From Standards to Rubrics: Comparing Full-Range to At-Level Applications of an Item-Level Scoring Rubric on an Oral Proficiency Assessment

Troy L. Cox and Randall S. Davies

Introduction

Standards-based proficiency frameworks have become an integral part of the educational assessment landscape. These frameworks take complex, multidimensional competencies and attempt to represent them as a numerical value on a vertical scale that can be used by students, teachers, testing organizations, school admissions officers, employers, and others that want some certification of the proficiency of examinees. Framing performance in this way allows stakeholders to communicate and compare results. With some frameworks, like the Common Core (National Governors Association Center for Best Practices & Council of Chief State School Officers 2010), the vertical axis of the scale is based on grade levels. With language proficiency, the vertical axis of the scales is based on major-level descriptors which define what an individual should be able to do if they are to be certified as being proficient in that language at a specific level [see for example, the Common European Framework of Reference (Verhelst et al. 2009) and the American Council on the Teaching of Foreign Languages (ACTFL 2012)].

Rubrics are an essential component of any framework (Bargainnier 2004; Tierney and Simon 2004). Practitioners attempting to assess any performance must use rubrics that align with the standards. Students wanting to improve their performance need to understand how their score relates to the standards. Test developers needing to create equivalent test forms must have a way to ensure those forms are based on the standards. To be useful, the relationship between the rubric and the

T.L. Cox (✉)
Center for Language Studies, Brigham Young University,
3086-C JFSB, Provo, UT 84602, USA
e-mail: troyc@byu.edu

R.S. Davies
Instructional Psychology and Technology, Brigham Young University,
150-L MCKB, Provo, UT 84602, USA

Q. Zhang (ed.), *Pacific Rim Objective Measurement Symposium (PROMS) 2015 Conference Proceedings*,
DOI 10.1007/978-981-10-1687-5_14

standard should be transparent and the way in which raters use the rubrics must be consistent and appropriate.

Often when assessing speaking ability, some type of oral proficiency interview (OPI) is used. For example, with an ACTFL OPI, trained interviewers prompt examinees to respond to a wide variety of tasks (Buck et al. 1989). Each speaking task is designed to target a specific level on the scale and is intended to provide the examinee with an opportunity to demonstrate his or her ability to speak at that level. When an individual responds well to a specific prompt it provides evidence the individual is able to speak at that level. The assessment is designed to push the limits of the examinee and determine when the individual's speaking ability breaks down. Using the evidence they have gathered, interviewers then rate the performance against the proficiency standards and assign a score based on the scale being used. The score provides an estimate of the general speaking ability of the examinee. Since interviews are expensive to administer, many testing companies are transitioning to computer-administered speaking tests which in turn affects the way the test can be rated.

Rating can be done at the test-level or at the item-level. When raters assess speaking competency at the test-level, the rater listens to the entire performance and, rating it holistically, determines the overall speaking ability or level of the examinee. However, if computer-aided assessments are to be used or equivalent forms of a test are needed, item-level assessments are required. When rating at the item-level, raters listen to an individual's response to a specific task rating each it against the rubric. The individual item scores are combined to determine the overall speaking level of the examinee.

One problem human raters have when rating at the item-level is how to apply the rubric. Raters might be inclined to rate the performance using the full-range of the rubric as they would when rating holistically at the test-level. However, individual task prompts are not designed to provide that type of evidence. For example, a task designed to elicit evidence that an examinee can speak at an Intermediate level on the ACTFL scale would not likely provide evidence that the individual can speak at a higher level (e.g., superior). The individual would need to be prompted with another task designed to elicit that type of evidence. This study examined two ways of applying a rubric at the item-level—one that was directly tied to the full-range of the proficiency scale and another that used a restricted-range of that same scale.

Research Questions

1. How reliable is the full-range proficiency-based rubric when used at the item level?
2. How reliable is the at-level proficiency-based rubric when used at the item level?
3. Which rubric (full-range or at-level) most closely aligned with the expert-predicted item difficulty (EID) of each prompt?

Methods

This paper examined two ways to implement a proficiency-based rubric (full-range and at-level) when rating at the item-level. The item difficulty statistics calculated from the two rubrics were compared to the intended item difficulty of the item writers. Finally, a comparison was made between examinee test scores that were scored using a full-range and an at-level restricted-range application of the rubric.

Study Participants

The subjects participating in this study were students enrolled at an intensive English Program affiliated with a large research university that were taking their exit exams during winter semester 2012. There were 201 students who spoke 18 different languages (see Table 1). They were in the school to improve their English to the point at which they could successfully attend university, where the language of instruction was English. With the ACTFL guidelines, they ranged in speaking ability from Novice to Advanced.

Data Collection Instrument

The data collection instrument used in this study was designed to assess speaking ability at proficiency levels 2 through 6 (see Appendix A). It was assumed that after one semester of instruction, all the examinees participating in this study would have some ability to speak English yet none would be considered the functional equivalent of highly educated native speakers.

To determine to what extent the prompts on the instrument aligned with their expected difficulty level, a panel of expert raters was consulted. The rating rubric

Table 1 Composition of subjects by language and gender

Native language	Gender		Total
	Female	Male	
Spanish	54	34	88
Korean	21	16	37
[a]Other	17	18	35
Portuguese	13	15	28
Chinese	9	4	13
Total	114	87	201

[a]The following languages had five or fewer speakers: Arabic, Armenian, Bambara, French, Haitian Creole, Italian, Japanese, Mauritian Creole, Mongolian, Spanish, Tajik, Thai, Ukrainian, and Vietnamese

had been in use for six semesters so the maximum number of semesters a rater could have rated was six. The expert panel consisted of eight raters with an average of 4.75 semesters of rating experience and a range of 3 to 6 semesters. For each prompt, the raters used the speaking score rubric to predict the level of language an examinee would need to succeed with the prompt identified what objectives were being measured, and provided feedback on whether they felt the item would function as intended.

The results of the ratings assigned by the eight raters were used to obtain the expert-predicted item difficulty (EID). The experts were presented 15 prompts, and based on their feedback, 10 prompts, two for each of the targeted levels, were selected for inclusion on the test. The items were designed with varying amounts of preparation time and response time to meet the functions of the prompt.

The EIDs of the 10 selected items rose monotonically in that every Level 2 prompt was easier than every Level 3 prompt and every Level 3 prompt was easier than the Level 4 prompts, etc. and this was considered to be evidence that the prompts did reflect the scale descriptors. The EIDs were also used to examine the extent to which the estimated difficulty of the speech prompts ordered as expected with the full-range and at-level rubric. The speaking test was designed with the same framework as an interview-based test that progresses from easier to harder and then back down so that examinees would experience a full range of prompt difficulties.

Rating and Scoring Procedures

The assessment was administered to students as part of their final exams. The scoring rubric was based on an 8-level scale that roughly corresponded to the ACTFL OPI scale. The rubric addressed three axes: (a) text type (e.g., word and phrase length, sentence length, paragraph length, etc.), (b) content, and (c) accuracy. Each axis ranged from *no ability* to *high ability* (i.e., the functional equivalent of a well-educated highly articulate native speaker). The scale was intended to be noncompensatory so that a response that was native-like in one area (e.g., pronunciation) could not compensate for a weak performance in another area (e.g., a text type that was only word length). The full-scale rubric required raters to keep the full range in mind as they judged performances. The at-level scale allowed raters to focus on a restricted-range of five levels: far below level, below level, at-level, above level, and far above level (see Table 2). To ensure the results had the characteristics of interval data and fully justified the use of parametric statistics, both ratings (holistic and analytic) were converted from raw scores (typically used in classical test theory) to fair averages (based on Rasch modeling).

Table 2 Full-range speaking rubric to at-level scale conversion matrix

	At-level	Intended item difficulty level				
	Rating	2	3	4	5	6
Below by 2 or more levels	1					0
					0	1
				0	1	2
			0	1	2	3
		0	1	2	3	4
Below by 1 level	2	1	2	3	4	5
At-level	3	2	3	4	5	6
Above by 1 level	4	3	4	5	6	7
Above by 2 or more levels	5	4	5	6	7	
		5	6	7		
		6	7			
		7				

Rating Methods

The tests were rated by judges with ESL training that were working as teachers. All of the raters had been trained at various times to use the rubric for the regularly scheduled computer-administrated speaking tests. The existing rater training material was designed to train raters how to use the 8-level scale for test-level scoring. The raters had received over 3 h of training and completed a minimum of 12 calibration practice ratings to ensure sufficient knowledge of the rubric. The raters had a packet that contained a copy of the rubric and a printed copy of the exam prompts, the objective of the prompt, and the intended difficulty level of the prompt. Details on how the test-level rubric was adapted will be discussed below.

Item-level rating designs. To get the item-level statistics, each test had to be rated at the item level. There were two possible incomplete connected design possibilities that could have provided the requisite data. The first was an *incomplete, connected* design in which all the items on a single test were rated by raters, who were linked to other raters. While this design is more cost-effective than a fully-crossed design, examinee ability estimates can be biased if there is an "unlucky combination of extreme raters and examinees" (Hombo et al. 2001, p. 20). The second design possibility was an *incomplete, connected spiral* design. This design was differentiated from the prior by assigning individual items to raters and linking raters to other raters through shared item ratings (Eckes 2011). This design shared the cost-effectiveness of the incomplete, connected designs, but has some distinct advantages. First, when raters listen to the same item from different examinees, they can have a deeper understanding of the response characteristics needed to assign a rating. Second, the spiral design can minimize errors associated with the *halo effect*. Halo effect occurs when performance on one item biases the

Table 3 Incomplete spiral connected design for analytically rated speaking test by prompt

Students	Prompt	Raters			
		1	2	3	4
1–4	1, 2, 3, 4	X	X	X	X
5	1	X	X		
5	2		X		
5	3		X	X	
5	4		X		X
6	1	X		X	
6	2		X	X	
6	3			X	
6	4			X	X
7	1	X			X
7	2		X		X
7	3			X	X
7	4				X

rating given on subsequent prompts (Myford and Wolfe 2003). For example, if a rater listens to a prompt and determines the examinee to speak at a Level 4 based on the rubric, then the rater might rate all subsequent prompts at 4 even when the performance might be higher or lower. Finally, spiral rating designs have been found to be robust in providing stable examinee ability estimates in response to rater tendencies (Hombo et al. 2001).

For this design, each rater was assigned to rate a single prompt (e.g., rater 1 scored all of prompt 1, rater 2 scored all of prompt 2, etc.). To avoid having disconnected subsets, a subset of the same students was rated on each item by all the raters. To further ensure raters were familiar with the items, raters rated some additional tests in their entirety. Table 3 presents an example of an incomplete, spiral design representing seven examinees, four raters, and four prompts. For the actual study, the design included 201 students, 10 raters, and 10 prompts.

Full-Range scale. Since all existing training materials for the rubric were designed in rating tests as a whole, the raters had to be given special instructions. They knew the intended level of the prompt they were scoring, and were told to reference that as they applied the rubric. For example, when rating a prompt that was designed to elicit Level 2 speech samples (ask simple questions at the sentence level), a rater was able to use the entire range of categories in the rubric (0–7). Since a rating of 2 would be passing, the only way the higher categories would be used is if the examinee spontaneously used characteristics of those higher categories through the use of more extended discourse, more academic vocabulary, native-like pronunciation, etc.

At-Level scale. To compensate for the possibility that a restricted-range bias or misuse of the rating rubric impacted the scores, the ratings were recoded to a five-point scale referred to as the at-level scale. Since the raters knew the intended level of the prompt they were evaluating, the rating likely reflected whether the

student response was below the targeted prompt level, at-level or above level. Table 2 shows how each intended level's 8-point rubric was converted to the 5-point at-level scale.

For example, with a Level 2 prompt, a rating of 0 which indicated little or no language on the holistic scale would be transformed to a 1, a rating of 1 which indicated that the language elicited for the Level 2 prompt was still below level was transformed to a 2, a rating of 2 which indicated that the language elicited was at-level and was transformed to a 3, a rating of 3 which indicated that the language elicited fulfilled all the required elements of level 2 language and had characteristics of Level 3 language was transformed to a 4, and a rating of 4, 5, 6 or 7 which indicated that the language elicited had characteristics of those levels was transformed to a 5. Similar conversions were calculated for each of the prompts. Thus, if the response was deemed at level for that prompt the associated score would be a 3.

Data Analysis

To answer the questions on how reliable the two item-level scales functioned, the facets software was used to conduct Many Facets Rasch modeling (MFRM). In MFRM, the facets were modeled in such a way that person ability, item difficulty, and rater severity were measured conjointly with the rating scale.

Measurement invariance. Besides being interval data, another advantage of using Rasch scaling is that the parameter estimates for both persons and items have the quality of *measurement invariance* (Engelhard 2008). That is, when measuring a unitary construct, person ability estimates are the same regardless of the items that are presented to the examinees, and item ability estimates are the same regardless of the examinees who respond to them. Since the application of the findings of this study were directed for test developers in equating test forms, measurement invariance of the items was highly relevant. Beyond the advantage of measurement invariance, the Rasch analysis provided information on how well the scale functioned and the reliability of the test scores and test items.

Diagnoses of rating scales. To evaluate how well a scale functions with Rasch measurement, there are a number of diagnostics available including (a) category frequencies, (b) average logit measures, (c) threshold estimates, (d) category probability curves, and (e) fit statistics (Bond and Fox 2007). For category frequencies, the ideal is that there should be a minimum of 10 responses in each category that are normally distributed. For average logit measures, the average person ability estimate of each rating category should increase monotonically (Eckes 2011). The threshold estimates are the logits along the person ability axis at which the probability changes from a person being in one category to another. Those estimates should increase monotonically as well. In order to show distinction between the categories, they should be at least 1.4 logits apart and to avoid large gaps in the variable and the estimate should be closer than five logits (Linacre 1999). When looking at a graph of the category probability curves, each curve

should have its own peak, and the distance between thresholds should be approximately equal. If one category curve falls underneath another category curve or curves, then the categories could be disordered and in need of collapsing. Finally, fit statistics provide one more way to examine a rating scale. If the outfit mean squares of any of the categories are greater than 2.0, then there might be noise that has been introduced into rating scale model (Linacre 1999). Using these diagnostics through a FACETs analysis, a measurement scale can be analyzed.

Reliability analysis. Finally, Rasch scaling provides more tools in determining the reliability of test scores, especially when there are multiple facets. Reliability is defined as the ratio of the true variance to the observed variance (Crocker and Algina 1986). Unlike classical test theory which can only report reliability on the items of a test (e.g., Cronbach's alpha or Kuder–Richardson 20) or the agreement or consistency of raters (e.g., Cohen's kappa or Pearson's Correlation coefficient), Rasch reliability reports the relative reproducibility of results by including the error variance of the model in its calculation. Furthermore Rasch reliability provides estimates for every facet (person, rater, item) that is being measured. When the reliability is close to 1.0, it indicates that the observed variance of whatever is being measured (person, rater, item) is close or nearly equivalent to the true (and immeasurable) true variance. Therefore, when person reliability is close to 1, the differences in examinee scores are due to differences in examinee ability. If there are multiple facets such as raters, it might be desirable for a construct irrelevant facet to have a reliability estimate close to 0. If raters were the facet, a 0 would indicate the raters were indistinguishable from each other and therefore interchangeable. Any examinee would likely obtain the same rating regardless of which rater were assigned to them. Conversely, if the rater facet had a reliability estimate close to 1.0, then the raters are reliably different and the rating obtained by a given examinee is highly dependent on the rater. When the rater facet is not close to 0, it is necessary that an adjustment be made to the examinee score to compensate for the rater bias.

To obtain item-level ratings, the item speaking level was calculated using a FACETS analysis. The three facets for this analysis included examinees, raters, and prompts. Since the raters used (a) a full-range 8-point scale that could be converted to (b) a restricted-range 5-point at-level scale at the prompt level, the Andrich Rating Scale model was the most appropriate to use (Linacre and Wright 2009). One of the requirements for the use of Rasch MFRM is that the data be unidimensional and while it may appear that that a noncompensatory scale with three axes by definition does not have that characteristic, we argue that true unidimensionality rarely occurs in educational measurement and that essential unidimensionality can exist through shared understanding of construct definitions (see Clifford and Cox 2013) and the fit of the data (McNamara and Knoch 2012).

The rubrics used for the rating scale were (a) the same as the holistic rated speaking level scale but applied to individual prompts (or items) on the exam and the derived at-level scale (see Table 2). To see which rubric (full-range or at-level) most closely aligned with the expert-predicted item difficulty (EID) of each prompt, the item difficulty parameters from the two rubrics were correlated with the EIDs.

Results

To answer the research questions of this study on scale reliability required us to obtain scores at the prompt level. An 8-level full-range rubric that is typically used to rate at the test-level raters was applied to each prompt of the assessment. The ratings were then converted to an at-level scale that evaluated if the examinee performed below, at or above the intended difficulty level. A facet analysis was conducted with the ratings from the full-range and at-level scale. To do a comparison between the EID of the prompts and the actual difficulty based on ratings at the prompt level, a correlation was run.

Full-Range Scale Rasch Analysis. The items were rated with the full-range scale using an incomplete, spiral connected design and scale and reliability analyses were conducted.

Scale analysis. The eight categories of the full-range rubric functioned within acceptable parameters for the study (Table 4). With the exception of the category 0 (n = 3), the relative frequency of each category had a minimum of 10 in each category. The average measure increased monotonically from 1 to 7 without exception, as did the threshold estimates. The threshold estimates had the minimum recommendation of 1.4 logits between each category indicating that each category showed distinction, and none of the thresholds were over 5 logits apart, and the spacing of the categories was more evenly spaced than the human-rated holistic speaking level. An examination of the category probability distributions was indicative that each category functioned well (see Fig. 1).

For the fit statistics, the outfit mean squares of the categories did not exceed 2.0 with the exception of the 0 category which only had 3 responses. The only category that did not fit the guidelines of a good scale was 0. Since the 0 category is typically reserved for little or no production and since the students had one semester of instruction, this category could be combined with category 1 if it were only used as an end of instruction scale. However, since the scale is used for placement testing as well, all eight categories were retained. The full-range rubric functioned within acceptable parameters to be used in the analysis.

Table 4 Full-range rubric scale category statistics

Category	Absolute frequency	Relative frequency (%)	Average measure	Outfit	Threshold	SE
0	3	0	−3.04	2.0		
1	54	2	−3.72	1.1	−7.13	0.59
2	345	12	−2.39	1.0	−5.06	0.77
3	933	33	−0.23	1.0	−2.27	0.27
4	916	32	1.62	0.9	0.77	0.16
5	449	16	2.91	1.0	3.00	0.17
6	142	5	3.88	1.1	4.55	0.23
7	24	1	4.93	1.0	6.14	0.49

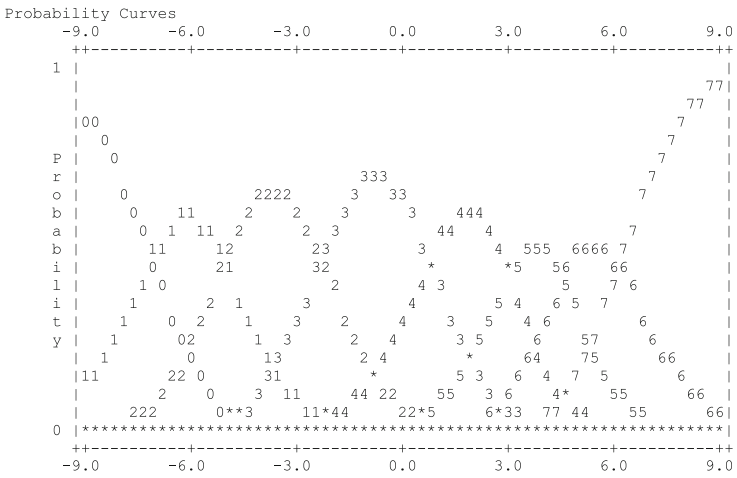

Fig. 1 Full-range rubric rating category distribution

Reliability analysis. The reliability statistics with the full-range rubric found that all three facets were reliably separated. Figure 2 is a vertical scale map that shows the logit in the first column, the examinee ability level in the second column, the rater severity in the third column, the empirical item difficulty in the fourth column and the scale equivalency in the fifth column. The 0 in the middle of the vertical scale is tied to the means of the rater and item ability estimates or logits. An examinee with an ability logit of 0 (the second column) would have a 50 % chance of being rated in category 3 (the fifth column), by raters R5 or R9 (the third column) on item L4-2 or L6-2.

Figure 2 showed that the examinee ability ranged from category 2 to 6. The examinees separation reliability was 0.94, thus we can be confident of the different ability levels of the examinees. For rater reliability, there was a separation reliability coefficient of 0.96. In Fig. 2, we can see that the raters R7 and R10 were the most generous and rater 4 was the most severe. Even though the raters rated the items differently than one another, the fit statistics were indicative of high internal consistency with an average mean outfit square of 1.0 and an average mean infit square of 1.0. Thus, the rater severity error could be mathematically modeled out of the examinees' scores through the use of the fair average.

The item facet had a reliability of 0.89 indicating that items could not be used interchangeably without compensating for their difficulty level. In Fig. 2, the third column represents the intended level and item number. The easiest item was L6-1 (i.e. EID Level 6, Item 1) and the most difficult item was L5-1. While it was expected that the prompts would have varying item difficulties and that some kind of item equating would need occur to create equivalent test forms, it was unexpected that the item difficulty means did not order in their intended levels. This could be due to a restricted-range error in which raters were unwilling to use the extremes of the rubric. It was also notable that the prompts clustered around

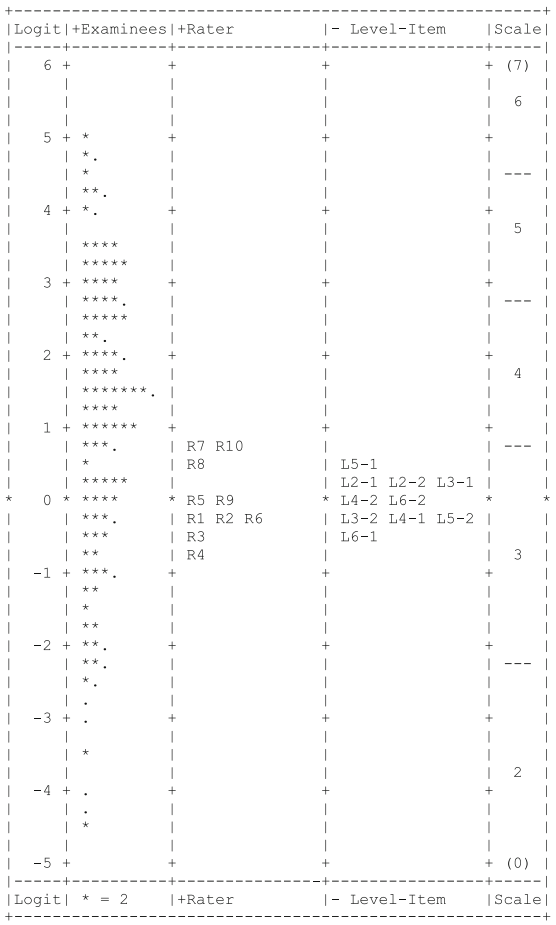

Fig. 2 Analytic full-range speaking level vertical scale

category 3 (SD = 0.27) and had a narrower range than the raters (SD = 0.48). The prompt fit statistics were indicative of high internal consistency with an average mean outfit square of 1.0 and an average mean infit square of 1.0.

At-Level Scale Rasch Analysis. The item ratings from the full-range scale were converted to the at-level scale and analyzed with FACETS.

Scale analysis. The at-level scale functioned within the parameters needed for a reliable scale (see Table 5). The relative frequency of each category had a minimum of 10 in each category. The average measure increased monotonically without exception, as did the threshold estimates. The threshold estimates had the minimum recommendation of 1.4 logits between each category indicating that each category showed distinction, and none of the thresholds were over 5 logits apart. Furthermore, the spacing between the thresholds was more evenly spaced than the full-range scale (see Fig. 3). The outfit mean squares of the other categories did not exceed 2.0.

Table 5 At-level scale rating scale category statistics

Category	Absolute frequency	Relative frequency (%)	Average measure	Outfit	Threshold	SE
1	770	27	−6.14	1.0		
2	553	19	−2.52	0.6	−3.86	0.08
3	589	21	−0.06	1.0	−1.39	0.07
4	552	19	2.38	1.2	1.19	0.07
5	402	14	5.22	1.1	4.07	0.09

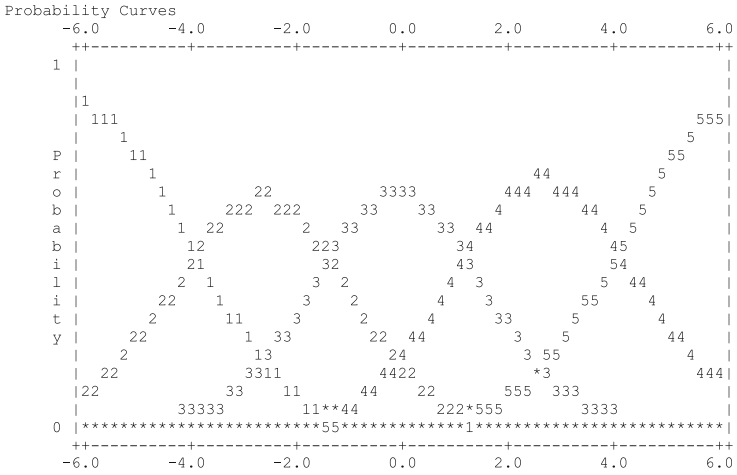

Fig. 3 At-level scale rating category distribution

Reliability analysis. The reliability statistics on the at-level item scoring found that all three facets (examinees, raters, and items) were reliably separated. In Fig. 4, we can see that the examinees have a range from categories 1–5. Note that the significance of these categories did ***not*** signify the levels of the speaking rubric facets of examinee, but rather whether how well they performed the task at its intended level. This analysis found that the separation reliability between the examinees was 0.93 and that the examinees could be separated reliably into different groups.

The raters had a reliability of 0.96 the same as the full-range analysis with the standard deviation being slightly larger than the at-level scale analysis (full-range rubric SD = 0.55 compared to at-level scale SD = 0.49). The raters still exhibited different levels of severity with R7 being the most generous and raters R3 and R4 being the most severe. Comparing the raters in Figs. 2 and 4 we see that the ordering of the severity and generosity of the raters is very similar with a high correlation ($r = 0.78$, $p < 0.05$) between the rating scales. The fit statistics were indicative of high internal consistency with an average mean outfit square of 1.0 and an average mean infit square of 1.0.

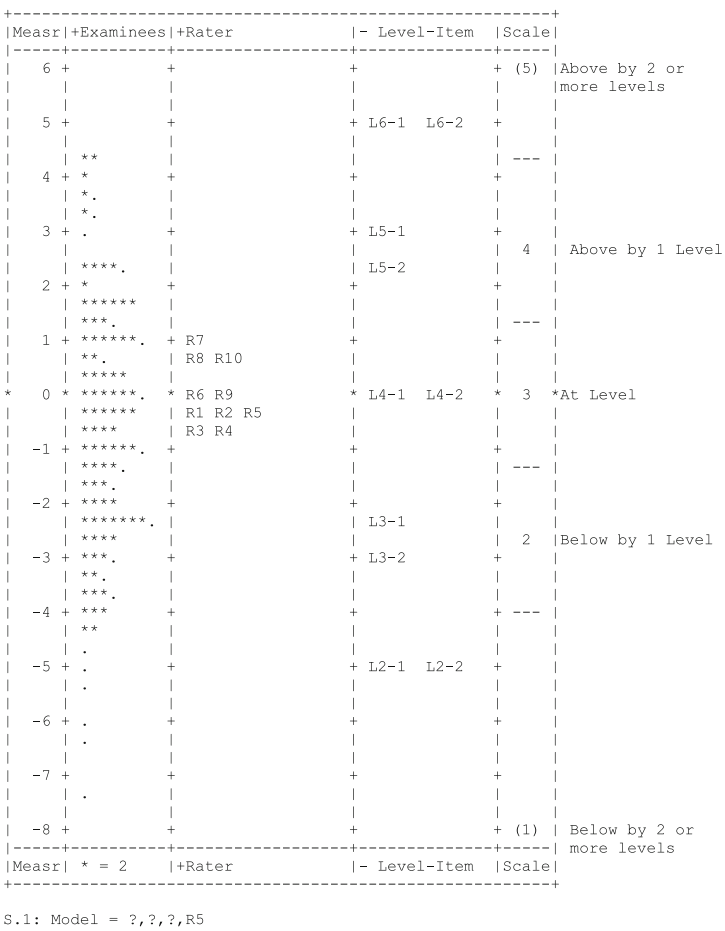

Fig. 4 At-level vertical scale of examinees, raters, and items

Most noteworthy though was the fact that the at-level scale item facet jumped to a reliability of 1.00 indicating that it would be virtually impossible to have the same score with the different items. Furthermore, the differences in difficulty aligned closely with the EID (see Fig. 5). The prompts that were intended to elicit Level 6 language (L6-1 and L6-2) were the most difficult while the prompts intended to elicit Level 2 language were the easiest (L2-1 and L2-2). Such high item separation reliability is in part due to the at-level scale being honed into the level that the prompt was eliciting. In other words, the only way for a prompts targeted at 2 subsequent levels to be conflated would be for the lower level prompt to have a preponderance of 4 and 5 ratings and the higher level prompt have a preponderance of 1 and 2 ratings. Table 6 illustrates the manner in which the item difficulty parameters increase monotonically as the intended difficulty increased.

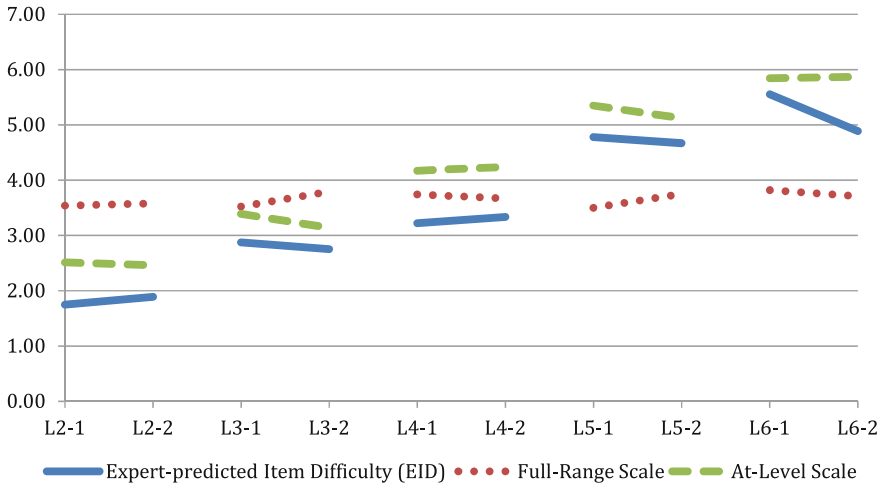

Fig. 5 Means of item difficulty measures

Table 6 At-level scale item statistics in order of measure

Item	Fair average	Logit	[a]Conversion to full-range rubric
L2-1	4.49	−4.86	2.51
L2-2	4.54	−5.04	2.46
L3-1	3.61	−2.22	3.39
L3-2	3.86	−2.93	3.14
L4-1	2.83	−0.15	4.17
L4-2	2.76	0.02	4.24
L5-1	1.65	2.92	5.35
L5-2	1.88	2.27	5.12
L6-1	1.16	4.88	5.84
L6-2	1.13	5.10	5.87
Mean	2.79	0.00	4.21
S.D.	1.30	3.74	1.30

[a]The conversion to the full-range rubric consisted of subtracting the Fair Average from seven, which was the highest category level of the full-range rubric

Comparing Prompt Difficulty

To determine the extent to which the expert-predicted item difficulty aligned with the full-range and at-level rubrics, the item difficulty statistics for the ratings of both rubrics were calculated. The full-range scale spanned from 0 to 7, and the averages of all three measures were in the middle of the scale with the ratings range from

Table 7 Comparison of speaking level item statistics

Item	[a]Expert predicted item difficulty	Full-range rubric fair average by prompt	At-level rubric converted fair average
L2-1	1.86	3.54	2.51
L2-2	2.00	3.58	2.46
L3-1	2.71	3.52	3.39
L3-2	2.71	3.79	3.14
L4-1	3.13	3.74	4.17
L4-2	3.38	3.67	4.24
L5-1	4.63	3.50	5.35
L5-2	4.75	3.75	5.12
L6-1	5.00	3.82	5.84
L6-2	5.50	3.71	5.87
Mean	3.57	3.66	4.21
S.D.	0.71	0.11	1.30

[a]Based on average from eight different expert raters

Table 8 Post-study correlations between item difficulty measures

	At-level rubric item difficulty	Full-range rubric item difficulty
Expert-predicted item difficulty	0.98	−0.43
At-level rubric item difficulty		−0.42

Note Correlations greater ±0.77 are significant at $p < 0.01$ (2-tailed)

3.50 to 3.82 (see Table 7), while the at-level scale had item ratings range from 2.51 to 5.87 (see Fig. 5).

A Pearson Product moment correlation was run between all of the item measures (see Table 8): the expert-predicted item difficulty, the full-range scale and the at-level scale. The highest correlation was between the at-level scale and the expert-predicted difficulty ($r = 0.98$, $p < 0.001$), but the full-range scale had a slight inverse relationship ($r = -0.43$) with the EID.

Discussion and Conclusions

Both the full-range and at-level rubrics functioned well as rating scales. With the exception of the 0 category with the full-range scale, each of the categories of both scales were used with enough frequency that there was no need to combine adjacent categories. Both of the scales resulted in high separation statistics between examinees, raters and items. The separation between examinees is desirable, but the

separation between raters could be cause for concern. The decision on how to treat rater disagreements depends if they are to be treated as "rating machines" that are interchangeable with one another or if they are to be viewed as "independent experts" that are self-consistent but whose ratings must be mathematically modeled and transformed to provide examinees with a fair score (Linacre 1999). It is often assumed that enough training can force raters to act as machines, however, without careful follow-up, that may not be the case (McNamara 1996) and acknowledging the disagreements of independent experts through mathematical modeling can more fully reflect real-world rating circumstances (Eckes 2011).

While both the full-range and at-level scales resulted in item difficulty statistics that were statistically separated, the ordering of the full-range rubric difficulties did not align with the experts' predictions. There are a number of possible explanations as to why the first could be that the descriptors in the scale upon which the items were written were flawed. While possible, the scale was based on the well-established ACTFL scoring rubric that has been in use for over 30 years, and after the 1999 revision was validated in inter-rater consistency in over 19 languages (Surface and Dierdorff 2003). Second, there is the possibility that the items did not adequately reflect the scales' descriptors. The raters that evaluated the prompts to determine their intended difficulty levels had a minimum of 3 semesters of rating experience with the average number of semesters being 4.75 semesters. These raters felt that the items did align with the rubric they used for rating.

Another possibility is likely the existence of a pervasive restricted-range error in using a full-range scale to rate a single item. When an item is targeted at Level 6, and the rater knows it is targeted at Level 6, that rater might be hesitant to give scores on the lower end of the scale (0, 1, and 2) even if the respondent language is characteristic of those levels. Similarly an item targeted at Level 2 might result in ratings that are not in the higher part of the range (5, 6, and 7) because the prompt did not elicit language in that upper range. This range restriction in scoring could have resulted in fair item averages that clustered close to the mean of all the items. One piece of evidence of this possibility is the fact that full-range ratings had the smaller fair average standard deviation (see Table 7) of the two scoring methods.

One additional explanation is that that raters did not adhere to the rubric when scoring responses provided by individual at the individual prompt level. Rather raters may have scored the response to each prompt based on how well they answered and not whether the response provided evidence that the examinee was able to speak well at that level. For example, the content of an examinee's response may have been very interesting, yet the language produced did not have the requisite abstract vocabulary and command of more complex grammar needed for a higher level rating. In this instance, the rater may have awarded a higher score than justified by the defined categories of the rubric. This may be the case as it is unlikely that a prompt intended to elicit a response demonstrating the examinees ability to speak at a basic level would consistently provide evidence that the examinee was able to speak at a higher level.

The use of the at-level scale, however, allowed for the EIDs to align with empirical item difficulties. Another benefit of this scale was that it gave information on prompts that were intended to elicit language at the same level. With the items in this study, for example, the prompts at Levels 2, 4 and 6 could be used interchangeably with the other prompts at those intended levels because their item difficulty parameters had comparable values. The prompts at Levels 3 and 5 however were not comparable. Item L5-1 was more difficult than Item L5-2 and similarly item L3-1 was more difficult than Item L3-2. If there were more prompts, then test developers could chose those that would create equivalent test. The prompt fit statistics were indicative of high internal consistency with an average mean outfit square of 1.0 and an average mean infit square of 1.0.

Discussion and Review of Findings

The research question this study addressed explored how the application of a scoring rubric (full-range and at-level) affected the reliability of the results as well as how the two rubric applications compared with the expert-predicted item difficulties. The implications of these findings impacts how to best create equivalent test forms for speaking exams. In creating a speaking proficiency test that is tied to a set of standards, an item writer would try to elicit a specific proficiency level of speech in the construction of the prompt. Consider the following prompts: (1) Describe your house and the neighborhood you live in, and (2) What is the impact of government subsidized housing on the quality of life in a neighborhood?

In the first prompt, the intent is to elicit speech at the ACTFL Intermediate level, whereas in the second prompt, the intent is to elicit speech at the Superior level. If those intended prompt difficulties do not align with the empirical human rating, there are important implications for item writers attempting to create parallel test forms. The determination of item equivalence from one test form to the next needs to be justified by demonstrating that item writers can reliably write prompts to an intended difficulty level.

Training raters on a full-range scale would be ideal for many reasons. First, they would have an understanding of the entire range of a set of standards and see how any performance relates to those standards. Feedback on examinee performance could be easily provided to examinees, teachers and other stakeholders and it could occur independent of the task presented to the examinee. Unfortunately the ratings of examinee responses at the prompt level using a full-range scale did not align with the EID levels. First, the item difficulty statistics had very little variance (SD = 0.27). In fact, Fig. 2 illustrated that the difference in the raters was greater than that of the items (SD = 0.48). Furthermore, the correlation between the EID levels and the full-range item fair averages were not statistically significant and inversely correlated ($r = -0.43$). The incongruence of trained raters (a) being able to predict differences in prompt difficulty yet (b) being unable to find performance differences from the prompts leads one to question the full-range rubric rating

approach. Using the holistic 8-level scale on each item seemed to have introduced a restricted-range error. This could be an artifact of telling the raters the intended level of the prompt they were rating, but it could also be failure to use the rubric properly. Raters more likely scored the responses for each item based on how well they provided evidence the examinee responded to the prompt.

The full-range rating scale spanned a range of language possibilities from simple sentences on familiar topics (Level 2) to extensive, complex speech on abstract academic topics (Level 6), but each of the individual prompts was aimed at only one of those levels (i.e., each prompt was intended to elicit evidence of speaking ability at a specific level and not higher). This misalignment created challenges and perhaps even cognitive dissonance for the raters. For instance, it would be difficult for a prompt targeted at a Level 2 task to elicit a speech sample much higher than a Level 3 or at most a Level 4, even if the examinee did respond with more extensive speech. The rater might be reticent in awarding a rating that was more than 2 levels higher than the prompt's intended difficulty level. Conversely, when a rater was scoring a failed attempt at a prompt targeted at Level 6 task, it might be difficult to know why an examinee was failing to perform at that level and there might be little evidence about what level the examinee could accomplish. The failure to offer an academic opinion on complex topics could mean the examinee was a beginning speaker with almost no speaking ability or it could be an intermediate speaker suffering linguistic breakdown because of the increased cognitive load. Raters might not know how low to rate such breakdown and may be reticent to assign a rating more than 2 or 3 levels below the prompt's intended difficulty level. Thus the ratings for all of the prompts judged with the full-range rubric clustered around the mean (mean = 3.66, SD = 0.11).

Using an at-level scale for each item (through the conversion of the holistic rubric ratings) functioned much better from a measurement perspective. First, there was a wider dispersion of the prompt difficulty means (mean = 2.79, SD = 1.30). Second, the Rasch analysis showed the categories had the most uniform distribution so the categorical differences in ratings examinees received were the most equidistant (see Fig. 3). Finally, there was a much stronger relationship ($r = 0.98$, $p < 0.01$) with the expert-predicted difficulties (see Table 8) than there had been with the holistic scale ratings. Therefore, the low relationship established through using the full-range scale at the item level could be more indicative of scale misuse than the inability of the raters to differentiate performance when judging the different prompts. The result seems to indicate that either responses obtained from lower level prompts did provide some evidence of the examinee's ability to speak at a higher level or that raters tended to rate the respondents overall quality of the response on an 8-level scale. Either way, an analysis of the at-level scale data verifies that the intended prompt difficulty did affect the overall assessment of speaking ability. From the result of this analysis it was also noted that those prompts intended to elicit evidence of speaking ability at Levels 2 (L2-1, L2-2), 4 (L4-1, L4-2), and 6 (L6-1, L6-2) were of equal difficulty within level, but the prompts at Levels 3 (L3-1, L3-2) and 5 (L5-1, L5-2) were not of equal difficulty. Analyzing prompts in this way can provide evidence of test equivalence when

attempting to create parallel forms of an assessment. It also provides an item-level statistic of difficulty that could be used when creating item banks. Finally, since the at-level scale is a subset of the full-range, it can maintain many of the advantages of the full-range scale through simple mathematical conversion.

Limitations and Future Research

In this study, the raters who judged the item responses had been trained to rate overall performances with a holistic scale. They were not given any instruction or exemplars on how to apply the scale at the item level, and the task may have been untenable, as the rubric was not designed for use at the micro level of item. This design weakness might have been overcome if there had been more rater training that focused on the item level. Through the training, the challenge of implementing a holistic scale at the item level could have emerged and a change to the design could have been implemented at that time. Fortunately, the existing holistic scale could be converted after the fact so a more accurate analysis could still be made. Were the research to be done again, it would be better to (a) initially design the scale at the item level and (b) trial the scale to ensure it functions as intended. This would have avoided the step of needing to conduct a post hoc analysis.

Treating a rubric with three distinct axes (text type, content, and accuracy) as a unidimensional construct could have affected the rating as well. Raters making expert judgments of performance have a cognitive load placed upon them that could be simplified by letting them focus only on one aspect at a time. Then, if a multidimensional IRT model had been applied, the findings might have been different as well.

Conclusions

Using full-range rubrics to rate individual items that are targeted at specific levels is problematic and should be done only with caution and verification that the ratings are free from rater error (central tendency, range restriction or logical errors). In this study, a 5-point at-level scale derived from a full-range application of the rubric but targeted at the intended level of the prompt yielded much better results. Using this scale, prompts that were targeted to elicit speech at the same level were more likely to represent their intended empirical difficulty levels. There was a clear separation in the scoring based on the intended prompt difficulty levels which would allow for these data to be used when creating equivalent forms of a test or selecting items for adaptive testing.

Appendix A

Speaking Rubric

Level	Text Type	Accuracy	Content
7—leaving Academic C	Exemplified speaking on a paragraph level rather than isolated phrases or strings of sentences. Highly organized argument (transitions, conclusion, etc.). Speaker explains the outline of topic and follows it through.	• Grammar errors are extremely rare, if they occur at all; wide range of structures in all time frames; • Able to compensate for deficiencies by use of communicative strategies—paraphrasing, circumlocution, illustration—such that deficiencies are unnoticeable; • Pausing and redundancy resemble native speakers; • Intonation resembles native-speaker patterns; pronunciation rarely if ever causes comprehension problems; • Readily understood by native speakers unaccustomed to non-native speakers.	• Discuss some topics abstractly (areas of interest or specific field of study); • Better with a variety of concrete topics; • Appropriate use of formal and informal language; • Appropriate use of a variety in academic and non-academic vocabulary.
6—starting Academic C	Fairly organized paragraph-like speech with appropriate discourse markers (transitions, conclusion, etc.) will not be as organized as level 7, but meaning is clear.	• Grammar errors are infrequent and do not affect comprehension; no apparent sign of grammatical avoidance; • Able to speak in all major time frames, but lacks complete control of aspect; • Pausing resembles native patterns, rather than awkward hesitations;	• Uses appropriate register according to prompt (formal or informal); • Can speak comfortably with concrete topics, and discuss a few topics abstractly; • Academic vocabulary often used appropriately in speech.

(continued)

(continued)

Level	Text Type	Accuracy	Content
		• Often able to successfully use compensation strategies to convey meaning.	
5—starting Academic B	Simple paragraph length discourse.	• Uses a variety of time frames and structures; however, speaker may avoid more complex structures; • Exhibits break-down with more advanced tasks—i.e. failure to use circumlocution, significant hesitation, etc; • Error patterns may be evident, but errors do not distort meaning; • Pronunciation problems occur, but meaning is still conveyed; • Understood by native speakers unaccustomed to dealing with non-natives, but 1st language is evident.	• Able to comfortably handle all uncomplicated tasks relating to routine or daily events and personal interests and experiences; • Some hesitation may occur when dealing with more complicated tasks; • Uses a moderate amount of academic vocabulary.
4—starting Academic A	Uses moderate-length sentences with simple transitions to connect ideas. Sentences may be strung together, but may not work together as cohesive paragraphs.	• Strong command of basic structures; error patterns with complex grammar; • Pronunciation has significant errors that hinder comprehension of details, but not necessarily main idea; • Frequent pauses, reformulations and self-corrections; • Successful use of compensation strategies is rare; • Generally understood by sympathetic speakers accustomed to speaking with non-natives.	• Able to handle a variety of uncomplicated tasks with concrete meaning; • Expresses meaning by creating and/or combining concrete and predictable elements of the language; • Uses sparse academic vocabulary appropriately.

(continued)

(continued)

Level	Text Type	Accuracy	Content
3—starting Foundations C	Able to express personal meaning by using simple, but complete, sentences they know or hear from native speakers.	• Errors are not uncommon and often obscure meaning; • Limited range of sentence structure; • Intonation, stress and word pronunciation are problematic and may obscure meaning; • Characterized by pauses, ineffective reformulations; and self-corrections; • Generally be understood by speakers used to dealing with non-natives, but requires more effort.	• Able to successfully handle a limited number of uncomplicated tasks; • Concrete exchanges and predictable topics necessary for survival; • Highly varied non-academic vocabulary.
2—starting Foundations B	Short and sometimes incomplete sentences.	• Attempt to create simple sentences, but errors predominate and distort meaning; • Avoids using complex/difficult words, phrases or sentences; • Speaker's 1st language strongly influences pronunciation, vocabulary and syntax; • Generally understood by sympathetic speakers used to non-natives with repetition and rephrasing.	• Restricted to a few of the predictable topics necessary for survival (basic personal information, basic objects, preferences, and immediate needs); • Relies heavily on learned phrases or recombination of phrases and what they hear from interlocutor; • Limited non-academic vocabulary.
1—starting Foundations A	Isolated words and memorized phrases.	• Communicate minimally and with difficulty; • Frequent pausing, recycling their own or interlocutor's words; • Resort to repetition, words from their	• Rely almost solely on formulaic/memorized language; • Very limited context for vocabulary; • Two or three word answers in responding to questions.

(continued)

(continued)

Level	Text Type	Accuracy	Content
		native language, or silence if task is too difficult; • Understood with great difficulty even by those used to dealing with non-natives.	
0—starting foundations prep.	Isolated words.	• May be unintelligible because of pronunciation; • Cannot participate in true conversational exchange; • Length of speaking sample may be insufficient to assess accuracy.	• No real functional ability; • Given enough time and familiar cues, may be able to exchange greetings, give their identity and name a number of familiar objects from their immediate environment.

References

ACTFL Proficiency Guidelines 2012 (n.d.). In *American Council on the Teaching of Foreign Languages Proficiency website*. Retrieved from http://actflproficiencyguidelines2012.org/.

Bargainnier, S. (2004). Fundamentals of Rubrics. In D. Apple (Ed.), *Faculty Guidebook*. Lisle, IL: Pacific Crest Inc.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah, NJ: Lawrence Erlbaum Associates.

Buck, K., Byrnes, H., & Thompson, I. (1989). *The ACTFL oral proficiency interview tester training manual*. Yonkers, NY: ACTFL.

Clifford, R., & Cox, T. (2013). Empirical validation of reading proficiency guidelines. *Foreign Language Annals*, *46*(1), 45–61. Retrieved from Google Scholar.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Fort Worth, TX: Harcourt Brace.

Eckes, T. (2011). *Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments*. Frankfurt, Germany: Peter Lang.

Engelhard, G., Jr. (2008). Historical perspectives on invariant measurement: Guttman, Rasch, and Mokken. *Measurement, 6*(3), 155–189.

Hombo, C. M., Donoghue, J. R., & Thayer, D. T. (2001). *A simulation study of the effect of rater designs on ability estimation [Research Report (RR-01-05)]*. Retrieved from Google Scholar: Educational Testing Service.

Linacre, J. M. (1999). *A user's guide to FACETS*. Chicago: MESA press.

Linacre, J. M., & Wright, B. D. (2009). *A user's guide to WINSTEPS*. Chicago: MESA press.

McNamara, T. F. (1996). *Measuring second language performance*. London, UK: Longman.

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.

National Governors AssociationCenterfor Best Practices & Council of Chief State School Officers. (2010). *Common core state standards*. Washington, DC: Authors.

Tierney, R. & Simon, M. (2004). What's still wrong with rubrics: focusing on the consistency of performance criteria across scale levels. *Practical Assessment, Research & Evaluation*, 9(2). Retrieved from http://PAREonline.net/getvn.asp?v=9&n=2.

Verhelst, N., Van Avermaet, P., Takala, S., Figueras, N., & North, B. (2009). Common European framework of reference for languages: Learning, teaching, assessment. ISBN:0521005310.