AN EVALUATION OF HELP'S WRITING PLACEMENT TEST RUBRIC USING MANY-FACET RASCH ANALYSIS

YU-TZU CHANG, ANN TAI CHOE, DAN HOLDEN, AND DAN ISBELL
University of Hawai'i at Mānoa

ABSTRACT

This study is part of a larger project that explores the ongoing revisions of a writing rubric used in the Hawai'i English Language Program (HELP) at the University of Hawai'i at Mānoa. Previously, HELP revised its writing placement test, changing from an 8-category writing rubric with a 4-point scale ("old rubric," 2010–2016) to a 4-category writing rubric with a 20-point scale ("new rubric," 2017–2021). Because the rationale for revising the old rubric was based on anecdotal evidence (see Rock, 2016), whether the revision was justified from an empirical standpoint remains unknown. In this study, we aim to investigate the empirical support for the motivations and outcomes of the rubric revision using many-facet Rasch measurement (MFRM). Findings revealed that the old rubric was reliable in distinguishing test taker writing ability, but there were several critical issues with its functioning: (1) a redundancy of scale points, (2) rating criteria redundancy, and (3) a high number of misfitting raters. These findings lend support for HELP's previous decision to revise the old rubric. Implications for rubric revision are discussed.

INTRODUCTION

Rubric revision and validation for performance-based language assessments are motivated by concerns toward reliability, measurement, and uses of scores for decision-making (Knoch & Chapelle, 2018). Compared to large-scale assessment contexts, however, test developers in local settings often have fewer resources and rely more on expert intuition or anecdotal evidence to inform revision decisions. Although the stakes are not as high as most large-scale tests, the consequences of local tests are not trivial, and may involve decisions such as placement in a

language program. In accordance with these stakes, validity evidence must not be in short supply (Kane, 2013). This study reports on efforts to validate a local writing placement test used in the Hawai‘i English Language Program (HELP) at the University of Hawai‘i at Mānoa (UHM). In 2016, HELP's writing placement exam underwent a substantial revision, sparked by input from and efforts of program administrators and teachers (who also served as raters; see Rock, 2016). As researchers affiliated with the university and HELP, we followed up on those test development efforts by carrying out systematic validation research using many-facet Rasch measurement (MFRM). Our aim is to show how rigorous validation techniques can be used to evaluate revisions made to local tests and motivate further changes.

BACKGROUND

Decision Making in Rubric Design

Test developers have a myriad of choices when designing or revising a test. In the context of performance-based assessments, these choices may include selecting the type of rating scale (holistic or analytic) and deciding the number of evaluation categories and score points to appear in the rubric. These choices concerning different aspects of the rating scale design must be carefully considered in order to ensure that the scoring procedure is consistent and appropriate for the assessment's purpose and the needs of the program. Take writing placement exams in local assessment contexts as an example. If practicality is a major concern due to a shortage of time or human resources, a holistic rating scale is more advantageous than an analytic rating scale because raters only need to assign a single score rather than parsing an essay several times to focus on different aspects of writing quality. Choosing a holistic rating scale would therefore reduce a significant amount of time needed to score the essays and place students into their corresponding placement levels. As Weigle (2002) cautioned, however, there are notable disadvantages of holistic rating scales, including the fact that a single score does not generate sufficient diagnostic information about a candidate's ability (e.g., grammar, vocabulary, and organization), and that the score may be less interpretable because raters may draw on different criteria in assigning the score. On the other hand, while analytic scoring may be more time-consuming, it can provide rich information about various aspects of a candidate's writing ability. As areas of second language (L2) writing tend to develop at different rates, analytic scoring is

potentially useful for making placement and pedagogical decisions in language programs. Another advantage of using analytic scoring is that reliability tends to increase when an essay is assigned with multiple scores (Weigle, 2002). Therefore, if reliability and accuracy are the primary concerns, adopting an analytic scoring rubric would generate reliable results about various aspects of a student's writing ability to inform placement or pedagogical decisions.

Rubric Revision and Validation

Across both large- and small-scale assessment contexts, the quality of the test needs to be carefully validated. Validation is an iterative process as test developers accumulate new evidence from different stages of development to make revision decisions (Bachman & Palmer, 2010; Kane, 2013). Motivation for rubric revisions can arise from concerns raised in validation research related to the construct, measurement qualities, and the impact of using the rubric for decision making. Validation studies adopting the argument-based approach (Kane, 2006, 2013; Knoch & Chapelle, 2018) have conceptualized these concerns in terms of inferences, warrants, and assumptions. Each inference is specified by a claim, and each claim is evaluated based on warrants and a set of associated assumptions. The assumptions should be supported by evidence known as backing, which could be quantitative (e.g., statistical analysis, many-facet Rasch analysis) and/or qualitative (e.g., interviews with raters, analysis of rating sessions).

In our study, we focus on two inferences for rater-mediated assessments: *evaluation* and *generalization*. According to Knoch and Chapelle (2018), the evaluation inference (also known as "scoring," following Kane, 2006, 2013) "is justified if observations on assessment are evaluated using procedures that provide observed scores with intended characteristics" (p. 482). Concerns regarding the construct and measurement qualities can thus be conceptualized under the evaluation inference, with underlying warrants pertaining to rubric properties and measurement-related practices such as rater consistency and scale functioning. When evaluating the properties of an analytic scale, assumptions associated with the warrant include examining whether the scale steps are adequate in distinguishing among ability levels; assumptions to support the warrant on scoring could include raters' use of the scale in identifying distinct levels of performance, and rater consistency, especially when scores cannot practically be adjusted. Whereas the evaluation inference centers on scale properties and rater performance at the task level, the generalization inference focuses these issues at the test level and is warranted when the

scoring is consistent across different raters. Assumptions associated with the warrant include, for instance, assigning the same ratings at the test level and having a sufficient number of raters to generate a reliable score. To back the assumptions for each inference, validation studies have widely applied MFRM, which creates diagnostic information about the quality of a test and rater performance based on fit statistics (see details in Methods section).

THE CURRENT STUDY

Context

The current study draws on four years of operational testing data to investigate a major rubric revision for a writing placement test used in HELP at UHM. As part of the Department of Second Language Studies (SLS) at UHM, HELP is an intensive academic language program that prepares non-native speakers of English for academic work at UHM or other English-language institutions of higher education. The program's curriculum is divided into four levels (100 through 400). Students may enroll in an 8-week term during the spring or fall, or a 4- or 6-week term in the summer. Students who successfully complete two terms at the 400-level are eligible for conditional admission to UHM, raising the stakes of the outcomes associated with the institute's placement exam (i.e., higher placement can result in quicker university matriculation).

In many language programs, the placement tests are either bought from commercial publishing houses, adapted from other English as a second language (ESL) programs, or reproduced from current textbooks (Brown, 1996). Students enrolled in HELP are placed in their course levels on the intake day before classes begin, and their placement is based on a battery of tests including the Michigan English Placement Test (EPT), a writing sample, and an oral interview. While the University of Michigan EPT is a multiple-choice listening test, the writing sample and speaking test rely on raters for scoring. The writing and speaking tests are scored by trained teacher-raters and program administrators at HELP.

The old rubric (in use between 2010–2016) under investigation was developed locally at HELP by a group of writing instructors with guidance from the administration. The writing instrument asked incoming students to produce a hand-written, multi-paragraph essay in response to one of three prompts within a 30-minute timeframe:

1. Who is the most important person in your life? Why?
2. Write about an interesting tradition or holiday in your country.
3. Write about a time in your life when you had a problem. What was the problem? How did you find a solution to the problem?

The prompts aimed to elicit descriptive writing, and there was not any guidance in terms of paragraph organization. Students were encouraged to make a list of ideas before writing. They were also explicitly instructed to raise their hand if they have a question or need more paper, write legibly, and not to use a dictionary or talk to peers during the test. In terms of scoring, two instructors were asked to score eight categories (see Figure 1) according to numerical values that coincided with the four course levels at HELP (i.e., 100, 200, 300, or 400). Additionally, raters were asked to compute an average across all categories and make an overall level recommendation for the student.

In 2016, HELP decided to revise its rubric due to concerns raised by the teacher-raters. In this instance, the prompts did not elicit writing that was reflective of university writing tasks, such as formulating or evaluating an argument. Also, some teacher-raters were dissatisfied with this system as there had been confusion as to whether their scores should be considered analytic or holistic (Rock, 2016). If raters were working with analytic scores in mind, each of the eight categories should have been scored independently based on students' abilities; however, if raters chose to view the rubric holistically, they would either select uniform scores or scores that work out to a numerical average in line with their intended final placement score. Among raters who attempted to score each rubric category independently, some raters reported difficulty in distinguishing among eight different aspects of writing quality, some of which were seen as overlapping (e.g., "Organization" and "Flow/Cohesion"). Related to this potential for inconsistency in scoring among raters, it was found that some raters preferred to create their own 'desire paths' in placing the students by recording half scores (i.e., 150, 250, and 350) when they were unwilling to commit to providing a whole score.

Based on input from the teachers and administrators, in 2016, HELP revised the 8-category old rubric with a 4-point scale to a 4-category new rubric with a 20-point scale. The writing prompts were changed from eliciting descriptive writing to argumentative writing. Although aspects of the new rubric have been examined using MFRM in a pilot study (Rock, 2016), to this day, no prior research has investigated the quality of the old rubric and its rater performance.

There is therefore no point of reference for comparing the quality of the old rubric versus the new rubric and their impact on rater performance from an empirical standpoint.

Figure 1

Old Writing Placement Test - Scoring Sheet & Rubric

Writing Placement Scoring Guidelines

	100	200	300	400
Fluency	Very little writing Only several sentences in 30 min			Very fluent language use Writes 1 ½ -2 pages in 30 min
Organization *(not all essays will have paragraphs, even at high levels)*	No intro/conclusion No thesis statement May not have paragraphs	Some paragraphs have a clear main idea	Many paragraphs w/clear main idea TOEFL-like, 5 paragraph organization	Most paragraphs have clear main idea
Flow / Cohesion	Few connectors (or none at all) Ideas are choppy (not related)	Some simple connectors (first, next, finally)	Frequent use of connectors, more advanced, start sentences w/ adj clause	Sophisticated use of connectors
Sentence Complexity	Short, simple or incomplete sentences (4-5 words)	A few different sentence and clause types Primarily writes using "I" as subject	Some use of complex sentence patterns	Variety of complex sentence patterns and embedded clauses
Grammar/ Mechanics	Basic grammar (SVO) May not capitalize/punctuate correctly	Errors often interfere w/comprehension		Errors seldom interfere w/ comprehension
Vocabulary	Limited/simple vocabulary Frequently repeats same words	Limited vocabulary	Uses a variety of general vocabulary items with some low frequency words	Uses very specific, academic, or technical vocabulary
Focused Topic/ Support	Topic is not focused or slightly focused Few (if any) supporting details and Examples	Somewhat focused w/ few details	Focused w/many details to support paragraph main ides	Very focused w/many details
Depth of Content	Lacks ability to express basic idea of exam question	Expresses ideas in very little detail; likely to address the question on a surface level	Expresses some advanced ideas surrounding the question	Expresses many advanced ideas surrounding the question clearly

Highlighted = Most distinguishing features of each writing level. Used to make borderline decisions (see back side of handout)

Writing Placement Test – Scoring Sheet

Student name: _____

Rater name/initials: _____

Category	100	200	300	400	Level
Fluency					
Organization					
Flow/ Cohesion					
Sentence Complexity					
Grammar/ Mechanics					
Vocabulary					
Focused Topic/ Support					
Depth of Content					

Total_____/8

Average _____

Recommended Level _____

The current study aims to address this gap in research by investigating the quality of the old rubric and its rater performance. Specifically, the study is guided by the following research questions:

RQ1. In the context of the HELP writing placement test, how well was the rubric functioning?

RQ2. How did the raters perform using this rubric?

METHODS

Data

The research team accessed archived placement test data from 325 unique students who had been admitted to HELP between the Spring semester in 2012 and Fall semester in 2016.[1] Students who typically attend HELP tend to be college-aged (19–22 years old) and the vast majority tend to come from East Asian countries (e.g., Japan, Korea, and China), with a small minority of students from countries in the Middle East, Europe, and Southeast Asia. For the latter groups of students, the ages tended to skew higher into their 30s and 40s, and there were a few examples of students in their late 50s and early 60s.

A total of 53 teacher-raters were included in the Spring 2012 to Fall 2016 data. The raters at HELP are composed of a mix of longtime administrators, part/full time instructors, and newer instructors, some of whom also serve as graduate teaching assistants. These raters come from a variety of backgrounds, ages, and are equally mixed in gender. Some had joined the program with a long history of L2 English teaching experience, while others were relatively more novice. As each term at HELP was typically held for an 8-week period, there was an increase in instructor turnover throughout the year, particularly with the graduate teaching assistants. This also meant that raters varied in their experience scoring the writing placement test, and each administration featured a different pool of raters. Due to limited resources, especially given that placement decisions have to be made almost immediately after the students have taken the test, training opportunities for new raters are limited.

[1] Although the old rubric was implemented between 2010–2016, the research team was only able to obtain archived data between 2012–2016.

Analysis

An MFRM analysis was used to examine how well the old rubric of the placement writing task functioned. The MFRM measurement investigates multiple measurement facets (i.e., students, raters, categories, and prompts) and enables direct comparisons among them in log-odds units (*logits*). Diagnostic information on (a) the quality of measurement elements such as prompt and students fitting the model and (b) the interactions across facets (e.g., how raters rate different prompts) are also provided. MFRM creates two measures of fit statistics: Infit Mean Square (IMS) and Outfit Mean Square (OMS). Since IMS is more sensitive to non-extreme unexpected ratings because it is weighted by the variance of ratings, we focused on IMS and set the acceptable infit value range as between 0.60 and 1.30 (Myford & Wolfe, 2003). IMS values that are higher than the acceptable range indicate underfitting of the raters; in contrast, values lower than the acceptable range signal overfit[2]. MFRM also allows for detailed analysis of rating scales. Logit-scaled thresholds between score points and probability curves for each score point can be estimated to evaluate the functioning of scales. Please refer to Eckes (2011) for more information on MFRM.

To analyze the data, the FACETS software (version 3.83.6, Linacre, 2021b) was used. A Four-Facet Rating Scale model (RSM) consisting of measurement facets (i.e., students, raters, categories, prompts), were constructed for the old rubric with 7-point scales (original 4 points with consideration of half points used by the raters). Unidimensionality of the datasets was checked using infit values, variance explained by Rasch measurement, eigenvalues, and standardized model residuals (see Linacre, 2021a). As the test developers intended subscales for rubric criteria to be applied similarly, the RSM was chosen for performing the MFRM rather than partial-credit hybrid models. There was an issue with connectivity that was discovered in the prompt facet: Prompt three ("Write about a time in your life when you had a problem. What was the problem? How did you find a solution to the problem?") was less selected by the students than the other two prompts, resulting in a small subset of unconnected data. To resolve the problem, this prompt was specified as a 'dummy' facet (i.e., not estimated) in the RSM, as

[2] A misfit is indicated in MFRM if violations of the theoretical expectations of the model exceed a certain degree of deviations. Underfit means the data is unpredictable and not fitting the model's expectations (e.g., unexpected ratings from the rater). Overfit is the case in which the deviation from the model's assumption is less than expected. For example, when a rater's scores are too predictable (e.g., when generally using a narrow portion of the rating scale), they are considered to be overfitting.

prompts appeared to differ very little in terms of difficulty based on observed score averages and partial MFRM calibrations of the data.

RESULTS

In the following, we present the results of the MFRM analysis of the rubric.

Many-facet Rasch Model fit

The infit values of the measurement model ranged from 0.89 to 1.10 and 0.95 to 1.03 for the category facet and the prompt facet, respectively, indicating unidimensionality of the data based on Smith (1996). Additional evidence for unidimensionality was that 63.71% of variance was explained by the Rasch measure. According to Reckase (1979), above 20% of variance explained by Rasch measure can suggest unidimensionality. Furthermore, a principal components analysis of Rasch model residuals (PCAR) was used to evaluate unidimensionality, and an eigenvalue greater than 2.0 for a contrast was considered as potential evidence of non-unidimensionality (Linacre, 2016). Our results returned a first contrast eigenvalue of 0.30. Tallying standardized model residuals showed that 2% of residuals were over |2.0| and none were over |3.0|, showing good data-model fit (Linacre, 2002).

Measurement Summary

A Wright map is presented to display the summary of the model with all measurement facets (Figure 2). The leftmost column (Measr) of the figure is a ruler indicating each element's associated logits in every facet. The first facet, student, had Rasch-estimated writing ability between -7 and +4 logits, where higher logits suggest greater writing ability. The estimate of Person separation reliability of the student facet was .95, showing the writing test reliably measured students' ability. With a separation reliability of .96, the rater separation index was 6.43, suggesting that raters could be reliably separated into at least 6 different strata in terms of severity. In addition, the maximum estimate of severity was 2.82 logits from the most severe rater, while the minimum estimate of severity, -2.65 logits, came from the most lenient rater; this range of 5 logits could yield a difference of up to 4 score points (of the 7-point scale; namely, 100, 150, 200, 250, 300, 350, 400) on a given category. In other words, for rater, the higher the

value of the logits, the stricter the rater is. The third column of Figure 2 corresponds to the 8 scoring categories, where higher logits indicate greater difficulty to earn a higher score on the category. "Vocabulary" was the most difficult category to receive higher scores (0.31 logits), whereas "Fluency" was the easiest (-0.31 logits). However, the difficulty range of rubric categories was not large, and several categories had nearly identical difficulties. The last column provides information on the scale of scoring points. Each horizontal line in the last column indicates the threshold in logits, that divides adjacent score points on the rating scale. For example, it was more likely for a student with an ability of 1 logit to receive 300 points in each category.

Rating Scale Analysis

Table 1 shows that the (unsanctioned) half points 150, 250, and 350 were used far less often than adjacent rubric-allowed score points. These three score points were half points assigned by the raters, which were not provided in the rubric. Although the average measure and the threshold of score points on the scoring scale seems to be in order (Table 1), each score point did not reflect a distinct range of student ability. Score points 150, 250, and 350 were wholly subsumed by an adjacent point(s) (see category probability curves in Figure 3).

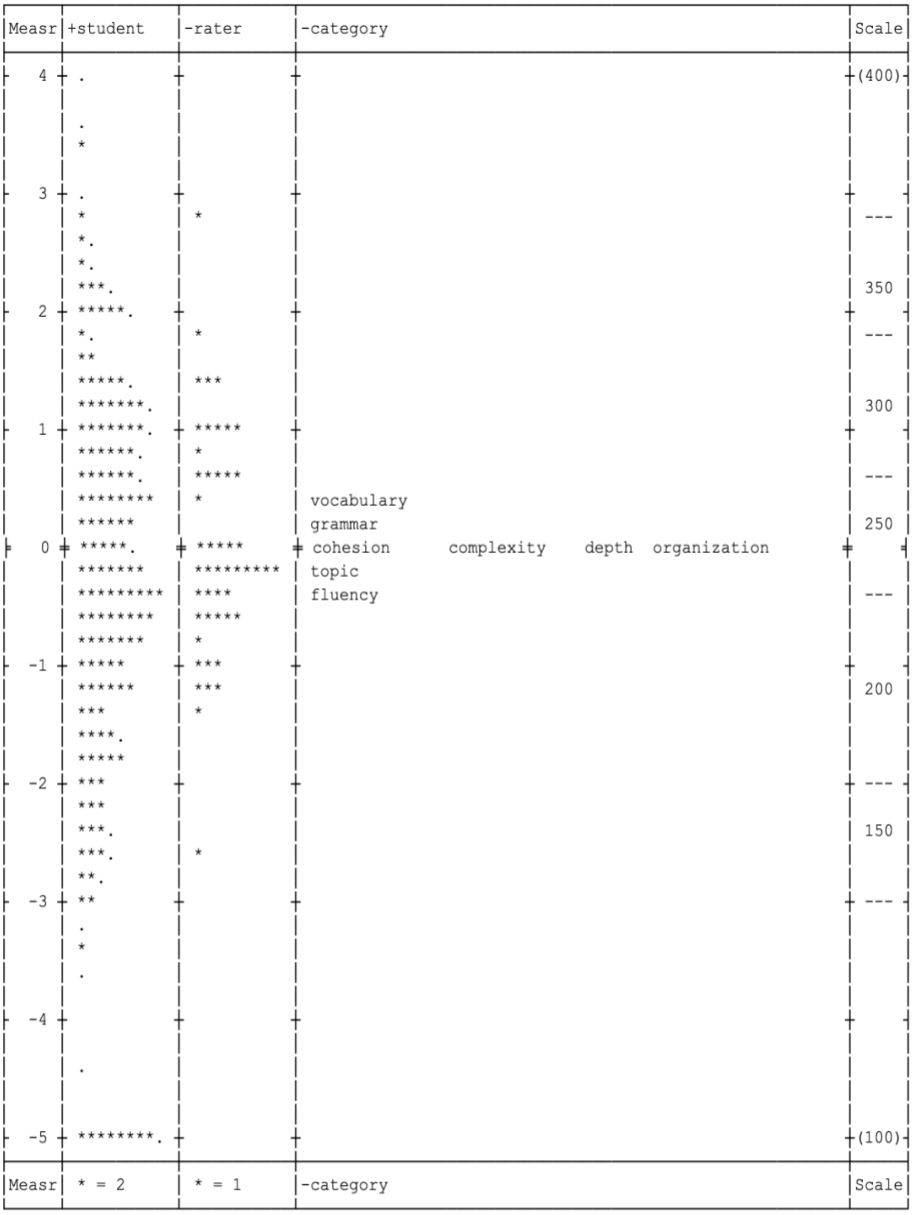
Table 1

Score Point Summary – Old Rubric

Score	% of all scores	Average measure	Outfit Mean square	Rasch-Andrich Thresholds
100	9%	-2.53	1.0	
150	2%	-1.88	0.8	-4.13
200	42%	-0.76	1.2	-3.97
250	5%	0.05	0.8	-1.41
300	32%	0.85	1.0	-0.67
350	1%	1.88	0.4	0.2
400	9%	2.35	1.0	1.92

Figure 2

Variable Map of the Writing Placement Test using the Old Rubric



Rater Fit

Figure 4 illustrates the consistency of individual raters based on their infit and outfit statistics. Ten out of 49 raters (20%) had IMS values outside of the acceptable range, including three raters with values > 2.0, indicating an extreme level of misfit (Linacre, 2016). A somewhat larger proportion of raters had OMS values outside of the acceptable range, including five that were near or exceeding 2.0.

Figure 3

Score Probability Curves for the Old Rubric

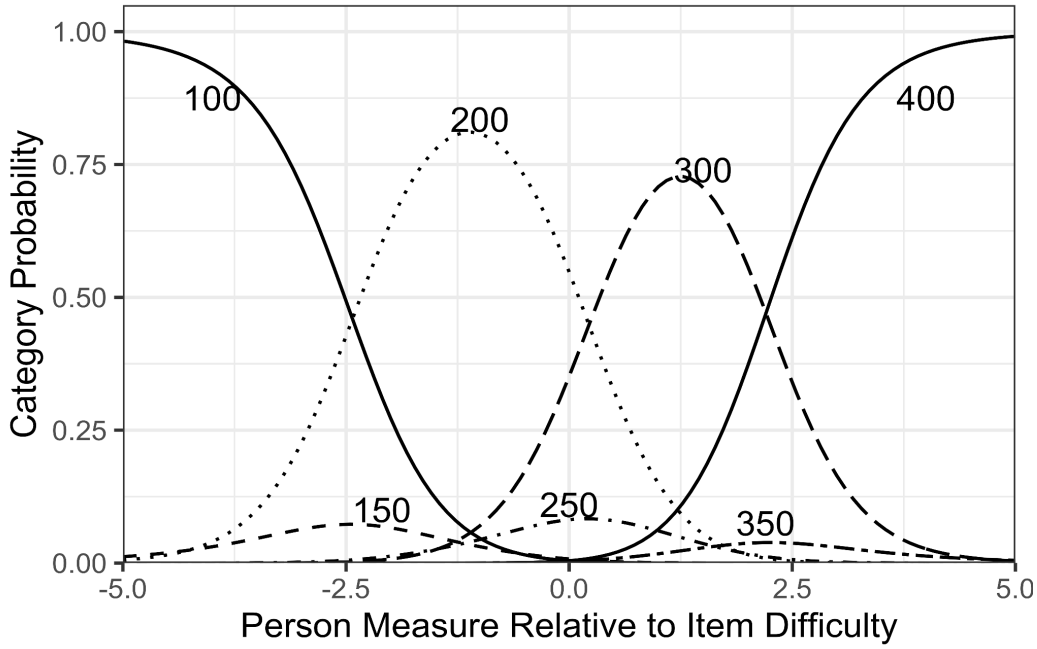
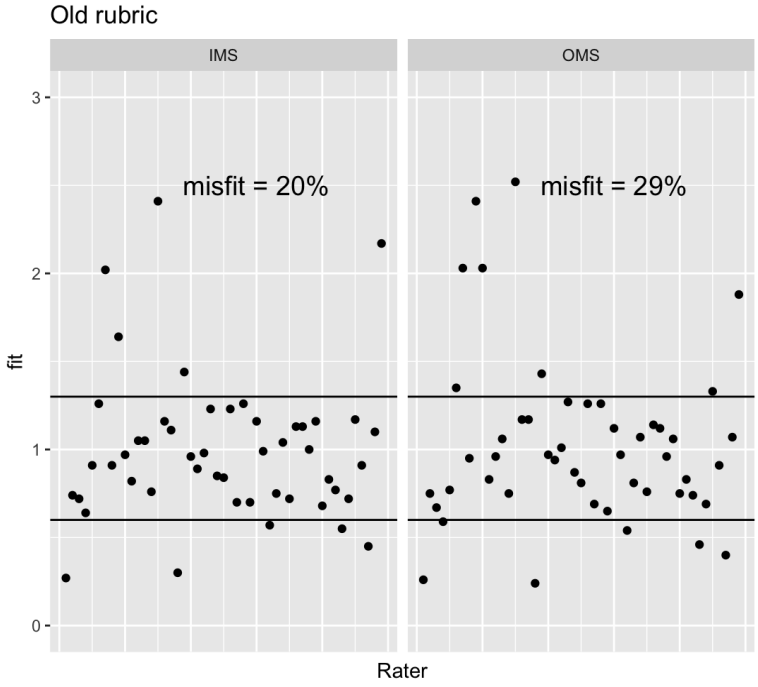


Figure 4

Rater Fit of the Old Rubric



Note. IMS = infit mean square, OMS = outfit mean square.

DISCUSSION AND CONCLUSION

When taking into consideration that the purpose of the HELP writing exam is meant to divide students into the four distinct levels associated with the core curriculum, the old rubric was actually functioning well from a broad perspective. However, when looking more closely at the results of the MFRM analysis, there were potential issues regarding the eight categories and sporadic use of half points that needed to be addressed.

In support of the anecdotal evidence reported in Rock (2016), it seems as though the teacher-raters had a difficult time making a meaningful distinction between the categories of "Flow/Cohesion," "Sentence Complexity," "Depth of Content," and "Organization." Looking more closely at "Sentence Complexity" and "Flow/Cohesion," for instance, the rubric descriptors for the 300-level stating "some use of complex sentence patterns" and "more advanced, starts sentences with adjective clauses" would likely be intrinsically linked, particularly when "some" and "more advanced" are open for interpretation from the teacher-raters. This seems to be the case, in reviewing the Rasch difficulty estimates for this pair, they were treated at almost exactly the same level. If the raters were unable to treat these categories as being distinct, their continued inclusion as part of the rubric would not have resulted in meaningful placement decisions. Additionally, it may be worth noting that the "Fluency" category may have been seen as the "easiest" category to score as it was entirely related to sentences/page length, and interestingly was rather unspecified on the rubric, lacking descriptors for both the 200 and 300 levels; considering this, it is unclear how teacher raters would treat these categories for students who fell into the middle of the range at the 200 or 300 level.

The other potential issue was the use of half points by a few teacher-raters in the data set. It was not such a wide-spread phenomenon that it impacted the ordering of the scoring scale, but rather, the appearance of half points is a reflection of the indecision of particular teacher-raters regarding each category's description. Because the 100 to 400 level on the *Writing Scoring Guidelines* (see Figure 1) were visualized as a range rather than distinct levels, there is the possibility that some teacher raters felt more confident to mark a score that fit within that range. However, since students are ultimately placed within one of the four curriculum levels, these half points were not useful in accomplishing this goal and ultimately undermined consistent criteria and procedures for assigning scores to writing performances.

In regard to how the teacher-raters were performing using this rubric outside of the half point issue, the majority of teacher raters were demonstrating good fit; however, there was a large group of raters that were misfitting. As previously explained, HELP typically has an increased likelihood of turnover rate regarding their instructors each semester and regularly-scheduled placement training sessions were limited when compared to large-scale tests. Considering this, there is evidence to suggest that some teacher-raters may have had difficulties in making meaningful distinctions among the eight categories of the rubric, as well as had difficulties in connecting the descriptors of the categories to specific levels of performance/ability aligning with placements without knowledge of the curriculum. This may have resulted in either the use of half points or the decision to mark the same score across multiple categories (e.g. marking "300" for both "Sentence Complexity" and "Flow/Cohesion").

Therefore, while the initial motivation for revising HELP's rubric in 2016 was based on anecdotal evidence, our empirical analysis also suggests that a revision was necessary and justified. While there is evidence that the rubric used from 2012–2016 was functioning well and the majority of teacher raters were using it as intended, there were notable issues related to likely confusion surrounding the distinctiveness of the eight categories, the potential for the use of half points, and the higher number of misfitting raters that undermine the consistency and meaningfulness of individual scores.

REFERENCES

Bachman, L., & Palmer, A. (2010). *Language assessment in practice.* Oxford University Press.

Brown, J. D. (1996). *Testing in language programs.* Prentice Hall Regents.

Eckes, T. (2011). *Introduction to many-facet Rasch measurement*. Peter Lang D.

Kane, M. (2006). Validation. In R. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). American Council on Education and Praeger.

Kane, M. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1–73. https://doi.org/10.1111/jedm.12000

Knoch, U., & Chapelle, C. A. (2018). Validation of rating processes within an argument-based framework. *Language Testing*, *35*(4), 477–499. https://doi.org/10.1177/0265532217710049

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean. *Rasch Measurement Transactions*, *16*(2), 878.

Linacre, J. M. (2016). Winsteps® (Version 3.92.1) [Computer Software]. Beaverton, OR: Winsteps.com. Retrieved from http://www.winsteps.com/

Linacre, J. M. (2021a) Fit diagnosis: infit outfit mean-square standardized, Retrieved on December 6, 2021 from https://www.winsteps.com/winman/misfitdiagnosis.htm

Linacre, J. M. (2021b). Winsteps® (Version 5.1.1) [Computer Software]. Portland, Oregon: Winsteps.com. Available from https://www.winsteps.com/

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

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational and Behavioral Statistics*, *4*(3), 207–230. https://doi.org/10.3102%2F10769986004003207

Rock, K. (2016). *Development and analysis of a writing rubric for an IEP.* Department of Second Language Studies, University of Hawai‘i (Unpublished manuscript).

Smith, R. (1996). A comparison of methods for determining dimensionality in Rasch measurement. *Structural Equation Modeling*, *3*(1), 25–40. https://doi.org/10.1080/10705519609540027

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