Available online at www.elixirpublishers.com (Elixir International Journal)

Social Sciences

Elixir Soc. Sci. 79 (2015) 30260-30266

ISSN: 2229-712X

Investigating the relationship between writing components and the quality of writing: a case of construct validation using Structural Equation Modeling (SEM)

Mohammad Naghi Keramati[1] and Mohammad Khatib[2]

[1]Islmaic Azad University, Science and Research Campus, Tehran.

[2]Allameh Tabatabei University, Tehran.

ARTICLE INFO

Article history:

Received: 19 November 2014;
Received in revised form:
20 January 2015;
Accepted: 4 February 2015;

Keywords

Structural Equation Modeling (SEM),
Scoring scales,
Scoring types,
Evaluation criteria.

ABSTRACT

This study is an attempt to investigate empirical justification for selection and number of components used in evaluation criteria based on which scoring scales are developed and score is assigned; a need which is felt by many scholars. Following the literature eight evaluation criteria was selected and a scoring scale was developed. 124 scripts (totally 248) for each task of letter writing and argumentation were scored analytically with the selected eight components. Then the scores were tested through a number of statistical procedures (Exploratory and Confirmatory Factor Analysis) to empirically examine the underlying components of writing against which quality of writing was assessed. The results indicate that it is possible to have a standard to empirically justify the selection and number of components. However, this finding is not conclusive and there are still some uncertainties which warrant further studies.

Introduction

Performance assessment has widely been used in assessing writing ability of the test takers in second language. In large scale testing situations, these tests take the modal of "timed, impromptu writing task" (weigle, 2002) which are different from process writing (Weir, 2005). The use of performance assessment in communicative approaches is contrasted with discrete item and indirect tests in that these assessments enjoy greater validity (Kane, Crooks, and Cohen, 1999).

However, there are some challenges associated with the scoring of these types of assessments. Score is the result of a number of factors including evaluation criteria (i.e., content, organization, grammar etc.). Intuitive selection and the kind of these criteria which are used in scoring scales can create scoring variances which may weaken the validity (Fulcher, 2003; Schoonen, 2005).

This study aims to address these issues to the extent possible. Using advanced statistical procedures, this study intend to see if any empirical justification can be proposed for the selection and number of components of writing.

Literature Review

To measure the writing ability of the test takers we have to get them to write (Haughes,1989).This is called direct measure of writing and is in contrast with the so-called multiple-choice type tests or indirect tests. Assessment of writing ability directly is one kind of performance based tests in which the test taker responds to the task (s) designed and developed by the test constructor. In such instances, the tasks should resemble as closely as possible a communicative situation in a non-testing situation and the test takers need to respond to the tasks by producing relevant information which is linguistically correct and socially appropriate (Clark, 1972 cited in Bachman, 1990; Bachman and palmer, 1996; schoonen, 2005). These tests are regarded as the most 'construct valid' form of writing assessment in comparison to standardized multiple choice test, because performance in these tests are closer to the performance in real life situation and the inference made based on the scores in these tests are likely to be more trustworthy (Bachman and Plamer, 1996). Assessing the quality of writing is a very complex issue and a true challenge, because writing ability is a multi-dimensional construct. The score given to a written text which further becomes the basis of judgment made about the test taker's language ability is the result of a number of factors (i.e, Bachman, 1990; Elder, Barkhuizen, Knoch, and Randow,2007; Hamp-Lyons,1990; McNamara, 1996; schoonen, 2005):

- Communicative ability of test taker (language proficiency)
- Test method facets (time, tasks, discourse mode or genre (i.e., argumentative, description etc), topic, and the writing mode (handwritten or word-processed)
- Personal characteristics of test takers (gender, nationality, age etc.)
- Unsystematic factors (fatigue, administration conditions etc.)
- The rating scales (rating types, i.e, analytic or holistic, and evaluation criteria such as content, organization etc.)

A reliable and valid score is the cornerstone of writing performance, because score is the basis of inference made about test takers' writing ability and any incorrect inferences about the test takers' writing ability will weaken validity. The focus of this study is only on the components or traits (evaluating criteria) based on which the quality of writing is assessed. It has been tried to control other factors influencing the quality of writing to the extent possible so that the finding of the study could be attributed only to evaluation criteria.

Evaluation criteria (i.e., content, language use, task realization, etc.) are the basis for the development of scoring scales and descriptors which are used by the raters to score the scripts. Raters perception which could interact with the evaluation criteria and rating scales have been investigated extensively in the literature (Lumely, 2002, 2005; Cumming,

Tele:
E-mail addresses: kaveh_j2004@yahoo.com

1990; Milanovic, Saville, and Shuhung, 1996; Vaughan,1991: Cohen 1994; Elder et.al., 2007).To identify which evaluating criteria were most frequently focused on by the raters, researchers (Cumming, 1990; Wolfe-Quintaro, Inagaki and Kim, 1998; Vaughan, 1991; Milanovic et al. 1996) used " think aloud protocol", a technique for data collection where raters are required to speak their thought out when they are scoring the scripts. The followings are the studies where the scholars used this technique to indentify evaluation criteria.

In a very important study where nine raters were used to rate a number of scripts,Vaughan (1991) identified the following evaluation criteria which were the most frequently used to evaluate the quality of writing of the test takers: (a) quality of content, (b) legibility and hand writing,(c) tense and verb problem, (d) punctuation and capitalization error, (e) quality of introduction, and (f) morphology/word form error. It is interesting to note that quality of content was the most frequently focused criterion that raters concentrated on.

While Vaughan identified six components which are mostly used by raters to assess the scripts, Cohen (1994) proposed ten evaluation criteria which could logically be used: (a) content, (b) rhetorical structure, (c) organization, (d) register, (e) style, (f) economy, (g) accuracy of meaning, (h) appropriateness of meaning, (i) reader's understanding, and (j) reader's acceptance.

Milanovic, et. al., (1996), in a different study, used two large scale proficiency tests of First Certificate in English (FCE) and Certificate of Proficiency in English (CPA) with a large group of test takers. The scripts were scored by 16 raters with different backgrounds. The researchers asked the raters to report the evaluation components they focused on most when rating the scripts. They found that (a) length, (b) legibility, (c) grammar, (d) structure, (e) communicative effectiveness, (f) tone, (g) vocabulary, (h) spelling, (i) content, (j) task realization, and (k) punctuation were most often used as evaluation criteria by the raters. They grouped the essays into high and intermediate essays and found that the focus of the raters with high-level essays were mostly on content and vocabulary, but, with the intermediate level essay, the raters concentrated on task realization and communicative effectiveness. This can be an indication that the interaction of the rater perception with the evaluation components might be influenced by the proficiency level of the test takers.

Other studies by institutions like IELTS and TOFEL as well as individuals like Elder et.al.(2007); Jacob, Zinkgraf, Wormuth, Hartfiel, and Hughey, (1981); Grant and Ginther (2000); and Bae and Bachman (2010) identified three to five evaluations criteria against which quality of writing was assessed. Scholars like Elder et.al.(2007), who identified three global components of (a) content, (b) fluency, (c) form to assess the writing quality, further divided each of the above-mentioned global categories into three subcategory. For example, content was sub-categorized as (a) data description, (b) interpretation, and (c) development of ideas or form as a global category was divided into (a) sentence structure, (b) vocabulary and spelling, and (c) grammatical accuracy. For the raters of institutions like IELTS that used four evaluation criteria to assess the quality of writing, some hidden components like length and spelling were also regarded important, though it was not explicit in the rating scales.

The above findings, however, indicate that there is a general agreement among the raters when examining and scoring aspects of L2 writing ability. Content, language use, and organization are three typical criteria which are consistently used by the scholars. That is, these components are the most influencing criteria in defining writing quality.The findings also indicate that, the scholars refer to the same concepts to describe evaluation criteria but use different terminologies. For example where some scholars (i.e., Jacob et.al., 1981; Vaughan 1991; Cohen 1994; Elder et.al., 2007) referred to a component as content or quality of content, Milanovic et al., (1996) and some institution like TOFEL or ILTS named the same concept as task realization, or task fulfillment. The same holds true with organization and development, cohesion and coherence, sentence structure and grammar, vocabulary and lexical use etc(*seeappendix I*).

Apart from the above findings, what is evident is that to assess quality of writing, educators and institutions use different traits and number of components (explicit or hidden) which can vary from three to twelve or even more. Despite the general consensus on the most frequently focused components (i.e., content, organization and language use), it does not seem the scholars fully agree with one another as to what constitutes the construct of writing. Hamp-Lyons (2002) is right in asserting that scholars do not share a construct of writing quality, because they assign different score to a single script. Much later, Schoonen (2005) also agreed that "the *selection of traits, and number of them* would influence the raters who rate the scripts and assign scores" (p. 3. Italic added by the researchers). One possible reason might be the lack of consensus on such questions as "whether separable comprehension sub-skills exists, and what such subskill might consist of and how they might be classified" (Alderson, 2000, p.10) in writing and speaking skills.

Fulcher (2003) shows that the existing scales are usually based on intuitive methods which means they are either adopted from the already existing scales or they are based purely on common sense of the test developer. Bae and Bachman (2010) claim that there are relatively few studies that design and test a factor model of what constitutes writing ability. Cumming et al., (2005) is right in noting that:

Although educators around the world regularly work with implicit understandings of what constitutes effective English Writing, no existing research or testing programs have proposed or verified a specific model of this, such as would be universally accepted (p.27)

To respond to the plea of the above scholars, this study intends to construct validate the underlying components of writing ability using a more complicated statistical procedure. Specifically, this study intend to know if a standard can be set for *selection of traits*and *number* of them used in evaluation criteria based on which scoring scales are developed.

Purpose of the study

This study intends to see whether or not any empirical justification can be proposed for the selection and number of traits used in the rating scale. And, whether or not there is any single component which can best predict writing ability of second language learners.

1. Is there any empirical justification for selection and number of components used in evaluation criteria, based on which scoring scales are designed and developed, and quality of writing is assessed? In other words, how many components should an evaluation criteria include, three, four, five or more?
2. Is there any statistically significant relationship between holistic and features of analytic ratings?

Method

Participants

Participants of this study were all Iranian MA students, majoring in Teaching English as a Foreign Language (TEFL) at different universities in Tehran. English is, therefore, regarded as a second language (L2) for all of them. The students were both males and females at 1st and 2ndyears of study in their universities with different age ranging from 22-31. They are coming from various universities in Tehran.

MA students were intentionally selected for this study, because they were proficient enough to produce a reasonable written text in response to a “timed writing task” which is a prevalent method in testing context internationally. Moreover, students at this level are assumed to have more or less similar language proficiency level, because they are screened by a nationwide entrance exam for their program. Besides, at this stage, they have already had one or two courses of writing during their studies. Therefore, the assumption that the proficiency level of the subjects is similar is controlled to a great extent.

The sample size in this study is 124 MA students. The subjects are randomly selected based on cluster sampling. That is, depending on the number of MA classes at different levels in the universities in Tehran, one or two classes have been randomly selected from each university.

Instrumentation

To collect data, two paper and pencil prompts were used: an argumentative essay where the test takerswere expected to present a written argument or case to an educated reader, and (b) a letter to a friend or family member in which they were required to write about their problemsof studying in another town other than their hometown and the way they overcame the problem.

Following the guidelines suggested in the “Into Europe, the writing handbook, Tanko (no date), p. 47”, theprompts used in this study were checked against the suggestions and then administered.

The prompt which the test takers required to write ‘a letter’ was constructed by the researchers taking the above guidelines into consideration. For ‘the argumentative task’, however, qualitative data was collected. Ten a priori prompts were first selected from a pool of 400 TOFEL prompts. To make sure that the task was authentic, realistic, and plausible; the prompts were distributed among a class of twenty MA students and the most frequently selected prompt was selected and administered to the target group of students along with the letter prompt (*appendix II* presents both prompts).

Procedure

The procedure of data collection which includesevaluation criteria, scoring scales, scoring procedure, and data analyses are detailed as follow:

Inspired by Bae and Bachman (2010) a componential scoring procedure was used at eight point scoring scale (1-8). All scripts in both letter writing and argumentation were scored against eight components /traits: content, organization, grammar, cohesion, vocabulary, spelling, length and hand-writing (*Appendix III and IV* shows the band scores and discriptors)

In this study, both methods of scoring were used to control the rating methods and to compare the methods with one another.All the scripts (N=124 by 2) were first evaluated and scored analytically, by the first rater, with eight components of content, organization, cohesion, grammar, vocabulary, spelling, length, and hand writing at 8 band levels according to a scoring scale (Bae and Bachman, 2010) (see appendix III). For practical consideration (time and expenses)a sample of 32 scripts (N=32) was then selected randomly on stratified basis from among the 248 scripts of argumentation and letter (N=124 by 2) to determine the inter-rater reliability. Scripts(N=32) were then scored analytically by the second rater. To establish inter-rater reliability the result was correlated by the scores of the same scripts which were earlier scored analytically by the first rater.All of the components of argumentative writing show significant inter-rater reliability (ranging from .68 -0.83) between the two raters. The same also holds true with letter writing which shows significant inter-rater reliability (ranging from .59 to .79). See appendix IV for detailed information.

Holistic scoring method makes it possible for the rater to evaluate the overall effectiveness of a script. Here, descriptors are given for each band level, but instead of focusing on specific features of the scripts (content, organization, cohesion etc), raters evaluate general writing ability and award only one score to the script on the basis of overall impression. In analytic scoring method, on the other hand, raters evaluate a script on several evaluation criteria such as content, organization, grammar, cohesion etc. Descriptors are given for each criterion at different levels and candidates receive scores on each assessment criterion. To determine whether or not there is any significant relationship between the holistic analytic scoring, the same sample of 33 scripts were also scored holistically and the scores were correlated with each scores on each components which were scored analytically, as well as composite score of all 8 components of analytic scoring.(*see appendix V and VI*).

Data analysis

Having established the normality of the data, descriptive statistics were produced. Then a series of factor analyses were run to probe the underlying constructs of the eight components of argumentative and letter writing tasks. To respond to the first question exploratory and confirmatory analyseswere runto find the relationship between the observed and latent variables. To answer the second question, correlation analyses were run between the holistic scoring and different components of analytic scoring. The results are reported in the following section.

Conclusions and implications

The findings of this study indicate that there is an empirical justification for the selection and the number of evaluation components against which writing quality is assessed. The results also indicate that evaluation criteria in different prompts (i.e., letter, argumentation etc.) are the same, though they might not have the same order in terms of effectiveness or importance. In other words, where the most important criteria in argumentation is cohesion, content, organization, and vocabulary, letter writing can be best predicted by grammar, organization, content, vocabulary, and cohesion. The findings indicate that raters agree more or less on similar components when examining and scoring aspects of L2 writing ability, but the weight and importance they place on the components for each task is different. This will imply that different prompts may need different evaluation criteria and the best indicator of writing ability for different tasks might be different.

The results of both EFA and CFA analyses indicate that cohesion in argumentation task and grammar in letter writing task are the most important evaluation criteria. The implication of this finding is that we may not need to go through the trouble of scoring a script with a number of components. If we assess writing ability of 2nd language writers only with one criterion (grammar, or cohesion) we will probably score the scripts fairly validly. Hence, we can make fair inference based on the score assigned to the script. The implications of this finding can probably help SLA teachers to improve the quality of writing of their students by concentrating only on these components. The practical value of this finding can potentially be immense in terms of time, expenses, and manpower.

Test length was not found to have any contribution to the quality of writing in both tasks of this study. Unlike Cumming (2002), length cannot be an indicator of overall writing ability, at least for students at this level (MA) of proficiency.

The findings of Pearson Correlation of holistic and analytic scoring of this study indicate that there is a high degree of correlation between holistic and analytically (all components of both tasks as well as average score of all components). Holistic scoring has been regarded as being faster, more practical and even more valid (White, 1984, 1985). Therefore, in SLA context this method of scoring seems to produce similar results while being more practical.

Despite these findings, the results of CFA in this study, though supports the findings of EFA to a great extent, is not encouraging enough. Therefore, further research can shed more light on the finding of this research. The followings are suggestion for further research:

- Sample size where SEM is used is very important. General recommendation is that the number of sample should be 100-400 (Kunnan, 1998). However, in studies where SEM was used, larger sample produced better the results (Purpura, 1998). Similar studies can be conducted using larger sample.
- The measurement model used in this study can be re-specified according to findings of this study (the way components predict quality of writing) to investigate if it results differently.
- And, finally, handwriting has been found to influence the quality of writing in this study, and hence can be regarded as an evaluation criterion of writing. This finding is in par with the claims made by a number of scholars that hand writing is an important writing feature (Sasaki, 1999; Vaughan, 1991; Shaw and Weir, 2007). Shaw and Weir even complain about the paucity of studies examining the influence of hand writing in the assessment of 2nd language writing. Therefore, further studies to investigate the influence of handwriting in writing assessment will be warranted.

References

Alderson, J. C. (2000). *Assessing reading*. Cambridge: CUP.

Bachman, L. F. (1990). *Fundamental consideration in Language testing*. New York: OUP. Bachman, L. F. (2005). *Statistical Analysis for Language Assessment (2nd ed.)*.NY: CUP.

Bachman, L.F. & Palmer, A.S. (1996*). Language testing in practice.* Oxford: OUP.

Bae, J. & L.F. Bachman (2010). An investigation of four writing traits and two tasks across two languages. *Language testing*,27(2), 213-234.

Cohen, A. D. (1994). *Assessing Language ability in the classroom.*(2nd ed.). Boston, MA. Heinle and Heinle.

Cumming, A. (1990). Expertise in 2nd language composition. *Language testing,*7, 31-51.

Cumming, A. (2002). Assessing L2 writing: An alternative constructs and ethical dilemmas. *Assessing writing, 8*, 73-83.

Cumming, A., Kantor, R., Baba, K., Erdosy, U., Eouanzoui, K., & James, M. (2005).

Differences in written discourse in independent and integrated prototype tasks for next generation TOEFL. *Assessing Writing, 10*, 5–43.

Elder,C., Barkhuizen, B., Knoch, U., and Randow, J. V. (2007). Evaluating rater responses to an online training program for L2 writing assessment. *Language Testing.* 24 (1), 37-64.

Fulcher, G. (2003). *Testing second language speaking*. London: Pearson Longman.

Grant, L., and Ginther, L. (2000). Using computer-tagged linguistic features to describe L2

writing differences. *Journal of Second Language Writing, 9*, 123–145.

Hamp-Lyons, L. (1990). Second language writing: Assessment issues. In B. Kroll (ed.), *Second language writing: research insights for the classroom* (pp. 69-87).NY: CUP.

Hamp-Lyons, L. (2002). The scope of writing assessment. *Assessing writing 8*, 5-16.

Hughes, A. (1989). *Testing for language teachers*. Cambridge: CUP.

Hughes, A. (2003). *Testing for language teachers*. Cambridge: CUP.

Jacob, H., Zinkgraf, S., Wormuth, D., Hartfiel, V. & Hughey, J. (1981). *Testing ESL composition: A practical approach.* Rowley, MA: Newbury House.

Jafarpour, A. (1991). Cohesiveness as a basis for evaluating composition. *System, 19*, 459-465.Kane, M., Crooks, T., and Cohen, A. 1999. Validating measures of performance. *Educational measurement: Issues and practice,* 18 (2), 5-17.

Kunnan, A. J. (1998). An introduction to structural equation modelling for language assessment research. *Language testing*, 15(3), 295-332.

Lumely, T. (2002). Assessment criteria in a large-scale writing test; what do they really mean to the raters? *Language Testing,*19 (3), 246-276.

Lumely, T. (2005). *Assessing second language writing;* the raters' perspective. Frankfurt: Pter Langt.

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

Milanovic, M., Saville, N., and Shuhung, S. (1996). A study of the decision making behaviour of composition makers. In M. Millanovic and N. Saville (Eds.), *Studies in language testing 3: performance testing cognition and assessment* (pp. 92-111). Cambridge: CUP.

Purpura, J. E. (1998). Investigating the effects of strategy use and second language test performance with high- and low – ability test takers: a structural equation modelling approach. *Language testing*, 15 (3), 333-379.

Sasaki, M. (1999). Development of an analytic rating scale for Japanese L1 writing. *Language testing*, 16 (4), 457-478.

Shaw, S. D., & Weir, C. J. (2007). *Examining Writing.* Cambridge. CUP.

Schonon,R. (2005). Generalizability of writing scores: an application of structural equating modelling. *Language testing*, 22(1) 1-30.

Tanko (no date). *Into Europe, the writing handbook.* Retrieved on July 2011: http://www.lancaster.ac.uk/fass/projects/examreform/into_europe/writing.pdf

Vaughan, C. (1991). Holistic assessment. What goes on in the rater's mind? In L. Hamp Lyons (Ed.). *Assessing second language writing in academic contexts* (pp.111-125). Norwood, New Jersey: Ablex Publishing Corporation.

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

White, E.M. (1984). *Holisticism*. College Composition and Communication 35(4), 400-409.

White, E. M. (1985). *Teaching and assessing writing*. Sanfrancisco, CA: Jossey-Bass.

White, E. M. (1995). An apologia for the timed impromptu essay test. *College composition and Communication,46*, 30-45.

Wolfe-Quintero, K., Inagaki, S., & Kim, H-Y. (1998).*Second language development in writing: Measures of fluency,* *accuracy and complexity*. Honolulu, HI: University of Hawaii at Manoa.

Appendices

Appendix I

1. Content, Quality of Content, Task fulfillment, Task realization, development of Idea, data description, substantive content
2. Organization, Coherence, fluency, rhetorical Organization, Use of detail, support
3. Cohesion
4. Grammar, Sentence structure, Language use, linguistic Accuracy, tense and verb problems, accuracy of meaning
5. Vocabulary, Syntactic variety, appropriacy of word choice,. Morphology, word form, register, appropriate meaning
6. Spelling, mechanics, punctuation and capitalization, appropriateness of the convention
7. Length, essay length
8. Hand-writing and neatness

Appendix ll

Borrowing the idea from Bae and Bachman (2010), the researcher constructed the following generic scale of ability

1_____2_____3_____4_____ 5_____6_____7_____8

(1) Very limited to evaluate, (2), Limited, (3) not sufficient at all, (4) somewhat insufficient, (5) good, (6) good but not very good, (7) almost perfect, (8) perfect for 2nd language learner at this level

Appendix III

Scoring guides: including band levels and descriptors

1. Very limited to evaluate.

2. Not thorough at all (only 15–30% of the content was expressed). Serious distortion of the picture content/task, or large segments of the content missing. Does not communicate. no organization, or not enough to evaluate. Frequent critical errors. Extensive minor errors. Few sentences. A sample with length <100/<70words (for argumentation or letter) is considered limited unless the writing contains complex grammatical features. Essentially translation. Little knowledge of English Vocabulary, idioms, word form, or not enough to evaluate. No mastery of conventions, dominated by errors of spelling, punctuation, capitalization, paragraphing. Words are comprehensible if the reader makes an effort to comprehend the incorrectly written words.

3. Insufficient and irrelevant and inaccurate content. Traces of organization can be found in the study; i.e. it has somewhat a topic, or it has somewhat a few supports. Has one or two occurrence of complex or compound sentence structure. Limited range of vocabulary with frequent errors in usage, idioms, forms, with confused meaning. Somewhat disconnected and choppy. Frequent spelling, punctuation, capitalization, paragraphing errors with obscured meaning. Words eligible but handwriting is not totally looking pleasant. 135-50 words / 78-100 words

4. Somewhat insufficient in content. Somewhat irrelevant/inaccurate/not thorough as a whole or locally. Or a couple of sentences per major scene with mere (literal) descriptions. Not fluent. Ideas confused or disconnected. Lacks logical sequencing and development. Some critical errors. Frequent minor errors. +0.5 if the writing sample is long limited range. Frequent errors or words/idiom forms, choice, usage. Meaning confused or obscured. Frequent errors of spelling, punctuation, capitalization, paragraphing. Meaning obscured/ confused. Words legible with frequent minor errors (e.g. 4-5 errors) with or without occasional critical errors. Long as many as 130-150 words of argumentation or 80-100 words for letter

5. Good but the writer needs to polish his/her language add more details etc. Somewhat choppy, loosely organized, main idea does not stand out, limited support, and not very logical. Few critical errors, occurrence of 3-5 complex and compound patterns. Good connections but need to be polished. Adequate range of vocabulary some errors in form, choice and usage. Words completely legible with 3-4 minor errors. 175-200 / 111-125 words.

6. The argument/letter is complete and thorough in general. Accurate/relevant in general. In general, *fine*, but elaboration and sophistication not observed. Descriptions good (literal) but not *impressive*. Or, descriptions somewhat insufficient; however, some impressive, relevant elaboration observed locally. Somewhat choppy,. Loosely organized but maim ideas stand out. Limited support. Logical but incomplete sequencing. Complex/compound connection observed but some critical errors, or less than *N* (7) complex/compound sentences but no errors No or few critical errors. Occasional (1–2) minor errors with a few occurrences of complex or compound sentence patterns. Adequate range of vocab. Occasional errors of word/idiom form, choice, usage. But meaning not obscured. Occasional errors of spelling, punctuation, capitalization, paragraphing. Meaning is not obscured or confused. Completely legible but not that beautiful with few errors. Length is around 201-225 / 126- 140 words.

7. The story/letter is complete and thorough in general. Accurate/relevant in general. In general fine but elaboration is needed. Sophistication not observed in all parts of the content. Not very impressive, though description is good. Fluent expression. 1 irrelevant ideas. Not satisfactorily succinct. Complex and compound connection observed but some error. Fully connected and fluent. Not sophisticated range nor errors of words, idioms, choice, usage. Demonstrating mastery of conventions. Very few spelling errors, capitalization, punctuation, paragraphing. 226-240 words / 126-140 words.

8. An Essay at this level has wonderful descriptions of the situations/events. Very thorough. No irrelevance whatsoever. Creative. Persuasive. Convincing. Impressive. Fluent expression. Ideas clearly stated or supported. Well organized. Logical sequencing. Excellent flow of language with excellent cohesion. Complete control of grammar (Native like). A variety of grammatical use with good number of complex and compound sentences. Sophisticated range of words. Effective word/idiom choice and usage. Word form mastery. Appropriate register. Demonstrate mastery of conventions. Few errors of spelling, punctuation, capitalization, paragraphing. Completely legible and beautiful by the look and adequate length (250 words argumentation or 150 words letter)

Assessment criteria in this research are: content, organization, grammar, vocabulary, cohesion, spelling, length and handwriting. The followings are statements about the criteria which define the levels of assessment criteria

Content: Content refers to substantive ideas. Freedman (1979) explains content as “development and logical consistency between the ides”, p. 161. Content according to Bae, and Bachman (2010) refers to the relevance of written text to a given task, as well as thoroughness, persuasiveness, impressiveness, and creativity of ideas consistent with the task expectation.

Organization: Organization in this study refers to the overall shape of the composition and the internal pattern – supported argument.

Summary of components used in most research project

Evaluation criteria	1.Content	2.Organization	3.cohesion	4.Grammar	5. Vocabulary	6. Spelling	7. length	8. Hand Writing
DELNA	xx	xxx		xx	x	x		
IELTS	x	x	x	x	x			
TOFEL	x	x		x	x			
Jacob et al.	x	x		x	x	x		
Vaughan	x	x		x	x	x		x
Milanovich	xxx			xx	x	xx	x	x
Cohen	x	x		xx	xx			
Cumming	x	x		x	x			
Schoonen	x	x		x	x			
Elder et al.	x	x		x	x	x		

Table 1: Pearson Correlation Holistic Scoring with Components of Argumentative Writing

		HOLARG
ARGHW	Pearson Correlation	.216
	Sig. (2-tailed)	.227
	N	33
ARGLEN	Pearson Correlation	.320
	Sig. (2-tailed)	.069
	N	33
ARGSP	Pearson Correlation	.542**
	Sig. (2-tailed)	.001
	N	33
ARGCOH	Pearson Correlation	.708**
	Sig. (2-tailed)	.000
	N	33
ARGORG	Pearson Correlation	.685**
	Sig. (2-tailed)	.000
	N	33
ARGGRAM	Pearson Correlation	.462**
	Sig. (2-tailed)	.007
	N	33
ARGVOC	Pearson Correlation	.368*
	Sig. (2-tailed)	.035
	N	33
ARGCONT	Pearson Correlation	.682**
	Sig. (2-tailed)	.000
	N	33
**. Correlation is significant at the 0.01 level (2-tailed).		

Table 2: Pearson Correlation Holistic Scoring with Components of Letter Writing

		HOLLET
LETHW	Pearson Correlation	.136
	Sig. (2-tailed)	.451
	N	33
LETLEN	Pearson Correlation	.446**
	Sig. (2-tailed)	.009
	N	33
LETSP	Pearson Correlation	.338
	Sig. (2-tailed)	.054
	N	33
LETCOH	Pearson Correlation	.569**
	Sig. (2-tailed)	.001
	N	33
LETORG	Pearson Correlation	.696**
	Sig. (2-tailed)	.000
	N	33
LETGRAM	Pearson Correlation	.565**
	Sig. (2-tailed)	.001
	N	33
LETVOC	Pearson Correlation	.641**
	Sig. (2-tailed)	.000
	N	33
LETCONT	Pearson Correlation	.747**
	Sig. (2-tailed)	.000
	N	33
**. Correlation is significant at the 0.01 level (2-tailed).		

It refers to the logical paragraphing, topic sentence, and reader engagement, according to ELTA EU ORG's resources.

Grammar: Grammar refers complexity and variety of rules used for writing to form phrases, and sentences. The range of structure used by the writer the type of errors committed (minor, somewhat serious and serious).

Cohesion: Cohesion is connectedness of the sentences and paragraphs and the proper use of cohesive ties - overall flow of the answer. Inspired by Tanko (no date) in "the handbook of writing", the term cohesion is operationalized as:

> … grammatical and lexical relationship between the elements of the text, for example reference realized by personal/possessive pronouns, demonstratives, and comparatives; substitutions /ellipsis (the replacement of an item by a different one or by nothing); conjunction realized by conjunctions, adverbial connections (*and, or, firstly, secondly, moreover, therefore, in conclusion*); or lexical repetition (same word, synonym/near synonym, general word) p. 304-6

Vocabulary: Refers to the accurate and appropriate use of vocabulary. It also refers to the degree of sophistication of lexical features; limited, adequate or sophisticated use of lexical features.

Spelling: Spelling in this study is defined as the ability to spell individual letters of a word correctly in terms of form and order (Collings Cobuild, 1996). Spelling is determined by the recognizability of letters in words and the type (critical or minor) and number of errors in spelling. Repeated errors (recieve, and recieve) is regarded as one error. In this study, mechanics (punctuation and capitalization) is included in spelling.

Length: Text length is selected in this study because it is believed that test length is an indication of overall writing ability according to many studies. The operational definition of text length in this study is the total number of words written for the task within the time given for each task (i.e. 250 words)

Handwriting: In this study, handwriting, as a divisible skill is operationalized as "tidy" and "legible" handwriting with adequate and "consistent spaces" between the words which renders to legibility and comprehensibility (Pollock et. al., 2009).

Appendix V

Are there any significant relationships between the components of argumentative writing and its holistic score?

As displayed in the following Table the holistic score shows significant correlations with six components of the argumentative writing as follows;

Spelling (R = .54, P = .001 < .05),
Cohesion (R = .70, P = .000 < .05),
Organization (R = .68, P = .000 < .05),
Grammar (R = .46, P = .007 < .05),
Vocabulary (R = .36, P = .036 < .05),
Content (R = .68, P = .000 < .05).

However it does not show any significant relationships with;
Hand-writing (R = .21, P = .227 > .05),
Length (R = .32, P = .069 > .05).

Are there any significant relationships between the components of argumentative writing and its holistic score?

As displayed in the following Table the holistic score shows significant correlations with six components of the argumentative writing as follows;

Length (R = .44, P = .009 < .05),
Cohesion (R = .56, P = .001 < .05),
Organization (R = .69, P = .000 < .05),
Grammar (R = .56, P = .001 < .05),
Vocabulary (R = .64, P = .000 < .05),
Content (R = .74, P = .000 < .05).

However it does not show any significant relationships with;

Hand-writing (R = .13, P = .451 > .05),
Spelling (R = .33, P = .054 > .05).