Tabaran Institute of Higher Education
Iranian Journal of Language Testing
Vol. 5, No. 1, March 2015

ISSN 2251-7324

Examining the Psychometric and Psychological Unidimensionality of Writing in English as a Foreign Language

Masood Siyyari[1], *Negar Siyari*[2]

Abstract

The distinction between psychometric and psychological unidimensionality is a very debatable issue in tests of psychological constructs including language ability tests. In the language testing literature, psychometric unidimensionality has been controversially found not to necessarily guarantee psychological unidimensionality and vice versa. This issue provided impetus for this study to see what the analysis of psychometric unidimensionality of writing reveals about its psychological unidimensionality from an *a priori* construct validity perspective. In this study, Principal Component Analysis, as a frequently used technique in language testing, was employed to investigate unidimensionality. The results of the analysis on a heterogeneous sample of writings by 135 EFL learners demonstrate that writing is both psychometrically and psychologically unidimensional only in terms of its linguistic rather than mechanical aspects; however, it does not mean that the mechanical aspects of writing should not be considered in assessing writing.

***Keywords**: multi-dimensionality, principal component analysis, psychometric/ psychological unidimensionality, writing skill*

1. Introduction

Unidimensionality has always been one of the main concerns in psychological tests including language tests since ignoring it results in lower validity of test constructs, test scores, and test-score-based decisions. In simple terms, unidimensionality is to do with the extent to which the test and its items test only one single trait. Evidently, this definition bears many relationships with the definition of test validity; that is to say, a test is valid if it tests what it really purports to test. What a test as a whole and its items in particular purport to test is what can be called the main and only dimension of a test. However, if a test tests something other than its claim, another dimension is naturally added to the test, which makes unidimensionality as a prerequisite for validity questionable. Similarly in language tests, and specifically tests of writing, when a test

[1] *Science and Research Branch, Islamic Azad University, Tehran, Iran. E-mail: m.siyyari@srbiau.ac.ir*

[2] *Shahid Beheshti University, Tehran, Iran. E-Mail: negar_siyari@yahoo.com*

is designed to test the writing ability, it must unidimensionally test only the writing ability, and any contamination of the results by the inclusion of items to test other abilities must be avoided.

The investigation of unidimensionality first starts with a hypothesis as to the unidimensionality of the construct which is going to be tested. At this theoretical/psychological level, it is attempted to explain the unidimensionality of the construct (i.e., psychological unidimensionality) as logically and credibly as possible based on the available theories related to the construct. This level is called a priori construct validity or face validity which very much depends on the professional and theoretical expertise of the test developers (McNamara, 1996). The psychological unidimensionality could be further investigated via techniques of investigating psychometric unidimensionality which, as the name suggests, are statistical by nature. Examples in this regard are Rasch fit analysis, Bejar method (Bejar, 1980), and factor analysis, to name the most common ones at least in language testing literature (McNamara, 1996; Henning, Hudson, & Turner, 1985; Brown & Hudson, 2002). The Bejar method, as a less known method among the others, involves a comparison of two difficulty calibrations, one for the item as a part of the total test and the other for the item included in its subtest items only (See Spurling, 1987 for a review and critique).

If psychometric investigation leads to multi-dimensionality, the data analysis results are fed back to the test developers to revisit the items and test construct definition. This might lead to removal and addition of items or modification of theory if response validity is observed, or in other words, construct irrelevant factors are ensured to have been removed (e.g., fatigue, anxiety and test wiseness to name a few). According to McNamara (1996), the reverse of this scenario might also happen when the data analysis results in a single measurement dimension while it has been argued in the construct definition of the test that multiple psychological dimensions are involved in the construct. Assuming that valid data have been employed in the analysis, the reason for this issue might be that there is indeed a multiplicity of separate constructs; however, these different constructs are correlated for one reason or another, which can be further investigated by the test developers and construct theorizers. In this regard, McNamara (1996) points out the fact that, "After all, every human activity is bound to be made up of countless differ- ent sub—skills, not all of which it is desirable to measure separately (indeed, it may not be possible even to conceptualize them). Given then that psychometric and psychological unidimensionality do not map onto one another in any simple one-to-one fashion" (p. 273).

In classical test theory, unidimensionality is usually examined as an a priori requirement for test analysis for instance via factor analysis; however, in Rasch analysis, unidimensionality is hypothesized (not taken for granted) for the test, and then this hypothesis is investigated as the analysis of the data is carried out (McNamara, 1996). These two are usually done to validate the results of one another too. An example in this regard is Henning et al.'s (1985) study in which they investigated and triangulated the results of UCLA's ESLPE via Rasch analysis, Bejar method, and factor analysis.

2. Method

2.1. Participants

In the present study, the unidimensionality of writing was investigated by scrutinizing paragraph writing performances of 135 male and female undergraduate English major students studying at

Allameh Tabatabee University and Islamic Azad University in Tehran, Iran, while taking the course *Advanced Writing*.

2.2. Instrumentation

The investigation in this study required a writing measurement scale. If this scale is validly developed and administered on a sufficient sample, the extent to which it shows psychometric unidimensionality in measuring paragraph writing ability can determine the extent to which paragraph writing skill is psychometrically unidimensional, and thereby support or reject the hypothesis as to the psychological unidimensionality of writing. Description of the construction stages of this scale is presented under the following headings.

2.2.1. Scale content analysis

After reviewing the literature in the area of writing assessment, no rating scale specifically developed for paragraph writing assessment was found; almost every scale which has gone under acceptable development procedures is an essay rating scale, for instance the scales by Brown and Bailey (1984), Jacobs, Zingraf, Wormuth, Hartfiel, and Hughey (1981), Weir (1990), Hamp-Lyons (1991), and the ones used for the Test of Written English, and IELTS. Of course, there are many common points between different writing tasks as far as the mechanical, lexical, syntactical, and content-wise issues are concerned; however, many differences exist between a paragraph writing tasks and other writing tasks. In a nutshell, the main difference between the two lies in the fact that an essay contains several paragraphs which are linked to one another with a thesis statement which runs through the whole essay. This is in contrast to a single paragraph, as the concern of this study, which is considered as a complete piece of discourse in its own right, whose topic is introduced with a topic sentence, while thesis statement does so in an essay, which is supposed to be presented in a quite different way. Also, the essay includes a topic sentence for each body paragraph which is again different from the topic sentence in a single paragraph. The way the support sentences are presented in a single complete paragraph is also different from the way the body paragraphs are presented in an essay. Moreover, the content should be presented in a more concise way than an essay since the word limit in paragraphs does not allow for much detailed exposition by the writer.

To develop the needed scale of this study, scales based on holistic, primary and multiple trait, and analytic scoring could be developed. However, based on the arguments of many scholars in the field such as Jacobs et al. (1981), Perkins (1983), Hamp-Lyons (1991), Weir (1993), Genesee and Upshur (1996), to name a few, analytic scoring was preferred to other types of scoring for its accuracy and positive washback effect.

To begin the development of this scale, it was necessary to define the paragraph writing ability as the construct this scale was supposed to measure. This meant that the necessary features and components of a well-written paragraph needed to be defined operationally. To do so, the above-mentioned writing scales (i.e., by Jacobs et al., 1981; Brown & Bailey, 1984; Hamp-Lyons, 1991; Weir, 1993), as well as several textbooks and manuals on writing such as Arnaudet and Barrett's *Paragraph Development* (1990), Kane's *Oxford Essential Guide to Writing* (1988), Hinkel's *Teaching Academic ESL Writing* (2004), and *Publication Manual of the*

American Psychological Association (2001) were consulted. At this point, a pool of writing components was made; however, whatever component which had to do with an essay was modified to make it applicable to a paragraph. For instance, based on recommendations by Hartfiel, Jacobs, Zinkgraft, Wormuth, and Hughey (1985), the components 'introductory paragraph' and 'body paragraphs' were replaced with 'topic sentence' and 'supporting sentences' respectively. Finally, all these modified components were classified under the following headings: 1) Organization, 2) Content, 3) Grammar, 4) Vocabulary, and 5) Mechanics.

To ensure the relevance and adequacy of the scale content, a list of all these components accompanied by their definitions and examples was sent to 12 instructors experienced in teaching *Advanced Writing*. The instructors were supposed to rate the degree to which these micro components were relevant and needed to be considered in rating a paragraph on a scale ranging from "completely irrelevant" (1) to "completely relevant" (5). Each micro component was fully defined in a separate attached file. Besides, if the instructors had any further comment about their ratings, or they thought any micro component or macro component needed to be added to the scale, they could mention it in the space provided below the scale. Almost every instructor agreed on the components since the average ratings for the relevance of the micro components to the content ranged from 3.83 to 5.00 (see Appendix A for a table of average ratings); however, some modifications were suggested in terms of the wording of the components. The outcome of this stage, resulting in 20 micro components, classified under 5 macro components, are presented in Table 1.

After defining paragraph writing in terms of the above micro and macro components, it was necessary to develop a scoring system and descriptors. According to Weir (1993) and Weigle (2002), it is of utmost importance to develop the descriptors based on real performances of learners rather than based on intuition; therefore, a heterogeneous sample of writing performances covering almost every point on the continuum of writing ability ranging from very weak to very strong needed to be prepared. Evidently, the larger and the more random in nature the sample is, the more precise and comprehensive data one could gather about a measure; however, for practical reasons this was not possible to pursue. The solution for this shortcoming was to come up with a heterogeneous sample whose scores may cover all the points on the writing performance continuum or scale. To come up with such a sample, it was needed to find a heterogeneous sample in terms of paragraph writing performance, at least consisting of a lower-ability group and a higher-ability group. The more of these groups were identified in a sample, for example a sample consisting of five ability groups rank-ordered from high to low, the more heterogeneous the sample could be claimed to be.

The parallel criterion chosen for measuring the paragraph writing performance was the holistic scoring of the paragraphs written by some university students majoring in English literature and translation. This choice was justified on the grounds that holistic scoring, apart from its shortcomings, is a proper choice for differentiation across the levels of performance and rank-ordering purposes (Brown, 2004). Following this point, paragraphs written by more than 200 hundred students in the current and previous semesters were surveyed, and out of them 135 paragraphs were selected. In fact, these 135 paragraphs were divided into five groups (i.e. each group with 27 paragraphs) based on the holistic scores ranging between 1 to 5. All holistic scorings were done by a trained native-like bilingual English speaker who has taught different English language skills, specifically writing, for both general and academic purposes at institute and university levels. These ratings were correlated with another trained rater's ratings for inter-rater reliability purpose. The resulting correlation between the two sets of scores was .88. Assuming that these paragraphs were correctly rank-ordered based on the holistic impression they had given to the raters, this sample could be considered almost heterogeneous having two identifiable pairs of high and low ability subgroups on either ends of the continuum. This way of coming up with a heterogeneous sample was done drawing on Embretson and Reise's (2000, p. 123) suggestion that in order to have a heterogeneous sample, "item responses" (i.e. performances) which correspond to each "response category" (i.e. score or scale band) on the rating scale should be included in the data. Next, the paragraphs were rated by three trained raters as regards all the micro components come up with in the above, on a scale of 1 to 5, 1 standing for "very poor" and 5 for "excellent". Pearson correlation coefficient among the three raters' ratings for each micro component was also calculated to investigate the interrater reliability. In a nutshell, all these correlation coefficients were significant at the 0.01 level (2-tailed), ranging from 0.60 to 0.98.

To determine the final score out of the three awarded scores by the raters, for the sets consisting of two similar and one different score, the median or the more frequent score was chosen. If all the scores were different with a maximum range of two, the average of the three scores was chosen.

Table 1. Primary Macro Components and Micro Components of a Paragraph

Macro components	Micro components
1 Organization	1. Title 2. Topic sentence 3. Conclusion 4. Development and organization of points & supporting sentences, using a particular sequence or method of development such as exemplification, description & details, facts & statistics, or anecdotes, process or chronological enumeration, descending/ascending-order enumeration, cause & effect, comparison/contrast, definition,
2 Content	5. Unity & relevance of supporting sentences 6. Thoughtful content & understanding of the subject 7. Effective repetition of key words, phrases and ideas without losing concision 8. Fluent expression of ideas by (a) transition elements to link ideas & (b) substitution, referencing, repetition & deletion with no under/overuse to create cohesion with implicit/explicit use of cohesive devices
3 Grammar	9. Syntactic complexity & variety 10. Accuracy & acceptability in using structures to fully communicate ideas 11. Matched function & form in passive/active sentences
4 Vocabulary	12. Precise, unambiguous, & familiar vocabulary & collocations 13. Lexical variety 14. Emphatic word placement 15. Use of Prefixes, suffixes, roots, compounds & parts of speech 16. Register choice (formal/informal)

5 Mechanics	17. Margins & indentation 18. Punctuation 19. Spelling & capitalization 20. Neatness & legibility

The reason for not choosing the average of the three ratings was the fact that averaging produced decimal numbers which were not defined in the scoring system of the writing samples, that is to say, the defined scores were all whole numbers ranging from 1 to 5. After finalizing the analytic scores, in order to check the correlation between these analytic scores and the holistic scores, a Pearson correlation coefficient between the two sets was calculated (.95), which could be considered acceptable.

2.2.2. Producing descriptors

To develop the descriptors for each band score, all the sample writings by the participants were classified. For instance, all the papers which had received the score 1, 2, 3 … for the micro component of "punctuation" were put aside so that the three raters write their qualitative descriptions of the discourse characteristics of those particular papers in terms of that particular micro component. Later these descriptions were discussed in a group discussion to arrive at the final descriptors. The final outcome of this phase was a scale with 5 macro components, and 20 micro components, each macro component having 5 descriptors for band scores ranging from 1 to 5, 1 meaning very poor and 5 excellent. The total score is calculated by summing the scores for the macro components making the total score 25 (See Appendix B for a copy of the scale and its descriptors).

2.2.3. Reliability analysis

In order to compute the reliability of the scale, the very 135 students' paragraphs, consisting of 27 paragraphs corresponding to each holistic score (ranging between 1 and 5), were scored once more analytically by three trained raters, including the researcher, based on the generated rating scale. The interrater reliability for every macro component was then calculated, the values of which ranged from 0.75 and 0.96. A correlation was also calculated between the holistic scores and the analytic scores which turned out to be 0.9. Moreover, to estimate the test reliability depending on the homogeneity of item variance as a sign of internal consistency (Henning, 1987), Cronbach's alpha was also calculated which was sufficiently high (0.83).

3. Results of Unidimensionality Analysis

As explained in the previous sections of this study, the subskills or micro components of writing were classified under five different headings, that is organization, content, grammar, vocabulary, mechanics. Now if it is assumed that these macro components can theoretically stand independently from one another, then one is right to claim that writing is multi-dimensional from an a priori construct validity perspective. However, if it is not the case; that is to say, it is assumed that grammar could not stand independently from vocabulary, or if the organization is weak, both content and even grammar might be overshadowed negatively, then one is right to claim that writing is unidimensional. This is exactly what I claim, since based on Halliday's concept of lexicogrammar (1978), not much distinction is made between grammar and vocabulary, and there is a lot of overlap between these two.

To investigate the unidimensionality of writing, Principal Component Analysis, a variant of factor analysis, was performed on the analytically-rated paragraphs of the initial participants of the study (n = 135). It should be noted that several approaches to test unidimensionality exist in the literature none of which provides satisfactory indices (Brown & Hudson, 2002); however, factor analysis was employed here since it is more straightforward and more often employed in the area of language testing (Henning et al., 1985; Brown & Hudson, 2002).

After checking the assumptions of factor analysis by employing factorability indices (i.e., Bartlett's test of sphericity and the Kaiser-Meyer-Olkin value), Principal Component Analysis was run as follows. Based on the factor analysis results, if one dominant factor emerges, one is on safe grounds to claim unidimensionality, but if more than one dominant factor emerges, unidimensionality is still open to question. With regard to these points, Principal Component Analysis was employed two times, once on the micro components and the other time on the macro components. As regards the micro components' factor analysis, Table 2 demonstrates that almost all the components have loaded on the first factor, thus meaning that unidimensionality is almost observed. A closer look reveals that the micro components to do with "title', as well as "margins and indentation" and "neatness/legibility" together have loaded on different factors from the dominant factor.

Table 2. Component Matrix[a] for the Micro components

Scale's Micro Components	Component					
	1	2	3	4	5	6
Title		-.41	.86			
Topic sentence	.89					
Conclusion	.89					
Development method	.91					
Unity	.85					
Thoughtful content	.80				-.44	
Repetition & concision	.93					
Cohesion	.89					
Syntactic complexity/variety	.92					
Accuracy	.92					
Passive/active	.96					

Lexical precision & collocations	.85					
Lexical variety	.89					
Lexical emphatic placement	.92					
Affixes, roots, compounds, parts of speech	.91					
Register	.92					
Margins	.46	.77				
Punctuation	.82			-.44		
Spelling	.91					
Neatness/legibility		.79				

Extraction Method: Principal Component Analysis.

[a] 6 components extracted.

As regards the macro components' factor analysis, Table 3 shows that three factors have emerged, the first of which could be safely considered as the dominant factor since all the macro components have loaded on it with no exception.

Table 3. Component Matrix[a] for the Macro components

Scales'	Component		
Macro Components	1	2	3
Organization	.92	-.32	.12
Content	.92	.29	.21
Grammar	.96	.05	-.05
Vocabulary	.97	-.08	.01
Mechanics	.94	.05	-.29

Extraction Method: Principal Component Analysis.

[a] 3 components extracted.

4. Discussion

This study started with developing an operational definition of the writing skill which was done based on writing expert group opinion. To do so, experts enumerated the components which they believed were related to writing skill.

Next, those components were statistically investigated to see if they were correlated with one underlying factor which is presumably the writing construct. Table 2 above showed that on the micro component level three mechanical components (i.e., margins, indentation, and neatness/legibility) loaded on factors different from the factor on which the other micro components (i.e., the micro components of content, organization, vocabulary, and grammar), which were more linguistic by nature, loaded. What could be argued is that this finding is justifiable since no good content could be demonstrated to the audience (e.g. raters) without any good command of vocabulary or grammar, or even organization. However, it is indeed possible to find a person whose expressive aspect of language is good while his handwriting, neatness, margins and indentations,which are more to do with the surface features of the text rather than the linguistic aspects, are weak.

The results of factor analysis also indicated that the other micro components of the mechanical macro component of writing, that is spelling, capitalization and punctuation (albeit classified as mechanical micro components, loaded on the factor on which all the linguistic micro components had loaded. What could be argued to justify this finding is that although spelling, capitalization and punctuation are less linguistic than the other micro components, they are more linguistic by nature than handwriting, neatness, margins and indentation. No one can deny the fact that punctuations are meaningful in texts, and their appearance signals particular meanings or functions. It is very probable that the more these linguistic and semantic functions are identified, the more linguistically proficient that person could be conceived. In addition, one can certainly only deal with spelling and capitalization when he has had enough contact with language and as a result that person could be imagined to be linguistically more proficient too.

Given the above arguments, now the question is whether these micro components (i.e., margins, indentation, and neatness/legibility) are really not related to the construct of writing. Apparently not; however, by having a reformulation of the definition of writing construct, the results may sound viable. In fact, the writing construct has a definition from psychological perspective which necessitates the construct being unidimensional by nature. This is certainly too idealistic and theoretical to achieve since "examinee performance is confounded with many cognitive and affective test factors such as test wiseness, cognitive style, test-taking strategy, fatigue, motivation and anxiety. Thus, no test can strictly be said to measure one and only one trait" (Henning et al., 1985, p. 142). However, there is an actual or practical definition of the writing construct which is to do with what happens in reality, that is actual paper-based writing (if computer-based writing could be ignored).

The actual definition of writing skill needs to consider issues of mechanics which are less to do with the expressive aspects of writing and more to do with the surface editing of the text. This non-expressive aspect of writing is what is not included in the psychological definition of writing as a unidimensional construct, and that is exactly why margins, indentation, and neatness/legibility, which are even less linguistic than other mechanical micro components of punctuation and spelling, have loaded on different factors.

Also, if remembered from the analysis of the relevance and adequacy of the scale content in the initial development stage of the scale, these two micro components received the least ratings in terms of their relevance to the content. However, it is conventionally of significance to observe the mechanical aspects if a good final impression is pursued in actual writing.

Finally, it should be noted that when margins, indentation, and neatness/legibility were considered as the micro components of a larger component called 'mechanics', which incorporates other micro components as well, the whole macro components of mechanics was found correlated with other macro components. This is probably due to the fact that scores on these micro components clustered with other micro components under the macro component of mechanics, and then together they showed the more holistically the construct is looked at, the less dimensionality becomes an issue. Similar results to the above findings are cited by McNamara (1996) from O'Loughlin (1992) which can corroborate the results and interpretations above. O'Loughlin designed a comprehension-based writing task the Rasch analysis of which showed that items or components to do with productive writing skill and components to do with comprehension skill represented different dimensions. Following this emergence of multi-dimensionality, the writing components scores were analyzes separately from the comprehension-based items. This time, spelling and punctuation were found misfitting or representing a different dimension from the main one representing other writing components. While justifying that these components of writing were misfitting since they are concerned with surface editing of the text rather than the expressive aspect of writing, it is further added that these components were not misfitting in the original analysis including the comprehension-based items since

> this more subtle degree of multi-dimensionality was masked by the grosser level of multi-dimensionality apparent in the first analysis … scores on spelling and punctuation clustered more with the other writing skill scores rather than with the comprehension scores. When the comprehension scores were removed, the diversity within the 'writing' cluster became clearer." (McNamara, 1996, p. 278)

What finally remains to be discussed is the issue of 'title' which has also loaded on a different factor. Obviously, giving a good title to one's written work is an important step which is linguistic by nature as most of other micro components are. However, no learner in this study was supposed to write a topic for his or her writing since they were already assigned the topics. So all the participants got the same score for this micro component no matter what they had done in terms of other micro components, and as a result different factor loading occurred for this micro component.

5. Conclusions

The analysis above resulted in finding psychometric multi-dimensionality in the construct of writing skill. This multi-dimensionality was interpreted in the light of views on the nature of the micro components involved in the writing skill, which would justify the observed multi-dimensionality. In sum, the plausible conclusion of this finding is that writing is both psychometrically and psychologically unidimensional only as far as its components to do with the expressive and linguistic aspects are concerned; however, it does not mean that the mechanical aspects of writing should not be taken into account in assessing writing performance.

What should also be added is that unidimensionality especially at psychological and psychometric level is a matter of theory and practice whose gap always remains to be bridged. Based on these results, it seems to be viable to maintain that linguistic constructs especially at the skill level could be thought of as theoretically unidimensional constructs which might turn out to be dimensional in actual practical representation. This dimensionality does not matter if its source is identified and justified as it was the case in this study.

Dimensionality could be problematic in measurement and construct validation if there is either any shortcoming with the definition of the construct at the theoretical level or if the results are contaminated by external factors resulting in low response validity (Henning, 1987).

To mention one of the main implications of this study, the gap between theory and practice in psychological and psychometric kinds of unidimensionality should be bridged by further research and modification of practice. One measure that can help bridge the gap between the theory of writing as a unidimensional construct and its practice as a multi-dimensional construct is probably devising writing test procedures free of mechanical/conventional issues. Computer-based writing could be one solution to this issue by means of which neatness, legibility, margins, and indentations could be by default excluded from the testing procedure.

For this study, further research could be recommended in terms of triangulation of unidimensionality investigation with other techniques such as Rasch fit analysis and Bejar technique. Also, investigation of unidimensionality in essay writing and other modes of writing could be done to compare the results with the findings of this study.

Acknowledgements

I would like to thank the reviewers especially Dr. Purya Baghaei, the editor in chief of Iranian Journal of Language Testing, for their insightful and constructive advice on the earlier drafts of this paper.

References

American Psychological Association (2001) *Publication manual of the American Psychological Association* (5th ed.). Washington, DC: American Psychological Association.

Arnaudet, M. L., & Barrett M. E. (1990) *Paragraph development: A guide for students of English (2nd ed.).* New Jersey: Prentice Hall

Bejar, I. (1980) A procedure for investigating the unidimensionality of achievement tests based on item parameter estimates. Journal of Educational Measurement, *17*, 283-96.

Brown, H. D. (2004) *Language assessment: Principles and classroom practice*. New York: Pearson education, Inc.

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

Brown, J. D., & Hudson, T. (2002) *Criterion-referenced language testing*. Cambridge: Cambridge University Press.

Educational Testing Service. (2008) *The Test of Written English*. New Jersey: Educational Testing Service.

Embretson, S. E., & Reise, S. P. (2000) Item response theory for psychologists. Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc.

Fulcher, G. (1996) Does thick description lead to smart tests? A data-based approach to rating scale construction. *Language Testing, 13*, 208-238.

Genesee, F. and J. Upshur. (1996) *Classroom-based evaluation in second language education.* Cambridge: Cambridge University Press.

Hamp-Lyons, L. (1991) Scoring procedures for ESL contexts. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 241-276). Norwood, NJ: Ablex.

Hartfiel, V. F., Jacobs, H. L., Zinkgraf, S. A., Wormuth, D. R., & Hughey, J. B. (1985) *Learning ESL composition*. Rowley, MA: Newbury House.

Henning, G. (1987). *A guide to language testing: development, evaluation, research.* Massachusetts: Newbury House Publishers.

Henning, G., Hudson, T., & Turner, J. (1985) Item response theory and the assumption of unidimensionality. *Language Testing, 2*, 141-54.

Hinkel, E. (2004) *Teaching Academic ESL Writing*. Lawrence Erlbaum Associates Publishers.

Jacobs, H. J., Zingraf, S. A., Wormuth, D. R., Hartfiel, V. F., & Hughey, J. B. (1981). *Testing ESL composition: A practical Approach*. Massachusetts: Newbury House.

Kane, T. S. (1988). *Oxford essential guide to writing*. New York: Oxford University Press.

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

Perkins, K. (1983) On the use of composition scoring techniques, objective measures, and objective tests to evaluate ESL writing ability. *TESOL Quarterly*, *17*, 651-71.

Spurling, S. (1987) Questioning the use of the Bejar method to determine unidimensionality. *Language Testing*, *4*, 93-95.

University of Cambridge Local Examinations Syndicate (2007) *IELTS Handbook*. Cambridge: Cambridge University Press.

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

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

Weir, C. (1993) *Understanding and developing language tests*. New York: Prentice Hall.

Appendix A: The average content relevance ratings of the micro components

Micro components	Average relevance ratings in descending order
Topic sentence	5.00
Accuracy & acceptability	5.00
Conclusion	4.92
Support & development method	4.92
Unity & relevance of supporting sentences	4.83
Thoughtful content & understanding of the subject	4.67
syntactic complexity & variety	4.58
Precise & familiar vocabulary/idioms & collocations	4.58
Fluent expression of ideas	4.50
Use of Prefixes, suffixes, roots, compounds & parts of speech	4.50
Register choice (formal/informal)	4.50
Effective & concise repetition of key words, phrases, and ideas	4.42
Lexical variety	4.42
Matched function & form in passive/active sentences	4.33
Emphatic word placement	4.33
Punctuation	4.08
Title	4.00
Spelling & capitalization	4.00
Neatness & legibility	3.92
Margins & indentation	3.83

Appendix B: Paragraph writing scale

	Micro components	1 (Very poor)	2 (Very poor to weak)	3 (Weak to fair)	4 (Fair to good)	5 (Good to excellent)
ORGANIZATION	1. Title	Absent or totally irrelevant & non-sense title, absent or quite unclear & irrelevant topic sentence to the title & content, absent or non-sense conclusion, no apparent organization & showing no knowledge of any development method, too short with illogical supporting sentences	Confusing & vague title, not showing the features of a topic sentence with little relevance to the topic or content, conclusion present but shows no attempt to be fully relevant & logical, weak outline with insufficient length & confused method of development	Too general or too narrow & ambiguous title, relevant topic sentence to the title but does not adequately narrow down the content, somewhat irrelevant conclusion not giving the paragraph a sense of completeness, difficult to outline by the reader, identifiable development method but not meeting all the features	1 Acceptable title but with minimal irrelevance, ambiguity & focus, topic sentence, somewhat logical & relevant but could be better narrowed down, simple & relevant conclusion but could be more sophisticated, outlinable organization with an identifiable method of development but missing some points	Clear, relevant, concise & focused title, clear topic sentence with full relevance to the title & effectively narrowing down the content , logical conclusion with a sense of completeness without starting a new topic or undermining the argument, fully developed & organized points & supporting sentences using a particular method of development
	2. Topic sentence					
	3. Conclusion					
	4. Development & organization of points & supporting sentences, method of development					
	5. Unity & relevance	No unity, full of irrelevance, content reflecting no understanding or	Shaky unity with frequent irrelevance, content reflecting little thought &	Unity at times threatened by irrelevance, some ideas signal the misunderstanding or hurriedness of	Unity almost observed with little digression, the subject is well-thought but minor misunderstan	Unified paragraph, no digression, thoughtful content, full understanding of the subject, effective
	6. Thoughtful content					
	7. Repetition of key words, phrases &					

CONTENT	ideas, concision	thought by the writer, repetitive &illogical content showing no familiarity with how to achieve concision, no cohesion with wrong or no use of cohesive devices signaling lack of competence in this regard	attempt to deeply develop the ideas, mostly sounding unnecessarily repetitive, reflecting not much competence with effective repetition & concision, weak cohesion with numerous errors in the use of cohesive devices, not showing much knowledge of substitution, referencing, repetition & deletion	the writer about the subject, repetition is at times unnecessary at the cost of concision, cohesion at times at risk with under/over/misuse of substitution, referencing, repetition & deletion	ding of the subject is seen, somewhat balanced repletion & concision but could be even better, substitution, repetition, deletion, referencing & cohesion observed & acceptable with negligible under/over/misuse	repetition to emphasize the main ideas without losing concision via effective techniques of concision e.g. ellipsis & parallelism, fluent expression of ideas by accurate & effective use of (a) transition elements to link ideas & (b) substitution, referencing, repetition & deletion with no under/overuse to create cohesion with implicit/explicit cohesive devices
	8. Fluency of expression by (a) transition elements (b) substitution, referencing, repetition & deletion to create cohesion with cohesive devices					
GRAMMAR	9. Complexity & variety in Sentence types & length	Grammatical knowledge is too little to include any complexity & variety, numerous incomplete &	Mainly of simple sentences with numerous errors in sentence formation, wrong passive forms	Simple but accurate sentence formation but with not much complexity & variety, some erroneous passive formation with little	Somewhat adequate complexity & variety in sentence types & length, correct passive/active forms but functions are not	Native-like complexity & variety in sentence types & length, matched function & form in passive/active sentences,
	10. Accuracy & acceptability in the use of					

	structures to communicate ideas	inaccurate sentences leading to communication failure	showing no awareness of passive/active functions, frequent inaccuracies at times leading to communication failure	observation of passive/active function, numerous inaccuracies but still conveying the message	adequately observed, some minor inaccuracies in grammar	almost no inaccuracy & unacceptability in the use of structures to fully communicate ideas
	11. Passive/active					
VOCABULARY	12. Precise & familiar vocabulary & collocations causing no ambiguity or misunderstanding	Wrong choice of words & collocation causing communication failure, vocabulary knowledge is too little to show any variety or emphatic placement & use, widespread mistakes showing no command of parts of speech, confused or wrong register choice reflecting no awareness of register	Many mistakes in word choice & collocations that hardly communicates anything, most attempts at variety are failed, no particular emphatic word placement is seen showing little awareness of that, parts of speech are mostly wrong with little awareness of roots & compounds, no particular	Some wrong collocations & word choices but the message is still conveyed, some words could have been replaced with better synonyms to achieve precision & variety, emphasis is not much accomplished by word placement, frequent wrong prefixes, suffixes, roots, compounds, & parts of speech, wrong or confused register,	Acceptable word choice & collocations causing little ambiguity, adequate variety, emphatic word placement is more or less achieved, prefixes, suffixes, roots, compounds, & parts of speech have very minor mistakes, register choice somewhat acceptable	Precise & familiar vocabulary & collocations causing no ambiguity or misunderstanding, native-like variety in the vocabulary use, efficient word placement to give emphasis, effective & accurate use of prefixes, suffixes, roots, compounds, & parts of speech, flawless register choice
	13. Lexical variety					
	14. Emphatic word placement					
	15. Prefixes, suffixes, roots, compounds, & parts of speech					
	16. Register					

			register is identified showing little knowledge of that			
MECHANICS	17. Margins & indentation 18. Punctuation 19. Spelling & capitalization 20. Neatness & legibility	Wrong or unclear margins & indentations showing no awareness of them, absent or frequently wrong punctuations & capitalization, illegibility & wrong spellings leading to communication failure	Non-standard or sketchy margins & indentation, with few accurate uses of punctuation & capitalization, distracting illegibility & without much care for neatness, noticeable errors in spelling causing difficulty to reader,	Margins & indentation not fully observed, some erroneous punctuation, spellings & capitalization, not quite neat or legible but communication is not impeded	Somewhat correct margins & indentation, acceptable punctuation with minor mistakes, clean & legible, acceptable spelling with few errors in capitalization	Correct margins & indentation, precise punctuation, neat, legible with nice handwriting, Accurate spelling & capitalization