TESL-EJ Teaching English as a Second or Foreign Language
The Electronic Journal for English as a Second Language

Investigating Writing Sub-skills in Testing English as a Foreign Language: A Structural Equation Modeling Study

Vahid Aryadoust
National Institute of Education, Singapore
arya2004v@yahoo.com

Abstract

This study investigates the validity of a writing model proposed as the underlying structure of the writing skill in English as a foreign language (EFL). Four writing prompts were administered to 178 Iranian EFL learners. The scripts were then scored according to writing benchmarks similar to the IELTS Writing criteria but narrower in scope. After inter- and intra-rater reliability analysis, a three-factor model was posited for validation. Structural modeling of the sub-skills revealed the two sub-skills of Idea Arrangement and Communicative Quality are psychometrically inseparable, but the Vocabulary and Grammar sub-skills proved to have good measurement properties. Using parcel indicators, a two-factor model was then evaluated which had the best fit and parsimony. The researcher concludes Idea Arrangement and Communicative Quality appear to have similar conceptual and theoretical foundations and should be considered the elements of one measuring criterion. Further research is required to support this finding. [1]

Introduction

Measurable sub-skills of second language (L2) essay writing in analytic approaches have been extensively researched to the present day. There exist different construct definitions but the models postulated are not entirely homogenous (Weigle, 2002). Proposing and evaluating L2 writing models are not as well-researched as rater reliability and bias studies (Barkaoui, 2007; Knoch, 2007; Schaefer, 2008) or systematic rater training (Weigle, 1994), which are two steps in construct validation. In this light, the present study seeks to investigate the underlying structure of the writing skill and its measurable sub-skills.

Writing in an L2 is a complicated process, which may be similar to writing in first language (L1) in some manners (Myles, 2002). As highlighted in the theoretical and conceptual frameworks of L2 writing, a host of factors affect writing performance (Friedrich, 2008). For example, Mickan, Slater, and Gibson (2000) contended that

syntax, lexicon, and task objectives affect L2 text writing. Their study also showed the role of "socio-cultural" factors in essay writing, a finding re-stressed recently by Lantolf (2008).

Research also shows whereas external variables can directly affect the writing style and performance (Ballard & Clancy 1991; Lantolf, 2008), the effective underlying factors considered in writing assessment have not exceeded a handful such as vocabulary, grammar, cohesion, and coherence (Leiki, 2008; Ferris, 2002). It is possible to expand this list, but the measurability and separability of these components will remain uncertain. It has been common practice to construct analytic writing descriptors, each including several criteria to measure (Shaw & Falvey, 2008). An example of lengthy lists to measure writing sub-skills is Weir's (1990) list which has seven subcategories and an instance of a shorter (perhaps more practical) list is Astika's (1993) three proposed rating benchmarks.

Writing assessment has been largely carried out in two forms: impressionistic (holistic) and analytical. "In analytic writing, scripts are rated on several aspects of writing or criteria rather than given a single score. Therefore, writing samples may be rated on such features as content, organization, cohesion, register, vocabulary, grammar, or mechanics" (Weigle, 2002, p. 114). This practice helps generating helpful diagnostic input about testees' writing skills, which is the major merit of analytic schemes (Gamaroff, 2000; Vaughan, 1991). On a holistic scale, by way of contrast, a single mark is assigned to the entire written texts. The underlying assumption is that in holistic marking raters will respond to a text in the same way if a set of marking benchmarks are to guide them in marking (Weigle, 2002, p. 72).

In relation to the analytic assessment of the writing skill, Aryadoust, Akbarzadeh, and Nasiri (2007) discussed three criteria based on which to score the text, that is, Arrangement of Ideas and Examples (AIE), Coherence and Cohesion (CC) or Communicative Quality (CQ), and Sentence Structure and Vocabulary (SSV). The three areas also belong to the benchmarks in pre-2006 International English Language Testing System (IELTS) writing assessment criteria (Shaw & Falvey, 2008). These criteria were modified in 2008 and the current rating practice in the IELTS Writing test is based on a new exposition of writing performance and assessment (Shaw & Falvey, 2008); for example, it was agreed to separate the SSV criterion into vocabulary and grammar. Also, the CC was found to be the most difficult area for raters to score. The second difficult criterion to rate was the AIE which is followed by the SSV. Shaw and Falvey (2008) capitalized on the similarity of CC and AIE, which could cast doubts on the inseparability of these sub-skills in writing. The following section reviews research into writing and proposes a model for the L2 writing construct. The model will be validated via structural equation modeling.

Nature of Second Language Writing

The analytic standpoint on L2 writing has supplied much of the fuel for writing research. According to Hedge (2005), one can construct a list of "crafting skills",

which comprise such components as lexis, syntax, spelling, and communicating ideas in assessing writing and yet expand on the list in analytic writing. Writing researchers have articulated other crafting skills influencing writing performance, that is, overall effectiveness, intelligibility, fluency, comprehension, appropriateness, and resources which influenced writing performance the most (McNamara, 1990, 1996); control over structure, organization of materials, vocabulary use, and writing quantity (Mullen, 1977); relevance and adequacy of content, compositional organization, cohesion, adequacy of vocabulary, grammar, punctuation, and spelling (Weir, 1990); content, language use, organizing ideas, lexis, and mechanics (punctuations and spelling) (Jacobs, Zinkgarf, Wormuth, Hartfiel, & Hughey, 1981); and sentence structure, vocabulary, and grammar (Daiker, Kerek, & Morenberg, 1978).

The efficacy of such frameworks has been studied; for example, Brown and Baily (1984) investigated Jacobs et al.'s (1981) and Mullen's (1977) frameworks. They found using an analytic framework of organization, logical development of ideas, grammar, mechanics of writing, and style is a sound practice in assessing writing performance. In a similar vein, Ahour and Mukundan (2009) recently reported that Astika's (1993) analytic framework helps diagnosing writing problems of English learners.

Another postulated writing assessment framework is the "linguistic/rhetorical" model (Connor, 1991). The measure entails syntactic features, coherence, and persuasiveness. Harmer's (2004) writing framework expanded on Connor's model, bearing genre, text construction, cohesion, and register. Likewise, Moore and Morton (1999, 2005) stressed rhetorical functions alongside genre and the source of information in writing assessment.

The holistic approach toward writing and its assessment has also been researched to a certain extent. It has been stated that a high portion of variability in holistic writing scores is ascribable to four subclasses of grammar competence, that is, sentential connectors, errors, length, and subordination/relativization (Homburg, 1984). Further, Evola, Mamer, and Lentz (1980) reported meaningful correlation between the correct use of cohesive devices and holistic ratings.

Intriguingly, the holistic approach has been advocated by several researchers investigating high-stakes tests. Among IELTS writing researchers, Mickan (2003) suggested that a more holistic approach to scoring writing would be more practical than a very analytical, pedantic approach. Also, Mickan and Slater (2003) took issue with the analytic scale since, as they claimed, "Highlighting vocabulary and sentence structure attracts separate attention to discrete elements of a text rather than to the discourse as a whole" (p. 86). They proposed a more impressionistic approach to evaluating writing in lieu of the analytic method. But their assumption was undermined in later research on writing. Contrary to Mickan and Slater's (2003) study, recent investigations into the writing indicated that vocabulary and grammar accuracy appear to be complementary and are possible to be classified under a single rubric (Banerjee, Franceschina, & Smith, 2007). Such a proposal is supportive of the

assumption that similarities between writing sub-skills make it possible to have composite sub-skills where two or more categories are accommodated into a single rubric.

On the other hand, Banerjee et al. (2007) deemed it practical to reduce the rating criteria by accommodating several rating criteria into more unifying headings. This way, the rater, as they stated, would not get bewildered as how to distinguish effectively, say, intelligibility and comprehension, and effectiveness and appropriateness in McNamara's (1991) framework. In this light, the present study seeks to explore the convergence and separability of sub-skills of a writing construct model including grammar and lexis, cohesion and coherence, and arrangement of ideas. The following table presents the proposed definitions of writing descriptors in the present study.

Table 1. Criterion and Descriptors to Assess and Score L2 Writing Samples

Criterion (sub-skill)	**Description and elements**
Arrangement of Ideas and Examples (AIE)	1) presentation of ideas, opinions, and information 2) aspects of accurate and effective paragraphing 3) elaborateness of details 4) use of different and complex ideas and efficient arrangement 5) keeping the focus on the main theme of the prompt 6) understanding the tone and genre of the prompt 7) demonstration of cultural competence
Communicative Quality (CQ) or Coherence and Cohesion (CC)	1) range, accuracy, and appropriacy of coherence-makers (transitional words and/or phrases) 2) using logical pronouns and conjunctions to connect ideas and/or sentences 3) logical sequencing of ideas by use of transitional words 4) the strength of conceptual and referential linkage of sentences/ideas
Sentence Structure Vocabulary (SSV)	1) using appropriate, topic-related and correct vocabulary (adjectives, nouns, verbs, prepositions, articles, etc.), idioms, expressions, and collocations 2) correct spelling, punctuation, and capitalization (the density and communicative effect of errors in spelling and the density and communicative effect of errors in word formation (Shaw & Taylor, 2008, p. 44)) 3) appropriate and correct syntax (accurate use of verb tenses and independent and subordinate clauses) 4) avoiding use of sentence fragments and fused sentences 5) appropriate and accurate use of synonyms and antonyms

In summary of the table, the AIE is defined as an aspect of writing which concerns the appropriate tone of the text and genre, appropriate exemplification, efficient arrangement of ideas, completeness of responses to the prompt, and relevancy. Therefore, it was made explicit to students in the study that the reader of the text would be a university professor or an educated individual. In relation to the SSV, the use of appropriate vocabulary, correct spelling, punctuation, and syntax was considered. The CC (or CQ) encompasses elements of argument where components of causality and coherent presentation of ideas are essential. Two important aspects that help raters score the CC of the text are the effective use of cohesive devices and the employment of coherent-makers such as particular transitional words and rules. Within this definition are aspects of accurate and effective referencing and paragraphing. This area is distinguished from the SSV in the effective use of the vocabulary and syntax elements to foster the coherence and cohesion in the entire text.

Research Questions

1. What measurable sub-skills underpin the writing skill?
2. Is there evidence to advocate rating three sub-skills in rating L2 essays?

Method

Participants

Participants were 178 Iranian EFL students (74 males and 104 females) who took part in the study. They ranged in age from 19 to 34 (M = 25; SD = 3.34), and Persian was their mother tongue. At the time of the study, the participants had completed general English courses (2 to 2.5 years of learning English) and were either applying for IELTS preparation courses or were recently enrolled in the course. The general English courses offered at the institute where the study was carried out were based on a curriculum which highlighted the communicative needs of the students in four language skills: listening, reading, writing, and speaking. Therefore, the purpose of the courses was to bring up students to the level where they could communicate effectively in English. The main materials used in these courses were Interchange series by Richards, Hull and Proctor (2004), which include three textbooks and additional materials such as videos and audio programs. The textbooks were replaced by IELTS materials when students completed them, so that students were involved in more communicative practices and activities. Writing was an indispensable section of both stages (Interchange textbooks and IELTS), which was instructed by the teacher.

Materials

After Lougheed (2004), Aryadoust et al. (2007) classified essay prompts into four main categories:

(a) Agreement-disagreement (AD)

(b) Stating a Preference (SP)
(c) Giving Explanation (GE
(d) Making Arguments (MA)

This classification is not made according to the responses to the prompt or manuscripts; rather it is centered on the wording and requirements of the prompts. Table 2 presents the sample wordings representing these prompt types. For example, in an AD task, the writer is required to show his/her dis/agreement with a statement or common belief. It is also important to underscore there is a fuzzy border between some prompt classes which makes it difficult for researchers decide on the task type (Aryadoust et al., 2007).

Table 2. Definitions of Four Tasks Based on Their Prompts

Prompt	Sample Wording
Agreement-disagreement	To what extent do you agree or disagree?
Stating preferences	Which one do you prefer?
Explanation	Explain what you would do? Explain you reasons.
Argumentation	To what extent would you say this can be true?

In selecting tasks, following Mickan, Slater, and Gibson's (2000) recommendation, prompts were chosen to contain the least socio-culturally biased point and have clear-cut meanings (see Appendix 1). In so doing, I presented 12 prompts to four experts who agreed on the clarity and objectivity of four prompts. The selected tasks were administered to the testees in the same order as in Table 2. Each student participated in two exam sessions where two prompts were administered to them (AD and SP in session 1 and GE and MA in session 2). There was a 10-minute interval between each two tasks in each session. Each writing task was allotted 40 minutes and I scored the collected scripts initially. Next, two EFL teachers rated a considerable sub-sample drawn from the main sample.

To help participants have a clear idea of the possible readership of their text, I used the instructions similar to the ones formerly used in the IELTS Writing test. The instructions read: "*write an essay in response to the following question/statement for a university professor or educated person. Use specific reasons and examples to support your answer* [italics added]." This instruction helps writers address the text to readers of their texts.

Scoring

Two major rounds of scoring were conducted. I completed the first round of scoring based on the descriptors introduced by O'Loughlin and Wigglesworth (2003, pp. 100-113) and Hamp-Lyons (1991a, 1991b, 1991c) as summarized in Table 1. Other sets of useful materials were also used to further study the structure of scoring system and

benchmarks in IELTS since a 10-point scale (0-9) like the IELTS Writing rating benchmarks was used, e.g. Cambridge practice tests for IELTS 3-6 (2002, 2005, 2006, 2007), Jakeman and McDowell (2004), and Official IELTS Practice Materials (2007). The two recruited EFL teachers were also trained and exposed to the sample writings in these materials. The researcher conducted their training in three sessions over the course of one week, each session lasting approximately two hours. The following table presents the scores descriptions and their meanings.

Table 3. Band Score Definitions of IELTS Used in the Present Study

Band score	Title	Definition
1	**Non user**	Essentially has no ability to use the language beyond possibly a few isolated words.
2	**Intermittent user**	No real communication is possible except for the most basic information using isolated words or short formulae in familiar situations and to meet immediate needs. Has great difficulty understanding spoken and written English.
3	**Extremely limited user**	Conveys and understands only general meaning in very familiar situations. Frequent breakdowns in communication occur.
4	**Limited user**	Basic competence is limited to familiar situations. Has frequent problems in understanding and expression. Is not able to use complex language.
5	**Modest Users**	Can communicate and understand the general meaning in most situations but are likely to make a lot of mistakes.
6	**Competent Users**	Can generally communicate effectively but will still make some mistakes and have some misunderstandings. They can use and understand some complex language.
7	**Good Users**	Can communicate effectively, using and understanding complex language. They will still make occasional mistakes, however, and have misunderstandings in some situations.
8	**Very good user**	Has fully operational command of the language with only occasional unsystematic inaccuracies and inappropriacies. Misunderstandings may occur in unfamiliar situations. Handles complex detailed argumentation well.
9	**Expert user**	Has fully operational command of the language: appropriate, accurate and fluent with complete understanding.

Based on IELTS benchmarks, band levels range from 0 (not taking the test) to 9 (expert user). Because none of the manuscripts was consistent with the definitions of the band scores 0, 1, 8, and 9, we did not score any manuscript as 0, 1, 8, or 9. Each text was marked in three areas as displayed in Table 1. On the whole, 178 participants wrote on four prompts, which totals 712 essays ($178 \times 4 = 712$).

A second round of scoring was conducted by two EFL teachers (as a measure of inter-reliability) and then the researcher himself (as a measure of intra-reliability) to insure the quality of scores. Due to time constraints and other commitments of the two assistant raters, the researcher had to randomly draw 240 writing samples out of the manuscripts marked (60 writing tasks in response to each prompt). Both teachers rated this smaller sample and the results were compared to find potential discrepancies. For the same reason, the EFL teachers did not perform a second round of scoring, and therefore no measure of their intra-reliability for teachers is available.

Results

Inter-rater and Intra-rater Reliability

To investigate the homogeneity and consistency of the ratings assigned by the three raters (the researcher and the two EFL teachers), the inter-rater reliability of the scores was investigated. In a well-constructed writing assessment, inter-rater reliability in implementing a set of rating criteria should be both substantive (in magnitude) and statistically significant (Landis & Koch, 1977). In this light, I employed the Cohen's Kappa, ranging from -1.0 to +1.0, which provides substance and significance of the inter-reliability. Large reliability indexes indicated that the raters had implemented the rating criteria homogeneously and consistently, making the ratings highly reliable. Indexes close to zero and below suggested that observed performances of the raters could be attributable to chance or intervening variables which significantly influenced the ratings, such as inconsistent rater severity or leniency. According to Landis and Koch (1977), Cohen's Kappa values from 0.40 to 0.59 are moderate, 0.60 to 0.79 are substantial, and 0.80 and above are outstanding. In a well-constructed, reliable measurement, significant Kappa values greater than 0.60 ($p < 0.05$ or 0.01) are desirable.

SPSS for Windows (version 16, SPSS Inc., Chicago, IL) software package was used to calculate the Kappa coefficients ($p < 0.01$). Composite scores were constructed to report the performance of each participant on each sub-skill. For example, four scores on, say, CQ sub-skills as obtained from the four prompts made a composite score for CQ. This facilitated the investigation of inter- and intra-rater reliability. Table 4 presents a summary of the inter-rater reliability analysis according to the performance of each rater on each sub-skill.

Table 4. Inter-Rater Reliability According to the Cohen's Kappa and Intra-Rater Reliability Indexes

	Variable	Kappa Values								
		First rater			Second rater			Third rater		
		Cq	aie	ssv	cq	aie	ssv	cq	aie	ssv
First rater	cq	0.89			0.67			0.80		
	aie		0.92			0.81			0.76	
	ssv			0.95			0.72			0.82
Second rater	cq	0.75						0.88		
	aie		0.81						0.77	
	ssv			0.71						0.74

Note. All indexes are significant at 1% ($p < 0.01$).
Cq = communicative quality. Aie = arguments, ideas, and evidence. Ssv = sentence structure and vocabulary.
Italicized figures report the Kappa coefficients. Bold figures present the interclass correlation coefficients (ICC) for rater 1 (researcher).

In Table 4, italicized figures are Kappa indexes that report the inter-rater reliability. As we observe, these indexes range from 0.67 (substantial) to 0.88 (outstanding) ($p < 0.01$). I also used interclass correlation coefficients (ICC) to evaluate intra-rater reliability coefficients. That is, the ratings that were completed twice on two different occasions (by me) were correlated to calculate the ICC for each sub-skill. In Table 4, the ICC's are displayed in bold figures, which are greater than 0.85 ($p < 0.01$). For example, the ICC for CQ was 0.89 ($p < 0.01$). In this study, Kappa and ICC indexes lent strong support to the inter- and intra-reliability of the ratings assigned by the three raters.

Structural Equation Modeling

In this study, a Structural Equation Modeling (SEM), using LISREL computer program, Version 8.8 (Jöreskog & Sörbom, 2006) was performed. The SEM programs provide a model summary and fit statistics. Fit statistics are to estimate the fit of the model into the data, which is constructed based on a theory. For example, in the present study, the models presented in Figure 1 are based on the literature review reported above. According to McDonald and Ho (2002), the most common fit statistics reported in SEM studies are:

(a) Degrees of freedom (*df*) reported together with the chi-squared (χ^2) statistic, and the ratio of χ^2/df. For large sample sizes, the χ^2value tends to be significant; therefore, other fit indexes have been developed to investigate the fit of the postulated model.

(b) Tucker-Lewis Index (TLI), also known as the Non-normed Fit Index (NNFI), which depends on the correlation among variables in the model. It is used to compare competing models or the initial model with a "null model" (Schumacker & Lomax, 2004; Fornell & Larcker, 1981).

(c) Comparative fit index (CFI), which is an index similar to TLI. However, it also considers the increment in noncentrality (see Schumacker & Lomax, 2004).

(d) Root mean square error of approximation (RMSEA), and standardized root mean residual (RMSR), which is used to compare two postulated models for a set of data. These fit statistics show the "badness of fit" (Schumacker & Lomax, 2004). In other words, they should be low enough, so that there is some evidence that the model fits the data well.

The first model (M1) on the left side of Figure 1 comprised three correlated latent traits (factors) as three big ellipses, for example, Argument, Ideas, and Evidence (AIE), Communicative Quality (CQ), and Vocabulary and Sentence Structures (SSV). Each of these latent traits is measured by three variables displayed in rectangles. One-headed arrows run from each ellipsis to rectangles, meaning the observed variance in each sub-skill (rectangle) is mainly attributable to (or caused by) the hypothesized latent trait. Latent traits are hypothetically correlated. Therefore, two-headed arrows have connected them. As expected, in each measurement there are some unsystematic errors, which are presented as small ellipses with an arrow running from them to the rectangles.

According to Table 5, the first proposed model (M1) did not capture a good fit since the χ^2 was significant, the TLI and CFI values were below the tenable constraints, and the RMSEA and SRMR indexes showed the model had high badness-of-fit statistics ($\chi^2 = 296.755$ ($p < 0.05$); $df = 51$; $\chi^2/df = 5.82$; TLI = 0.87; CFI = 0.90; RMSEA = 0.144; SRMR = 0.059).

LISREL 8.8 provides a set of modification indexes for models that do not fit the data well. Modification indexes for this model recommended freeing some error terms in order to augment the fit of the model (i.e., covary errors of measurement from different indicators). Applying modifications to the model needs to be theory-driven (Geldhof, Selig, & McConnell, 2008) and should not override the theory.

Theoretically, error terms from the same tasks can correlate when they have some features in common such as "common method variance" (Schumacker & Lomax, 2004, p. 170). Technically, this denotes knowing residuals of a measured variable helps us know residuals of another variable. For instance, the Halo effect is suspect to have affected individuals answering items on a questionnaire that surveys their social status, that is, they may be inclined to overestimate themselves. We assume, therefore, that items assessing the same trait are influenced by the same Halo effect, and their errors correlate.

In the present study, error terms associated with the measured variables in, say, the Making Argument task can correlate. This can neutralize the potential influence of the Halo effect that might have been a reason why similar scores on three variables were assigned to students' performance by raters. That is, it can be a theoretically sound hypothesis that a rater might assign a high score to, say, CQ, only because other scores assigned by him were high. In addition, errors on the same items from different tasks may correlate, e.g. SSV in two task types. Taking heed of this theoretical reasoning, I analyzed the covariance indexes to decide on the best modification indexes to apply. For example, the large modification index generated from freeing the corresponding error covariance parameter between the error terms of GECQ and GEAIE could improve the fit very well and decrease the chi-squared statistics. This modification is also theoretically sound since both of these error terms belong to the Giving Explanation task and performance in one area, say AIE, can correlate with performance in CQ. Model 2 (M2) is the modified form of M1.

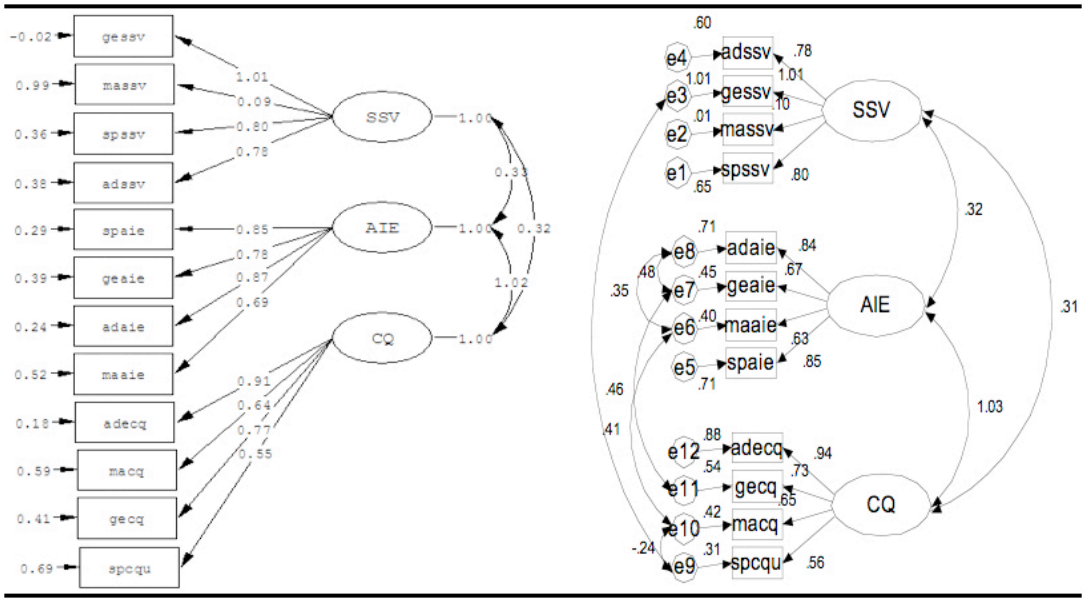

Figure 1. Model 1 (M1) and the Modified Model (M2) with Standardized Parameters.

Model 2 (M2) displayed a better fit to the data ($\chi^2 = 136.77$ ($p < 0.05$); $df = 47$; $\chi^2/df = 2.91$; TLI = 0.907; CFI = 0.937; RMSEA = 0.076; SRMR = 0.071). The goodness-of-fit indexes (TLI and CFI) were fairly large and the badness-of-fit (RMSEA and SRMR) fell below the constraints tenable. These constraints were proposed by Hair, et al. (2006) who recommended cut-offs according to the sample size. Nevertheless, the χ^2 index was significant, which can be attributed to the relatively large sample size.

Table 5. Fit Indices of the Models Postulated in the Study

Model	χ2	df	χ2/df	TLI	CFI	RMSEA	SRMR
M1	296.755*	51	5.82	.87	.90	.144	.059
M2	136.77*	47	2.91	.908	.937	.076	.071
Constraint tenable	Non-sign.	—	≤ 3	≤ .90	≤ .95	≤ .08	≤ .08

Note. *Significant at $p < 0.05$.
RMSEA = Root Mean Square Error of Approximation. GFI = Goodness of Fit Index. TLI = Tucker Lewis Index. SRMR = Standardized Root Mean Residual. CFI = Comparative Fit Index. *Df* = degrees of freedom
M1 = three-factor model or model 1. M2 = M1 modified.

Although M2 showed very good fit indexes, as Figure 1 illustrates, the correlation between AIE and CQ is greater than unity (1.03). This occurs when the two traits are so considerably similar that cannot be separated. Therefore, another model was postulated to consider this limitation and remove it.

A Limited Study to Evaluate other Models of Writing

Vocabulary and grammar proved to be the elements of one measuring criterion, yet the statistical separability of AIE and CQ was not established. Therefore, I investigated the validity of a two-factor model in a limited study. Accordingly, parcel scores were constructed from AIE and CQ by aggregating scores from AIE and CQ (researcher correction) and dividing the sum by two to get the arithmetic average ([AIE + CQ]/2 = new variable). This measure was taken to help explore the features of a model comprising two factors (SSV and AIE + CQ) and compare it with the previous models. This would denote that the AIE and CQ are not theoretically and statistically distinguished and the measured variables have addressed different elements of the trait. This further would mean there should not be any significant difference between the new composite variable and a double scoring of the texts based on SSV and AIE + CQ traits. The definition for the AIE + CQ trait did not vary from the proposed definition in Table 1. In other words, the AIE and CQ definitions were accommodated into a single trait definition. Next, 60 texts were randomly selected to score. Due to time and budget limitations, I managed to recruit only one of the assistants (teacher 1) to help rescoring the texts. At the end, there were eight measured variables in the new model.

To investigate the differences between the parcel AIE + CQ and the rescored AIE + CQ within and between raters, a series of *t*-tests (within and between designs) were performed. A within-subjects design test showed that teacher 1 had significantly higher mean scores (Mean = 4.98) than did the researcher (R) (M = 4.87) in ADAIE + CQ (t (103) = 3.22, $p < 0.05$) only. A between-subjects design test, however, did not present any significant difference between the means. Then, the following model was generated based on the rescored manuscripts, which is displayed in Figure 2.

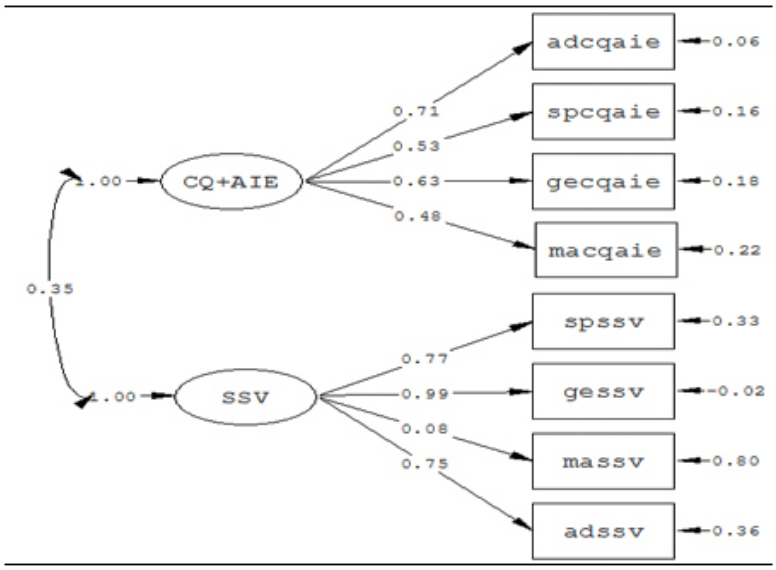

Figure 2. A Two-Factor Model of Writing with Standardized Parameters

The hypothesized model in Figure 2 shows significant goodness-of-fit indexes and also captures good parsimony ($\chi^2 = 39.10$; $df = 19$; $\chi^2/df = 2.05$; TLI = 0.975; CFI = 0.983; RMSEA = 0.07; SRMR = 0.065). There is also a moderate correlation between the latent traits (0.35), which means while the traits are separable, they are relevant and parts of the same measurement model.

Discussion

This study set out to investigate the validity of a writing model. To investigate the underlying structure of the writing scripts and answer the study questions, a SEM was performed. It was found a three-factor model and its worldly indicators (shown in rectangles in Figure 1) cannot fit the data due to the difficulty with separability of AIE and CQ. This is in part due to the low discriminant validity of the model. Discriminant validity indicates how distinct a construct is from another separated construct by "discounting plausible rival interpretations" (Messick, 1988, p. 13). A discriminant validity criterion in SEM models is that the correlation coefficients should not be too high to be considered inseparable (Hair et al., 2006). Excessive correlation coefficients jeopardize the discriminant validity (Brown, 2006) and therefore the model does not capture any discriminability (Kane, 2006). Because the correlation coefficient between two latent traits was greater than unity in M1, the nomological validity, which is "the *degree that the summated scale makes accurate predictions of other concepts in a theoretically based model*" (Hair et al., 2006, p. 136; emphasis in original), is at stake. M1 and M2 failed to show good features of discriminability in terms of their traits.

This observation concurred with the Shaw and Taylor's (2008) assumption that Argument, Ideas, and Evidence (AIE) and Communicative Quality (CQ) are very similar and may prove to be non-separable. It may be due to the structure of the AIE

which can assume a subcategory of coherence and cohesion under its heading. For example, to arrange ideas, information, and examples, it is necessary to use cohesive devices to make the movement within and through sentences of a text smooth. Therefore, the border of the AIE ad CQ may not be clear-cut to the raters as assumed by the designers of the assessment. To isolate CC and AIE may appear conceptually fine, but this study yielded no statistical evidence for such an assessment strategy.

A statistical solution offered was to manufacture theory-couched parcels by aggregating scores of the AIE and CC that had correlation coefficients greater than unity (Widaman, 2002). Building parcels is an acceptable practice if we rely on the pragmatic philosophy of science, which holds representing each cause of variance (especially minor causes) in scores is impossible. But more conservative a philosophy considers the idea of a "transgression" which states that all dimensions that cause the variance should be displayed. This is difficult in social sciences and troublesome in language assessment where the range of skills in performance is very extensive but their separability and measurability may be neither desirable nor possible.

Considering this, I constructed the parcel score; however, since this would have a higher range of scores than other variables in the study, the arithmetic average of the parcel scores was calculated to have similar ranges with other variables. Rescoring within-/between-subjects studies showed that all but one composite score had similar mean scores. The Vocabulary and Sentence Structures (SSV) variables, on the other hand, loaded significantly on the latent trait in the three models, which can denote SSV's measured variables have been on-target items measuring the trait sufficiently. Research shows that using grammar or vocabulary as a criterion in writing can produce constant results (Banerjee et al., 2007), although the range of vocabulary is not always consistent with the band scores and their definitions. Therefore, in analytical rating in L2 writing, it seems theoretically and statistically plausible to rate two major areas of the scripts: sentence structure and vocabulary and arrangement of ideas and examples including the cohesion and coherence of the text. As the present study showed, this strategy can explain a significant amount of variance in scores and ease the process of scoring and decision making.

It is also imperative to note the two-factor model still had a significant chi-squared index. There are different viewpoints how to interpret this index. Kline said of such observations:

> There are two problems with the chi square statistic as a fit index. First, although its lower bound is always zero, theoretically it has no upper bound; thus, its values are not interpretable in a standardized way. Second, it is very sensitive to sample size. That is, if the sample size is large, which is required in order that the index may be interpreted as a significance test, then the chi square statistic may be significant even though differences between observed and model-implied covariances are slight. (1998, p. 128)

Schumacker and Lomax (2004, p. 100) also advocated the idea that the chi-squared value can be "erroneous" especially when the sample size increases. Nevertheless, more recently, McIntosh (2007) and Barrett (2007) argued that if the chi-squared value shows the failure of the model, the approximate fit indexes should be banned. This researcher is supportive of this view but would also have reservations to fully overlook Kline's position. Therefore, for a more in-depth analysis of the findings from this study, the use of a larger sample size and integrated writing criteria which divide the underlying construct into two major parts is deemed useful. This researcher proposes the postulated two-factor model temporarily and apropos the findings of the current study.

Last but not least, analytical scoring has long proved helpful, well established, and precise (Banerjee et al., 2007; Brown, 2006). To illuminate this area further, it is recommend that grammar/lexicon and the merged criterion of AIE + CQ, which I refer to as Idea Arrangement and Task Fulfillment (IA-TF), should be further researched in future studies. The issue of statistical and psychometric separability of all proposed criteria is of a paramount importance in investigations into the construct validity of the proposed models.

Conclusion and Implications

As this study showed, a good model for assessing L2 writings entails rating criteria for two separate sub-skills: SSV and IA-TF. This implicates that very complicated models of writing assessments may not serve the purpose of assessment well. Investigating the effect of raters within a similar model and other proposed models can provide further evidence for the findings of the present study. It is also helpful for the L2 writing teachers to focus on such skills as SSV and IA-TF in writing courses by providing different types of exercise for their students because, as it appears, a good underlying model for L2 writing must bear at least these two components (see Coxhead & Byrd, 2008).

Note

[1] A previous version of this article was presented in the ICELT 2009 Conference in Malaysia, Melacca.

About the Author

S. Vahid Aryadoust is a PhD candidate in applied linguistics from National Institute of Education, Nanyang Technological University, Singapore. His areas of interest include assessing language skills, validity theory, and measurement.

References

Ahour, T., & Mukundan, J. (2009). Analytic assessment of writing: Diagnosing areas of strength and weakness in the writing of TESL undergraduate students. *Iranian Journal of Language Studies, 3*(2), 195-208.

Aryadoust, S. V. (2009). *Validity arguments in the context of high-stakes tests of second language listening: A quantitative and qualitative study*. Unpublished confirmation report. Nanyang Technological University, National Institute of Education, Singapore.

Aryadoust, S. V., Akbarzadeh S., & Nasiri, E. (2007). *IELTS writing tutor: Writing task1, academic module*. Tehran: Jungle Publication.

Archibald, A. (2002). Managing L2 writing proficiencies: Areas of change in students' writing over time. *International Journal of English Studies, 1*(2), 153-174.

Astika, G. G. (1993). Analytical assessment of foreign students' writing. *RELC Journal, 24*(1), 371-389.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Ballard, B., & Clancy, J. (1991). Assessment by misconception: Cultural influences and intellectual traditions. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 19-36). Norwood, NJ: Ablex Publication Corporation.

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).

Barkaoui, K. (2007). Rating scale impact on EFL essay marking: A mixed-method study. *Assessing Writing, 12,* 86-107.

Barrett, P. (2007). Structural equation modeling: Adjudging model fit. *Personality and Individual Differences, 42*(5), 815–824.

Brown, J. D., & Baily, K. M. (1984). A categorical instrument for scoring second language writing skills. *Language Learning, 34*(4), 21-42.

*University of Cambridge Local Examinations Syndicate. (*2002*). Cambridge practice tests for IELTS 3*. Cambridge: Cambridge University Press.

*University of Cambridge Local Examinations Syndicate. (*2005*). Cambridge practice tests for IELTS 4*. Cambridge: Cambridge University Press.

*University of Cambridge Local Examinations Syndicate. (*2006*). Cambridge practice tests for IELTS 5*. Cambridge: Cambridge University Press.

*University of Cambridge Local Examinations Syndicate. (*2007*). Cambridge practice tests for IELTS 6*. Cambridge: Cambridge University Press.

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.

Coxhead, A., & Byrd, P. (2008). Preparing writing teachers to teach the vocabulary and grammar of academic prose. *Journal of Second Language Writing, 16*, 129–147.

Daiker, D., Kerek, A., & Morenberg, M. (1978). Sentence combining and syntactic maturity in freshman English. *College Composition and Communication, 19*(1), 36-41.

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: Newbery House.

Ferris, D. (2002). ***Treatment of error in second language student writing.*** Ann Arbor: University of Michigan Press.

Fornell, C., & Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research, 48*, 39–50.

Friedrich, P. (2008). ***Teaching academic writing.*** NY: Continuum.

Gamaroff, R. (2000). Rater reliability in language assessment: The bug of all bears. *System, 28*(1), 31-53.

Geldhof, G. J., Selig, J. P., & McConnell, E. K. (2008). *Interpreting LISREL output*. Retrieved December 5, 2008, from http://www.quant.ku.edu/pdf/Interpreting%20LISREL%20Output.pdf

Hair, J.F, Jr., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R.L. (2006). *Multivariate analysis*. NJ: Pearson Prentice-Hall, Englewood Cliffs.

Hamp-Lyons, L. (1991a). The writer's knowledge and our knowledge of the writer. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 51-70). Norwood, NJ: Ablex Publication Corporation.

Hamp-Lyons, L. (1991b). Basic concepts. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 5-15). Norwood, NJ: Ablex Publication Corporation.

Hamp-Lyons, L. (1991c). Reconstructing "academic writing proficiency". In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 127-154). Norwood, NJ: Ablex Publication Corporation.

Harmer, J. (2004). *How to teach writing*. Essex, UK: Longman.

Hedge, T. (2005). *Writing*. Oxford, UK: Oxford University Press.

Homburg, T. J. (1984). Holistic evaluation of ESL composition: Can it be validated objectively? *TESOL Quarterly, 18*, 87-107.

Jacobs, H. L., Zinkgarf, S. A., Wormuth, D. R., Hartfiel, V. F., & Hughey, J.B. (1981). *Testing ESL composition: A practical approach*. Rowley, MA: Newbery House.

Jakeman, V., & McDowell, C. (2004). *Set up to IELTS*. Cambridge, UK: Cambridge University Press.

Jöreskog, K. G., Sörbom, D. (2006). LISREL 8 (Version **8.8**) [Computer Software]. Chicago, IL: Scientific Software International Inc.

Kane, M. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed.) (pp. 17-64). Westport, CT: American Council on Education, Praeger Series on Higher Education.

Kline, R. B. (1998). *Principles and practices of structural equation modeling*. New York, NY: Guilford.

Knoch, U. (2007). Little coherence, considerable strain for reader: A comparison between two rating scales for he assessment of coherence. *Assessing Writing, 12,* 108-128.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159-174.

Lantolf, J. P. (2008). ***Sociocultural theory and the teaching of second languages.*** London: Oakville, Conn., Equinox Publishing.

Lougheed, L. (2004). *Barron's how to prepare for the Computer-Based TOEFL essay*. NY: Barron's Educational Series, Inc.

McDonald, R. P., & Ho, M. H. R. (2002). Principles and practice in reporting structural equation analyses. *Psychological Methods, 7*(1), 64-82.

McNamara, T. F. (1990). Item response theory and the validation of an ESP test for health professionals. *Language Testing, 7*, 52-75.

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

Messick, S. (1988). *Meaning and values in test validation: The science and ethics of assessment*. Princeton, NJ: Educational Testing Service, Princeton.

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).

Moore, T. & Morton, J. (1999). *Authenticity in the IELTS academic module writing test*. (Research Report No. 2, IELTS Australia).

Moore, T. & Morton, J. (2005). Dimensions of difference: A comparison of university writing and IELTS writing. *Journal of English for Academic Purposes, 4*, 43–66.

Myles, J. (2002). Second language writing and research: The writing process and error analysis in student texts. *TESL-EJ*, *6*(2), A-1. Retrieved January 31, 2009, from http://tesl-ej.org/ej22/a1.html.

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).

University of Cambridge ESOL Examinations. (2007). *Official IELTS practice materials*. Cambridge: Cambridge University Press.

Schaefer, E. (2008). Rater bias patterns in an EFL writing assessment. *Language Testing, 25*(4), 465-495.

Schumacker, R. E., & Lomax, R. G. (2004). *A beginner's guide to structural equation modeling*. London: Lawrence Erlbaum Association.

Shaw, S., & Falvey, P. (2008). *The IELTS writing assessment revision project: Towards a revised rating scale*: Retrieved January, 08, 2009, from http://www.cambridgeesol.org/assets/pdf/research_reports_01.pdf.

Vaughan, C. (1991). Holistic assessment: What goes on in the raters' mind? In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts*, (pp. 111-126). Norwood, NJ: Ablex Publication Corporation.

Weigle, S.C. (1994). Effects of training on raters of ESL compositions. *Language Testing, 11,* 197-223.

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

Weir, C. (1990). *Communicative language testing*. NJ: Prentice Hall Regents.

Widaman, K.F. (2002). To parcel or not to parcel: Exploring the question, weighing the merits. *Structural Equation Modeling, 9*(2), 151–173.

Appendix 1

IELTS Writing Task 2

The main prompt for Task 2:
You should spend about 40 minutes on this task. Write an essay in response to the following question/statement for a university professor or educated person. Use specific reasons and examples to support your answer. Give reasons for your answer and include any relevant examples from your own knowledge or experience. Write at least 250 words.

1. Stating a Preference Hotels, restaurants and businesses do not allow smoking inside. In public places such as airports smoking is banned. This is a good idea, but it takes away freedom of choice. Some smokers do not like the bans. Do you agree or disagree with banning smoking in public places? — Mickan (2003, p. 130)

2. Agreement/Disagreement Some people like to travel with a companion. Other people prefer to travel alone. Which do you prefer? Use specific reasons and examples to support your choice. — Aryadoust et al. (2007, p. 74)

3. Giving Explanations Traffic is a very serious problem. Pedestrians and bicycle riders are facing more and more dangers. Many gardens are being sacrificed to build highways. What are the best ways to satisfy citizens? Explain. — Aryadoust et al. (2007, p. 63)

4. Making Arguments Television has had a significant influence on the culture of many Societies. To what extent would you say that television has positively or negatively affected the cultural development of your society? — Aryadoust et al. (2007, p. 74)