

Received: December 12, 2011

Accepted: February 1, 2012

## How Does “Sentence Structure and Vocabulary” Function as a Scoring Criterion Alongside Other Criteria in Writing Assessment<sup>1</sup>?

Vahid Aryadoust<sup>2</sup>

### Abstract

Several studies have evaluated sentence structure and vocabulary (SSV) as a scoring criterion in assessing writing, but no consensus on its functionality has been reached. The present study presents evidence that this scoring criterion may not be appropriate in writing assessment. Scripts by 182 ESL students at two language centers were analyzed with the Rasch partial credit model. Although other scoring criteria functioned satisfactorily, SSV scores did not fit the Rasch model, and analysis of residuals showed SSV scoring on most test prompts loaded on a benign secondary dimension. The study proposes that a lexico-grammatical scoring criterion has potentially conflicting properties, and therefore recommends considering separate vocabulary and grammar criteria in writing assessment.

**Keywords:** *Lexico-grammatical scoring, Rasch model, Partial credit model (PCM), L2 writing, Writing models*

### 1. Introduction

To accurately assess second-language (L2) writing, an approach is that the designers of assessment frameworks base them on socio-cognitive models that capture the linguistic, social, and cognitive features that influence L2 writing performance (Weir, 2005). Not all frameworks meet this criterion. Some are developed based on the views of test designers, and some might misrepresent the theoretical framework (Lynch & McNamara, 1998). For example, whereas multiple theoretical frameworks emphasize that the influence of rater’s background on test takers’ scores should be controlled, in reality this influence is sometimes difficult to partial out, as raters’ performance can become confounded by a halo effect (e.g., Aryadoust, 2012a; Engelhard, 1994; Fahim & Bijani, 2011). Yet, laudable efforts to develop and promote validated

<sup>1</sup> Portions of this article were presented at the Japan Association for Language Teaching (JALT, 2009) Conference and the Pacific Rim Objective Measurement Symposium 2009 (PROMS 2009) in Hong Kong.

<sup>2</sup> Centre for English Language Communication, National University of Singapore, Singapore. Email: [elcsva@nus.edu.sg](mailto:elcsva@nus.edu.sg)

models have been made by Weir (1990), McNamara (1996), Engelhard (1992, 1996), Kondo-Brown (2002), Schaefer (2008), and Saito (2008).

That writing assessment frameworks map properly onto a socio-cognitive model of performance is particularly important in high-stakes and foreign language (FL) tests (Weir, 2005). One such high-stakes test is the International English Language Testing System (IELTS), which assesses writing by assigning two tasks (expository and graph-description) to test takers and scoring performance based on analytical scoring criteria. As I argue below, although the test<sup>3</sup> generally maps well onto the socio-cognitive model, the validity of the Sentence Structure and Vocabulary (SSV) scoring criterion, which attempts to operationalize or measure both vocabulary and grammatical competencies as a single criterion, is as yet in question. A closer examination of the relevant literature shows that there is no scholarly consensus on the use of such a mixed criterion, and as I will argue below three lines of research have generated divergent evidence about its efficacy. The purpose of this study is to investigate the features of SSV. This article reviews the literature on models of writing and on the SSV criterion, then proceeds to the study and conclusions.

## 2. Literature Review

### 2.1 Writing Models

Research has proposed a number of linguistic, social, and cognitive features that influence second language (L2) writing performance. Hayes and Flowers (1980) developed a model of writing which detailed the interactive processes of writing as setting a goal, plan, revision, and overseeing the whole process by a “monitor,” but it fell short to explain the observed differences in trait levels of writers (Hyland, 2002). Shortly after this framework, Bereiter and Scardamalia (1987) developed a “knowledge-telling” and “knowledge-transforming” model in first language (L1) which captured the differences in the trait level of writers. The knowledge-telling component concerned the ability of expert writers in planning, revising, and writing based on their “internal resources” and knowledge-transforming component addressed “how skilled writers use the writing task to analyze problems and set goals” (Bereiter & Scardamalia, 1987, p. 28). Like Hyland and Flowers’ (1980) model, this model puts a focus on the cognitive processes of writers but lacks explicit attention to the social and contextual aspects of the skill (Hyland, 2002).

Hayes (1996) proposed a model of social aspects of writing. The socio-physical environment in the Hayes model was regarded as an important factor in completing writing tasks; both the audience of the writings and the text produced shape writer’s composition. Additionally, individuals’ motivation and cognitive processes are recognized in the Hayes model. On the one hand, writers’ goals, predispositions, attitudes, and cost estimates, and on the other, text interpretation, reflection, and production influence writing style. Although this model details many underlying processes and skills in writing, it lacks “specificity in defining the situational variables involved in writing” (Weigle, 2002, p. 28).

---

<sup>3</sup> This statement and other citations of the IELTS writing test throughout the article refer to the test structure before the fairly recent revisions of some of the scoring criteria of the test (see Shaw & Falvey, 2008).

Adopting Chapelle, Grabe, and Berns' (1997) communicative language use model, Grabe and Kaplan (1996) addressed the shortcomings of the previous models by focusing on social and cognitive issues in writing. In this model, task environment entails participant, context, task, text, and topic. Grabe and Kaplan employed the theory of Bachman's (1990) language knowledge theory<sup>4</sup>—supported by a number of researchers exploring the constituent linguistic structure of writing (Evola, Mamer, & Lentz, 1980; Connor, 1991; Hamp-Lyons, 1991; Hedge, 2005; Homburg, 1984)—and proposed that linguistic knowledge, discourse knowledge, and sociolinguistic knowledge contribute to L2 writing performance.

Finally, based on (socio-cognitive) models of writing, a number of writing assessment frameworks have been developed (Murphy & Yancey, 2008 [in L1]; Shaw & Weir, 2007). For example, Murphy and Yancy (2008) argued that theory-informed writing assessment framework should address scoring criteria, test takers' characteristics, raters, and contextual variables.

## 2.2 Assessment Frameworks

An important decision to be made in developing writing assessment frameworks is which type of scoring criteria to use: holistic scoring, where a text is assigned a single grade intended to represent the overall writing skills of the test taker, or analytic scoring, where scripts are rated separately on multiple criteria, such as organization, cohesion, coherence, vocabulary, grammar, and mechanics (Weigle, 2002).

Research shows that analytic criteria are more reliable and appropriate in assessing L2 writing and can provide diagnostic feedback (Aryadoust, 2012a; Kim, 2011; Knoch, 2009; Knoch & Elder, 2010; Weigle, 2002) although they can introduce tension and inaccuracy when a single scoring criterion is a mixture of two or more underlying writing skills, as when grammar and vocabulary are combined into a single criterion (Mickan, Slater, & Gibson, 2000; Mickan & Slater, 2003; Brown, 2006). Though expensive and time-consuming, analytic scoring is commonly used in high-stakes tests of writing such as the International English Language Testing System (IELTS) and the Test of English for Educational Purposes (TEEP).

Based on previous research (e.g., McNamara, 1990, 1996; Chapelle, Grabe, & Berns, 1997; Grabe & Kaplan, 1996; Hayes, 1996; Riazi & Rezaii, 2011), Weir (2005) proposed a general analytic writing assessment model which recognizes four elements in the assessment environment: (a) writers' knowledge of expectations of the text reader and the appropriate genre of the text; (b) the objective of the writing; (c) "the conventions relating to the text type or genre" the writer is involved with. McNamara refers to this feature as "intertextuality" of the text; and (d) linguistic knowledge comprising knowledge of vocabulary, syntax, language functions, and pragmatics (see Bachman, 1990).

---

<sup>4</sup> Bachman's (1990, p. 80) model "attempts to characterize the processes by which the various components interact with each other and with the context in which language use occurs", and includes language competence, strategic competence and psychophysiological mechanisms. Further explanation of the model falls out of the scope of this article. Readers are referred to Bachman's work (see also Motallebzadeh & Baghaee Moghaddam, 2011).

Analytic scoring criteria in the IELTS writing module and similar tests seem to meet Weir's (2005) framework concerns except, as I view it, in the case of scoring vocabulary and grammar. There are three major scoring criteria in the IELTS academic writing Task 1, which is a graph description task: Sentence Structure and Vocabulary (SSV), Task Fulfillment (TF), and Coherence and Cohesion (CC); and three in Task 2, an essay writing task in response to a prompt: SSV, Argument, Ideas and Evidence (AIE), and Communicative Quality (CQ) in Task 2. These two tasks are scored on slightly different scoring criteria due to their nature. In Task 1 clear and concise communication of necessary information to explain graphs or tables is important, whereas Task 2 is intended to measure articulating personal ideas, offering evidence to support oneself, and communicating thoughts are more important in Task 2.

In this framework, TF, CC, AIE, and CQ fit with Weir's (2005) general description of a writing assessment framework; TF addresses the objectives of the writing, the text content, and its relation to the task; CC and CQ address cohesion, coherence, and organization in the text, which are important elements of linguistic knowledge; and AIE addresses the skillful management of paragraphs and writers' knowledge of expectations of the text reader. Yet, there are two major concerns about SSV although it is intended to operationalize linguistic knowledge: (a) as literature shows, the evidence of validity and usefulness of this criterion is controversial; and (b) mixing grammar and vocabulary is not resonant with the theory of language knowledge proposed by Bachman (1990, p. 87). Bachman divided the grammatical competence into four components: vocabulary, morphology, syntax, and phonology/graphology, indicating that "the choice of words" is a different component than "their arrangement...to express propositions." For instance, a student is asked to write a short description of a picture showing a man and a woman with different heights. The student demonstrates his/her vocabulary knowledge by choosing appropriate words to address the picture. His/her knowledge of syntax is demonstrated through arranging the words in a grammatical order (Bachman, 1990). That his/her knowledge of vocabulary may be more extended than syntax can obscure the meaning of a score assigned to the SSV scoring criterion.

### 2.3 Sentence Structure and Vocabulary as a Scoring Criterion

Sentence structure and vocabulary (SSV) has been a commonly used scoring criterion in the academic and general modules of the International English Language Testing System (IELTS) for a long time. This criterion measures both "sentence structure," the use of grammar in the text, and "vocabulary," the range and appropriateness of vocabulary. SSV is regarded as a significant constituent of some of the writing assessment models, as it determines the level of performance (as opposed to competence) and proficiency of the test takers in using grammar and vocabulary in FL writing. Further, using SSV accurately in writing would indicate the satisfaction of some of the concerns of the socio-cognitive writing assessment model (Weir, 2005). For example, it can point to students' knowledge of expectations of the text reader, the appropriate use of vocabulary, and extended linguistic knowledge.

Several studies have evaluated SSV as a scoring criterion in assessing writing, but no consensus on its functionality has been reached; instead, three lines of research emerge from the literature. First, a number of researchers have criticized SSV as a mixed criterion with the

potential to puzzle raters (Mickan, Slater, & Gibson, 2000; Mickan & Slater, 2003; Mickan, 2003). For example, Mickan (2003) found it difficult to determine the “lexico-grammatical” features of texts belonging to different levels, and suggested a more holistic or impressionistic approach to scoring writing; he particularly highlighted the SSV criterion as having the potential to cause confusion among raters. This conclusion was further supported by Mickan and Slater (2003). Relatedly, Read and Nation (2006) showed that vocabulary knowledge associates with scores only in lower-ability writers; this relationship does not hold at higher ability levels. As Read and Nation (2006, p. 229) wrote, “Whereas broad distinction can be identified across [IELTS] band score levels, we find considerable variation in vocabulary use by candidates within levels”. If such variation also exists at the grammar level, then attempts to score grammar and vocabulary together within the stringent boundaries of a single lexico-grammatical scoring criterion might face serious challenges.

The second line of research does not specifically examine the SSV scoring criterion, but indicates that when grammar and vocabulary are scored as two separate scoring criteria, the scale functions optimally. McNamara (1990, 1996) used a six-component category specification in a writing study, comprising overall effectiveness, intelligibility, fluency, comprehension, appropriateness, and vocabulary resources. He found that these factors contribute significantly to writing performance. In McNamara’s model, along with Engelhard (1992, 1994, 1996), Kondo-Brown (2002), Schaefer (2008), and Saito (2008), grammar and vocabulary are regarded as two separate scoring criteria. In these studies, the criteria functioned in accord with the stringent expectations of the Rasch model, indicating the scores were not confounded by unmodeled factors, such as the interaction of raters and scoring criteria. These frameworks operationalized Bachman’s (1990) grammatical competence as separate components of vocabulary and grammar.

The third line of research supports combining grammar and vocabulary into a single test criterion. Eckes (2005) found that grammar and vocabulary were tested accurately under “linguistic realization” as a general heading alongside other scoring criteria, and Banerjee, Franceschina, and Smith (2007) found that vocabulary and grammar correctness appear to complement each other and can be classified under a single rubric. In an earlier study, Weigle (1998) constructed a scoring criterion comprising grammar, vocabulary, register, and mechanics. This criterion fitted the Rasch model very well, and appeared to account for a considerable amount of the variance in examinees’ scores. Other studies in this research direction have produced controversial findings. Gennaro (2009) developed a combined vocabulary-grammar scoring criterion, called “grammatical control”. Gennaro (2009, p. 552) argued that the criterion might not have identified the differences in students’ grammatical and vocabulary skills precisely, and called for “analyzing more precise categorical differences in examinees’ grammatical production”.

This study further investigates the quality of the SSV scoring criterion alongside other criteria. The study presents the results of a series of tests that rated students using either of two sets of three criteria: SSV, Task Fulfillment (TF), and Coherence and Cohesion (CC); and SSV, Argument, Ideas and Evidence (AIE), and Communicative Quality (CQ).

To investigate the psychometric features and fit of these scoring criteria, the rating scale, and students’ performance, these data were subjected to Rasch model analysis using the Rasch partial

credit model (RPCM), an expansion of the logistic Rasch model which is used to measure the psychometric features of polytomous rating scales. I use the model to fulfill the following objectives:

To measure the difficulty of each scoring criterion, the level and range of examinee abilities, and their fit to the Rasch model. To evaluate the rating scale used in the tests. Response categories should increase monotonically (Bond & Fox, 2007); otherwise, they and their thresholds do not function satisfactorily. This objective is specifically concerned with the functionality of the scale and scoring criteria.

### 3. Methodology

#### 3.1 Participants

The data in this study were collected from the performance of 182 EFL students aged between 19 and 28 ( $M = 24.04$ ;  $SD = 2.5$ ), who took several exams including the placement test for International English Language Testing System (IELTS) preparation courses and class tests. Each participant took a placement test and two follow-up tests (see Procedures).

#### 3.2 Materials

Two major prompt types were assigned to students: expository writing and graph description, which is according to the curriculum guidelines. According to Loughheed (2004) and Aryadoust, Akbarzadeh, and Nasiri (2007a, b), expository writing tasks fall into at least four major categories: (a) Agreement-Disagreement (AD), (b) Stating a Preference (SP), (c) Giving an Explanation (GE), and (d) Making Arguments (MA). "Expository tasks" and "Task 2" are used interchangeably in this article because these tasks were similar to the IELTS Task 2 prompts in timing (40 minutes), desired minimum word count (250 words), wording, and structure (the reader of the text, time limit, and desired minimum word count are articulated in the prompt).

Following O'Loughlin and Wigglesworth (2003), I selected two graph description tasks, one highly information-loaded and the other less complex (Appendix 1) to explore whether the load of information in each prompt affects the difficulty level of scoring criteria. These tasks were similar to IELTS Writing Task 1 in their wording and timing, and are referred to as "Task 1."

A panel of experienced English teachers analyzed several Task 2 prompts and judged four to be appropriate for the purpose of the present study; the teachers selected the prompts based on the curriculum guidelines, which mandate that prompts contain no culturally unknown features.

#### 3.3 Procedures

Participants attended up to three testing sessions. The first session consisted of one hour for responding to the prompts and 15-minute interval, and was a placement test comprising an

Agreement-Disagreement (AD) Task 2 prompt and the less complex Task 1 prompt. This session was compulsory for all applicants to IELTS preparation courses, and determined their placement. The two remaining tests were administered within 15 days of placement. The second testing session consisted of one hour for responding to the prompts and 15 minutes of interval, and was designed to investigate which of students' writing skills needed particular attention. This session was also compulsory, comprising a Stating a Preference (SP) Task 2 prompt and the more information-loaded Task 1 prompt. The third testing session took 80 minutes for responding to the prompts and 15 minutes of interval between prompts, was optional, and consisted of a Giving an Explanation (GE) and a Making Arguments (MA) Task 2 prompt. Most students chose to participate in this testing session. Table 1 describes these three testing sessions.

**Table 1.** Descriptions of Testing Sessions in the Study

Session	Duration	Tasks	Remarks
Session 1	60 minutes	(a) less information-loaded graph description (Task 1) (b) A/D expository task (Task 2)	Compulsory placement test
Session 2	60 minutes	(a) more information-loaded graph description (Task 1) (b) SP expository task (Task 2)	Compulsory test to assess writing skills and set the writing class schedule
Session 3	80 minutes	(a) GE expository task (Task 2) (b) MA expository task (Task 2)	Voluntary

In all, each student responded to a total of four Task 2 prompts and two Task 1 prompts, so there were 1092 prompt responses overall, 728 of which correspond to Task 2 prompts and 364 to Task 1 prompts. Although the possible scores ranged from 0 to 9, the effective region of the scale was anticipated to range from 2 to 7. However, the first testing session required the full 10-point scale, since it was a placement test; and to maintain consistency in scoring, the full scale was used in all three testing sessions.

### 3.4 Scoring Procedure

Tests were scored by the researcher using a 10-point rating scale similar to that used in the IELTS writing test (O'Loughlin & Wigglesworth, 2003, pp. 100-113; Hamp-Lyons, 1991; Aryadoust et al., 2007a, 2007b; Jackeman & McDowell, 1996, 2000; Cambridge Practice Tests for IELTS, 2002, 2005, 2006, 2007; Official IELTS Practice Materials, 2007). As I will argue below, recruiting more raters was not economically viable in this study; so, the main part of

scoring was performed by the researcher. Possible scores on this scale range from 0 (no attendance) to 9 (highly competent). The scale is illustrated in Table 2.

**Table 2.** Description of Testing Sessions and Writing Tasks

Scoring category	Levels and description	Task
1. Task Fulfillment (TF): Does the text bear any relation to and answer the task adequately? (or is there any redundant and irrelevant information in the text?) Does the text comprise at least 250 words?	0: No performance	Task 1
	1: Very poor performance	
	2: Poor performance	
	3: Barely adequate	
	4: Adequate	
	5: Fair performance	
	6: Good performance	
	7: Very good performance	
	8: Excellent performance (minor flaws)	
9: Highly competent user		
2. Cohesion and Coherence (CC): How satisfactory is the cohesion of the text? How effective is the use of cohesion in communication? Are the text and its content	0: No performance	Task 1
	1: Very poor performance	
	2: Poor performance	



---

well-organized? How evident, coherent, and	3: Barely adequate	
adequately are the organizational skills?	4: Adequate	
Does the text follow a sequence?	5: Fair performance	
	6: Good performance	
	7: Very good performance	
	8: Excellent performance (minor flaws)	
	9: Highly competent user	
3. Argument, Ideas and Evidence (AIE):	0: No performance	Task 2
How skillfully is the paragraphing managed?	1: Very poor performance	
How well information and ideas are	2: Poor performance	
sequenced and meet the expectations of the	3: Barely adequate	
reader?	4: Adequate	
	5: Fair performance	
	6: Good performance	
	7: Very good performance	
	8: Excellent performance	

---

---

	(minor flaws)	
	9: Highly competent user	
4. Communicative Quality (CQ): How	0: No performance	Task 2
satisfactory is the cohesion of the text? How	1: Very poor performance	
effective is the use of cohesion in	2: Poor performance	
communication? Are the text and its content	3: Barely adequate	
well-organized? How evident, coherent, and	4: Adequate	
adequately are the organizational skills?	5: Fair performance	
Does the text follow a sequence?	6: Good performance	
	7: Very good performance	
	8: Excellent performance	
	(minor flaws)	
	9: Highly competent user	
5. Sentence Structure and Vocabulary (SSV)	0: No performance	Tasks
How accurate are the grammatical patterns?	1: Very poor performance	1 and
How extended is the range of grammatical	2: Poor performance	2
structures used? Do the grammatical errors	3: Barely adequate	
cause misunderstanding? How adequately,		
correctly, and appropriately was the		

---

---

vocabulary used for intended	4: Adequate
communication? Is the word choice	5: Fair performance
appropriate for the intended message? How	6: Good performance
extended was the lexicon? Is there any	7: Very good performance
circumlocution in the text?	8: Excellent performance (minor flaws)
	9: Highly competent user

---

As shown in Table 2, Task 1 scripts were graded on Task Fulfillment (TF), Cohesion and Coherence (CC), and SSV, and Task 2 scripts were graded on Argument, Ideas and Evidence (AIE), Communicative Quality (CQ), and SSV. For example, while TF in Task 1 concerns such questions as “Does the text bear any relation to and answer the task adequately?”, SSV pertains to a different set of questions such as “How accurate are the grammatical patterns? How extended is the range of grammatical structures used?”. Each script was assigned one score per scoring criterion, as well as a total band score calculated by averaging these three scores.

### 3.5 Inter-Rater Reliability

To investigate scoring consistency, I contracted a second rater, an experienced English teacher who had been teaching English courses for approximately five years. After we met to discuss the rating procedures, meanings, and implications of the scoring criteria and rating scales, I gave the rater 40 Task 2 texts and 20 Task 1 texts randomly selected for rescoring. (Complete rescoring of all texts was not possible because of time and budget constraints.) After rescoring, I used SPSS for Windows (Version 16, SPSS Inc., Chicago, IL) to calculate inter-rater reliability coefficients using Cohen’s Kappa index. This index ranges from -1 to +1; values between 0.40 and 0.59 indicate moderate correlations between two raters’ performance, values between 0.60 and 0.79 are substantial, and values above 0.80 are outstanding (Landis & Koch, 1977). Kappa values greater than 0.60 and significant at  $p < 0.05$  indicate a well-constructed measurement. The funding for recruiting raters was critically limited, so only this rater could be recruited.

### 3.6 Internal Consistency and Descriptive Statistics

To measure the overall internal consistency of student scores on the test prompt/scoring criterion, I calculated their Cronbach's Alpha index. This internal consistency measure is an expansion of Kurder-Richardson 20 (KR-20). To measure the consistency of each scoring criterion with the others, I also calculated the corrected item-total correlation of each criterion, and the Cronbach's Alpha index of the remaining criteria when each was excluded. Finally, I calculated descriptive statistics, including mean, standard deviation (SD), skewness, and kurtosis, for all scoring criteria, to summarize the features of data in a clear and understandable way.

### 3.7 Rasch Model

To investigate the psychometric features of the test data, I used the WINSTEPS computer program, Version 3.64 (Linacre, 2009a) to fit the data to an expansion of the Rasch model known as the Rasch partial credit model (RPCM). I used this model to investigate the psychometric qualities of steps (in this case, scoring categories) and their thresholds. The RPCM is expressed, as follows:

$$\log \left[ \frac{P_{nij}}{P_{ni(j-1)}} \right] = B_n - D_i - F_{ij}$$

where:

$P_{nij}$  is the probability person  $n$  being placed in scoring category  $j$  on test item  $i$ ,

$B_n$  is the person ability measure,

$D_i$  is the difficulty measure of item  $i$ , and

$F_{ij}$  is the difficulty of category  $j$  as compared with category  $j-1$ .

I performed a number of statistical analyses to investigate the quality of the test data and their fit to this model. To measure student performance and criterion difficulty, I created a Wright map, which presents both measures on the same logit scale. To investigate the fit of the test data to the Rasch model, I calculated infit and outfit mean square values for both students and test criteria. Fit indices hint at the dimensionality of assessment tools (see R. M. Smith, 1991, 1996, 2000, for a review of fit functions) and represent the degree of congruence between the expectations of the Rasch model and observed data. For example, a fit value of 1.20 indicates that 20% of the observed variance is attributable to noise or unpredicted variables. Bond and Fox (2007) regard 0.6 to 1.4 as the acceptable (yet stringent) range of fit indices in polytomous data, and Linacre (2009b) recommends a more liberal range of infit and outfit mean square values between 0.5 and 1.5. (Infit mean square values are sensitive to unexpected inliers, and outfit mean square values are sensitive to outliers) (see also Baghaei, 2008, and Baghaei & Amrahi, 2009). Values falling outside these boundaries indicate variation which is not predicted by the model.

Finally, to determine the reliability of the data and the presence of defined strata of student performance and criterion difficulty, I calculated both reliability and separation indices.

### 3.8 Rasch Investigation of the Rating Scale

To investigate the psychometric features of the rating scale used in the study, I conceptualized the RPCM using probability curves. The intersections between probability curves give the points which have equal probability indices between categories. These points are thresholds (Linacre, 2009b), and are known as Rasch-Andrich (also Step Difficulty and Step Calibration) thresholds. Thresholds are important because they measure examinee performance. A Rasch-Andrich threshold plot should look like a chain of hills with clear peaks in each response category curve; a plot without peaks indicates disordered transition calibrations.

### 3.9 Unidimensionality and Local Independence

Because the RPCM is used to answer the questions of the study, two prerequisites to latent trait models were probed<sup>5</sup>: unidimensionality and local independence (Wright, 1996). Psychometric unidimensionality occurs when a scoring criterion addresses a single trait—or, arguably, a group of interrelated traits. For unidimensionality to be supported, fit indices must fall within the range between 0.5 and 1.5 and principal component analysis of linearized Rasch residuals (PCAR) should not extract the dimensions that explain more than two eigenvalues (see Hattie, 1985; E. V. Smith, 2002; R. M. Smith & Maito, 1994; for a review of dimensionality analysis). If the extracted dimension explains more than two eigenvalues, it is likely that the data are tainted by a construct-irrelevant factor (Bond & Fox, 2007), which would compromise the construct validity of the instrument. In this analysis, I took the Rasch model as the major dimension of variance in the data and examined residuals which were not accounted for by the Rasch model to identify possible additional dimensions in the data.

I further investigated local independence. Local independence occurs when no datum influences another datum, and is demonstrated by investigating residual correlations. Local independence is important because if performance on a test item relies on successful performance on a prior test item, then failure to answer that prior item would affect the performance on the current item.

## 4. Results

### 4.1 Overview

---

<sup>5</sup> This general statement applies only to unidimensional models such as Rasch and items response theory models. Multidimensional models (including cognitive diagnostic assessment and multidimensional Rasch models) would not require adherence to unidimensionality assumption unless the researcher wishes to investigate the unidimensionality assumption in each of the dimensions separately.

The Rasch model was used to examine the difficulty measures, fit, scoring criteria, thresholds, and reliability indices. Specifically, the SSV scoring criteria range in difficulty and fit were probed. The following sections discuss the results.

#### 4.2 Inter-Rater Reliability

Table 3 presents the results of the inter-rater reliability analysis by rater and scoring criterion. As shown in the table, Kappa coefficients range from a low of .67 in the Task 2/SSV scoring variable to a high of .83 in the Task 1/TF variable. All coefficients are significant at  $p < 0.01$ . Notably, the SSV scoring criterion produced the lowest index (0.67), indicating the least rater agreement.

**Table 3.** Inter-rater Reliability according to the Cohen's Kappa Indices

		First rater					
		Variable	AIE	CQ	SSV	TF	CC
Second rater	AIE	0.78					
	CQ		0.82				
	SSV			0.67			
	TF				0.83		
	CC					0.79	
	SSV						0.71

*Note.* All indices are significant at  $p < 0.01$ .

AIE = Argument, Ideas and Evidence. CQ = Communicative Quality. SSV = Sentence Structure and Vocabulary. TF = Task Fulfillment. CC = Cohesion and Coherence.

#### 4.3 Internal Consistency and Descriptive Statistics

The overall Cronbach's Alpha value when all 21 prompt/criterion scoring variables were included was .919. Table 4 presents corrected item-total correlations for each variable, and the Cronbach's Alpha estimate of the remaining variables when that variable is removed.

**Table 4.** Summary Statistics of Writing Scoring Criteria in Task 1 and 2

Variable	Mean	SD	Skewness	Kurtosis	Corrected item-total correlation	Cronbach Alpha if item is deleted
AD-CQ	4.76	0.85	-0.395	0.652	0.835	0.909
AD-AIE	4.80	0.84	-0.263	0.001	0.855	0.908
AD-SSV	5.05	0.97	-0.320	-1.437	0.420	0.920
SP-AIE	4.80	0.76	-0.418	1.235	0.618	0.914
SP-CQ	4.92	0.94	-0.485	0.345	0.759	0.911
SP-SSV	4.96	0.95	-0.218	-0.991	0.457	0.919
GE-CQ	5.30	0.76	-0.502	0.589	0.788	0.911
GE-AIE	5.18	0.97	-0.445	-0.721	0.766	0.910
GE-SSV	4.95	0.99	-0.222	-1.265	0.509	0.917
MA-CQ	5.03	0.76	0.844	2.171	0.632	0.914
MA-AIE	5.17	0.81	0.363	0.660	0.744	0.911
MA-SSV	4.82	0.90	-0.126	-0.985	0.192	0.925
T11-CC	4.99	1.04	-0.251	-0.131	0.799	0.909
T11-SSV	4.92	0.92	0.234	-0.842	0.199	0.924
T11-TF	4.79	0.84	-0.263	0.001	0.855	0.908
T12-SSV	4.93	0.85	-0.185	-1.005	0.275	0.923
T12-CC	5.07	0.61	-0.040	-0.330	0.422	0.919
T12-TF	5.18	0.97	-0.445	-0.721	0.766	0.910

*Note.* AD = Agreement-Disagreement. SP = Stating a Preference. GE = Giving Explanations.

MA = Making Arguments. T11 = First Task 1 (less complex). T12 = Second Task 1 (more complex). AIE = Argument, Ideas and Evidence. CQ = Communicative Quality. SSV = Sentence Structure and Vocabulary. TF = Task Fulfillment. CC = Coherence and Cohesion.

Number of participants = 182. Task 2 Scripts = 728. Task 1 Scripts = 364.

Most corrected item-total correlations were quite high; however, the SSV criterion generated correlations below 0.3 on the MA Task 2 prompt and both Task 1 prompts (MA-SSV, T11-SSV, and T12-SSV), and below 0.5 on the SP and AD Task 2 prompts (SP-SSV and AD-SSV). Because of these low correlations, the Cronbach's Alpha values increase most when the SSV variables are removed.

Table 4 also presents descriptive statistics for all variables, including mean score, standard deviation (SD), skewness, and kurtosis. Observed scores ranged from 2 to 8 on the rating scale, slightly wider than the expected range from 2 to 7. The GE-CQ variable had the highest mean score (5.30; SD = 0.76), and AD-CQ had the lowest mean score (4.76; SD = 0.89). As shown in Table 4, skewness and kurtosis coefficients fell well within the range from -2.5 to +2.5 expected in a normal distribution (Field, 2005), indicating a good approximation of univariate normality.

#### 4.4 Fitting the Data to the Rasch Model

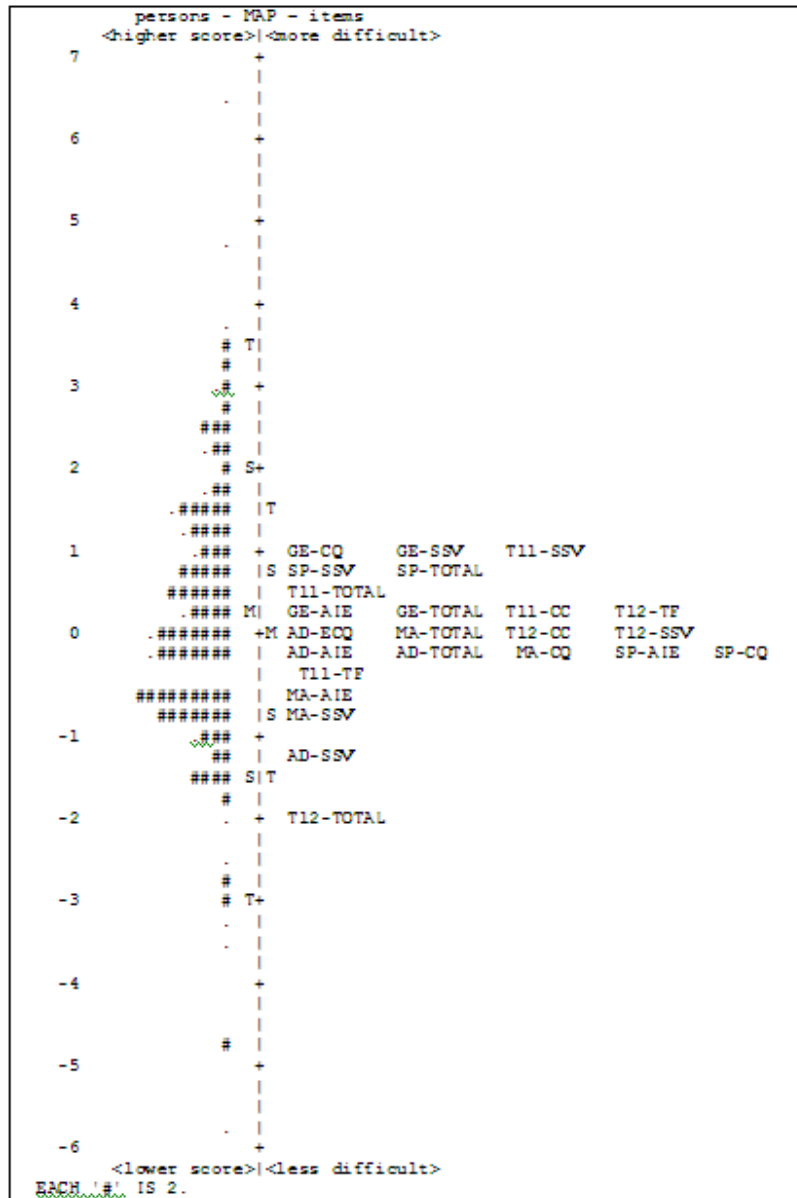
##### 4.4.1 Objective one: Measures, fit, and reliability

In an attempt to investigate the difficulty measures of scoring criteria and their function, I fitted the data into the Rasch model and analyzed item and person measures. Figure 1 is an item-person map (also known as a Wright map), which presents students' measured performance on the left and the measured difficulty of test criteria on the right. The most difficult criteria and highest-performing students land at the top of the map. The units of the map, shown in the leftmost column, are known as logits (log odd units). These provide an interval representation of both criterion difficulty and student performance. The M on the map is the mean score for both items and persons; S represents one standard deviation from the mean; and T represents two standard deviations from the mean.

The most difficult scoring criterion was GE-CQ (difficulty = 1.07), followed by the SSV criterion in the less complex Task 1 prompt (T11-SSV; difficulty = 1.01), the GE-AIE (difficulty = 0.99), and SP-SSV Task 2 (difficulty = 0.79) prompts. Performance on the three scoring criteria in T11 varied relatively more widely than on those in T12, suggesting the latter were more similar in difficulty than the former. Notably, the SSV criteria range rather widely across the difficulty spectrum, indicating that their cognitive demands were different depending on the type of the task, although all but MA-SSV and AD-SSV are of above average difficulty.



**Figure 1.** Item-person map with student performance on the left and scoring criterion difficulty on the right.



Next, I investigated the fit of the test data to the Rasch model by calculating infit and outfit mean square values for all students and test criteria. As Table 5 illustrates, most scoring criteria fit the Rasch model well. Average item infit and outfit mean square indices were 0.99 and 1.13, with respective standard deviations of 0.54 and 0.89. Interestingly, only the SSV scoring criteria were underfits (greater than 1.5); that is, outfit and infit mean square indices for the AD-SSV, SP-SSV, MA-SSV, T11-SSV, and T12-SSV scoring criteria fell outside Linacre’s (2009b) recommended boundaries; and the outfit and infit indices for the GE-SSV criterion, which are 0.49 and 0.51 respectively, are close to the borders of the acceptable fit range (i.e., 0.5—1.5).

The rightmost column in Table 5 gives point-measure correlations for each scoring variable; these are correlations between test scores, with Rasch difficulty measures replacing total scores. Negative or zero values would indicate a problem in the data, such as item or person misfit (Wright & Stone, 1999); however, this is not observed in the table. Most students' performance fit the Rasch model well: average student infit and outfit mean square indices were 1.05 (SD = 0.44) and 1.12 (SD = 1.02). Nine students (4.9%) were misfits.

**Table 5.** Item Measures, Mean Square Fit Indices, and Point Measures

#	Score	Measure	SE	Infit MNSQ	Outfit MNSQ	Scoring criterion	PT- Measure
1	868	0.12	0.12	0.58	0.57	AD-CQ	.85
2	873	-0.21	0.12	0.52	0.52	AD-AIE	.87
3	919	-1.35	0.11	1.63	2.44	AD-SSV	.49
4	884	-0.26	0.13	0.57	0.55	AD-TOTAL	.85
5	896	-0.33	0.13	0.72	0.69	SP-AIE	.79
6	894	0.79	0.11	1.51	2.10	SP-SSV	.53
7	965	-0.30	0.11	1.14	1.16	SP-CQ	.66
8	894	0.76	0.13	0.65	0.64	SP-TOTAL	.81
9	965	1.07	0.13	0.64	0.59	GE-CQ	.80
10	943	0.13	0.11	0.64	0.51	GE-TOTAL	.77
11	902	0.99	0.11	1.41	1.49	GE-AIE	.58
12	934	0.26	0.13	0.51	0.49	GE-SSV	.84
13	916	-0.36	0.14	0.97	0.93	MA-CQ	.70
14	941	-0.50	0.13	0.71	0.69	MA-AIE	.79
15	877	-0.68	0.12	2.23	4.18	MA-SSV	.26
16	909	0.07	0.16	0.81	0.76	MA-TOTAL	.73
17	909	0.35	0.11	0.66	0.62	T11-CC	.82
18	873	-0.21	0.12	0.52	0.52	T11-TF	.87
19	896	1.01	0.11	2.01	2.17	T11-SSV	.37

20	892	0.47	0.13	0.56	0.54	T11-TOTAL	.84
21	898	0.06	0.12	2.17	2.38	T12-SSV	.28
22	943	0.13	0.11	0.64	0.51	T12-TF	.77
23	923	-0.05	0.15	1.28	1.24	T12-CC	.48
24	913	-1.97	0.15	0.78	0.76	T12-TOTAL	.72
Mean	905.8	0.00	1.25	0.99	1.13	NA	NA
SD	25.7	0.70	0.14	0.54	0.89	NA	NA

Note. *SE* = standard error of measurement. *MNSQ* = mean square. *PT-Measure* = point-measure correlation.

AD = Agreement/Disagreement. SP = Stating a Preference. GE = Giving Explanations. MA = Making Arguments. T11 = First Task 1 (Less complex). T12 = Second Task 1 (More complex). AIE = Argument, Ideas and Evidence. CQ = Communicative Quality. SSV = Sentence Structure and Vocabulary. TF = Task Fulfillment. CC = Coherence and Cohesion.

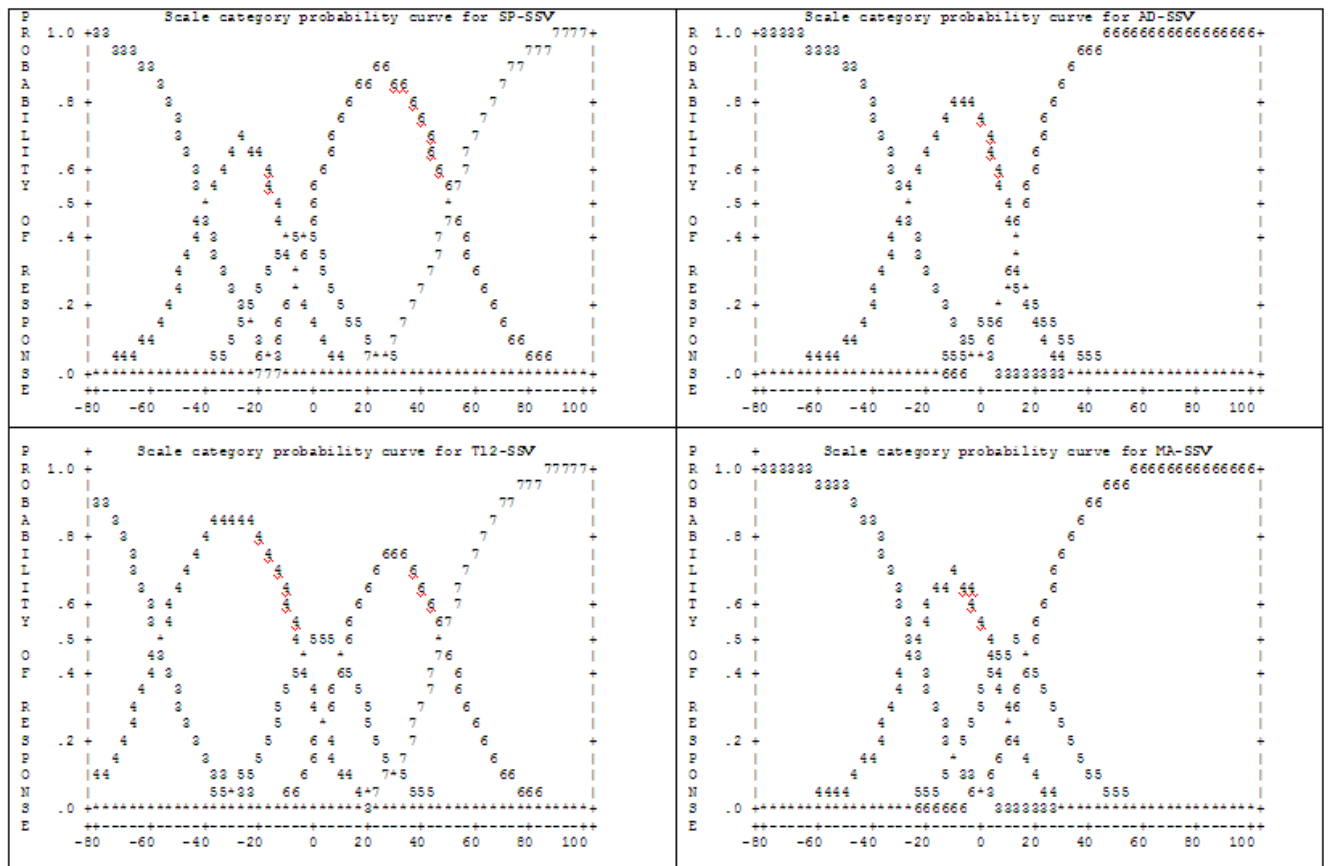
Finally, I calculated reliability and separation indices for students and scoring criteria. Reliability and separation indices for scoring criteria were 0.96 and 5.00, respectively, indicating high confidence in the discrimination power of the scale and five identifiable difficulty levels (or strata) among criteria. Student reliability and separation indices were 0.95 and 4.21, respectively, indicating that four ability levels were differentiated well, and that the scoring criteria employed have successfully measured these levels.

To summarize, this Rasch model yielded the following measure, fit, and reliability results: the SSV scoring criteria ranged widely in difficulty but were typically among the most difficult, and failed to fit the Rasch model; the other scoring criteria ranged little in difficulty and fit the model without exception; and the Rasch reliability analysis showed well-differentiated strata in both scoring criteria and student performance.

#### 4.4.2 Objective two: Rating scale evaluation

To investigate the efficacy of the 10-point scale and its thresholds, I used a probability curve conceptualization of the RPCM: band scores are presented as curves, which intersect at thresholds where there is an equal probability of being awarded either of two adjacent scores. Figure 2 displays probability curves for scoring criteria by prompt. (Because of space considerations, only some scoring criteria's curves are displayed.)

**Figure 2.** Scale category probability curves for AD-SSV, SP-SSV, MA-SSV, and T12-SSV.



In general, SSV criteria interacted problematically with the rating scale. By definition, band scores 1, 8, and 9 are expected to be less frequent than other band scores; however, the probability curves for AD-SSV and MA-SSV in Figure 2 indicate that only band scores between 3 and 6 were assigned for those criteria.

In AD-SSV, the scale thresholds ascend monotonically from -27.09, through 17.93, to 9.15 (i.e., the thresholds preserve their decreasing trend). In the AD-SSV plot, the thresholds for band scores 4, 5, and 6 are very close together (i.e., they were different by less than one logit), indicating the scale in AD-SSV is problematic.

RPCM analysis of AD-SSV, SP-SSV, MA-SSV, T11-SSV, and T12-SSV produced similar scale category probability curves and similar problems for these variables: although their thresholds increased monotonically, they did not fall within the desired range between 1.4 and 5.00, and band scores 4, 5, and 6 were too close together.

Table 6 summarizes this information. Many band scores in many SSV variables display problematic fit values; the most prominent are band scores 3 in SP-SSV; 3, 4, and 6 in MA-SSV; 2 and 5 in T11-SSV; and 3, 4, 6, and 7 in T12-SSV. The CQ, AIE, CC, TF, and TOTAL variables had overfitting and underfitting values, as well, but on the whole they functioned better than the SSV variables.

**Table 6.** Scale Category Probability Standardized Fit Indices for Variables

Score	AD-CQ	AD-AIE	AD-SSV	AD-TOTAL	SP-CQ	SP-AIE	SP-SSV	SP-TOTAL	GE-CQ	GE-AIE	GE-SSV	GE-TOTAL
2	0.20	0.35	NA	0.34	0.19	0.35	NA	NA	NA	0.33	NA	NA
3	0.71	0.27	0.98	0.34	0.97	0.63	1.94	0.35	0.24	0.13	1.71	0.41
4	0.62	0.57	1.70	0.73	1.10	0.73	1.39	0.80	0.81	0.69	1.29	0.60
5	0.60	0.60	0.86	0.61	1.42	0.77	1.32	0.70	0.63	0.76	1.29	0.32
6	0.55	0.54	1.83	0.51	0.95	0.64	1.57	0.52	0.58	0.61	1.47	0.53
7	0.29	0.29	NA	0.27	3.00	1.36	.66	1.16	NA	NA	1.18	0.25
8	NA	NA	NA	NA	NA	NA	NA	NA	0.18	0.18	NA	NA
Score	MA-CQ	MA-AIE	MA-SSV	MA-TOTAL	T11-CC	T11-TF	T11-SSV	T11-TOTAL	T12-CC	T12-TF	T12-SSV	T12-TOTAL
2	NA	NA	NA	NA	0.20	0.35	2.87	NA	NA	0.33	NA	NA
3	0.26	0.25	2.41	0.56	0.63	0.27	1.87	0.31	NA	0.13	2.91	0.41
4	0.96	0.99	2.33	0.89	0.68	0.57	1.78	0.66	1.26	0.69	2.46	0.67
5	0.88	0.54	1.79	0.67	0.81	0.60	2.01	0.56	1.20	0.76	1.38	0.81
6	1.07	0.69	2.30	0.78	0.67	0.54	0.86	0.51	1.34	0.61	2.09	0.90
7	1.12	0.69	NA	1.07	NA	0.29	NA	0.34	NA	NA	2.93	NA
8	NA	NA	NA	NA	.40	NA	NA	NA	NA	0.18	NA	NA

As shown in Table 6, standardized fit indices in band scores 7 in SP-QC, 3, 4, and 6 in MA-SSV, 2 and 5 in T11-SSV, and 3, 4, 6, and 7 in T12-SSV were greater than 2.00, indicating unsatisfactory functionality in these categories. To remedy the problem, many SSV response categories were collapsed. Collapsing these categories resolved many problematic SSV thresholds and fit indices; interestingly, moreover, it led to the development of more distinct peaks in scale category probability curves for all variables. Problematic variables in Figure 2 had unclear high points, but collapsing response categories alleviated this problem, and also improved other variables' curves.

In summary, two problems with the rating scale's response categories and thresholds emerged in several scoring criteria: fit indices of many response categories did not fall within the acceptable boundaries, and some scales, especially of SSV variables, had large gaps. Collapsing the scales of the SSV variables solved these problems to a considerable extent, even for other scoring criteria.

#### 4.4.3 Unidimensionality and local independence

As earlier noted, fit indices indicate the extent to which the data are unidimensional. I further investigated the adherence of data to unidimensionality by conducting a principal components analysis of Rasch (linearized) residuals (PCAR). In addition, I examined the local independence of items by analyzing the correlation of Rasch residuals.

Initially, I took the Rasch model to be the major dimension of variance in the data, and searched for residuals which were not accounted for by the model to identify a possible second dimension (which, depending on its explanatory power, might be either benign or adverse). In this analysis, the total number of units of randomness equals 24, the total number of scoring variables in the study.

The PCAR extracted a few components within the residuals. Factor 1 extracted 5.1 units out of the 24, and accounted for 10.1% of the variation in data; the variance explained by the Rasch model was 66.5%. Further, the signal-to-noise ratio (the ratio of the variance explained by the Rasch model to all the randomness in the data) is 10.4, meaning the Rasch model was 10.4 times stronger than the first component in the residuals. The empirical model based on the study data did not show a sharp contrast with the Rasch-modeled data: the measures in these two models explain 66.5% and 71.8% of total variance, respectively.

To determine whether an observed component in residuals is a substantive dimension or simply noise, I examined data points at both extremes of this component to look for patterns. I observed that the absolute values of the loadings of GE-SSV, SP-SSV, AD-SSV, T11-SSV, and MA-SSV on the first component were .71, .58, .58, .45, and .38, respectively. Three other variables—AIE-CQ, AD-AIE, and T11-TF—loaded on this dimension, but no pattern is observed in them.

Next, the data were examined for local independence by an analysis of standardized residual correlations; the lower this index, the stronger the evidence to support local independence. Therefore, if we find high correlations of residuals for a great portion of variables, we have evidence that these variables are locally dependent.

**Table 7.** Standardized Residual Correlations

Residual correlations	Correlated variables	Correlated variables
.67	SP-SSV	GE-SSV
.58	MA-CQ	MA-TOTAL
.56	AD-SSV	GE-SSV
.55	MA-AIE	MA-TOTAL
.55	AD-CQ	T11-CC
.52	AD-CQ	SP-AIE
.50	AD-CQ	AD-AIE

.50	AD-CQ	T11-TF
-.50	AD-CQ	GE-SSV

Table 7 presents the largest standardized residual correlations. All pairs of correlated variables had correlations below 0.7, meaning that they shared less than half of their variance ( $0.7 \times 0.7 = 0.49$ ). In general, this analysis provides support for the assumption that the data exhibit local independence.

## 5. Discussion

This study set out to investigate quality of the SSV scoring criterion alongside several others in assessing writing. Data for the study came from IELTS preparation course placement tests and follow-up writing tasks completed by 182 students.

Scoring consistency was evaluated through Cohen's Kappa index (Cohen, 1960), which shows the degree of agreement among raters. All but SSV scoring criteria had significantly high rater agreement, likely indicating that SSV scoring criterion is defective or there is not much homogeneity among raters. Some reasons might have led to this observation. It can be due to the confusion and tension that SSV imposes on raters; if it is true, then it provides supporting evidence for the first line of SSV research noted earlier (Mickan, Slater, & Gibson, 2000; Mickan & Slater, 2003; Mickan, 2003). This result can also lend indirect support to the research that indicates that when grammar and vocabulary are scored as two separate scoring criteria, the scale functions optimally (Engelhard, 1992, 1994, 1996; Kondo-Brown, 2002; Schaefer, 2008; Saito, 2008); this can be an area of investigation in future research. As Bachman (1990) pointed out, vocabulary and grammar are divisible components of the grammatical competence; therefore, whereas, in scoring SSV, one rater may pay more attention to vocabulary, another may regard grammar as more significant, and another rater may stand somewhere between other two raters. The three viewpoints can compromise the homogeneity of SSV scores.

Cronbach's Alpha indicated strong correlations between variables and therefore good internal consistency (Field, 2005). However, Alpha value increased most when the SSV variables were removed, indicating a high overall degree of correlation between variables, but poor correlation between the SSV variables and the others.

Collectively, inter-rater reliability analysis and Cronbach's Alpha indices provide support for the efficacy of all scoring criteria except SSV in the study, implying that the present writing assessment framework, in concordance with Weir's (2005) socio-cognitive writing model, can provide reliable representations of test takers' writing skills but the problem lies in the operationalization of the grammatical competence, a component of Bachman' (1990) theory.

### 5.1 Measures, Fit, and Reliability

The observation of the difficulty measure of scoring criteria partially conforms to McNamara's (1990) finding that grammar is the most difficult category in writing tests. Students' performance ranged from a low of -5.50 logits to a high of +6.52 logits on the item-person map. Performance was well-calibrated, as evidenced by the lack of large gaps in the distribution of scores (McNamara, 1996). Although most scoring criteria fitted the Rasch model well and had difficulty measures close to the mean, the SSV scoring criteria ranged widely in difficulty, and many were among the most difficult criteria in the dataset; they also did not satisfactorily fit the Rasch model. This provides further evidence that SSV is a limited scoring criterion, a finding which conforms to the inter-rater reliability results; that the SSV criteria had a wide range of difficulty points to their inefficacy because we expect that they have similar difficulty for students, but SSV in different tasks had different demands for students and, depending on which tasks students were assigned to, students' measure of this writing skill differed. For example, whereas SSV in an Advantage/Disadvantage (AD) task is an easy criterion, in a Giving Explanation (GE) task, it is the most difficult. Differences in the difficulty measures of SSV across tasks should urge the tester and teacher to scale and adjust the differences if students are assigned different tasks. Equating these criteria is much more important in the context of high-stakes tests of writing. This also points to the demands of different tasks for students (Weigle, 2002). Whereas some tasks can be more familiar to test takers, others can introduce a challenge beyond the requirements of writing assessment (Bachman, 1990).

That SSV did not fit the model satisfactorily indicates that much of the variance in data was not explained by the Rasch model, and the possible presence of confounding variables. Linacre (2009b) recognizes this as a condition where "noise" in scores is more than the "music" it plays: SSV scores do not provide a clear representation of the level of hypothesized subskills in test takers. The noise can be attributed to different known factors as literature reveals (e.g., Kondo-Brown, 2002; Schaefer, 2008). It may be that, in a socio-cognitive framework, grammar and vocabulary in writing performance are best measured if they are regarded as separate subskills of writing as implied by Kondo-Brown (2002), Schaefer (2008), and Saito (2008). Separating them, as I view it, is then resonant with Bachman's (1990) communicative competence model which partitions language knowledge into grammar, vocabulary, language functions, and pragmatic knowledge; or it may be that SSV has introduced too much irrelevant factors, such as tension and perplexity (Mickan, Slater, & Gibson, 2000; Mickan & Slater, 2003; Mickan, 2003), so that the valid interpretation of SSV scores is somewhat compromised.

With a few exceptions, students' performance fit the Rasch model; however, since the number of misfits exceeds the acceptable value of 2% suggested by Pollitt and Hutchinson (1987), these students should be further examined to see if a common cause of their unexpected fit indices might be found. However, this would require comparing students' test scores with their personal characteristics (Wright & Stone, 1988, 1999), and the researcher did not have access to this data; this was a limitation of the study. This finding partly satisfies "person performance validity," which holds when measured performance is mainly a reflection of ability level and not other contaminating variables (Wright & Stone, 1999).

The high person and item reliability and separation indices indicated real differences in student ability and test criterion difficulty (Grabowski, 2008): four levels of student ability and five levels of criterion difficulty emerged from the data. This almost met the hypothesized



expectation of six levels of performance (the hypothesized effective region for the rating scale was from 2 to 7).

## 5.2 Rating Scale Evaluation

As the RPCM probability curve analysis showed, band scores in all but SSV variables functioned well. Observed scores increased monotonically, but in SSV variables the increment of this increase was far outside the desired size, indicating large gaps in the scale. Threshold estimates should only change by between 1.4 and 5.00 logits (Bond & Fox, 2007): smaller values indicate that scoring categories are poorly distinguished, and larger values indicate gaps in the scale. This problem was to a great extent remedied by collapsing SSV categories; in addition to highlighting the fit problems of the SSV variable, this suggests that the rating scale itself caused measurement problems, and that a narrower scale would likely function better. This finding is in agreement with Bachman's (1990) framework for assessing performance; it is difficult for raters to distinguish among all ability levels of test takers if the rating scale has too many trait levels in performance tests. The implications of this finding for second language writing assessment and high stakes tests of writing such as IELTS, TEEP, and TOEFL writing subtests is that "to persist in using" a large rating scale in a socio-cognitive writing framework "would give a false impression of precision not only to the raters, but to test users as well" (Bachman, 1990, p. 36). As a general finding in this analysis, the smaller the scale, the better its thresholds' fit, and the more acceptable their fit indices. (As previously mentioned, some of the scales ranged between 2 and 7, which is a narrower range than the expected).

As noted, an interesting ancillary observation was that collapsing SSV response categories seemed to resolve rating scale problems even in other scoring criteria. This may indicate that poorly fitting response categories in one scoring criterion can adversely affect other criteria when they are calibrated jointly; while the scale may interact appropriately with some scoring criteria, it may not do so with others, and these in turn may affect even those parts of the scale that are well designed.

## 5.3 Unidimensionality and Local Independence

One important reason why a scoring criterion may function poorly is the presence of a dimension which was not intended in the measurement (Linacre, 2009b). According to Linacre (2009b), if any identified component in Rasch residuals extracts more than three units, it is likely to be a second dimension, although the researcher may decide if this is so. Indeed, the PCAR analysis revealed that the poorly functioning SSV scoring criteria loaded substantively on a second dimension in the residual. This second dimension explained approximately 10% of the observed variance in the data; given that the Rasch model dimension explained 66.5% of the variance, the latter remained the dominant dimension. Because most of these variables were misfits and loaded on a second dimension, they appear to represent a benign dimension in the measurement (Aryadoust, 2012b; Aryadoust, Goh, & Lee, 2011; Linacre, 2009b). AIE-CQ, AD-AIE, and T11-TF also loaded on this dimension; since no pattern is observed in them, this is likely to be random loading. Based on this analysis, the observed second dimension in this dataset appears to

be a “conspicuous branch, but not a separate tree-trunk” (Linacre, 2009b, p. 216). I wish to reiterate here that this erratic pattern may have been due to one or more of the following causes:

It is possible that SSV variables were not scored consistently. This was observed in the inter-rater reliability and fit analyses; the secondary dimension may be pointing to this imprecision in measurement.

As discussed above, SSV has two components with conflicting properties. As the principal rater of the texts in this study, I encountered several examples where use of vocabulary was satisfactory, but grammar was not at the same proficiency level. Attempting to apply the SSV criterion to a student whose grammar and vocabulary knowledge are obviously different introduces confounding factors into the scoring process, and scores may be confounded by raters’ resulting uncertainty (Mickan, Slater, & Gibson, 2000; Mickan & Slater, 2003; Mickan, 2003). This mechanism may have led to the observed second dimension in this study. Further research is required to test these hypotheses.

Finally, correlation analysis of variables’ residuals found few strong correlations, supporting the assumption of local independence, a necessary condition in the application of latent trait models.

In summary, the dataset appears to meet the criteria for local independence, but not necessarily for unidimensionality. There is some evidence that the SSV variables represent a benign second dimension.

## 6. Conclusion

This study investigated the functionality of a number of scoring criteria in second language writing assessment which were developed according to a socio-cognitive assessment framework. The study finds that adopting a complex criterion comprising two different elements of grammatical competence can introduce confounding variables to the assessment of performance. Lynch and McNamara (1998) reported that most of the scoring criteria in performance assessment are developed according to the attitudes of test developers rather than research. I argue that if such decisions are not theory-informed, such as the case of the SSV variable, then this can affect score interpretations and utility even if the other scoring criteria can be mapped to a firm theoretical model such as the socio-cognitive model of assessment (Weir, 2005). In particular, this study recommends evaluating the measurement features of grammar and vocabulary as two separate scoring criteria in the future.

Future research should further investigate the findings of this study, using statistical models which can capture the behavior of different raters on different scoring criteria and test prompts. One such model is the many-facets Rasch model, which can evaluate bias in rater-task, rater-criterion, and rater-scale interactions (Aryadoust, 2012a; Engelhard, 1996, 2002; Engelhard & Anderson, 1998).

## References

- Aryadoust, V. (2012a/forthcoming). Evaluating the psychometric quality of an ESL placement test of writing: A many-facets Rasch study. *Linguistics Journal*.
- Aryadoust, V. (2012b). Differential item functioning in while-listening performance tests: The case of IELTS listening test. *International Journal of Listening*, 26(1), 40-60.
- Aryadoust, V., Akbarzadeh S., & Nasiri, E. (2007a). *IELTS writing tutor: Writing task1, academic module*. Tehran: Jungle Publication.
- Aryadoust, V., Akbarzadeh S., & Nasiri, E. (2007b). *IELTS writing tutor: Writing task1, general module*. Tehran: Jungle Publication.
- Aryadoust, V., Goh, C., & Lee, O. K. (2011). An investigation of differential item functioning in the MELAB listening test. *Language Assessment Quarterly*, 8(4), 361-385.
- Bachman, L. F. (1990). *Fundamental consideration in language testing*. Oxford: Oxford University Press.
- Baghaei, P. (2008). The Rasch model as a construct validation tool. *Rasch Measurement Transactions*, 22, 1145-1146.
- Baghaei, P., & Amrahi, N. (2009). Introduction to Rasch measurement. *Iranian EFL Journal*, 5, 139-154.
- Banerjee, J., Franceschina, F., & Smith, A. M. (2007). *Documenting features of written language production typical at different IELTS band score levels*. (IELTS Research Report No. 7, the British Council/University of Cambridge Local Examinations Syndicate).
- Bereiter, C., & Scardamalia, M. (1987). *The psychology of written composition*. Hillsdale, NJ: Lawrence Erlbaum.
- Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences*. London: Lawrence Erlbaum Associates.
- Brown, A. (2006). *Candidate discourse in the revised IELTS speaking test*. (Research Report No. 6, IELTS Australia).
- Cambridge Practice Tests for IELTS 3*. (2002). Cambridge: Cambridge University Press.
- Cambridge Practice Tests for IELTS 4*. (2005). Cambridge: Cambridge University Press.
- Cambridge Practice Tests for IELTS 5*. (2006). Cambridge: Cambridge University Press.
- Cambridge Practice Tests for IELTS 6*. (2007). Cambridge: Cambridge University Press.
- Chapelle, C., Grabe, W., & Berns, M. (1997). *Communicative language proficiency: definition and implications for TOEFL-2000*. TOEFL-2000 Monograph Series. Princeton, NJ: Educational Testing Service.

- Cohen, J. (1960). A coefficient for agreement for nominal scales. *Education and Psychological Measurement*, 20(1), 37–46.
- Connor, U. (1991). Linguistic/rhetorical measures for evaluating ESL writing. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 215-226). Norwood, NJ: Ablex Publication Corporation.
- Eckes, H. (2005). Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis. *Language Assessment Quarterly*, 2(3), 197–221.
- Engelhard, G., Jr. (1992). The measurement of writing ability with a many-facet Rasch mode. *Applied measurement in Education*, 5(2), 171-191.
- Engelhard, G., Jr. (1994). Evaluating rater accuracy in performance assessments. *Journal of Educational Measurement*, 33(1), 56-70.
- Engelhard, G., Jr. (1996). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement*, 33(1), 93-112.
- Engelhard, G., Jr. (2002). Monitoring raters in performance assessments. In G. Tindal and T. Haladyna (Eds.), *Large-scale assessment programs for ALL students: Development, implementation, and analysis* (pp. 261-287). Mahwah, NJ: Erlbaum.
- Engelhard, G., Jr., & Anderson, D. W. (1998). A binomial trials model for examining the ratings of standard-setting judges. *Applied Measurement in Education*, 11(3), 209-230.
- Evola, J., Mamer, E., & Lentz, B. (1980). Discrete point versus global scoring for cohesive devices. In J. W. Oller & K. Perkins (Eds.), *Research in language testing* (pp. 177-181). Rowley, MA: Newbury House.
- Fahim, M., & Houman, B. (2011). The effects of rater training on raters' severity and bias in second language writing assessment. *Iranian Journal of Language Testing*, 1(1), 1-16.
- Field, A. (2005). *Discovering statistics using SPSS*. (2nd ed.). London: Sage.
- Gennaro, K. D. (2009). Investigating differences in the writing performance of international and Generation 1.5 students. *Language Testing*, 26(4), 533–559.
- Grabe, W., & Kaplan, R. B. (1996). *Theory and practice of writing: An applied linguistic perspective*. New York: Longman.
- Grabowski, K. C. (2008). Investigating the construct validity of a performance test designed to measure grammatical and pragmatic knowledge. *Spain Fellow Working Papers in Second or Foreign Language Assessment*, 6, 131-179.
- Hamp-Lyons, L. (1991). Reconstructing “academic writing proficiency”. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 127-154). Norwood, NJ: Ablex Publication Corporation.

- Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement*, 9(2), 139-64.
- Hayes J. R., & Flower L. S. (1980). Identifying the organization of writing processes. In L. Gregg & E. R. Steinberg (Eds.), *Cognitive processes in writing* (pp. 3-30). Hillsdale, NJ: Erlbaum.
- Hayes, J. R. (1996). A new framework for understanding cognition and affect in writing. In C. M. Levy & S. Ransdell (Eds.), *The science of writing: Theories, methods, individual differences, and applications* (pp. 1-26). Mahwah, NJ: Erlbaum.
- Hedge, T. (2005). *Writing*. Oxford: Oxford University Press.
- Homburg, T. J. (1984). Holistic evaluation of ESL compositions: Can it be validated objectively? *TESOL Quarterly*, 18(1), 87-109.
- Hyland, K. (2002). *Teaching and researching writing*. London: Longman.
- Jackeman, V., & McDowell, C. (1996). *Cambridge practice tests for IELTS 1*. Cambridge: Cambridge University Press.
- Jackeman, V., & McDowell, C. (2000). *Cambridge practice tests for IELTS 2*. Cambridge: Cambridge University Press.
- Kim, Y.-H. (2011). Diagnosing EAP writing ability using the Reduced Reparameterized Unified model. *Language Testing*, 28(4), 509-541.
- Kondo-Brown, K. (2002). An analysis of rater bias with FACETS in measuring Japanese L2 writing performance. *Language Testing*, 19, 1-29.
- Knoch, U. (2009a). *Developing and validating a rating scale for diagnostic writing assessment*. Frankfurt: Peter Lang.
- Knoch, U. (2009b). Diagnostic assessment of writing: A comparison of two rating scales. *Language Testing*, 26(2) 275-304.
- Knoch, U., & Elder, C. (2010). Validity and fairness implications of varying time conditions on a diagnostic test of academic writing proficiency. *System*, 38(1), 63-74.
- Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33, 159-174.
- Linacre, J. M. (2009a). WINSTEPS Rasch measurement, Version 3.64 [computer program]. Chicago: Winsteps.com.
- Linacre, J. M. (2009b). *A users' guide to WINSTEPS® MINISTEPS Rasch-model computer programs*. Chicago: Winsteps.com.
- Lougheed, L. (2004). *Barron's how to prepare for the Computer-Based TOEFL essay*. New York: Barron's Educational series, Inc.

- Lynch, B., & 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: Longman.
- McNamara, T. F. (1990). Item response theory and the validation of an ESP test for health professionals. *Language Testing*, 7, 52-75
- Mickan, P. (2003). 'What's your score?': An investigation into language descriptors for rating written performance. (Research Report No. 4, IELTS Australia).
- Mickan, P., & Slater, S. (2003). *Text analysis and the assessment of Academic Writing*. (Research Report No. 4, IELTS Australia).
- Mickan, P., Slater, S., & Gibson, C. (2000). *A study of response validity of the IELTS writing subtest*. (Research Report No. 3, IELTS Australia).
- Motallebzadeh, K., & Baghaee Moghaddam, P. (2011). Models of language proficiency: A reflection on the construct of language ability. *Iranian Journal of Language Testing*, 1(1), 42-48.
- Murphy, S., & Yancey, K. B. (2008). Construct and consequence: Validity in writing assessment. In C. Bazerman (Ed.) *Handbook of research on writing: History, society, school, individual, text* (pp. 365-385). New Jersey: Lawrence Erlbaum.
- O'Loughlin, K., & Wigglesworth, G. (2003). *Task design in IELTS academic writing task 1: The effect of quantity and manner of presentation of information on candidate writing*. (Research Report No. 4, IELTS Australia).
- Official IELTS Practice Materials*. (2007). Cambridge: Cambridge University Press.
- Pollitt, A., & Hutchinson, C. (1987). Calibrating graded assessments: Rasch partial credit analysis of performance in writing. *Language Testing*, 4(1), 72-92.
- Read, J., & Nation, P. (2006). An investigation of the lexical dimension of the IELTS speaking test. In P. McGovern & S. Walsh (Eds.), *IELTS research reports* (pp. 207-231). Volume 6. Canberra: IELTS Australia.
- Riazi, M., & Rezaii, M. (2011). Teacher- and peer-scaffolding behaviors: Effects on EFL students' writing improvement. In A. Feryok (Ed.), *CLESOL 2010: Proceedings of the 12th National Conference for Community Languages and ESOL* (pp. 55-63). Retrieved from <http://www.tesolanz.org.nz/>
- Saito, H. (2008). EFL classroom peer assessment: Training effects on rating and commenting. *Language Testing*, 25(4), 553-581.
- Schaefer, D. (2008). Rater bias patterns in an EFL writing assessment. *Language Testing*, 25(4), 465-493.

- Shaw, S., & Falvey, P. (2008). *The IELTS Writing Assessment Revision Project: Towards a revised rating scale*. Retrieved from [http://www.cambridgeesol.org/assets/pdf/research\\_reports\\_01.pdf](http://www.cambridgeesol.org/assets/pdf/research_reports_01.pdf)
- Shaw, S., & Weir, C. J. (2007). *Examining writing: Research and practice in assessing second language writing*. Cambridge: Cambridge University Press.
- Smith, Jr., E. V. (2002). Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. *Journal of Applied Measurement*, 3, 205- 231.
- Smith, R. M. (1991). The distributional properties of Rasch item fit statistics. *Educational and Psychological Measurement*, 51, 541-565.
- Smith, R. M. (1996) Polytomous mean-square fit statistics. *Rasch Measurement Transactions*, 10, 516-517.
- Smith, R. M. (2000). Fit analysis in latent trait models. *Journal of Applied Measurement*, 1, 199-218.
- Smith, R. M., & Miao, C. Y. (1994). Assessing unidimensionality for Rasch measurement. In M. Wilson (Ed.), *Objective measurement: Theory into practice* (Vol. 2) (pp. 316-327). Norwood, NJ: Ablex.
- Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing*, 15(2), 263–287.
- Weigle, S. C. (2002). *Assessing writing*. Cambridge, UK: Cambridge University Press.
- Wright, B. D. (1994). Local dependency, correlations, and principal components. *Rasch Measurement Transactions*, 10(3), 509-511.
- Wright, B. D., & Stone, M. H. (1988). *Validity in Rasch measurement*. University of Chicago: Research Memorandum No. 55.
- Wright, B. D., & Stone, M. H. (1999). *Measurement essentials*. (2nd ed.). Wilmington, Delaware: Wide Range, Inc.