)	0.9	-8.87
2	Changing Water	111 (13%)	0.9	-12.17
3	Garden Surprise	212 (24%)	0.9	-3.19
3	Changing Water	233 (27%)	1.0	-6.64
4	Garden Surprise	188 (21%)	0.9	1.00
4	Changing Water	300 (35%)	1.0	-1.39
5	Garden Surprise	118 (13%)	0.9	5.22
5	Changing Water	146 (17%)	0.9	3.81
6	Garden Surprise	15 (2%)	1.1	9.57
6	Changing Water	49 (6%)	1.3	6.73
7	Garden Surprise	1 (0%)	1.0	11.90
7	Changing Water	3 (0%)	0.9	9.65

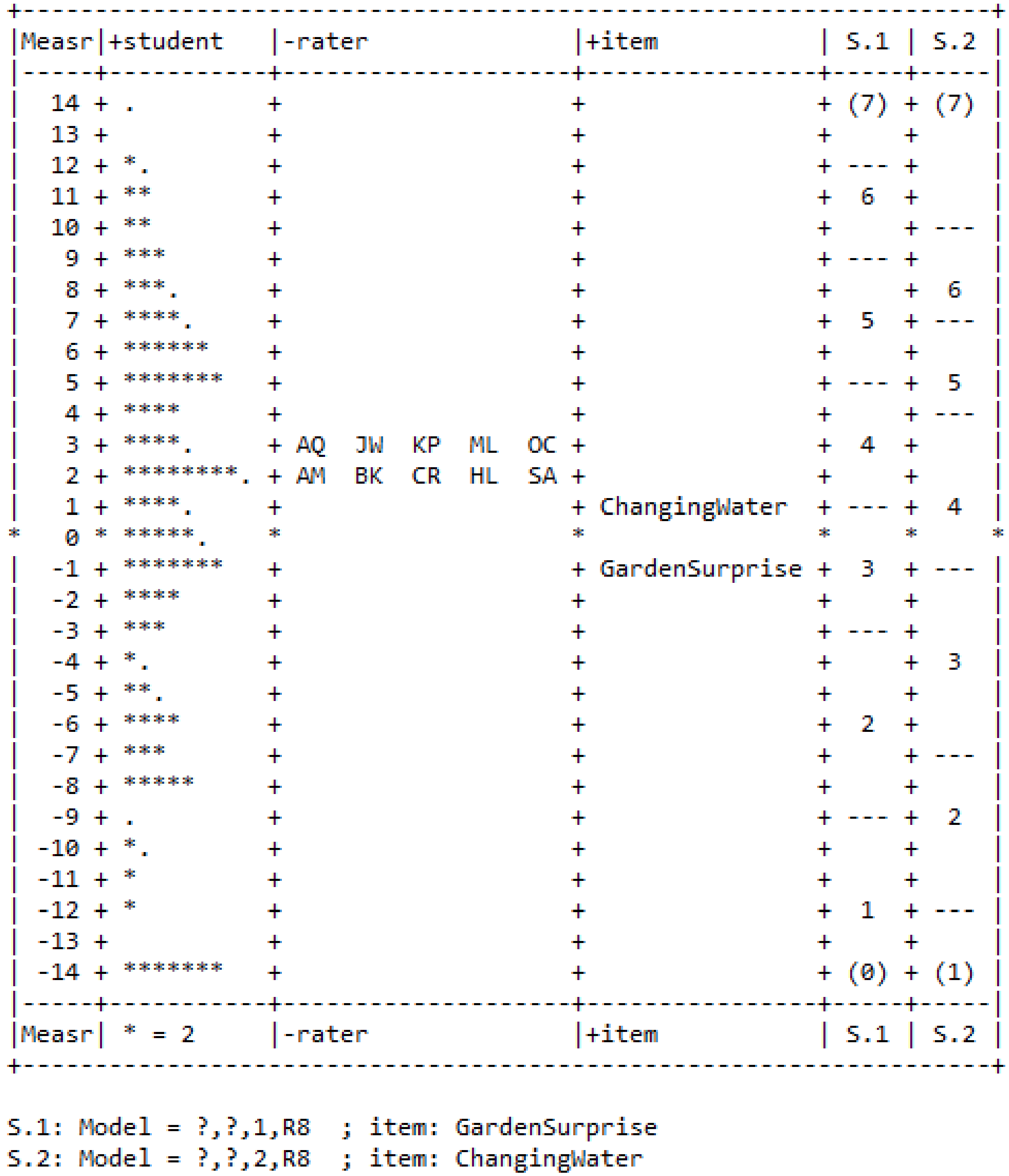
Note: *Garden Surprise* is Tier A; *Changing Water* is Tier B/C

Grade-level cluster 2–3 (PCM)

Figure 29 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.

Figure 29

Wright map for grade-level cluster 2–3 PCM



```
+-------------------------------------------------------------------+
|Measr|+student   |-rater              |+item           | S.1 | S.2 |
|-----+-----------+--------------------+----------------+-----+-----|
|  14 + .         +                    +                + (7) + (7) |
|  13 +           +                    +                +     +     |
|  12 + *.        +                    +                + --- +     |
|  11 + **        +                    +                +  6  +     |
|  10 + **        +                    +                +     + --- |
|   9 + ***       +                    +                + --- +     |
|   8 + ***.      +                    +                +     +  6  |
|   7 + ****.     +                    +                +  5  + --- |
|   6 + ******    +                    +                +     +     |
|   5 + *******   +                    +                + --- +  5  |
|   4 + ****      +                    +                +     + --- |
|   3 + ****.     + AQ  JW  KP  ML  OC +                +  4  +     |
|   2 + ********. + AM  BK  CR  HL  SA +                +     +     |
|   1 + ****.     +                    + ChangingWater  + --- +  4  |
*   0 * *****.    *                    *                *     *     *
|  -1 + *******   +                    + GardenSurprise +  3  + --- |
|  -2 + ****      +                    +                +     +     |
|  -3 + ***       +                    +                + --- +     |
|  -4 + *.        +                    +                +     +  3  |
|  -5 + **.       +                    +                +     +     |
|  -6 + ****      +                    +                +  2  +     |
|  -7 + ***       +                    +                +     + --- |
|  -8 + *****     +                    +                +     +     |
|  -9 + .         +                    +                + --- +  2  |
| -10 + *.        +                    +                +     +     |
| -11 + *         +                    +                +     +     |
| -12 + *         +                    +                +  1  + --- |
| -13 +           +                    +                +     +     |
| -14 + *******   +                    +                + (0) + (1) |
|-----+-----------+--------------------+----------------+-----+-----|
|Measr| * = 2     |-rater              |+item           | S.1 | S.2 |
+-------------------------------------------------------------------+

S.1: Model = ?,?,1,R8  ; item: GardenSurprise
S.2: Model = ?,?,2,R8  ; item: ChangingWater
```

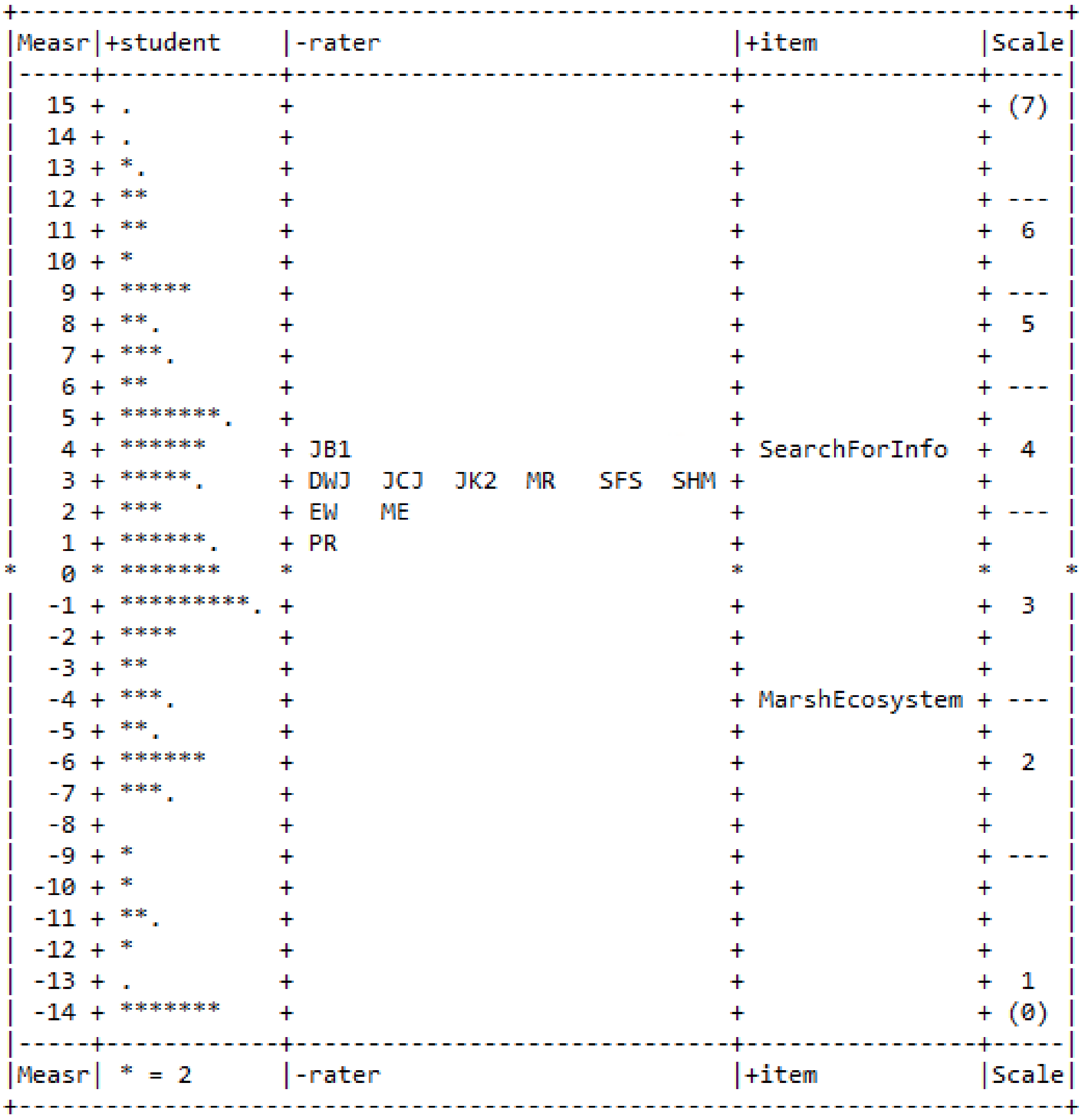
Grade-level cluster 4–5 (RSM)

Figure 30 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 30

Wright map for grade-level cluster 4–5 RSM



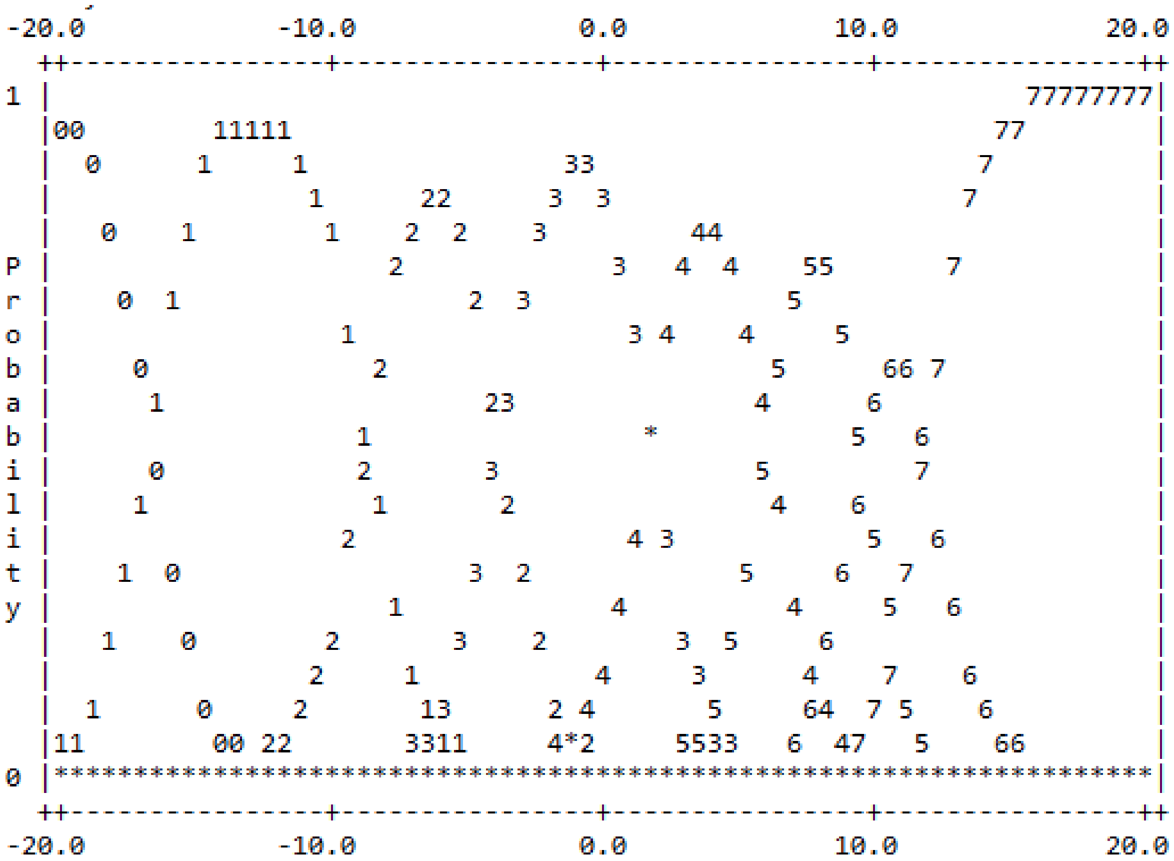
```
+-------------------------------------------------------------------------+
|Measr|+student     |-rater                        |+item           |Scale|
|-----+-------------+------------------------------+----------------+-----|
|  15 + .           +                              +                + (7) |
|  14 + .           +                              +                +     |
|  13 + *.          +                              +                +     |
|  12 + **          +                              +                + --- |
|  11 + **          +                              +                +  6  |
|  10 + *           +                              +                +     |
|   9 + *****       +                              +                + --- |
|   8 + **.         +                              +                +  5  |
|   7 + ***.        +                              +                +     |
|   6 + **          +                              +                + --- |
|   5 + *******.    +                              +                +     |
|   4 + ******      + JB1                          + SearchForInfo  +  4  |
|   3 + *****.      + DWJ  JCJ  JK2  MR   SFS  SHM +                +     |
|   2 + ***         + EW   ME                      +                + --- |
|   1 + ******.     + PR                           +                +     |
*   0 * *******     *                              *                *     *
|  -1 + *********.  +                              +                +  3  |
|  -2 + ****        +                              +                +     |
|  -3 + **          +                              +                +     |
|  -4 + ***.        +                              + MarshEcosystem + --- |
|  -5 + **.         +                              +                +     |
|  -6 + ******      +                              +                +  2  |
|  -7 + ***.        +                              +                +     |
|  -8 +             +                              +                +     |
|  -9 + *           +                              +                + --- |
| -10 + *           +                              +                +     |
| -11 + **.         +                              +                +     |
| -12 + *           +                              +                +     |
| -13 + .           +                              +                +  1  |
| -14 + *******     +                              +                + (0) |
|-----+-------------+------------------------------+----------------+-----|
|Measr| * = 2       |-rater                        |+item           |Scale|
+-------------------------------------------------------------------------+
```

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 among all of them, suggesting the need for more training to ensure similar harshness 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 14 shows the distribution of scores for grade-level cluster 4–5. The fit statistics were mostly around 1 and the thresholds increased monotonically. Probability curves for the rating scale (see Figure 31) reveal that most score levels have a clear peak.

Figure 31

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



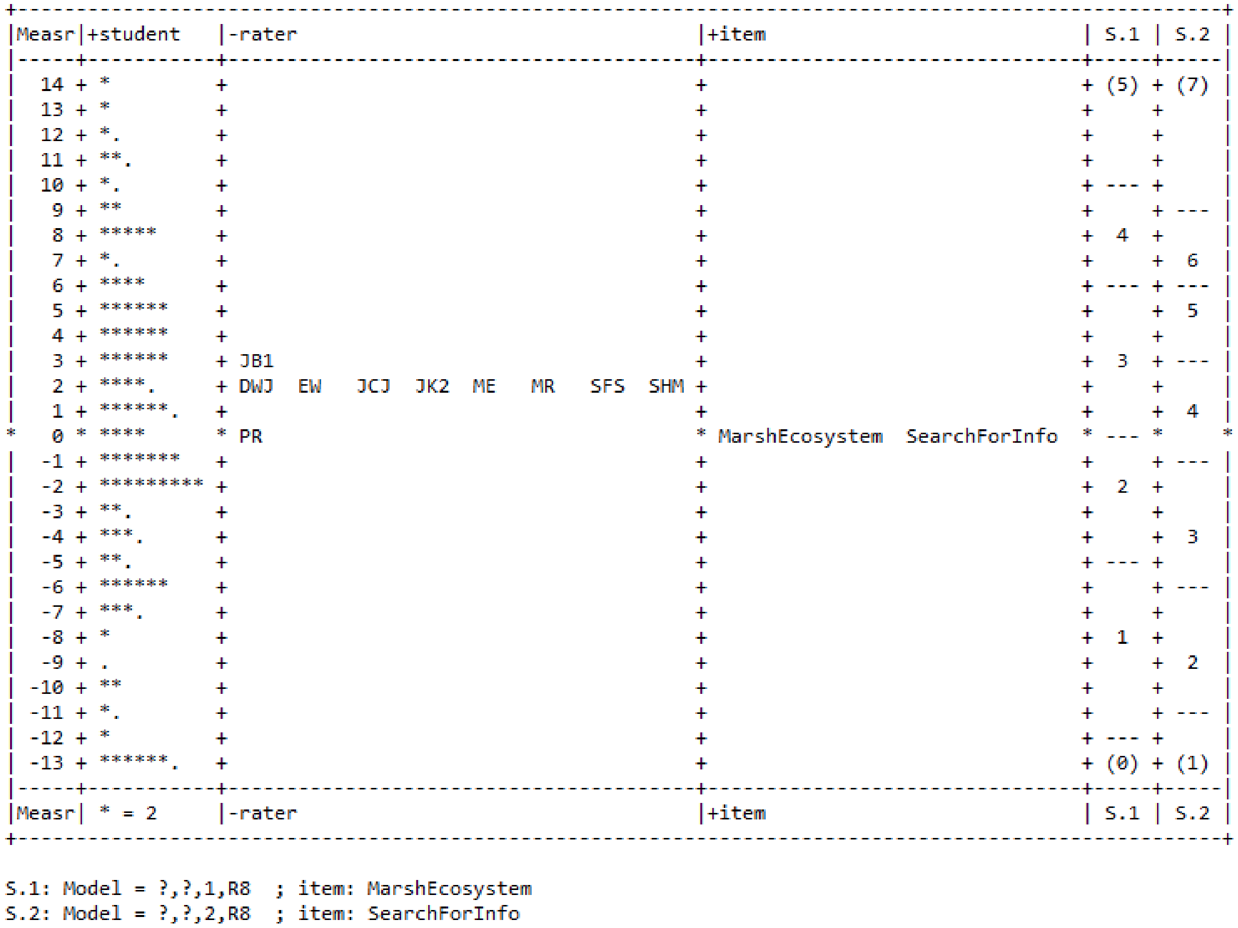
```
-20.0            -10.0              0.0             10.0             20.0
  ++----------------+----------------+----------------+----------------++
1 |                                                             77777777|
  |00        11111                                            77        |
  |  0      1     1                1                33                        7          |
  |                1      22      3  3                      7           |
  |   0    1        1    2  2    3         44                           |
P |                     2             3   4  4    55       7            |
r |    0  1                  2  3                5                      |
o |                  1                 3 4    4     5                   |
b |     0              2                        5      66 7             |
a |      1                    23               4      6                 |
b |                   1                 *            5   6              |
i |      0            2       3                5         7              |
l |     1              1       2                4    6                  |
i |                  2                 4 3            5   6             |
t |    1  0                  3  2             5     6   7               |
y |                     1             4          4     5   6            |
  |   1    0        2       3    2        3  5     6                    |
  |                2     1           4     3      4    7    6           |
  |  1      0     2       13      2 4       5     64  7 5    6          |
  |11        00 22       3311     4*2     5533   6  47   5    66        |
0 |*********************************************************************|
  ++----------------+----------------+----------------+----------------++
-20.0            -10.0              0.0             10.0             20.0
```

Grade-level cluster 4–5 (PCM)

Figure 32 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.

Figure 32

Wright map for grade-level cluster 4–5 PCM



```
+-------------------------------------------------------------------------------------------------------+
|Measr|+student   |-rater                                  |+item                           | S.1 | S.2 |
|-----+-----------+----------------------------------------+--------------------------------+-----+-----|
|  14 + *         +                                        +                                + (5) + (7) |
|  13 + *         +                                        +                                +     +     |
|  12 + *.        +                                        +                                +     +     |
|  11 + **.       +                                        +                                +     +     |
|  10 + *.        +                                        +                                + --- +     |
|   9 + **        +                                        +                                +     + --- |
|   8 + *****     +                                        +                                +  4  +     |
|   7 + *.        +                                        +                                +     +  6  |
|   6 + ****      +                                        +                                + --- + --- |
|   5 + ******    +                                        +                                +     +  5  |
|   4 + ******    +                                        +                                +     +     |
|   3 + ******    + JB1                                    +                                +  3  + --- |
|   2 + ****.     + DWJ  EW   JCJ  JK2  ME   MR   SFS  SHM +                                +     +     |
|   1 + ******.   +                                        +                                +     +  4  |
*   0 * ****      * PR                                     * MarshEcosystem  SearchForInfo  * --- *     *
|  -1 + *******   +                                        +                                +     + --- |
|  -2 + ********* +                                        +                                +  2  +     |
|  -3 + **.       +                                        +                                +     +     |
|  -4 + ***.      +                                        +                                +     +  3  |
|  -5 + **.       +                                        +                                + --- +     |
|  -6 + ******    +                                        +                                +     + --- |
|  -7 + ***.      +                                        +                                +     +     |
|  -8 + *         +                                        +                                +  1  +     |
|  -9 + .         +                                        +                                +     +  2  |
| -10 + **        +                                        +                                +     +     |
| -11 + *.        +                                        +                                +     + --- |
| -12 + *         +                                        +                                + --- +     |
| -13 + ******.   +                                        +                                + (0) + (1) |
|-----+-----------+----------------------------------------+--------------------------------+-----+-----|
|Measr| * = 2     |-rater                                  |+item                           | S.1 | S.2 |
+-------------------------------------------------------------------------------------------------------+

S.1: Model = ?,?,1,R8  ; item: MarshEcosystem
S.2: Model = ?,?,2,R8  ; item: SearchForInfo
```

Table 14

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	Marsh Ecosystem	52 (6%)	0.4	N/A
0	Search for Info	N/A	N/A	N/A
1	Marsh Ecosystem	190 (21%)	1.1	-11.84
1	Search for Info	28 (3%)	1.0	N/A
2	Marsh Ecosystem	286 (31%)	0.7	-4.71
2	Search for Info	128 (15%)	0.9	-10.88
3	Marsh Ecosystem	264 (29%)	0.8	0.33
3	Search for Info	263 (31%)	0.8	-6.40
4	Marsh Ecosystem	91 (10%)	1.4	5.86
4	Search for Info	184 (22%)	0.7	-0.89
5	Marsh Ecosystem	27 (3%)	2.5	10.36
5	Search for Info	123 (15%)	1.1	3.22
6	Marsh Ecosystem	N/A	N/A	N/A
6	Search for Info	70 (8%)	0.7	6.35
7	Marsh Ecosystem	N/A	N/A	N/A
7	Search for Info	41 (5%)	1.2	8.60

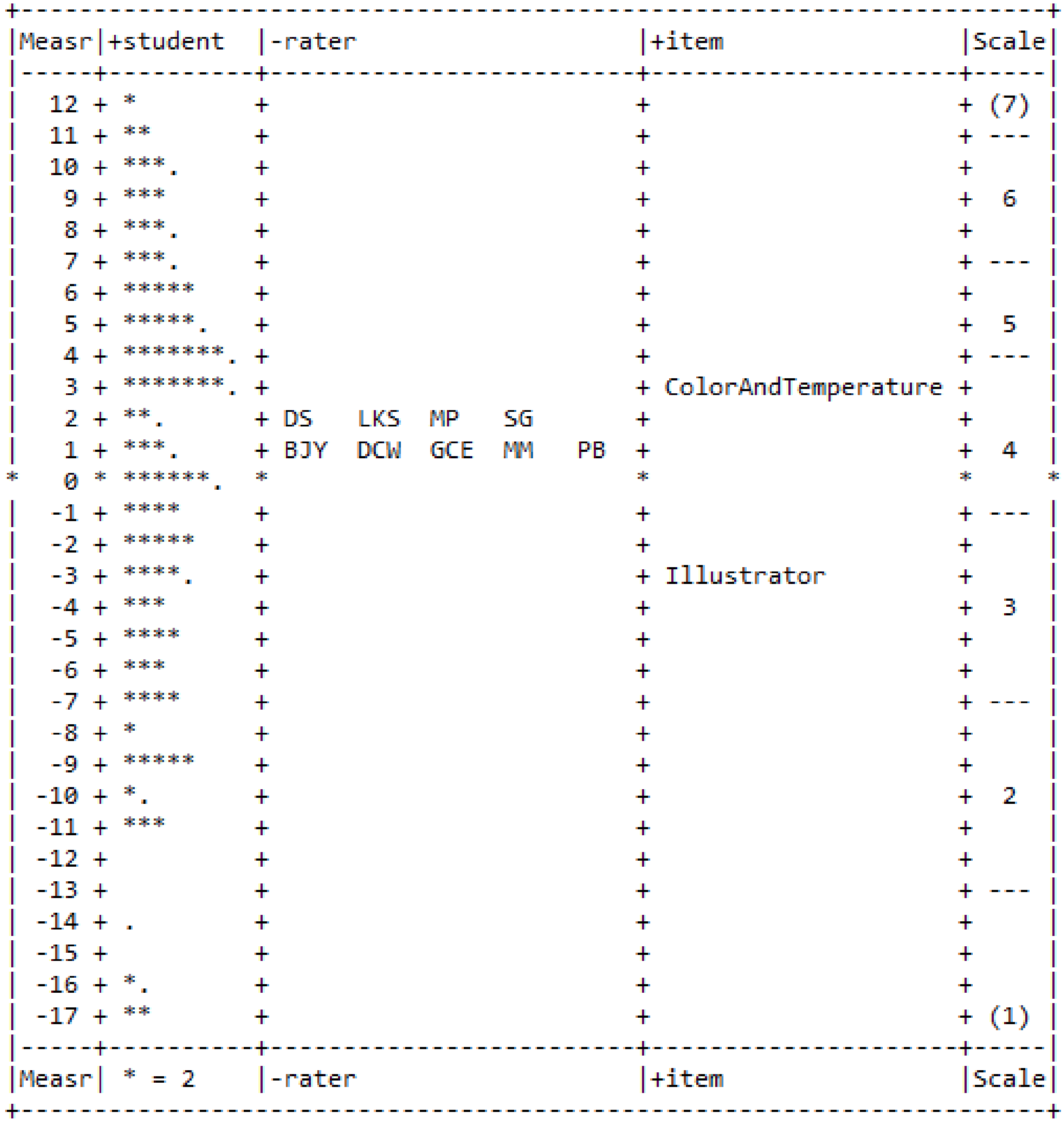
Note: *Marsh Ecosystem* is Tier A; *Search for Info* is Tier B/C

Grade-level cluster 6–8 (RSM)

Figure 33 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 33

Wright map for grade-level cluster 6–8 RSM

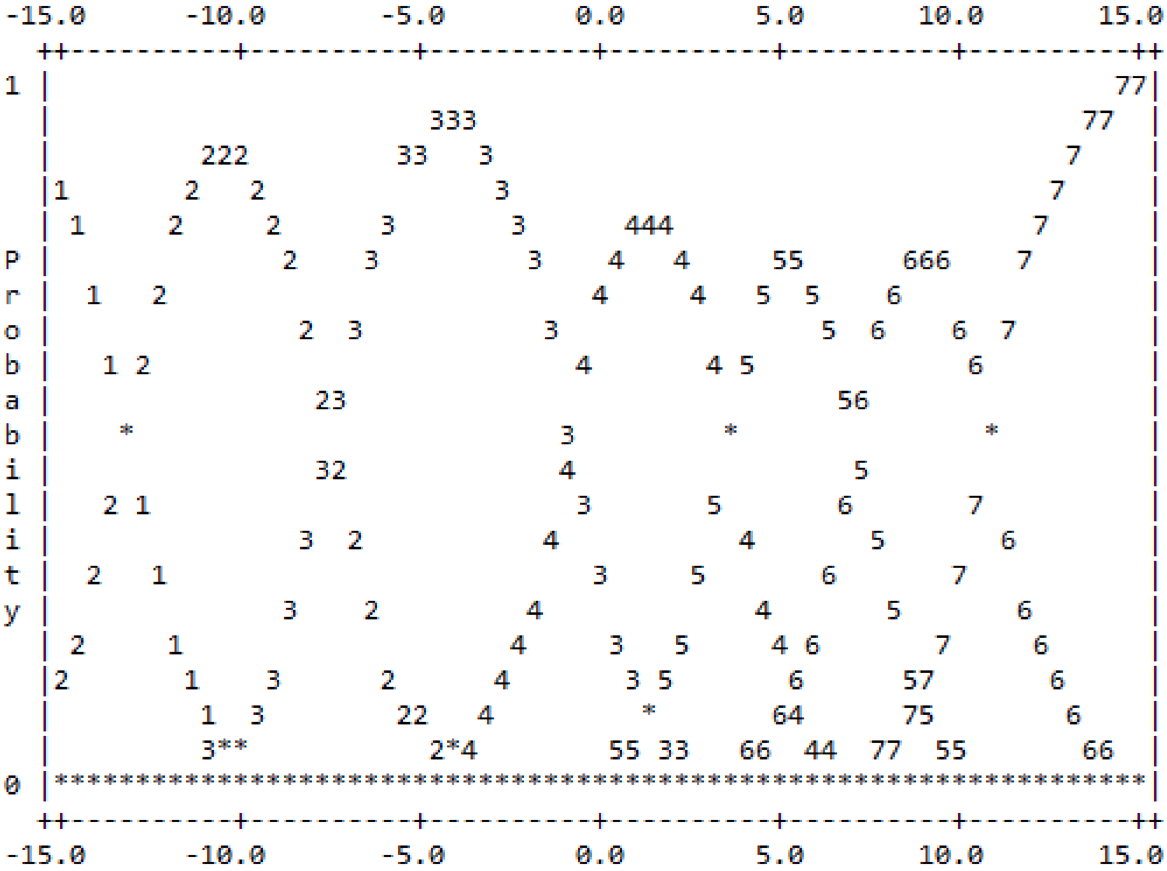


```
+-----------------------------------------------------------------------+
|Measr|+student  |-rater                    |+item                |Scale|
|-----+----------+--------------------------+---------------------+-----|
|  12 + *        +                          +                     + (7) |
|  11 + **       +                          +                     + --- |
|  10 + ***.     +                          +                     +     |
|   9 + ***      +                          +                     +  6  |
|   8 + ***.     +                          +                     +     |
|   7 + ***.     +                          +                     + --- |
|   6 + *****    +                          +                     +     |
|   5 + *****.   +                          +                     +  5  |
|   4 + *******. +                          +                     + --- |
|   3 + *******. +                          + ColorAndTemperature +     |
|   2 + **.      + DS   LKS  MP   SG        +                     +     |
|   1 + ***.     + BJY  DCW  GCE  MM   PB   +                     +  4  |
*   0 * ******.  *                          *                     *     *
|  -1 + ****     +                          +                     + --- |
|  -2 + *****    +                          +                     +     |
|  -3 + ****.    +                          + Illustrator         +     |
|  -4 + ***      +                          +                     +  3  |
|  -5 + ****     +                          +                     +     |
|  -6 + ***      +                          +                     +     |
|  -7 + ****     +                          +                     + --- |
|  -8 + *        +                          +                     +     |
|  -9 + *****    +                          +                     +     |
| -10 + *.       +                          +                     +  2  |
| -11 + ***      +                          +                     +     |
| -12 +          +                          +                     +     |
| -13 +          +                          +                     + --- |
| -14 + .        +                          +                     +     |
| -15 +          +                          +                     +     |
| -16 + *.       +                          +                     +     |
| -17 + **       +                          +                     + (1) |
|-----+----------+--------------------------+---------------------+-----|
|Measr| * = 2    |-rater                    |+item                |Scale|
+-----------------------------------------------------------------------+
```

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, the 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 (fitting too well to the model).

Table 15 shows the distribution of scores for grade-level cluster 6–8. All the fit statistics were around 1. Rasch-Andrich threshold measures increased monotonically. Probability curves for the rating scale (see Figure 34) indicate that most score levels have an outstanding peak and an evenly spaced hill.

Figure 34
Probability curves for the rating scale (grade-level cluster 6–8)



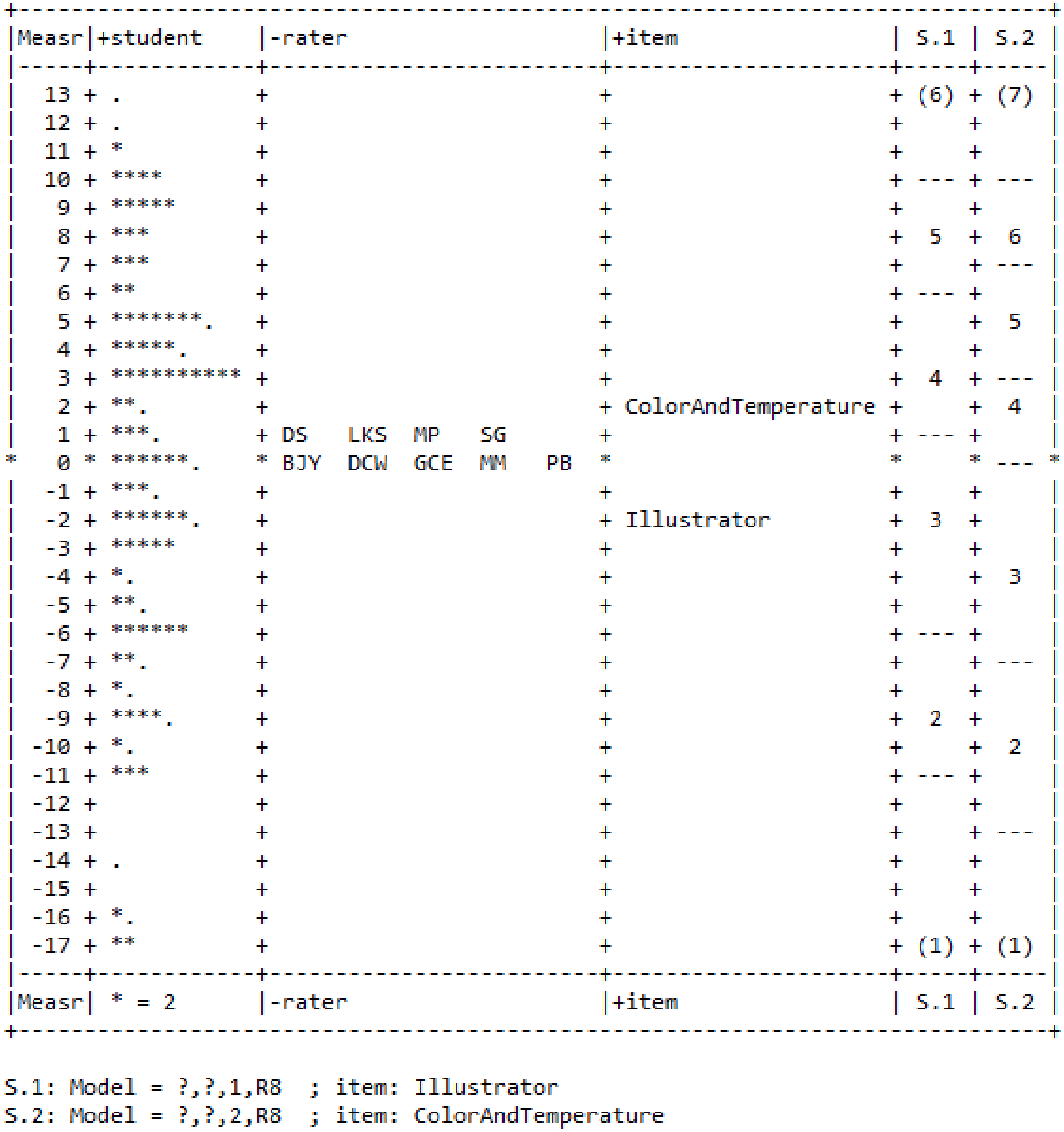
```
  -15.0      -10.0       -5.0        0.0        5.0       10.0       15.0
    ++----------+----------+----------+----------+----------+----------++
  1 |                                                                 77|
    |                       333                                     77  |
    |         222         33   3                                   7    |
    |1       2   2              3                                 7     |
    | 1     2     2      3       3      444                      7      |
  P |              2    3         3    4   4     55      666    7       |
  r |  1   2                          4     4   5  5    6               |
  o |               2  3           3                5  6    6  7        |
  b |   1 2                          4       4 5             6          |
  a |                23                              56                 |
  b |    *                          3         *               *         |
  i |                32             4                 5                 |
  l |   2 1                          3       5       6       7          |
  i |               3  2           4           4       5       6        |
  t |  2   1                          3     5       6       7           |
  y |              3    2         4             4       5       6       |
    | 2     1                    4     3   5     4 6       7     6      |
    |2       1    3      2      4       3 5       6      57       6     |
    |         1  3        22   4         *       64      75        6    |
    |         3**           2*4        55 33   66  44  77  55       66  |
  0 |*******************************************************************|
    ++----------+----------+----------+----------+----------+----------++
  -15.0      -10.0       -5.0        0.0        5.0       10.0       15.0
```

Grade-level cluster 6–8 (PCM)

Figure 35 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. Category statistics for individual tasks are in Table 15.

Figure 35

Wright map for grade-level cluster 6–8 PCM



```
+------------------------------------------------------------------------------+
|Measr|+student    |-rater                   |+item                | S.1 | S.2 |
|-----+------------+-------------------------+---------------------+-----+-----|
|   13 + .          +                         +                     + (6) + (7) |
|   12 + .          +                         +                     +     +     |
|   11 + *          +                         +                     +     +     |
|   10 + ****       +                         +                     + --- + --- |
|    9 + *****      +                         +                     +     +     |
|    8 + ***        +                         +                     +  5  +  6  |
|    7 + ***        +                         +                     +     + --- |
|    6 + **         +                         +                     + --- +     |
|    5 + *******.   +                         +                     +     +  5  |
|    4 + *****.     +                         +                     +     +     |
|    3 + ********** +                         +                     +  4  + --- |
|    2 + **.        +                         + ColorAndTemperature +     +  4  |
|    1 + ***.       + DS   LKS  MP   SG       +                     + --- +     |
*    0 * ******.    * BJY  DCW  GCE  MM   PB   *                     *     * --- *
|   -1 + ***.       +                         +                     +     +     |
|   -2 + ******.    +                         + Illustrator         +  3  +     |
|   -3 + *****      +                         +                     +     +     |
|   -4 + *.         +                         +                     +     +  3  |
|   -5 + **.        +                         +                     +     +     |
|   -6 + ******     +                         +                     + --- +     |
|   -7 + **.        +                         +                     +     + --- |
|   -8 + *.         +                         +                     +     +     |
|   -9 + ****.      +                         +                     +  2  +     |
|  -10 + *.         +                         +                     +     +  2  |
|  -11 + ***        +                         +                     + --- +     |
|  -12 +            +                         +                     +     +     |
|  -13 +            +                         +                     +     + --- |
|  -14 + .          +                         +                     +     +     |
|  -15 +            +                         +                     +     +     |
|  -16 + *.         +                         +                     +     +     |
|  -17 + **         +                         +                     + (1) + (1) |
|-----+------------+-------------------------+---------------------+-----+-----|
|Measr| * = 2      |-rater                   |+item                | S.1 | S.2 |
+------------------------------------------------------------------------------+

S.1: Model = ?,?,1,R8  ; item: Illustrator
S.2: Model = ?,?,2,R8  ; item: ColorAndTemperature
```

Table 15

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	Illustrator	N/A	N/A	N/A
0	Color and Temperature	N/A	N/A	N/A
1	Illustrator	49 (6%)	0.8	N/A
1	Color and Temperature	26 (3%)	0.9	N/A
2	Illustrator	203 (25%)	0.8	-11.29
2	Color and Temperature	80 (10%)	0.8	-12.91
3	Illustrator	246 (30%)	0.8	-5.73
3	Color and Temperature	163 (21%)	0.9	-6.86
4	Illustrator	207 (25%)	1.0	0.82
4	Color and Temperature	181	(24%)	-0.38
5	Illustrator	113 (14%)	0.9	5.73
5	Color and Temperature	158 (21%)	1.0	3.49
6	Illustrator	10 (1%)	1.1	10.47
6	Color and Temperature	101 (13%)	1.2	6.46
7	Illustrator	N/A	N/A	N/A
7	Color and Temperature	59 (8%)	1.1	10.20

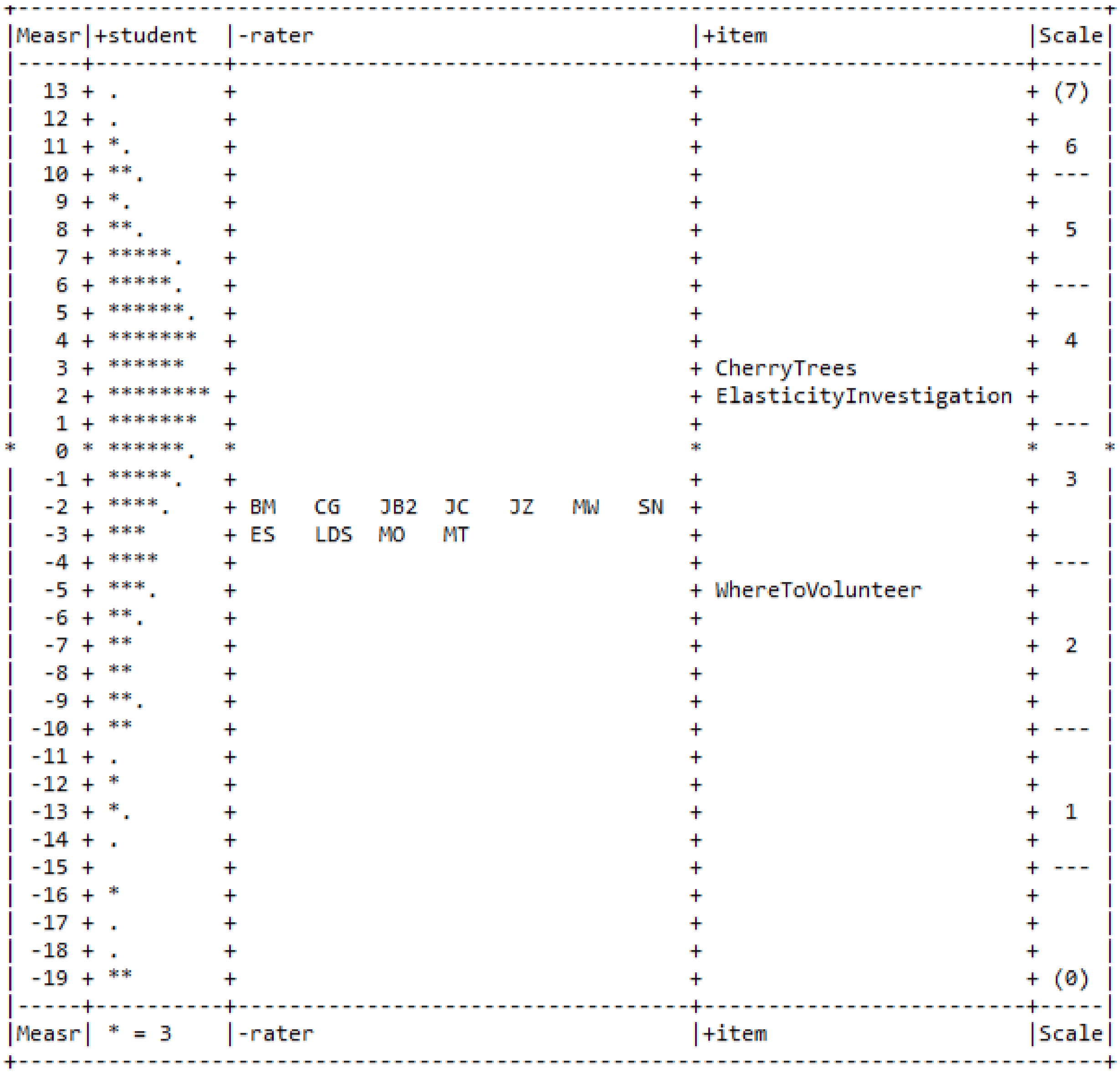
Note: *Illustrator* is Tier A; *Color and Temperature* is Tier B/C

Grade-level cluster 9–12 (RSM)

Figure 36 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 36

Wright map for grade-level cluster 9–12 RSM



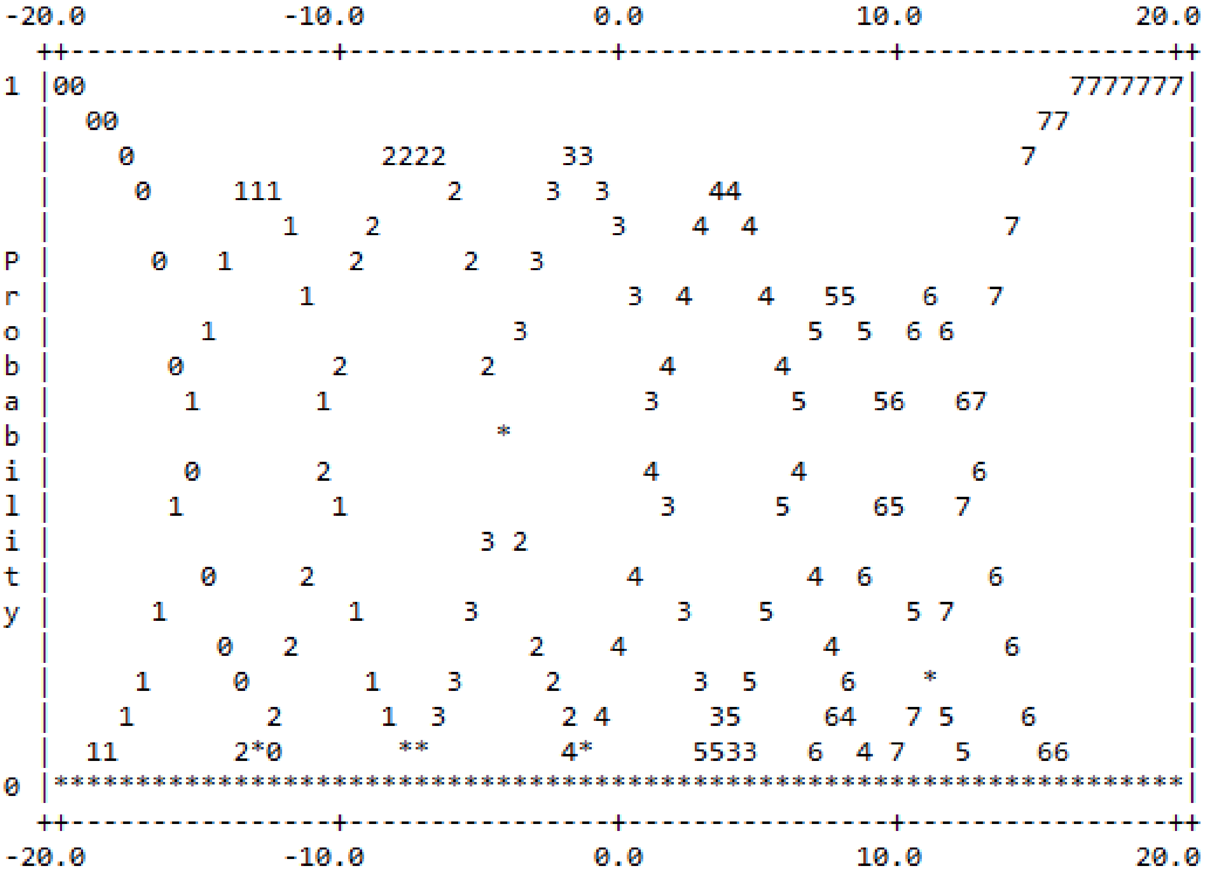
```
+-------------------------------------------------------------------------------------+
|Measr|+student  |-rater                              |+item                    |Scale|
|-----+----------+------------------------------------+-------------------------+-----|
|  13 + .        +                                    +                         + (7) |
|  12 + .        +                                    +                         +     |
|  11 + *.       +                                    +                         +  6  |
|  10 + **.      +                                    +                         + --- |
|   9 + *.       +                                    +                         +     |
|   8 + **.      +                                    +                         +  5  |
|   7 + *****.   +                                    +                         +     |
|   6 + *****.   +                                    +                         + --- |
|   5 + ******.  +                                    +                         +     |
|   4 + *******  +                                    +                         +  4  |
|   3 + ******   +                                    + CherryTrees             +     |
|   2 + ******** +                                    + ElasticityInvestigation +     |
|   1 + *******  +                                    +                         + --- |
*   0 * ******.  *                                    *                         *     *
|  -1 + *****.   +                                    +                         +  3  |
|  -2 + ****.    + BM   CG   JB2  JC   JZ   MW   SN   +                         +     |
|  -3 + ***      + ES   LDS  MO   MT                  +                         +     |
|  -4 + ****     +                                    +                         + --- |
|  -5 + ***.     +                                    + WhereToVolunteer        +     |
|  -6 + **.      +                                    +                         +     |
|  -7 + **       +                                    +                         +  2  |
|  -8 + **       +                                    +                         +     |
|  -9 + **.      +                                    +                         +     |
| -10 + **       +                                    +                         + --- |
| -11 + .        +                                    +                         +     |
| -12 + *        +                                    +                         +     |
| -13 + *.       +                                    +                         +  1  |
| -14 + .        +                                    +                         +     |
| -15 +          +                                    +                         + --- |
| -16 + *        +                                    +                         +     |
| -17 + .        +                                    +                         +     |
| -18 + .        +                                    +                         +     |
| -19 + **       +                                    +                         + (0) |
|-----+----------+------------------------------------+-------------------------+-----|
|Measr| * = 3    |-rater                              |+item                    |Scale|
+-------------------------------------------------------------------------------------+
```

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 them all, and that the exact agreement rate was the lowest, suggesting the need for more training to ensure similar harshness 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 16 shows the distribution of scores for grade-level cluster 9–12. The fit statistics were all around 1 and the thresholds increased monotonically. Probability curves for the rating scale (see Figure 37) reveal that most curves seem to be evenly spaced hills with an outstanding peak.

Figure 37

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



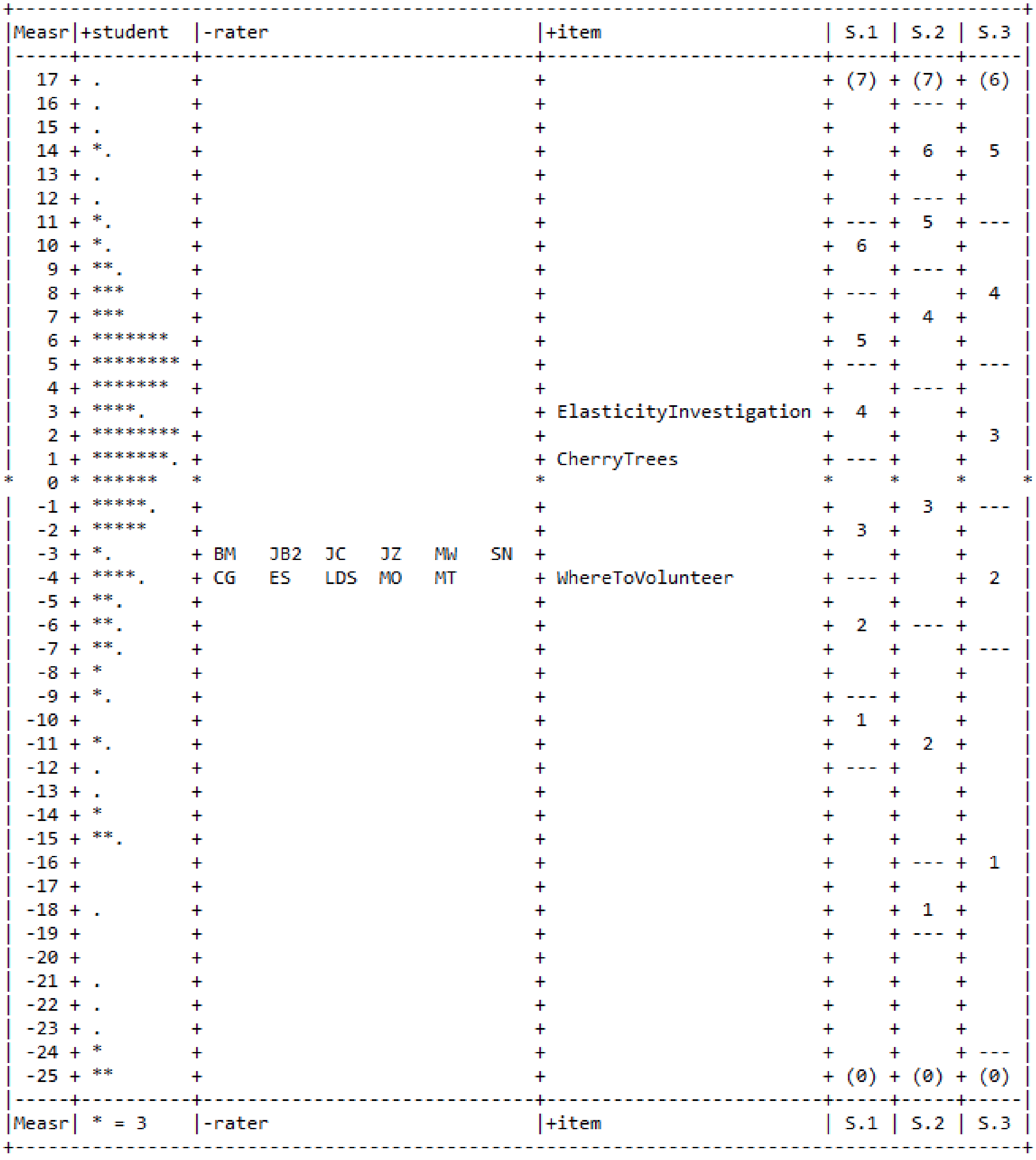
```
-20.0            -10.0              0.0             10.0             20.0
  ++----------------+----------------+----------------+----------------++
1 |00                                                            7777777|
  |  00                                                        77       |
  |    0               2222       33                          7         |
  |     0     111          2     3  3      44                           |
  |              1    2              3    4  4               7          |
P |      0   1       2      2   3                                       |
r |               1                   3  4    4   55    6   7           |
o |         1                  3                 5  5  6 6              |
b |       0         2        2          4      4                        |
a |        1       1                   3        5    56   67            |
b |                           *                                         |
i |        0       2                   4        4          6            |
l |       1         1                   3      5     65   7             |
i |                          3 2                                        |
t |         0     2                   4          4  6       6           |
y |      1           1      3            3    5        5 7              |
  |          0   2              2    4            4          6          |
  |     1     0       1    3     2        3  5     6    *               |
  |    1        2      1  3       2 4      35     64   7 5    6         |
  |  11       2*0       **        4*      5533   6  4 7   5    66       |
0 |*********************************************************************|
  ++----------------+----------------+----------------+----------------++
-20.0            -10.0              0.0             10.0             20.0
```

Grade-level cluster 9–12 (PCM)

Figure 38 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.

Figure 38

Wright map for grade-level cluster 9–12 PCM



```
+-------------------------------------------------------------------------------------------+
|Measr|+student  |-rater                        |+item                    | S.1 | S.2 | S.3 |
|-----+----------+------------------------------+-------------------------+-----+-----+-----|
|  17 + .        +                              +                         + (7) + (7) + (6) |
|  16 + .        +                              +                         +     + --- +     |
|  15 + .        +                              +                         +     +     +     |
|  14 + *.       +                              +                         +     +  6  +  5  |
|  13 + .        +                              +                         +     +     +     |
|  12 + .        +                              +                         +     + --- +     |
|  11 + *.       +                              +                         + --- +  5  + --- |
|  10 + *.       +                              +                         +  6  +     +     |
|   9 + **.      +                              +                         +     + --- +     |
|   8 + ***      +                              +                         + --- +     +  4  |
|   7 + ***      +                              +                         +     +  4  +     |
|   6 + *******  +                              +                         +  5  +     +     |
|   5 + ******** +                              +                         + --- +     + --- |
|   4 + *******  +                              +                         +     + --- +     |
|   3 + ****.    +                              + ElasticityInvestigation +  4  +     +     |
|   2 + ******** +                              +                         +     +     +  3  |
|   1 + *******. +                              + CherryTrees             + --- +     +     |
*   0 * ******   *                              *                         *     *     *     *
|  -1 + *****.   +                              +                         +     +  3  + --- |
|  -2 + *****    +                              +                         +  3  +     +     |
|  -3 + *.       + BM   JB2  JC   JZ   MW   SN  +                         +     +     +     |
|  -4 + ****.    + CG   ES   LDS  MO   MT       + WhereToVolunteer        + --- +     +  2  |
|  -5 + **.      +                              +                         +     +     +     |
|  -6 + **.      +                              +                         +  2  + --- +     |
|  -7 + **.      +                              +                         +     +     + --- |
|  -8 + *        +                              +                         +     +     +     |
|  -9 + *.       +                              +                         + --- +     +     |
| -10 +          +                              +                         +  1  +     +     |
| -11 + *.       +                              +                         +     +  2  +     |
| -12 + .        +                              +                         + --- +     +     |
| -13 + .        +                              +                         +     +     +     |
| -14 + *        +                              +                         +     +     +     |
| -15 + **.      +                              +                         +     +     +     |
| -16 +          +                              +                         +     + --- +  1  |
| -17 +          +                              +                         +     +     +     |
| -18 + .        +                              +                         +     +  1  +     |
| -19 +          +                              +                         +     + --- +     |
| -20 +          +                              +                         +     +     +     |
| -21 + .        +                              +                         +     +     +     |
| -22 + .        +                              +                         +     +     +     |
| -23 + .        +                              +                         +     +     +     |
| -24 + *        +                              +                         +     +     + --- |
| -25 + **       +                              +                         + (0) + (0) + (0) |
|-----+----------+------------------------------+-------------------------+-----+-----+-----|
|Measr| * = 3    |-rater                        |+item                    | S.1 | S.2 | S.3 |
+-------------------------------------------------------------------------------------------+

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

Table 16

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

Note: *Cherry Trees* and *Elasticity Investigation* are Tier B/C; *Where to Volunteer* is Tier A.

Discussion

The four hypotheses

Based on the results of the MFRA, 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 frequently, with a distribution rate of only 2.5%

across the entire dataset. That said, it was not expected nor required that this score point be used across all tasks and grade-level clusters. The expectation for score point 7 is that it be an attainable score point. It is more likely to be used in responses to Tier B/C tasks and particularly at the higher grade-level clusters. If the score point (7) is not used for response to Tier A tasks and only very sparingly for responses in the grades 1 and 2–3 cluster, this is less of a concern. 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 2–3 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) could reduce the possibility of leaving score points unused. Final decisions and future research into the new scoring rubric will be reported in the next section of the report.

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 great 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. The Rasch data reported here provide the first opportunity that WIDA has had to examine data on rater severity and leniency. The operational scoring data from ACCESS does not support the crossed design required for MFRA and interpreting these data is not straightforward as DRC raters were using the new rubric for the first time. WIDA will consider future research studies with DRC raters to examine severity and leniency when the raters are more familiar with the new rubric.

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 new rubric.

The rater reliability data require some interpretation. It is important to keep in mind that DRC raters were using the new writing scoring rubric for the very first time. In addition, these raters were provided with only rudimentary rater training materials. In comparison with the training materials typically provided to DRC raters for operational scoring, the materials provided with the new writing scoring rubric were quite sparse. So, it was to be expected that the raters in this study would find it challenging to achieve the reliability targets typical during operational scoring. The rater reliability reported above indicate that rater training will be required to improve the exact agreement rates between raters during operational scoring. The 2023–24 ACCESS administration will generate data on exact agreement rates from both the existing writing scoring scale (from operational scoring) and the new writing scoring rubric (from field test scoring). Comparison of these agreement rates will provide evidence for appropriate and realistic rater agreement expectations going forward.

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

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

Conclusions and limitations

The MFRA findings provide validity evidence for the quality of the new WIDA ACCESS for ELLs writing scoring rubric. The findings indicate that the new rubric is an improvement on the writing scoring scale in terms of variation of ratings. Score point 7 is not awarded frequently, but it is used by raters for responses in the higher grades and in response to tasks that target higher proficiency levels. Score distributions are less peaked than with the writing scoring scale. The findings indicate that the new scoring rubric discriminated well between students at different ability levels. There is work to do on training raters to apply the new scoring rubric reliably, but the reliability data reported from this study provides a good baseline from which to build as operational raters gain familiarity with the new scoring rubric and receive more thorough training.

While the study suggests the new rubric is overall well-functioning, there are 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 them rating unfamiliar grade-level clusters and writing tasks. This 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, the Rasch findings still offer important information regarding the technical quality of new scoring rubric.

Final rubric

After reviewing the mfra results, the teams at WIDA and CAL who collaborated on the development of the new rubric needed to decide whether to retain a 0–7 rubric or to shorten the number of score points to a 0–6 rubric. Neither the feedback gathered from reviewers nor the mfra data were conclusive in indicating whether a 0–7 or 0–6 rubric would be preferable. The project leads requested a final recommendation from test development and psychometric directors at WIDA and CAL. All four directors were unanimous in recommending that we maintain a 0–7 point rubric. There were several reasons cited; for example:

> *The ultimate decision hinges on the indispensability of the language criteria delineated by the 2020 Standards. This dictates that, based on the manifestation of this score point within clusters 6–8, and 9–12, the adoption of the 0–7 scale emerges as a plausible avenue. Furthermore, the potential merger of score points 6 and 7 remains an option contingent upon the psychometric insights gleaned from FT outcomes.*
>
> *If we had a 0–6 rubric, there may not be enough raw score points (only 12 raw score points) to make five cuts and WIDA may need to make arbitrary decisions about what PL corresponds to which proficiency levels. This happened with the Speaking assessment in the first generation of ACCESS, which only had 12 raw score points, when sometimes it was impossible to score into a PL on the basis of performance on the Speaking test. This was particularly true at the higher levels of performance. Having the possibility of 14 raw score points rather than 12 will decrease the risk of students not getting into a PL level on the basis of their performance.*
>
> *Since we know from the DRC study that there are indeed papers meriting a score of 7 in higher grades in the B/C task, if the scores are collapsed in field testing (meaning a reduction in the number of raw score points), they could never be separated if that is ultimately desired. If it turns out that there are too few performances of "7", scores 6 and 7 could be psychometrically collapsed in the future as we do now with 5, 5+, and 6. However, if only 6 score points were used in the field test, scores of 6 could not be divided into 6 and 7 for the operational program if later needed or desired.*
>
> *In my opinion, we have enough justification to proceed with 0–7 score points. While score point 7 was not used in grades 1 and 2–3, it was used in higher grades. It was also encouraging to see that the frequency with which score point 7 was used increased between G4–5 and G6–8 between G6–8 and G9–12. Finally, it doesn't appear that raters had difficulty differentiating between score points 6 and 7. So, I would recommend we keep the 0–7 score points, for the reasons above and the added benefit that this might potentially help address the confusion we have observed with having the same number of full score points as proficiency levels (e.g., the tendency to interpret score point 1 as equal to PL1).*

After the completion of all rounds of review described above and considering the MFRA results, the final version of the WIDA Writing Scoring Rubric Grades 1-12 was created. It is included in Appendix B.

In summary, the new writing scoring rubric offers a number of advantages when compared with the previously used writing scoring scale. The new rubric has a different number of raw score points from the reported proficiency level scores, hopefully alleviating the confusion that can arise with score interpretation. There is evidence from reviewers and DRC ratings that all the raw score points on the new rubric are attainable by the test population, particularly in the higher grades in response to the Tier B/C test tasks. The removal of the "+" score points between the solid score points, a key feature of the writing scoring scale, was positively received by all reviewers and should support the calculation and reporting of more transparent rater reliability data.

WIDA will continue to monitor the reliability of scores awarded by raters using the new rubric and will provide enhanced rater training materials to DRC raters to help implement the new rubric operationally. Finally, and perhaps most importantly, the new rubric has operationalized the WIDA ELD Standards Framework, 2020 Edition, incorporating the grade-level cluster specific approach to performance descriptions and the enhanced focus on discourse competence.

References

Becker, A. (2018). Not to scale? An argument-based inquiry into the validity of an L2 writing rating scale. *Assessing Writing*, *37*, 1–12.

Hamp-Lyons, L. (1991). *Assessing second language writing in academic contexts*. Norwood, NJ: Ablex Publishing Corporation. (clothbound: ISBN-089391-659-5; paperback: ISBN-0-89391-792-3).

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

Knoch, U. (2007). 'Little coherence, considerable strain for reader': A comparison between two rating scales for the assessment of coherence. *Assessing writing*, *12*(2), 108–128.

Linacre, J. M. (2017). FACETS: Computer Program for Many Faceted Rasch Measurement (Version 3.80.0). Chicago, IL: Mesa Press.

Linacre, J. M. (2002). Optimizing rating scale category effectiveness. *Journal of applied measurement*, *3*(1), 85-106.

Turner, C. (2000). Listening to the voices of rating scale developers: Identifying salient features for second language performance assessment. *Canadian Modern Language Review*, *56*(4), 555–584.

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.

Upshur, J. A., & Turner, C. E. (1995). Constructing rating scales for second language tests. *ELT Journal*, 49(1), 3–12.

Weigle, S. C. (2002). *Assessing writing*. Cambridge University Press.

WIDA. (2022). *Annual Technical Report for ACCESS for ELLs Online English Language Proficiency Test Series 503, 2021-2022 Administration*. Board of Regents of the University of Wisconsin System.

WIDA. (2023). *Annual Technical Report for ACCESS for ELLs Online English Language Proficiency Test Series 601, 2022-2023 Administration*. Board of Regents of the University of Wisconsin System.

WIDA. (2020). *WIDA English Language Development Standards Framework, 2020 Edition: Kindergarten-Grade 12*. Board of Regents of the University of Wisconsin System.

Appendix A: New WIDA Writing Scoring Rubric Grades 1–12 (Draft 1)

	Score Point 7 D: Ideas are coherently organized with a sense of purpose appropriate to the genre, using language that connects ideas together smoothly throughout the response S: Demonstrates use of complex clause structure, producing simple, compound, and complex sentences, which may not always be error free but meaning is consistently clear W: Uses a wide variety of words and phrases appropriately and with precision, making choices that are relevant to the content area
6+	
	Score Point 6 D: Ideas are well organized with a sense of purpose appropriate to the genre, using a variety of connectors with precision S: A wide variety of sentence structures, producing simple, compound, and complex sentences which are not always error free but meaning is almost always clear W: A flexible repertoire of words and phrases used with some precision, making choices that are increasingly relevant to the content area
5+	
	Score Point 5 D: Text that conveys an emerging sense of purpose appropriate to the genre with clear organizational pattern, using connectors with some precision and a variety of types of elaboration S: A variety of sentence structures, using both compound and complex structures that are sometimes successful with generally clear meaning W: An expanding repertoire of words and phrases used with some precision
4+	
	Score Point 4 D: Text that conveys intended purpose with increasingly clear organizational pattern, using a growing number of connectors and some elaboration S: Sentences with emerging use of clauses, attempts at complex sentences, some of which may be successful while others may be challenging to distinguish W: A growing repertoire of words and phrases beyond the stimulus and prompt, used with increasing precision

3+	
	Score Point 3 D: Text that shows developing organization with emerging use of simple connectors (e.g. and, then, next, but, because) and some elaboration S: Can produce several simple sentences and may attempt complex sentences, though sentence boundaries may be challenging to distinguish W: Emerging repertoire of words and phrases, though language is still commonly drawn from the stimulus and prompt
2+	
	Score Point 2 D: Text that represents an idea or ideas with simple elaboration S: Can produce a short, simple sentence, though attempts beyond short, simple sentences may be challenging to distinguish W: Some frequently used words and phrases, in addition to language drawn from the stimulus and prompt
1+	
	Score Point 1 D: Minimal distinguishable text that represents an idea or ideas S: Some attempted sentences that may be partially distinguishable, but primarily words, chunks of language, or phrases W: Some distinguishable English words, some words may be reformulated from the stimulus or prompt, others may be attempted yet challenging to distinguish, exhibiting creative spelling and letter formation **Nonscorable**: The response is blank; consists only of verbatim copied text; consists only of text that is completely off task (no discernible relevance to the prompt); or is entirely in a language other than English.

Appendix B: WIDA Writing Scoring Rubric for Grades 1–12 (Final)

WIDA™
UNIVERSITY OF WISCONSIN–MADISON

WIDA Writing Scoring Rubric for Grades 1–12

Use this rubric to score writing responses for WIDA ACCESS and WIDA Screener tests only. The score points described below are not equivalent to WIDA English language proficiency levels.

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

Off-Task & Off-Topic Responses

An **off-task** response shows no understanding of, or no interaction with, the prompt. It may be a memorized response, indicate refusal or inability to answer the prompt, or appear to answer another, unrelated prompt. A response that is entirely off-task is scored at **Score Point 0.** A response that is partially off-task is scored by ignoring the off-task portion of the response and scoring only the on-task portion using the scoring rubric.

An **off-topic** response shows a misunderstanding or misinterpretation of the prompt. It is related to the prompt in some way, but does not address it as intended. Whether entirely or partially off-topic, these responses are scored in their entirety using the scoring rubric.

WIDA™
UNIVERSITY OF WISCONSIN–MADISON

Additional Scoring Notes

Scoring Materials

In addition to this scoring rubric, it is critical to utilize the relevant set of anchor responses and explanations provided in WIDA training materials, which demonstrate how to apply the rubric to the specific grade-level clusters and tasks being scored.

Scoring Process

Responses should be assigned the score point whose descriptors provide the best holistic description of the response. For example, if a response corresponds to *most* of the Score Point 3 descriptors, but also one of the Score Point 4 descriptors, it should still be scored a 3, as that provides the best fit. Similarly, if the response corresponds to *most* of the Score Point 3 descriptors, but one of the Score Point 2 descriptors, it should still be scored a 3. Responses should be scored a 0 if **one or more** of the descriptors at Score Point 0 apply.

Writing Mechanics

Responses may contain issues with mechanics, such as inconsistent or absent capitalization, inconsistent or absent punctuation, typos, and creative spelling. These aspects of mechanics are not considered central to the evaluation of multilingual learners' writing. Responses should not be penalized, in terms of the score awarded, as a result of mechanical errors.

Sentence Boundaries

Responses may lack traditional sentence boundaries, which are typically marked by the use of punctuation. Raters should evaluate responses without undue concern for this absence. Responses should be evaluated for how ideas are connected together within the response. If the connection of ideas is discernible, even without clearly marked sentence boundaries, credit should be given. Responses should not be penalized, in terms of the score awarded, as a result of a lack of clear sentence boundaries.

Languages other than English

Responses may be written in English and languages other than English. Raters should award scores based only on the English language used in the response. However, responses should not be penalized, in terms of the score awarded, for using languages other than English.

Glossary

Errors

This refers to *language* errors. It does not refer to factual or mechanical errors.

Stimulus

This refers to the text that appears in front of the student, either in the test booklet or on the computer screen. It does not refer to language provided in the audio or test administrator scripting, or to elements depicted in the graphics.

Original words and phrases

These are any words or phrases that appear in a response, but which were not provided in the stimulus.

Sentence Types

There are three types of sentences: simple, compound, and complex.

- **Simple sentences** contain a single independent clause. Simple sentences can be short (e.g., *She sleeps.* or *The children got seeds.)* or long (e.g., *The hollow ball bounced highest on the wood floor in this experiment*).

- **Compound sentences** contain two or more independent clauses, often linked with coordinating conjunctions such as *and, so, but,* or *yet* (e.g., *The boy cut some flowers and he gave them to the teacher.*).

- **Complex sentences** contain multiple clauses. The relationships among the clauses are not equal in that one of the clauses is independent and the others are dependent. A complex sentence is useful for conveying intricate and detailed relationships among ideas (e.g., *The hollow ball will bounce lower if the floor is covered in carpet. The experiment shows that hard floors result in higher bounces. They wanted to grow flowers. When the seeds began to sprout, the students gave them water. The teacher was happy because he got a thoughtful gift.*).

  Clauses in complex sentences are often, but not always, joined by subordinating conjunctions, for example: *after, as a result of, as if, as long as, as well as, although, because, before, besides, despite, even if, except for, if, in case, instead of, like, since, that, unless, until, when,* or *while*.

Clauses

- **Independent clauses** can stand alone to communicate a complete idea, and form a complete sentence. An independent clause usually has a subject (a noun) and a predicate (a verb), unlike a dependent clause.
- **Dependent clauses** depend on an independent clause for meaning and cannot stand alone.

Connectors
Connectors include text connectives, coordinating and subordinating conjunctions, and linking phrases used to connect ideas within and across sentences and signal different relationships (see examples below). Connectors create cohesion and support the logical development of ideas across a text.

Purpose	Example Connectors
addition	*and, and then, in addition, furthermore, besides, again, along with*
cause/consequence	*because, so, despite, nevertheless, even though, therefore, consequently, due to, because of this, as a result*
comparison/contrast	*but, for example, instead, in other words, however, in fact, in that case, while, although, on the other hand, despite*
concession	*while, although*
condition	*if, unless*
purpose	*in order to, so*
sequence	*first, second, finally, in the first place, to start with, at this point, to get back to the point, in short, all in all, to conclude*
time	*when, then, next, after, afterward, after a while, at the same time, at this moment, meanwhile, previously, before that, finally*

Key Language Uses consist of prominent genre families across academic content standards.

- **Narrate:** language to convey real or imaginary experiences through stories and histories. Narratives serve many purposes, including to instruct, entertain, teach, or support persuasion.
- **Inform:** language to provide factual information. As students convey information, they define, describe, compare, contrast, organize, categorize, or classify concepts, ideas, or phenomena.
- **Explain:** language to account for how things work or why things happen. As students explain, they substantiate the inner workings of natural, human made, and social phenomena.
- **Argue:** language to develop claims and counterclaims, and to provide evidence to substantiate them. Argue is also used to evaluate issues, advance or defend ideas or solutions, change the audience's point of view, or bring about action.

WIDA™
UNIVERSITY OF WISCONSIN–MADISON

Technical Report

No. TR-2024-2
August 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