Assessing Writing 26 (2015) 51–66

Contents lists available at ScienceDirect

Assessing Writing

Building a better rubric: Mixed methods rubric revision

Gerriet Janssen[a,b,*], Valerie Meier[c], Jonathan Trace[b]

[a] Universidad de los Andes, Colombia
[b] University of Hawai'i, Mānoa, USA
[c] University of California, Santa Barbara, USA

ARTICLE INFO

Article history:
Received 8 January 2015
Received in revised form 6 July 2015
Accepted 16 July 2015
Available online 6 August 2015

Keywords:
Academic writing
Mixed-methods
Profile analysis
Rasch measurement
Rubrics

ABSTRACT

Because rubrics are the foundation of a rater's scoring process, principled rubric use requires systematic review as rubrics are adopted and adapted (Crusan, 2010, p. 72) into different local contexts. However, detailed accounts of rubric adaptations are somewhat rare. This article presents a mixed-methods (Brown, 2015) study assessing the functioning of a well-known rubric (Jacobs, Zinkgraf, Wormuth, Hartfiel, & Hugley 1981, p. 30) according to both Rasch measurement and profile analysis ($n = 524$), which were respectively used to analyze the scale structure and then to describe how well the rubric was classifying examinees. Upon finding that there were concerns about a lack of distinction within the rubric's scale structure, the authors decided to adapt this rubric according to theoretical and empirical criteria. The resulting scale structure was then piloted by two program outsiders and analyzed again according to Rasch measurement, placement being measured by profile analysis ($n = 80$). While the revised rubric can continue to be fine-tuned, this study describes how one research team developed an ongoing rubric analysis, something that these authors recommend be developed more regularly in other contexts that use high-stakes performance assessment.

1. Introduction

Scoring rubrics are important as they articulate the construct to be performed and measured. Rubrics "help explain terms and clarify expectations" (Crusan, 2010, p. 43). This is to say, the principled choice and use of rubrics is vital, as rubrics optimally link the task, the constructs developed by the task, and the assessment of these constructs. Weigle (2002) describes how the scoring process using rubrics can be particularly "critical because the score is ultimately what will be used in making decisions and inferences about writers" (p. 108). Rubrics can also help mitigate the long-recognized problem of rater variability (cf. Bachman et al., 1995; McNamara, 1996).

Recognizing the importance of rubrics, local program developers – when developing the *Inglés para Doctorados* (IPD; English for Ph.D. students) program and the corresponding IPD Placement Exam used to classify students into the program's courses – decided to use the analytic rubric developed by Jacobs, Zinkgraf, Wormuth, Hartfiel, and Hugley (1981, p. 30) for use with the performance writing component of the placement exam (Janssen et al., 2011). This rubric was adopted because of the strong construct validity it had in terms of proposed course goals and because Weigle (2002, p. 115), had characterized this rubric as being "one of the best known and most widely used analytic scales in ESL". Indeed, the Jacobs et al. (1981)

* Corresponding author at: Universidad de los Andes Depto. de Lenguajes y Estudios Socioculturales Cra 1 No. 18A-12 Bogotá, Colombia. Tel.: +57 1 339 4949x3248.
E-mail address: gjanssen@uniandes.edu.co (G. Janssen).

http://dx.doi.org/10.1016/j.asw.2015.07.002
1075-2935/

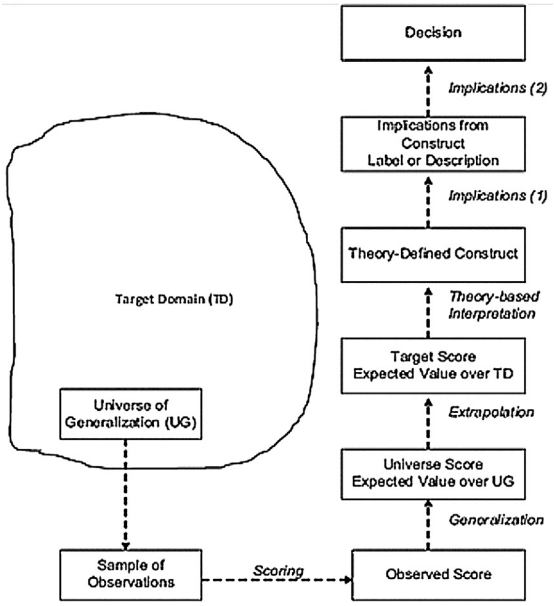

Fig. 1. Chapelle's (2012, p. 23) conceptualization of Kane's IUA.

rubric, in its original or modified form, has been used with some frequency (cf., Brown & Kondo-Brown, 2012; Delaney, 2009; East, 2009; Polio, 2001; Winke & Lim, 2015).

With time, several concerns arose concerning the IPD Placement Exam and its different uses. Though originally designed for use as a placement instrument, Ph.D. program directors began using different course level placements as one admission criterion for entrance into the university's Ph.D. programs. With this unforeseen high-stakes use, program developers began to intensively study different aspects of the exam's validation argument. Of relevance to this current study, in Janssen and Meier (2012) we first realized that the rubric chosen to score the performance writing section, while reliable, was not performing as expected. Indeed, the reliability reported for the Rasch model was 0.99, yet other indicators such as threshold distances (i.e., increments in difficulty) pointed to problems within the scoring bands of the rubric itself: increasing scores were not consistently representative of increased examinee ability (Janssen & Meier, 2012; Meier, 2013). Furthermore, interviews with exam raters, though generally positive in terms of the constructs the rubric represents, revealed other concerns with the rubric, specifically in relation to its ease of use when scoring. Thus, the current study: (a) considers the function of the original Jacobs et al. (1981) rubric; (b) proposes a reformulated rubric that addresses the scoring band problems and answers the raters' call for added simplicity; and (c) analyzes the functioning of this revised rubric.

2. Literature review and research questions

2.1. Validation

Following the work of Bachman and Palmer (2010), Chapelle (2008, 2012), and Kane (2006, 2013), IPD program developers have been building a validation argument for the uses of the IPD Placement Exam using an argument-based validity framework. Kane (2013) presents six sequential inferences that are typically addressed in the interpretation-use arguments (IUAs) for placement exams: scoring, generalization, extrapolation, theoretical, and two levels of implications. These inferences should be defended in the IUAs of most tests, though part of what makes the argument-based approach to validation so powerful is that the arguments claimed within each IUA should adjust themselves to the interpretations and uses found within the specific assessment context. Chapelle's (2012) helpful depiction of Kane's IUA has been included as Fig. 1.

In this paper, we focus on elements of the scoring inferences within the IUA. The scoring inference focuses on the scoring procedures and the application of these procedures to ensure that they are appropriate, accurate, and consistent (Kane, 2006, pp. 24, 34; Kane, 2013, p. 25). Clauser (2000) provides an in-depth description of several important components of the scoring inference of appropriacy that should be evidenced. Three key components to demonstrate include determining: (a) if the constructs developed within the rubric are appropriate to the larger construct being evaluated in this exam section; (b) whether the criteria used for evaluation are appropriate; and (c) if these are being applied in an appropriate fashion. The appropriateness of the rubric's constructs and criteria of evaluation can be evaluated by field experts; the appropriateness of application can be judged using Rasch measurement, which provides test developers with a variety of analyses (e.g., bias, fit/misfit, reliability, scale analysis) that can be done to help demonstrate how the test is functioning, and to what

degree the results are generalizable (Barkaoui, 2013; Bond & Fox, 2007). *Accuracy* and *consistency* can also be assessed using multi-faceted Rasch measurement.

2.2. Performance assessment

Kane, Crooks, and Cohen (1999, p. 7), write that the "defining characteristic of performance assessment is the close similarity between the type of performance that is actually observed and the type of performance that is of interest". Performance assessment is interesting to us as it – at its best – offers stakeholders a testing option that is both meaningful and transparent (Lane & Stone, 2006, p. 387), especially since it permits "direct alignment between assessment and instructional activities" (Lane & Stone, 2006, p. 389).

However, performance assessment can jeopardize the scoring inference of an exam's validation argument, since rater variability can introduce construct-irrelevant variance into the score (Lane & Stone, 2006; McNamara, 1996), with the result that the final measurement poorly reflects the original construct being measured (Messick, 1989). Barkaoui (2007) noted a "traditional concern with controlling for task and rater variability as 'sources of measurement error'" (p. 99), and rater effects have continued to be of central interest, with a number of recent studies attesting to rater variability. Eckes (2008) was able to organize raters of the TestDaF into significantly different categories, with scoring severity being one important factor in his categorization. Schaefer (2008) described how in his context, raters were more biased (either positively or negatively) toward more advanced writers compared with beginning writers. Winke et al. (2013) showed how having a mutually-shared L1 background may increase rater leniency during scoring, while Huang (2008, 2012) showed how the reliability of raters' scores was notably different when addressing the writing of ESL and native speaking students. Rater training can attempt to limit this variability, though research into the effectiveness of training has produced mixed results: while Lim (2011) reported that longitudinally, raters can gain increased precision, Knoch (2011) found otherwise.

2.3. Rubric development

When facing the task of developing a rubric for an assessment, many teacher practitioners employ the strategy of "adopt and adapt" (Crusan, 2010, p. 72), taking an intact rubric and modifying it for the local assessment context. This practice is likely to be adequate for most classroom assessments, yet with more high-stakes uses, rubric scoring criteria will be typically elaborated by field experts familiar with the assessment context (Clauser, 2000). Ideally, rubrics will reflect a progressive development of the skill in question (Lane & Stone, 2006); the validity of a score produced by the scoring rubric will be greatly increased according to the degree to which different levels of the rubric reflect different levels of proficiency, as based in current theories of learning (Kane, 2013) or theories of writing. While researchers have argued that rubric scales should be grounded in theory (Knoch, 2011; McNamara, 1996), adequate theories may not yet exist; Knoch (2011, p. 81), for example, has noted that "none of the theories/models available for this purpose are sufficient by themselves for a scale of diagnostic writing".

In terms of different approaches to rubric development – or in our case, revision – several classifications have been proposed. Fulcher (2003) described how rubric revision or development is based on either intuitive or quantitative processes. In his framework, intuitive processes are expert- or experience-based, while quantitative processes are data-based or data-driven. Hawkey and Barker (2004, pp. 128–129) reported the Common European Framework's slightly more refined conceptualization of three rubric development methodologies: intuitive, qualitative, and quantitative. Intuitive methodologies are when rubrics are based on other rubric samples, reflecting the experience of the rubric developers. Qualitative methodologies typically rely upon focus groups to provide information about the characteristic features of different levels of writing and how these should best be articulated in the rating scale. Quantitative methodologies rely upon empirical methodologies, such as Rasch measurement, to relate test taker proficiencies with rubric descriptors on an integer scale (CEF, 2001, in Hawkey & Barker, 2004, p. 128).

The literature reveals that a variety of different empirical methods have been used during rubric development. Very early on, Fulcher (1987) described how a rating scale should consist of data-based criteria. These data-based criteria can rely upon discourse analyses of texts and their important features. Studies by Knoch have looked at empirically grounding a rubric in features of topic structure analysis (Knoch, 2007) or discourse markers (Knoch, 2008). Zhao (2013) empirically developed a rubric for authorial voice, while Harsch and Martin (2012) described how they adapted a local scale to fit CEFR descriptors using a data-driven approach. Our rubric development methodology was predominantly quantitative, using Rasch measurement to evaluate the degree to which well the scale was functioning and in which parts; intuitive methods based on expert experience and the study of different text samples were used to help refine rubric descriptors.

2.4. Research questions

Given the above concerns about adapting a rubric for high-stakes contexts, the following study began with an analysis of the original Jacobs et al. (1981) rubric. This analysis suggested potential revisions to the scoring bands within each rubric category. To assess the effectiveness of these revisions, we examined the revised rubric in terms of how it both changed as

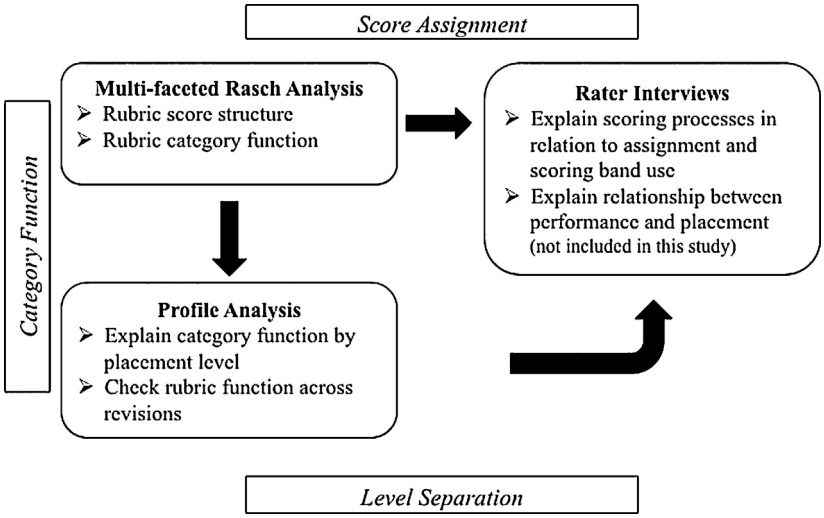

Fig. 2. Mixed-methods conceptual design model.

well as retained its original function as a measure of second language writing proficiency. To this end, the following research questions were posed:

(1) How were rubric category scales of the original Jacobs et al. (1981) rubric functioning when applied to the IPD Placement Exam writing section?
(2) How can the rubric category scales be restructured to be more accurate and efficient?
(3) How does the revised rubric function when classifying examinees who took the IPD Placement Exam writing section?

3. Methods

3.1. Mixed-methods approach

This study utilized a mixed-methods research (MMR) approach to gathering and reporting data through the complementary application of both quantitative and qualitative research methods. While MMR designs can take several forms, we adopted a quantitative mixed design (Brown, 2015; Johnson, Onwuegbuzie, & Turner, 2007), which "includes both qualitative and quantitative elements, but is predominantly quantitative" (Brown, 2015, p. 9). This study uses an explanatory design (Creswell and Plano Clark, 2007), as qualitative methods were used primarily to provide explanations for the initial quantitative results (also known as simultaneous triangulation; Morse, 1991).

Importantly, this design differs from multi-method or within-methods triangulation (Denzin, 1978; Hashemi, 2012; Johnson et al., 2007) because the methods produce "complementary strengths and nonoverlapping weaknesses" (Johnson et al., 2007, p. 127). Put simply, our goal was to support interpretations of the initial Rasch analyses by (a) using profile analysis to both confirm and preserve rubric function throughout the revision process and (b) draw on rater interviews to explain practically what the scoring processes had identified quantitatively. Following the advice of (Brown, 2015, p. 165), we present the conceptual design of the methods used in this study in visual form in Fig. 2.

3.2. Original rubric

The Jacobs et al. (1981) rubric scores writing performance according to five constructs (what we call rubric categories): Content, Organization, Vocabulary, Language Use, and Mechanics (Fig. 3 below presents a portion of this rubric). Jacobs et al. (1981) divided each construct-category into four broad ability bands or performance levels: excellent to very good; good to average; fair to poor; very poor. Accompanying each ability band are written descriptors and ranges of possible scores to be allocated according to the ability band. Each category has a different maximum score "to encourage readers (raters) to focus most of their attention on those aspects of the composition which reflect the writer's ideas" (Jacobs et al., 1981, p. 103). Each category also has a different range of possible scores: Mechanics has nine possible scores while Language Use has 21.

3.3. Rubric analysis, revision, and re-scoring overview

The functioning of the original Jacobs et al. (1981) rubric was first analyzed using multi-faceted Rasch measurement (MFRM) on operational test scores that had been collected over three years ($n = 524$). Given the wide range of possible scores

ESL COMPOSITION PROFILE

STUDENT DATE TOPIC

SCORE	LEVEL	CRITERIA	COMMENTS
CONTENT	30-27	EXCELLENT TO VERY GOOD: knowledgeable • substantive • thorough development of thesis • relevant to assigned topic	
	26-22	GOOD TO AVERAGE: some knowledge of subject • adequate range • limited development of thesis • mostly relevant to topic, but lacks detail	
	21-17	FAIR TO POOR: limited knowledge of subject • little substance • inadequate development of topic	
	16-13	VERY POOR: does not show knowledge of subject • non-substantive • not pertinent • OR not enough to evaluate	
ORGANIZATION	20-18	EXCELLENT TO VERY GOOD: fluent expression • ideas clearly stated/supported • succinct • well-organized • logical sequencing • cohesive	
	17-14	GOOD TO AVERAGE: somewhat choppy • loosely organized but main ideas stand out • limited support • logical but incomplete sequencing	
	13-10	FAIR TO POOR: non-fluent • ideas confused or disconnected • lacks logical sequencing and development	
	9-7	VERY POOR: does not communicate • no organization • OR not enough to evaluate	

Fig. 3. Screenshot of two subsections of the Jacobs et al. (1981) rubric. Note that each rubric category has four broad levels and a total of seven sublevels. These levels and sublevels oriented the restructuring of the rubric scale in this study.

described above, it is not surprising that this analysis revealed that frequently there were no clearly defined differences between neighboring scores on the scale. Adopting a data-driven rubric development perspective (Fulcher, 2003; Hawkey & Barker, 2004), we first focused on creating a more efficient scale structure by using MFRM to determine the optimal number of steps in each scale. After deciding upon the number of steps in the rubric scale, we used intuitive processes (Fulcher, 2003; Hawkey & Barker, 2004) to adapt the scoring descriptors to reflect the levels within each rubric category. We first considered a set of guidelines that local exam raters had developed by extending the descriptions found in the original Jacobs et al. (1981) rubric with reference to specific test samples from the local context. We then reworked these descriptors to omit features that appeared across multiple categories and sharpen distinctions across the different rubric categories and between the varied levels of performance. These refinements were made iteratively, with the contribution of all three authors and the feedback of the local exam raters.

To evaluate how the revised rubric functioned in comparison to the original rubric, 80 essays were randomly selected for re-scoring. These 80 essays were rated by two of the authors using both the original and the revised rubrics. The authors counterbalanced their rating to minimize potential treatment order effects: 40 essays were first rated using the original rubric, and the other 40 essays were rated using the revised rubric first. The authors independently rated approximately 10 essays and then discussed discrepant scores over Skype, as they were in different locations. With the original Jacobs et al. (1981) rubric and following local rating practices, all category scores that varied by more than two points were negotiated, while with the revised rubric, all category scores that did not exactly agree were negotiated. All 80 essays were rated by the author team over the course of approximately one month.

3.4. Multi-faceted Rasch measurement

We used FACETS (v3.67, Linacre, 2010) to conduct all multi-faceted Rasch measurements. Rasch models were constructed for (a) the original data set ($n = 524$) scored by local raters, (b) the 4-, 6-, and 7-point re-scaled rubrics based on the same data set ($n = 524$), and (c) the subset of essays ($n = 80$) scored by the authors using both the original and revised rubrics. All models incorporated three facets: raters, examinees, and rubric categories.

Rasch models permit the inclusion of many facets into one statistical model. Crucially, these facets can then be compared against each other (Bachman, 2004, pp. 141–142; Bond & Fox, 2007, p. 147) as they are mapped within a single graphic, the vertical ruler, which permits easy visual interpretation by experts and non-experts alike (see Eckes, 2008, 2011; Knoch, 2009; Schaefer, 2008; Sudweeks et al., 2005; Weigle, 1998 for studies on the rating of academic writing that present Rasch rulers). Other benefits of Rasch analysis include "sample-free item calibration, item-free person measurement, misfitting item and person identification, and test equating and linking" (Ellis & Ross, 2013, p. 1269). Lynch and McNamara (1998) write that "using the microscope as an analogy, (Rasch modeling) turns the magnification up quite high and reveals every potential blemish on the measurement surface" (Lynch and McNamara, 1998, pp. 176–177). For two excellent treatments of MFRM and assessment, see Barkaoui (2013) and Yen and Fitzpatrick (2006).

Table 1
Three rubric studies' category measures, fit statistics, separation values.

Rubric, Categories	Measure	SE	Infit MS	Separation	Reliability	χ^2
Jacobs et al. (1981) original raters (n = 542)						
Content	.05	.02	1.43			
Organization	−.08	.03	.86			
Vocabulary	−.03	.03	.66			
Language use	−.50	.02	.89			
Mechanics	.57	.03	1.16			
				13.07	.99	.00
Jacobs et al. (1981), author team (n = 80)						
Content	.85	.06	1.41			
Organization	−.32	.08	.91			
Vocabulary	.76	.07	.85			
Language use	−.40	.06	.63			
Mechanics	−.89	.10	1.16			
				8.95	.99	.00
Revised rubric, author team (n = 80)						
Content	−.23	.16	1.17			
Organization	.34	.17	.98			
Vocabulary	.56	.15	.90			
Language use	.83	.15	.66			
Mechanics	−1.5	.18	1.19			
				5.03	.96	.00

Note. The second and third data sets also considered in Meier, Trace, and Janssen (forthcoming).

3.5. Profile analysis

Profile analysis was used to examine the consistency of placement and variable function across different versions of the rubric. This analysis is a multivariate form of repeated measures ANOVA and is primarily used to determine patterns and differences (i.e., profiles) among multiple groups and variables. Profiles in this case refer to the descriptive statistical performances of different groups (e.g., the placement levels) "measured on several difference scales, all at one time" (Tabachnick & Fidell, 2013, p. 314). This form of analysis can be a useful tool for making comparisons about the performance profiles of different placement groups in the program. The benefit of this analysis is that it allows us to see how placement levels interact with each of the rubric categories, both as a way of examining expected differences as well as identifying similarities in performance across both the original and revised rubrics.

3.6. Qualitative analysis

Raters' discussions of discrepant scores were used to better understand how the rating scale structure of the original and revised rubric might influence raters' decision-making processes. 15 of 16 negotiation sessions were audio recorded and four sessions were selected for transcription and analysis. In two of these sessions, raters had rated the same 10 essays using first the original and then the revised rubric; in the other two sessions, raters had rated a different set of nine essays using first the revised rubric and then the original rubric. This made it possible to compare raters' discussions of the same essay using both rubrics and to explore broad similarities and differences across the raters' use of the different rubrics

4. Results

4.1. Multi-faceted Rasch measurement

4.1.1. Data fit

Good model fit is critical to establish, as all subsequent analyses depend upon the degree to which the data fit the model. Though there are a variety of available fit statistics (cf. van der Linden & Hambleton, 1997, p. 12), we judged model fit based on infit mean square (IMS) measures, which are an indicator "of how well each item fits within the underlying construct [model]" (Bond & Fox, 2007, p. 35). Should the different measures fit the model, "the output can be interpreted as interval level measures. . .. item estimations may be held as meaningful quantitative summaries of the observations" (Bond & Fox, 2007, p. 35).

For all three Rasch models, we judged rubric categories to adequately fit if their IMS values fell between 0.60 and 1.40, the values Bond and Fox (2007) propose are reasonable for rating scales (p. 243). Values of 1.00 indicate that there is an exact correspondence between the model and the data. Values higher than 1.40 are said to misfit (i.e., the measure falls outside of the expected model; there is 40% more variance in the data than the model predicted), whereas values lower than 0.60 are said to overfit (i.e., the measure is less chaotic than what is expected by 40%) (Bond & Fox, 2007, p. 310). Table 1 presents IMS values for all three models constructed for this paper. As can be seen, the IMS values are, broadly speaking, within the suggested range; reliability coefficients likewise show that the different models are highly reliable. Of interest is that the

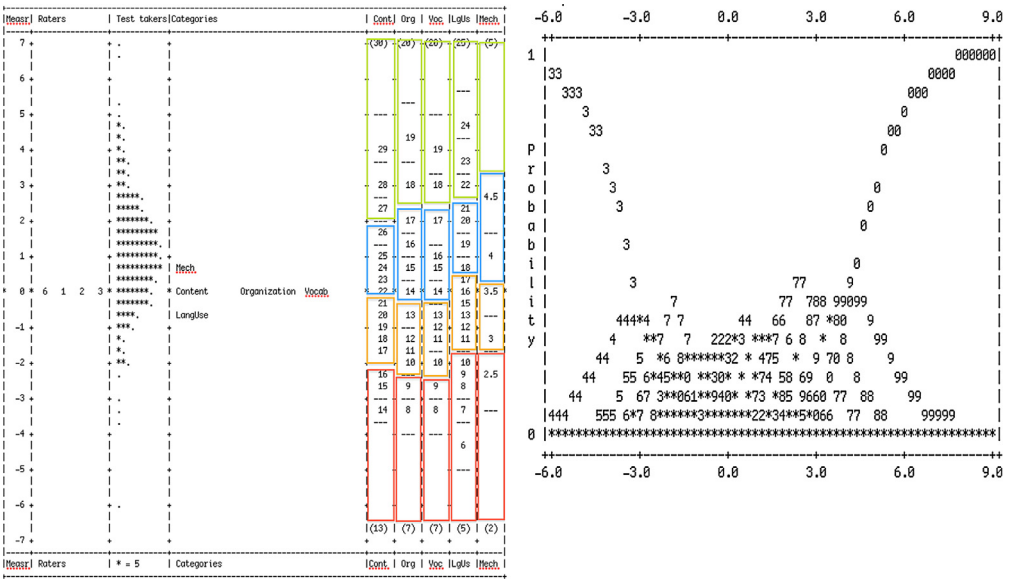

Fig. 4. Vertical ruler (left) and category curve responses (right) for the original data set ($n = 542$), modeled from scores assigned by the original exam raters. In the vertical ruler, each of the rows of boxes represents a broad level within the Jacobs et al. (1981) rubric. Note that the two middle score bands have narrow clusters of many possible scores, while the top and bottom score bands are more disperse. In the category curve response figure for Content (right), one can once again see how middle scores are seemingly indiscriminate.

original Jacobs et al. (1981) models have nearly the same IMS values for all categories, while the revised rubric (the third model) has IMS values that are somewhat more centered, indicating a slightly more stable model being produced by data gathered using the revised rubric.

Table 1 also presents an interesting trend concerning standard error (SE) values and separation. As can be seen, the original Jacobs et al. (1981) rubric, as scored by the original exam raters, has SE values of .02 or .03. Because of this high precision, the model then suggests that 13 statistically different levels of performance are able to fit within this data set (i.e., the separation calculation). In the second model (the author team using the original rubric), one can see higher SE values, between .06 and .10. Because the SE values are higher, fewer statistically different levels (8.95) are able to fit within what we assume to be an approximately similar data set. Finally, in the third model, (the author team using the revised rubric), the SE values are higher still, which means that only five statistically different levels of performance fit within this data set. Despite the increased SE values in the second and third models compared to the original model, the reliability coefficient is still above .95, which Kubiszyn and Borich (in press) state is "acceptable" for a standardized test (1993, p. 353).

4.1.2. Original rubric function

Despite the high reliability and the low SE values, other MFRM output indicated that the original rubric was not functioning adequately. To get an initial diagnosis of how the Jacob et al. (1981) rubric's rating scale was functioning in both the original and revised rubrics, we began by studying the vertical rulers produced by MFRM. The vertical rulers visually represent the average measure of difficulty, measured in logit units, for each score within a rubric category scale (see Fig. 4, left side); comparing these average measures is one of "the simplest way(s) to assess category functioning" (Bond & Fox, 2007, p. 222). An initial inspection of the vertical ruler revealed clusters of scores for each rubric category at nearly the same ability level (observe the two middle bands in Fig. 4), suggesting that there was redundancy within the rubric. This observation was substantiated by an inspection of the category response curves, in which we sought evenly spaced, well-defined peaks, with minimal overlap between adjacent scores. These peaks should not be "overshadowed and redundant" with other curves. Evenly spaced, well-defined peaks illustrate that "each (score) is indeed the most probable (score) for some portion of the measured variable" (Bond & Fox, 2007, p. 224). For example, Fig. 4 (right side) shows that there were no distinct category response curves for each score within Content, but instead an indistinguishable blur of scores.[1]

As a final step, we identified redundant scores based on the average step difficulty for adjacent scores and threshold distances between scores. The threshold distances (see Table 2) represent the distances between each step difficulty (i.e., the difficulty one score represents within the model) and are calculated by subtracting the next higher step difficulty measure from the previous one. Threshold distances should be large enough that "each step defines a distinct position on the variable" (Linacre, 1997); that is, each step should correspond to a distinct segment of test taker ability. Guidelines recommend that

[1] Given space considerations, we are not able to present response curves for all five rubric categories; here and throughout the paper, the results for Content illustrate problems found to a greater or lesser extent with all five categories.

Table 2
Scale structure: content.

Score	*n*	Ave. measure	Outfit MnSq	Step difficulty	SE	Threshold distance
16	10	−1.63	3.0	−2.23	.34	.33
17	24	−1.45	1.1	−2.49	.28	−.26
18	21	−.92	1.2	−1.14	.22	1.35
19	30	−.68	1.0	−1.31	.20	−.17
20	44	−.42	2.4	−1.03	.17	.28
21	54	−.03	1.7	−.57	.14	.46
22	90	.02	1.0	−.59	.12	−.02
23	97	.53	1.7	.14	.11	.73
24	129	.86	1.2	.27	.10	.13
25	117	1.14	1.5	1.02	.10	.75
26	126	1.46	1.5	1.26	.10	.24
27	138	1.87	1.5	1.71	.10	.45
28	83	2.60	1.0	2.86	.12	1.15
29	53	3.42	1.3	3.47	.16	.61

Note. This data was originally discussed in Meier (2013); $n = 524$.

these threshold distances be more than 1.4 but less than 5.0 logits (Bond & Fox, 2007, p. 224); however, as the step difficulties often have large standard errors, the threshold distances should not be interpreted too literally. Still, it is quite clear that for all five rating scales, very few threshold distances come close to meeting the recommended minimum of 1.4 logits. This suggests that there is not a clear psychological representation in the minds of the raters for each of the different scores, something that is certainly plausible when a rubric category has some 15 or 20 possible points. It was exactly this unruliness within the rubric that suggested to us that the scale be resized to include fewer possible scores. A full report of this is available in Meier (2013).

4.1.3. Rubric revision

Taken together, the vertical ruler, category response curves, and category score measures all pointed to the same interpretation: the rubric category scales contained too many possible scores. In this situation, Bond and Fox (2007, p. 222) suggest: "a general remedy is to reduce the number of response options (i.e., possible scores) by collapsing problematic categories (scores) with adjacent, better-functioning categories (scores), and then to reanalyze the data" (see also Eckes, 2011, p. 84). Consequently, to determine the optimum number of steps (i.e., scores) for each category scale, we applied MFRM to the original data and experimented with combining adjacent scores to produce 4-, 6-, and 7-point scales.

We began by constructing a 4-point scale, which reflected the four major ability levels of the original Jacobs et al. (1981) rubric, and a 7-point scale, which reflected the seven sub-levels of the original rubric (see Fig. 2). We then evaluated these new scales by inspecting vertical rulers and category response curves (see Fig. 5 below) as well as threshold distances. The 4-point scale was very stable, with clearly defined peaks for each score, but we were concerned that it erased possible distinctions between different proficiencies. On the other hand, the 7-point scale still seemed to contain too many scores for each one to correspond to a distinct segment of test taker ability; for example, the score "3" for Content is subsumed by adjacent scores in Fig. 5, an observation substantiated by the threshold distance values (these figures not reported here). Thus, we rescored the rubric according to a 6-point scale, which allowed us to maximize the number of different possible scores, without losing distinction between scores. From a psychometric standpoint a 6-point scale maximized the number of meaningful levels within each category, while maintaining broad qualitative distinctions between different performance levels.

4.1.4. Revised rubric function

Having decided upon a revised rubric with six possible scores per category (for Mechanics, we decided to continue with four possible scores, which maintained the original scoring structure), this new rubric was piloted using 80 randomly selected essays from the original sample. These essays were scored by two of the authors using the original Jacobs et al. (1981) rubric in addition to the revised rubric.

Fig. 6 presents vertical rulers that compare how the original rubric functioned (left) for 80 essays rated by these authors compared to the revised rubric (right). What is clear is that once again the original rubric results in many scores that are highly clustered, especially in the mid-score ranges. However, the revised rubric (right) presents scores that are more evenly spaced. A similar trend is visible in Fig. 7, which presents the category curve responses for one rubric category. Quite clearly, the category curves for Rasch models based on the original rubric (left) demonstrate heavy overlap in the mid-range scores, while models created using ratings from the revised rubric produce category curves that have distinct peaks and smaller areas of overlap. This indicates that in this test context, the revised rubric provides an improvement compared to the original Jacobs et al. (1981) rubric.

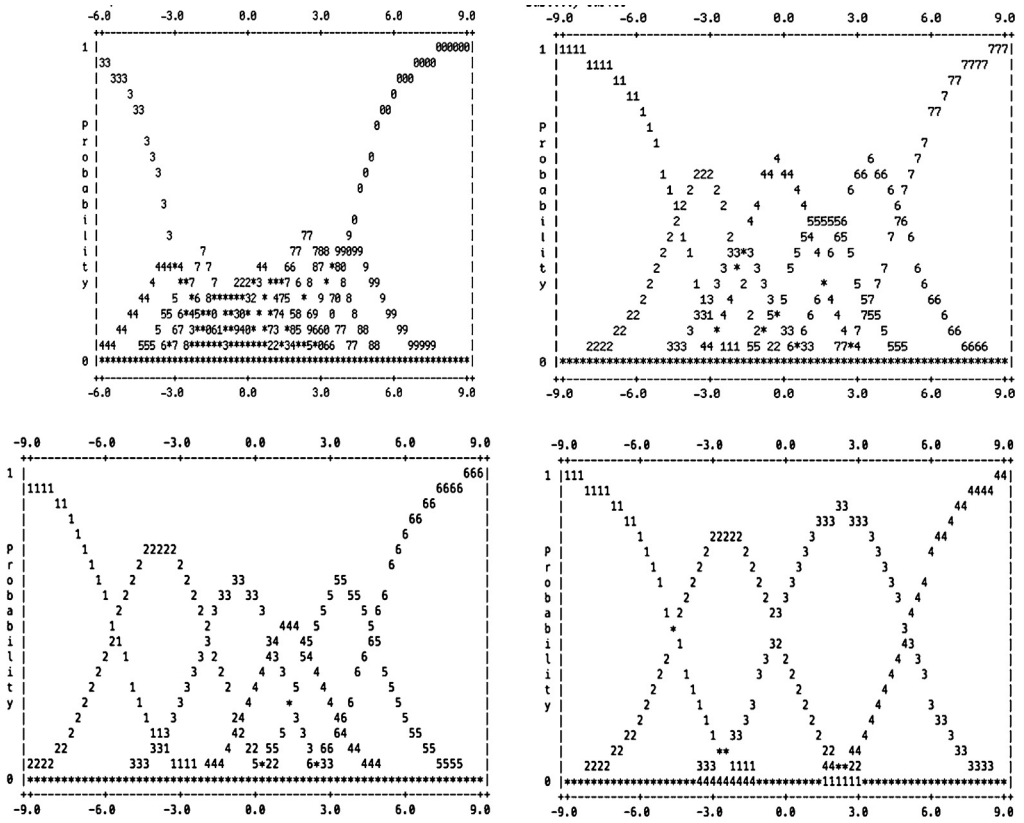

Fig. 5. Category response curves for Content. All curves formed using the original data set ($n = 542$). Top left, original Jacobs et al. (1981) rubric; top right, 7-point scale; bottom left, 6-point scale; bottom right, 4-point scale.

4.2. Profile analysis

Characteristics of placement and rubric category function were determined for both the original and revised rubrics using profile analysis. Mean comparisons in profile analysis are reported in terms of levels, flatness, and parallelism. Levels correspond to differences in placement levels, or between-subjects effects, while flatness describes differences in rubric categories (i.e., within-groups main effects). Finally, parallelism is related to how placement levels vary by rubric category (i.e., within-groups interaction effects).

4.2.1. Assumption checking

Before running the analyses, assumption checks were carried out for both the original ($n = 524$) and revised ($n = 80$) data sets. As a preliminary step, both sets of scores were converted to percentage scores for each rubric category. This was done because the original Jacobs et al. (1981) rubric has different score ranges for each category (e.g., Content scores range between 13 and 30, while Organization scores range between 7 and 20), while the revised rubric uses a 6-point scale for all categories. In order to make comparisons, percentage scores were used to place all categories on the same scale.

Sample size assumptions for profile analysis are similar to those for ANOVA in that there should be as many data points per group as there are dependent variables. As there were five dependent variables – one for each rubric category – all groups should have a minimum of $n = 5$. For both data sets, sample sizes per group were all above $n = 15$, with the exception of the Pre-IPD course level, which was $n = 4$ for the revised rubric. While ANOVA is quite robust against sample size violations (Tabachnick & Fidell, 2013), the fact that this one group does not meet the minimum requirements indicates that we should be careful making interpretations at the lowest level of the revised rubric.

To check assumptions about normality and outliers, descriptive statistics were run for both data sets (Table 3). None of the distributions were found to be markedly non-normal, and neither univariate nor multivariate outliers were located. Lastly, due to the fact that there were unequal sample sizes per placement level, a Box's M test was conducted to check the homogeneity of variance-covariance matrices. The results indicated that there were significant differences for the original ($F(60, 1109464.43) = 4.15$, $p = .00$) but not the revised rubric ($F(45, 10425.19) = 1.02$, $p = .44$).

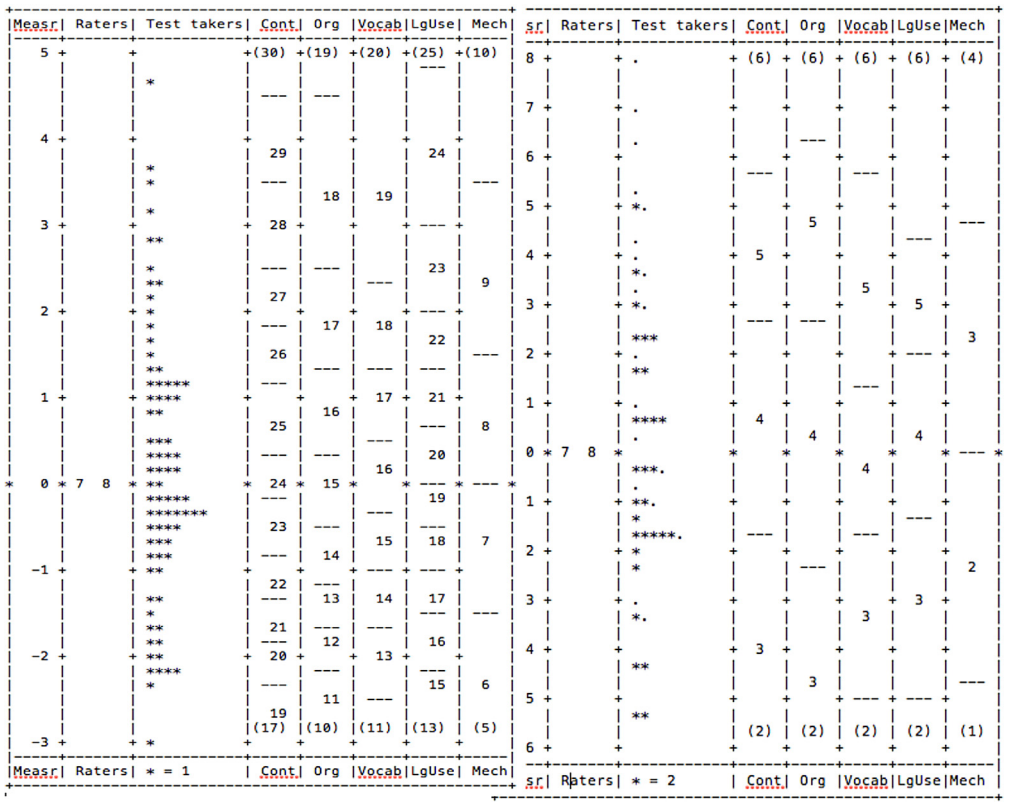

Fig. 6. Vertical rulers comparing the Jacobs et al. (1981) rubric (left) and the revised rubric (right). Both of these analyses were based on scorings of 80 randomly selected essays completed by this author team.

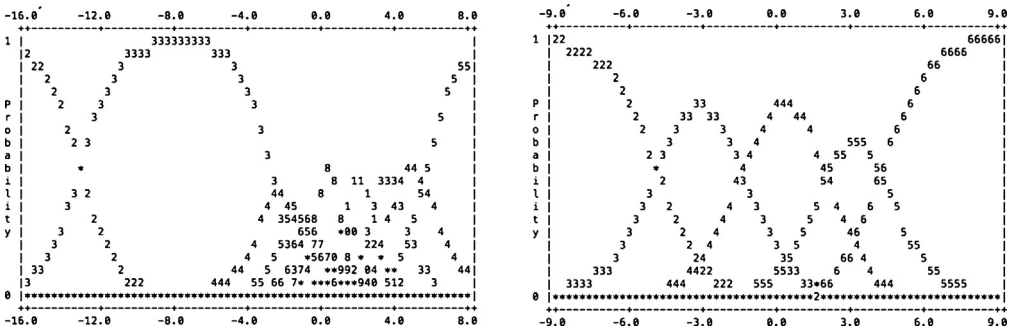

Fig. 7. Category response curves for the Jacobs et al. (1981) rubric (left) and the revised rubric (right). Both of these analyses were based on scorings of 80 randomly selected essays completed by this author team. It is worth noting that only five of six possible scores were assigned using the revised rubric.

Table 3
Descriptive statistics for the original and revised rubrics.

Original ($n = 524$)	M	SD	Min	Max	Skewness	SE of Skew
Content	24.34	3.45	13	30	−0.84	0.11
Organization	15.73	2.42	7	20	−0.88	0.11
Vocabulary	15.70	2.49	7	20	−0.84	0.11
Language Use	19.23	3.44	5	25	−1.09	0.11
Mechanics	3.74	0.70	2	5	−0.24	0.11
Revised ($N = 80$)						
Content	3.98	0.92	2	6	0.43	0.27
Organization	3.92	0.83	2	6	0.23	0.27
Vocabulary	3.89	1.04	2	6	0.00	0.27
Language use	3.73	1.07	2	6	0.37	0.27
Mechanics	2.83	0.77	1	4	−0.16	0.27

Table 4
Summary of profile analysis for the original rubric.

Source	SS	df	MS	F	p	η^2
Between groups						
Placement (Levels)	235984.15	4	58996.04	303.38	.00	.70
Error	100925.53	519	194.46			
Within groups						
Category (flatness)	9515.18	2.72	3495.14	53.77	.00	.09
Category × Placement (parallelism)	5363.95	10.89	492.58	7.58	.00	.06
Error	91846.93	1412.93	65.01			

Note. Due to a significant value for sphericity ($p = .00$), values represented for within-group tests are based on Huynh–Feldt methods. $n = 524$.

Table 5
Contrasts for adjacent levels for the original rubric.

IPD level		M difference	SE	P
0	1	14.38	1.21	.00
1	2	7.18	0.82	.00
2	3	5.96	0.79	.00
3	4	7.05	0.79	.00

Notes. $n = 524$. The course Pre-IPD is level 0 in this table.

4.2.2. Original rubric

A summary of the findings from a profile analysis of the original rubric ($n = 524$) is shown in Table 4. The results indicate that statistically significant differences were found for placement levels ($F(4,519) = 303.38$, $p = .00$), with a relatively large effect size of $\eta^2 = .70$. In order to determine where these differences were occurring, post-hoc contrasts were run using a Scheffe adjustment to account for multiple comparisons. Table 5 displays contrasts for adjacent placement levels only, as these are what we are most interested in, and we can see that significant differences were found at $p = .00$ between all levels. This indicates that in terms of scores on the rubric, levels are adequately separated and distinct, which is what we would ideally hope to see given the placement nature of this assessment tool.

Statistically significant differences were also observed for flatness and parallelism. Flatness, which is related to differences in rubric categories, was significant at $F(2.72, 1412.93) = 53.77$, $p = .00$, but the effect was quite minimal ($\eta^2 = .09$), indicating that whatever differences that exist were not overly large. Likewise, while parallelism was also significant ($F(10.89, 1412.93) = 7.58$, $p = .00$), the effect was limited to only $\eta^2 = .06$, again indicating that the degree of actual variance in placement by rubric category was quite inconsequential.

Profile analysis is perhaps most easily understood when presented visually, as we can see the individual profiles and variations among them. Fig. 8 displays the results of a profile analysis for the original rubric, with the topmost line indicating the highest level of placement. We can clearly see that each line is distinct from all of the others, and no line intersects one another indicating that, indeed, the placement levels in terms of mean scores seem well defined. Notice, too, that the

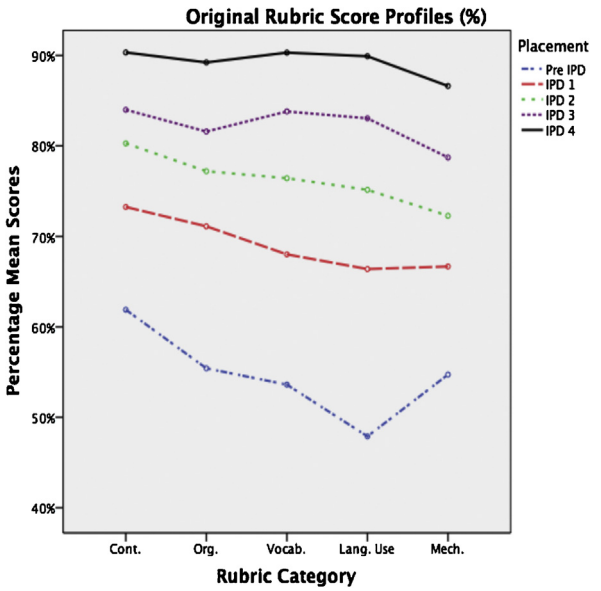

Fig. 8. Profiles for IPD course level placements using the original rubric with percentage scores, as rated by the original exam raters ($n = 524$).

Table 6
Summary of profile analysis for the revised rubric.

Source	SS	df	MS	F	p	η^2
Between groups						
Placement (levels)	48247.24	4	12061.81	30.21	.00	.62
Error	29947.03	75	399.29			
Within groups						
Category (flatness)	2743.41	4	685.85	7.06	.00	.09
Category × Placement (parallelism)	4239.60	16	264.97	2.73	.00	.13
Error	29132.985	300	97.11			

Note. $n = 80$.

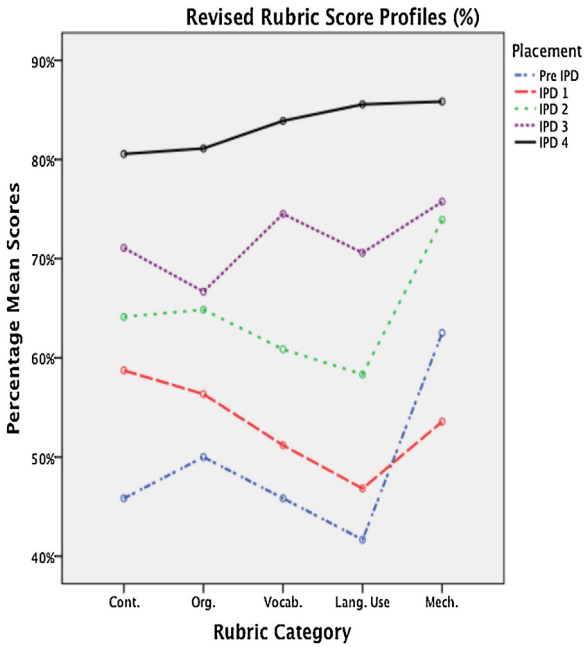

Fig. 9. Profiles for IPD course level placements using the original rubric with percentage scores, as rated by the authors ($n = 80$).

lines themselves are mostly flat (i.e., the slope of all lines is nearly horizontal without much variation), which explains why differences between the individual rubric categories appeared to be minimal at best. Differences do, however, appear to be more apparent at lower placement levels. Lastly, with the possible exception of the Pre-IPD group, the lines all follow a similar pattern (i.e., they are parallel), indicating that the profiles of the groups themselves are similar, and the majority of difference between the profiles appears in terms of levels alone.

4.2.3. Revised rubric

Table 6 provides summary statistics for the profile analysis of the revised rubric ($n = 80$). As above, statistically significant differences were found for each flatness ($F(4, 300) = 7.06$, $p = .00$), parallelism ($F(16, 300) = 2.73$, $p = .00$), and levels ($F(4, 75) = 30.21$, $p = .00$), and the analysis appeared very similar to the original rubric, albeit with some noticeable differences. Beginning with levels, this finding indicates that the rubric is distinguishing between the five placement levels in some way and that the effect is comparatively large ($\eta^2 = .62$). Fig. 9 displays these differences visually, and we can clearly see separation between the placements in most cases. Post-hoc contrasts were carried out to determine where these differences were specifically occurring (Table 7). Looking at adjacent levels only, we can see that only two statistically significant differences were observed in the data, and these were between course levels IPD 1 and 2 and course levels IPD 3 and 4. This

Table 7
Contrasts for adjacent levels for the revised rubric.

IPD placement		M difference	SE	P
0	1	4.17	4.88	.95
1	2	11.09	2.70	.00
2	3	7.30	2.86	.18
3	4	11.67	3.17	.01

Note. $n = 80$. The course Pre-IPD is level 0 in this table.

indicates that based on the writing subtest scores alone, the rubric seems to be unable to distinguish between Pre-IPD and IPD 1 examinees, or between IPD 2 and 3 examinees. It is possible that this can be explained because of the low sample size – especially for the Pre-IPD course level – or due to the fact that the writing subtest is only one of three subtests that actually determines final placement.

Looking at flatness and parallelism, differences can be found here as well, though both had a limited effect size (η^2 = .09 and .13 respectively), indicating that the differences for rubric categories and profiles by group were altogether small. Looking at Fig. 9, it is apparent that there is a considerable amount of variation among the rubric categories, which likely indicates again that none of the categories are markedly redundant. At the same time, some categories such as Organization (in the cases of IPD 2 and 3) and Language Use (in the cases of IPD 1 and 2) do seem very similar across placement levels, which might indicate that these categories are not effective in distinguishing levels. Given that these findings are not consistent across all placement levels, there is probably little need to be concerned at this stage. There do seem to be some potential issues for Mechanics in terms of similar scores for different placements (e.g., in the cases of IPD 2 and 3) and even higher scores for lower placements, as in the cases of Pre-IPD and IPD 1. These results are somewhat unsurprising, however, as even with the original rubric, Mechanics is the most difficult category to adequately identify multiple, different levels within.

4.3. Qualitative analysis

Analysis of the negotiation sessions not only helps elucidate how raters jointly arrived at their scoring decisions but also provides indirect evidence for the ways in which the structure of each rubric might have influenced raters' decision making process as they scored essays independently. One major difference seems to be that when using the original rubric, raters most frequently characterized examinee performance in terms of the adjectives Jacobs et al. (1981) use to define the four broad ability bands (i.e., excellent to very good; good to average; fair to poor; very poor), while when using the revised rubric, they described performance in terms of individual scores.

The ways in which raters ground their decisions in the band adjectives of the original rubric can be seen in the extract below, in which Rater 2, who had initially assigned a score of 13 for Vocabulary (representing the upper boundary of fair), negotiated with Rater 1, who had initially assigned a score of 17 (representing the upper boundary of good). Rater 2 acknowledges that this essay was better than "fair," more like a "low good," while Rater 1 considers lowering his score to reflect an "average" performance before ultimately deciding to choose a number that signals the essay is "good" but not a "high" good.

> R2: . . .fair is not warranted. I would say probably (8 s. pause) it's like a low good for me, maybe I would give it a 16.
>
> R1: what did I give it? a 17. hmm (5 s. pause) yeah, I don't know. I might move mine down to a, maybe an average, like a 15.
>
> (more discussion of essay)
>
> R1: maybe I would give it a 16 rather than a 17, which still keeps it as good but not quite so high.

Throughout the negotiation sessions in which raters used the original rubric they seemed primarily concerned with assigning numerical scores that preserve their qualitative judgments. One potential disadvantage of this approach is that if raters do not have a clear sense of what differentiates adjacent scores in the same band, they may assign scores somewhat impressionistically and inconsistently.

In contrast, when using the revised rubric, raters primarily described and evaluated test taker performance by referencing one of the six potential scores. This can be seen in the extract below, in which Rater 2 deliberates whether a score of three or four best captures the examinee's performance on Content, while Rater 1 evaluates the examinee's performance in terms of "what other fours look like" and concludes "I don't think this looked like a four."

> R2: I really went back and forth, the second paragraph is completely general and undeveloped, the first one is a bit more so, I think the only reason I gave it a three, sorry a four is that the rubric says "main/controlling idea is generally clear in its development" and the three says "main/controlling idea is underdeveloped and somewhat unclear" and I felt like ok, it's not unclear but it's definitely very underdeveloped, but I was really going back and forth.
>
> R1: yeah I was too, this is a case where I kind of went "what do other fours look like?" and I don't think this looked like a four and that's why I gave it a three, but it was in this kind of in between area for sure because (provides some examples). . .and so I felt that was underdeveloped in that sense or not specific in that sense but it wasn't confusing.
>
> R2: right, I think they were attempting to provide specific supporting information but it was at a very general level, I think that's true, ok I'm.
>
> R1: so why don't we, I don't know would this be a good situation to agree to disagree.
>
> R2: but I kind of did feel like I did want to give it a, I don't feel strongly that it's a four so I think I will change my score to a three.

In this and other instances, there is evidence that raters had internalized a standard of performance that corresponded to each of the numeric scores and thus were potentially assigning individual scores in a more principled and consistent manner.

However, the extract above also illustrates one perhaps foreseeable side effect of the restricted range of possible scores: namely that raters at times expressed trouble assigning scores for essays which seemed to fall in between the established levels of performance. Several times Rater 2 describes going "back and forth," and Rater 1 agrees that this essay was "in this kind of in between area for sure." Although they decided not to do so in this particular example, in other instances raters did use the "agree to disagree" strategy as one way to create a kind of intermediate score. While raters did not consistently express difficulty deciding between levels, they did so often enough that this issue warrants further attention. Further examination of additional rater negotiation sessions may reveal that raters had particular trouble differentiating between certain levels and/or for particular rubric categories, and that revision of performance indicators is in order.

5. Discussion

5.1. RQ1

How were rubric category scales of the original Jacobs et al. (1981) rubric functioning when applied to the IPD Placement Exam writing section?

MFRM analysis of our original data set showed that the Jacobs et al. (1981) rubric, though reliable, included too many scores for each category scale. This was visually apparent in the Rasch vertical rulers and the category response curves, which showed scores representing about the same ability clustering in the scale's mid-range, while at the scale's ends, few scores were available to represent broad tracts of ability. Threshold distance measurements confirmed that scores were not meaningfully different from each other and that the scale was too finely grained. This was corroborated by raters, who had expressed concerns about whether so many different scores were necessary.

This finding provides some evidence against the scoring inference within the exam's validation argument, as scoring procedures should be applied appropriately, accurately, and consistently (Kane, 2006, pp. 24, 34; Kane, 2013, p. 25). Though the Rasch model of the exam is very reliable using both the original and revised rubric, both the original internal raters and the external author team questioned how appropriate it was to have so many possible scores when each score is not distinguishable from neighboring scores, bringing into question the scoring inference of the validation argument.

5.2. RQ2

How can the rubric category scales be made more accurate and efficient?

This study showed how Rasch analysis could be used to guide the revision of the Jacobs et al. (1981) rubric scale in one context. We chose to first rescale the original data set onto 4- and 7-point scales, a theoretical decision that was taken as the original rubric specifies four broad levels of proficiency for each rubric category and seven sub-levels of proficiency. For our data set, threshold distance measurements of the 7-point scales revealed that there was still not enough meaningful difference between the scores in several rubric categories (e.g., Content), while a 6-point scale maintained as much distinction as possible while maintaining meaningful difference. In contexts that do not require so much distinction in score levels, our data indicates that a 4-point revision of the original rubric along broad ability bands would be adequate. No matter what scoring regimen is decided upon, we recommend piloting the revised scale before actual use and comparing placement using both the original and revised rubrics, as was done in this study.

5.3. RQ3

How does the revised rubric function when classifying examinees who took the IPD Placement Exam writing section?

In this context, adjustments to the original rubric scoring bands appear to have been successful in terms of both removing ambiguous or superfluous scores while also preserving the function of the rubric in terms of distinguishing placement levels and rubric category function. Importantly, where we previously saw overlapping scores for different rubric categories, the analyses seem to indicate that revisions have created a noticeable degree of separation between scores for all categories. Interestingly, where raters previously raised concerns about being able to assign scores reliably and validly at the level of detail required by the original rubric, raters for the revised rubric actually expressed concern that the new scoring bands were too broad at times, and indicated that half-scores might be useful in certain places. While this might be possible, it is typically acknowledged that people have difficulty reliably distinguishing between more than about seven levels (Miller, 1956), and fewer levels leads to more decision power (North, 2003). A more useful solution, then, may be to consider the performance indicators in the revised rubric and how they might be adjusted further to provide more precise classifications of different levels of performance—particularly in the middle levels where ratings become the most difficult.

In terms of function, the profile analyses were used as a confirmatory step to examine the revised rubric in comparison to the original rubric function for placement by category. While the actual profiles of the revised rubric are different than those of the original, the overall trends in the profiles are the same in that the levels are clearly separated, and this separation extends to almost all categories. While there are some areas of overlap between categories for placements (e.g., Mechanics),

this is likely due to a relatively limited sample size and we would expect these differences to be more powerful as more examinees are included in the analysis. That said, the results do help display areas where further revision might be beneficial.

6. Conclusion

As we move towards concluding, we would like to recognize first that calls for work on exam validation are frequent. Nevertheless, it is rather unusual for a mid-sized language program to support ongoing validity research concerning one of its language program's placement exams. This is probably because validation research is sensitive in nature, as it places researchers in contact with great quantities of confidential information, especially when the exam has high-stakes uses. In addition, most language programs can be described as being over-burdened and under-resourced, both in terms of human and financial capital. Because of these reasons, validation research such as this project is not prioritized and can often be put aside in favor of other issues. It is our express hope that both Ph.D. student researchers and language programs work to cultivate symbiotic relationships such as the one that culminated in this research product, so that both parties advance their respective agendas and that language exams can be used in ways that reflect the uses and interpretations specified in their validation argument.

This project also has provided us with visceral experience that validation projects are ongoing, iterative work, which requires the investment of human and financial capital for their completion. Though this project was able to propose an adapted rubric supported by empirical and interview data, this project is still not complete. Future steps should include the local piloting of the revised rubric, in addition to the continued discussion of the descriptors within the rubric's rating scale. This sort of continued discussion of placement exams does the work of creating a locally-shared understanding of the constructs found within language programs and also their assessment instruments. Convergence upon mutually-held interpretations is the first step in most standards setting methodologies (cf., Cizek & Bunch, 2007; Hambleton & Pitoniak, 2006), and it is worth noting that standards setting projects are also often ignored in many language programs. Because of the high-stakes nature, this standards setting discussion, however, should not be limited to the exam raters; representative stakeholders from the entire community affected by the language program should participate in the development of a common understanding of what the language program and its assessment instruments do. Thus, by continually reviewing their language program's high-stakes exam, this language program will be able to make advances in terms of having not only a placement exam whose uses are valid, but also a language program whose core constructs have been contemplated by the stakeholders who are affected by the program. This sort of discussion, ideally, will not be limited to the program's placement exam, but will consider and share information about all important exams within the program.

As our last remark, we believe our previous conclusion about ongoing empirical research and discussion creating a better shared understanding of what is happening in the language program applies to the broader context that concerns language programs in Colombia. We hope that as new educational policies become mandated and implemented, that they are not only supported by initial political fanfare, but also with ongoing empirical investigation concerning these programs' validity, and that there can be a shared discussion with all different local stakeholders about the constructs that are important to us all, as a community.

Acknowledgments

This research was only possible because of the convergence of several important factors. We are fortunate that Universidad de los Andes, where we conducted this research, has unconditionally opened its doors to us, and we are grateful to its directors, both past and present. As this paper's three authors are in Ph.D. programs, we would like to thank our professors and advisors for the training and support we have received, especially Dr. James Dean Brown. We hope that this product is worthy of their investment in us.

References

Bachman, L. (2004). *Statistical analyses for language assessment*. Cambridge, UK: Cambridge University Press.
Bachman, L., Lynch, B., & Mason, M. (1995). Investigating variability in tasks and rater judgements in a performance test of foreign language speaking. *Language Testing*, *12*(2), 238–257.
Bachman, L., & Palmer, A. (2010). *Language assessment in practice*. New York, NY: Oxford University Press.
Barkaoui, K. (2007). Participants, texts, and processes in ESL/EFL essay tests: A narrative review of the literature. *The Canadian Modern Language Review/La Revue Canadienne Des Langues Vivantes*, *64*(1), 99–134.
Barkaoui, K. (2013). Multifaceted Rasch analysis for test evaluation. In A. Kunnan (Ed.), *The companion to language assessment* (3) (pp. 1301–1322). Hoboken, NJ: Wiley Blackwell.
Bond, T., & Fox, C. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* ((2nd ed.)). New York, NY: Routledge.
Brown, J. D. (2015). *Mixed methods research for TESOL*. Edinburgh, UK: Edinburgh University Press.
Brown, J. D., & Kondo-Brown, K. (2012). Rubric-based scoring of Japanese essays: The effects of generalizability of number of raters and categories. In J. D. Brown (Ed.), *Developing, using and analyzing rubrics in language assessment with case studies from Asian and Pacific languages* (pp. 169–184). Honolulu, HI: University of Hawai'i, National Foreign Language Resource Center.
Chapelle, C. (2008). The TOEFL validity argument. In C. Chapelle, M. Enright, & J. Jamieson (Eds.), *Building a validity argument for the Test of English as a Foreign Language®* (pp. 319–352). New York, NY: Routledge.
Chapelle, C. (2012). Validity argument for language assessment: The framework is simple. . .. *Language Testing*, *29*(1), 19–27.
Cizek, G., & Bunch, M. (2007). *Standards setting*. Thousand Oaks, CA: Sage Publications.
Clauser, B. (2000). Recurrent issues and recent advances in scoring performance assessments. *Applied Psychological Measurement*, *24*(4), 310–324.

Creswell, J. W., & Plano Clark, V. L. (2007). *Designing and conducting mixed methods research*. Thousand Oaks, CA: Sage Publications.
Crusan, D. (2010). *Assessment in the second language writing classroom*. Ann Arbor, MI: The University of Michigan Press.
Delaney, Y. (2009). Investigating the reading-to-write construct. *Journal of English for Academic Purposes*, *7*(3), 140–150.
Denzin, N. K. (1978). *The research act: A theoretical introduction to sociological methods*. New York, NY: Praeger.
East, M. (2009). Evaluating the reliability of a detailed analytic scoring rubric for foreign language writing. *Assessing Writing*, *14*, 88–115.
Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing*, *25*(2), 155–185.
Eckes, T. (2011). *Introduction to many-facet Rasch measurement*. Frankfurt am Main, Germany: Peter Lang.
Ellis, D., & Ross, S. (2013). Item response theory in language testing. In A. Kunnan (Ed.), *The companion to language assessment* (3) (pp. 1262–1281). Hoboken, NJ: Wiley Blackwell.
Fulcher, G. (1987). Tests of oral performance: The need for data-based criteria. *ELT Journal*, *41*(4), 287–291.
Fulcher, G. (2003). *Testing second language speaking*. London, UK: Pearson Longman.
Hambleton, R., & Pitoniak, M. (2006). Setting performance standards. In R. Brennan (Ed.), *Educational measurement* ((4th ed.), 3, pp. 433–470). Westport, CT: American Council on Education/Praeger.
Harsch, C., & Martin, G. (2012). Adapting CEF-descriptors for rating purposes: Validation by a combined rater training and scale revision approach. *Assessing Writing*, *17*, 228–250.
Hashemi, M. R. (2012). Reflections on mixing methods in applied linguistics research. *Applied Linguistics*, *33*(2), 206–212.
Hawkey, R., & Barker, F. (2004). Developing a common scale for the assessment of writing. *Assessing Writing*, *9*, 122–159.
Huang, J. (2008). How accurate are ESL students' holistic writing scores on large-scale assessments?—A generalizability theory approach. *Assessing Writing*, *13*, 201–218.
Huang, J. (2012). Using generalizability theory to examine the accuracy and validity of large-scale ESL writing assessments. *Assessing Writing*, *17*, 123–139.
Jacobs, H., Zinkgraf, S., Wormuth, D., Hartfiel, V., & Hugley, J. (1981). *Testing ESL composition: A practical approach*. Rowley, MA: Newbury House.
Janssen, G., Ángel, C., & Nausa, R. (2011). *Informe de la investigación: El desarrollo de un currículo para la escritura de inglés nivel posgrado, según las necesidades y habilidades de los estudiantes (Proyecto IPD)*. Bogotá, Colombia: Universidad de los Andes (Internal document).
Janssen, G., & Meier, V. (2012). *IPD placement exam study*. Mānoa, Honolulu, HI: Department of Second Language Studies, University of Hawai'i (Unpublished manuscript).
Johnson, R. B., Onwuegbuzie, A. J., & Turner, L. A. (2007). Toward a definition of mixed methods research. *Journal of Mixed Methods Research*, *1*(2), 112–133.
Kane, M. (2006). Validation. In R. Brennan (Ed.), *Educational measurement* ((4th ed.), pp. 17–64). Westport, CT: American Council on Education/Praeger.
Kane, M. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1–73.
Kane, M., Crooks, T., & Cohen, A. (1999). Validating measures of performance. *Educational Measurement: Issues and Practice*, *18*(2), 5–17.
Knoch, U. (2007). 'Little coherence, considerable strain for reader': A comparison between two rating scales for the assessment of coherence. *Assessing Writing*, *12*, 108–128.
Knoch, U. (2008). The assessment of academic style in EAP writing: The case of the rating scale. *Melbourne Papers in Language Testing*, *13*(1), 34–67.
Knoch, U. (2009). Diagnostic assessment of writing: A comparison of two rating scales. *Language Testing*, *26*(2), 275–304.
Knoch, U. (2011). Investigating the effectiveness of individualized feedback to rating behavior—A longitudinal study. *Language Testing*, *28*(2), 179–200.
Kubiszyn, T., & Borich, G. (2015). *Educational testing and measurement: Classroom application and practice* ((11th ed.)). New York, NY: HarperCollins.
Lane, S., & Stone, C. (2006). Performance assessment. In R. Brennan (Ed.), *Educational measurement* ((4th ed.), pp. 387–431). Westport, CT: American Council on Education/Praeger.
Lim, G. (2011). The development and maintenance of rating quality in performance writing assessment: A longitudinal study of new and experienced raters. *Language Testing*, *28*(4), 543–560.
Linacre, J. (2010). *FACETS (Version 3.67.0)*. Chicago, IL: MESA Press.
Linacre, J. M. (1997). *Guidelines for rating scales and Andrich thresholds*. Retrieved from http://www.rasch.org/rn2.htm.
Lynch, B. K., & McNamara, T. F. (1998). Using G-theory and many-facet Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants. *Language Testing*, *15*(2), 158–180.
McNamara, T. F. (1996). *Measuring second language performance*. New York, NY: Longman.
Meier, V. (2013). Evaluating rater and rubric performance on a writing placement exam. *University of Hawai'i, Working Papers of the Department of Second Language Studies*, *31*(1), 47–100.
Meier, V., Trace, J., & Janssen, G. (in press). The rating scale in writing assessment. In J. Banerjee & D. Tsagari (Eds.), *Contemporary second language assessment.* New York, NY: Continuum.
Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement* ((3rd ed.), pp. 13–103). New York, NY: Macmillan.
Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *63*(2), 81–97.
Morse, J. (1991). Approaches to qualitative–quantitative methodological triangulation. *Nursing Research*, *40*, 120–123.
North, B. (2003). Scales for rating language performance: Descriptive models, formulation styles, and presentation formats. *TOEFL Monograph No. MS-24*, 24.
Polio, C. (2001). Research methodology in second language writing research: The case of text-based studies. In T. Silva, & P. Matsuda (Eds.), *On second language writing* (pp. 91–116). Mahwah, NJ: Lawrence Erlbaum & Associates.
Schaefer, E. (2008). Rater bias patterns in an EFL writing assessment. *Language Testing*, *25*(4), 465–493.
Sudweeks, R., Reeve, S., & Bradshaw, W. (2005). A comparison of generalizability theory and many-facet Rasch measurement in an analysis of college sophomore writing. *Assessing Writing*, *9*, 239–261.
Tabachnick, B., & Fidell, L. (2013). *Using multivariate statistics* ((6th ed.)). Boston, MA: Pearson.
Van der Linden, W., & Hambleton, R. (1997). *Handbook of modern item response theory*. New York, NY: Springer-Verlag.
Weigle, S. (1998). Using FACETS to model rater training. *Language Testing*, *15*(2), 263–287.
Weigle, S. (2002). *Assessing writing*. Cambridge: Cambridge University Press.
Winke, P., Gass, S., & Myford, C. (2013). Raters' L2 background as a potential source of bias in rating oral performance. *Language Testing*, *30*(2), 231–252.
Winke, P., & Lim, H. (2015). ESL raters' cognitive processes in applying the Jacobs et al. rubric: An eye-movement study. *Assessing Writing*, *25*, 37–53.
Yen, W., & Fitzpatrick, A. (2006). Item response theory. In R. Brennan (Ed.), *Educational measurement* ((4th ed.), pp. 111–153). Westport, CT: American Council on Education/Praeger.
Zhao, C. (2013). Measuring authorial voice strength in L2 argumentative writing: The development and validation of an analytic rubric. *Language Testing*, *30*(2), 201–230.

Gerriet Janssen is a Ph.D. candidate at the University of Hawai'i at Mānoa in the department of Second Language Studies. His dissertation evaluates the cut scores on one Colombian placement exam. His academic interests include language assessment, especially in terms of item response theory, and academic writing.

Valerie Meier is a Ph.D. student in the Education department of the University of California, Santa Barbara. Her research interests include academic literacies, bilingual education, and curriculum development.

Jonathan Trace is a Ph.D. candidate at the University of Hawai'i at Mānoa in the department of Second Language Studies. His interests include second language assessment, curriculum development, language for specific purposes, corpus linguistics, and mixed-methods research.