FACTORIAL VALIDITY AND INVARIANCE OF THE MUET ESSAY WRITING RATING SCALE: EMPIRICAL AND THEORETICAL CORRESPONDENCE

[1]Noor Lide Abu Kassim, Kamal Jamil Inbrahim Badrasawi, Mohd Sahari Nordin, Ainol Madziah Zubairi & Ratnawati Mohd Ashraf

Kulliyyah of Education, International Islamic University Malaysia, Malaysia

[1]Corresponding author: noorlide@iium.edu.my

Received: 2 January 2018 **Accepted:** 28 June 2018

ABSTRACT

Purpose – The validity of any performance assessment depends substantially on the rating scale used in the assessment procedure. It delineates the theoretical construct being measured and influences how performances are judged and interpreted. This study examined the factorial validity and invariance of the rating scale used for the assessment of the essay writing component of the Malaysian University English Test (MUET) in terms of the correspondence to empirical data and the theoretical construct of ESL writing.

Methodology – To determine the factorial validity of the MUET essay writing rating scale, a measurement model was developed based on the criteria description of the scale. Model-data fit was tested using CFA. The data comprised 392 essays, taken from a university English proficiency examination. Factorial invariance was tested through consecutively more restrictive models.

Results – A 3-factor structure model was found to produce the best fit to the data. Factorial invariance was tested by examining the comparability of the structure and values of parameters within the model across two groups. A more restrictive model was subsequently tested, where variances and covariances of the latent constructs together with factor loadings were constrained to be equal for the two groups. The resultant goodness-of-fit indices showed a better

fit to the data, providing support for the factorial invariance of the 3-factor structure model.

Significance – The results support a 3-factor structure model, congruent with the theory of ESL writing. However, further validation needs to be carried out across different ESL populations to ensure accurate representation and assessment of ESL writing performance.

Keywords: ESL writing; factorial validity; invariance; CFA; rating scale validation.

INTRODUCTION

Central to the validity of performance assessment is the rating scale used in the assessment procedure (McNamara, 1996). It delineates the operational definition of the theoretical construct that is being measured, thus exerting considerable influence on the way performances are judged (Weigle, 1994) and interpreted (Messick, 1995). Implications of the use of rating scales are therefore far too important to be taken lightly, particularly in high-stakes standardized testing. Hence, rating scales should be subjected to rigorous empirical validation to ascertain that they indeed reflect the theoretical construct they have been designed to measure (Bond & Fox, 2015; Cumming & Mellow, 1997).

As factorial validity relates to the correspondence between the structure of a measure or construct and its theoretical definition (Hoyle & Smith, 1994; Messick, 1995), its evidence is essential in establishing the validity of inferences from test scores (Messick, 1980, 1995). It is only when factorial validity is clearly demonstrated that the linear combinations of the indicators to form composite scores of latent variables, i.e., the theoretical construct, can be fully justified and the construct validity of inferences from test scores determined (Messick, 1980, 1995).

Another important concern in establishing the construct validity of inferences from scores on an instrument pertains to the factorial invariance of the instrument. If factorial invariance does not hold, differences in group performances are not reflective of true differences

in the measured latent construct but instead are attributable to variability in the measurement properties (Hoyle & Smith, 1994). Hence, it is critical that the factorial structure of the rating scale that is used and the values of its parameters be demonstrated to be comparable across subgroups before the interpretation of scores for different groups can have the same meaning.

In Malaysia, the level of students' English proficiency is an entry requirement into higher institutions, irrespective of whether English is the main medium of instruction or not. The entry English language requirement for international students into the higher institutions are usually based on the TOEFL (The Test of English as a Foreign Language) and/or IELTS (International English Language Testing System), while local or Malaysian students entry are based on the localized Malaysian English Test (MUET). According to the Malaysian Examinations Council (2006, 2011, 2015), MUET is designed to measure the English language proficiency of pre-university students for entry into tertiary education. MUET comprises all the four language skills: Listening (45 scores), Speaking (45 scores), Reading (120 scores) and Writing (90 scores), with an aggregated score range of zero to 300. The scores correlate with a banding system ranging from Band 1 (the lowest) to Band 6 (the highest). The writing skill is seen as the second important skill for academic purposes as it carries 30% of the total MUET score. The writing section has two parts: Report writing (40 marks) and Academic essay writing (50 marks). In the Report writing, a non-linear graph is presented and test takers are required to write an essay which has an introduction, a body and a conclusion based on the given prompt. On the other hand, the Academic essay requires test takers to write about their opinions and views on a given issue. The rating scale used for assessing the Academic essay comprises two components. Hence, the two latent variables underlying the rating scale are: (1) Task fulfillment (indicated by knowledge/ understanding of topic, ideas development and ideas presentation) and (2) Language (indicated by organization of ideas, linking of ideas 'i.e. coherence', complexity of sentence structure, language accuracy, effective sentence structures 'i.e. syntactic', appropriateness of vocabulary [idiom and word], and spelling).

Given that the MUET is a high stakes test and the writing component contributes 30% of the total score, it is important that the factorial

validity and invariance of its writing rating scale be rigorously examined. To test for factorial validity, a measurement model was developed based on the construct definition (criteria description) of the MUET essay rating scale and tested using AMOS, a data-fitting programme. Factorial invariance was determined through multiple group analysis and testing of baseline, metric, and factor variance/covariance models. As this study aimed to provide empirical evidence for the valid use of the MUET essay writing rating scale, the research questions that guided the intent of this study are as follows:

1. To what extent does the 2-factor structure model of the MUET essay rating scale correspond to empirical data?
2. To what extent does the best-fit model correspond to the theoretical construct of ESL writing?
3. To what extent is the best-fit model invariant across samples drawn from the same population?

LITERATURE REVIEW

In L1 writing, the development of analytic scales for assessing writing ability was pioneered by Diederich, French and Carlton in 1961 (see Sasaki & Hirose, 1999). In their study, Diederich et al. factor-analyzed 35 remarks given by English L1 readers on 300 compositions written by college students. From the analysis, they identified 5 major traits or dimensions of writing that raters focused on. These were identified as: Ideas, Form, Flavour, Mechanics, and Wordings. It was from this study that Diederich et al. constructed the first analytic scale for writing in English as L1 (Sasaki & Hirose, 1999). Though the resultant scale was criticized for being highly unreliable, and for focusing on aspects that were either too mechanical or difficult to quantify (see Hamp-Lyons & Henning, 1991), this study is of special significance because of its methodological approach in developing analytic scales (Sasaki & Hirose, 1999).

Another notable study on the development of L1 analytic writing rating scale was conducted by the International Association for the Evaluation of Educational Achievement (IEA) (see Hamp-Lyons & Henning, 1991; Sasaki & Hirose, 1999). The aim of the study was to develop and validate a common scoring scheme to evaluate

high school students' compositions across a number of writing tasks and across 14 different countries. Based on the review of related literature, comments given by the international reading team, and the results of several pilot studies, IEA developed a scoring scheme consisting of seven major traits or constructs that was felt to describe L1 writing – quality and scope of content, organization and presentation of content, style and tone, lexical and grammatical features, spelling and orthographic conventions, handwriting and neatness, and response of the rater. This study is also of particular significance as it represented "a modern sense of construct validation" (Sasaki & Hirose, 1999, p. 459), where the intended construct to be assessed or measured was first defined *a priori* and then validated *a posteriori* by collecting empirical evidence to support the adequacy and appropriateness of inferences based on performances elicited.

Sasaki and Hirose (1999) developed an analytic rating scale for the assessment of Japanese university students' expository writing. Information on the criteria that Japanese L1 teachers considered essential when evaluating expository writing was collected using a questionnaire survey. With the assistance of two Japanese L1 writing experts, 35 criteria of Japanese expository writing were outlined. Expert judgment was again used to categorize the descriptions into five major areas: expression, organization, content, appeal to the readers and social awareness. These descriptions were then piloted and reviewed several times before the final questionnaire was sent out to High-school Japanese teachers.

The construct validity of the rating scale was subsequently verified by examining (a) the relationship between the newly-developed analytic scale and a holistic scoring profile through regression analysis, and (b) the relationship between the newly developed scale and an existing scale by means of correlation coefficient and mean score. Factorial validity of the scale in terms of the factor-structure of the scale, however, was not investigated empirically. Division of the constructs and organization of the corresponding sub-constructs were done entirely through the use of expert judgment and their understanding of L1 writing.

In ESL writing, the empirical development of analytic rating scales and investigations of their construct validity have also been given attention. One of the most widely known ESL writing scale is the

ESL Composition Profile developed by Jacobs, Zinkgraf, Wormuth, Hartfiel and Hughey (1981). In this analytic scale, ESL writing ability is represented as comprising five dimensions or traits: content, organization, vocabulary, language use, and mechanics. This scale was validated by collecting evidence of reliability which included (1) reader reliability or interrater agreement (2) standard error of measurement (3) internal consistency (4) reliability of gains score, and (5) score reliability. Evidence of validity, on the other hand, was demonstrated through (1) face validity (2) content validity (3) concurrent validity, and (4) predictive validity. Construct validity was demonstrated by investigating gains in the construct (i.e., composition ability) and by comparing the mean score on pre- and post-tests of two cohorts: undergraduate and graduate students.

In another study, Brown and Bailey (1984) validated a rating scale developed by a team of ESL teachers at UCLA for scoring compositions written by students in an upper-intermediate ESL class. Aspects that were considered in the validation of the scale include: a) consistency of scoring across raters, b) identification of possible sources of error, and c) determination of raters' reaction to the scoring scheme. Findings of this study suggested that the rating scale was useful in measuring second language writing ability. The authors, however, noted that further investigations needed to be carried out to evaluate the validity of the instrument from the content and construct point of view.

Alongside investigations of ESL writing rating scales is the empirical validation of the theoretical constructs and sub-constructs of ESL writing. Cumming (1990) through the use of multivariate analyses, found that the processes and products of writing in a second language are characterized by two distinct but interrelated factors: writing expertise and second language proficiency. Writing expertise pertains to processes that are central in the mental activities and decision-making processes of the writer in producing and organizing the content appropriate to the writing task (Cumming, 1990). Second language proficiency, on the other hand, is considered as "an additive factor" (Cumming, 1990) and defined as control over the linguistic elements of a second language (Cumming, 1990).

In a subsequent study, which investigated whether novice and experienced raters implicitly distinguished students' writing expertise and second language proficiency, Cumming (1990) found

that both groups of raters did distinguish these two aspects in their ratings of students' compositions. The MANOVA results indicated no interaction effects between these two factors thus providing further empirical support that writing expertise and second language proficiency are separate distinct factors. This conception of second language writing is supported by other researchers (see Krapels, 1990) and has been widely accepted and embodied in many analytic ESL writing rating scales (Weigle, 1994). Another important study, which has helped to elucidate the relationship between constructs of writing ability, is by Roid (1994). Using an extensive series of cluster analyses, he found construct-related evidence in support of the multidimensional model of L2 writing ability, thus the use of analytic scales.

In recent years, rating scale validation has also taken a different path through the use of Rasch analysis. Curtis and Boman (2007) maintained that the application of the Rasch Measurement Model has made researchers and developers of survey and assessment instruments reconsider the construction of these instruments. This is because the Rasch Measurement Model provides "an alternative framework for understanding measurement and alternative strategies for judging the quality of a measuring instrument" (Kimberlin & Winterstein, 2008, p.2281). Good examples are the studies on the validation of rating scales which focus on rater judging behaviour. These studies looked at raters' interpretation and application of particular rating scales to particular tasks, rating criteria and candidates (e.g., Kondo-Brown, 2002; Wigglesworth, 1993); aspects related to raters' judging behaviour (e.g., Brown, 1995; Kobayashi & Rinnert, 1996; Schoonen, Vergeer & Eiting, 1997); and the effect of training on judges' rating (e.g., Lumley & McNamara, 1995; Weigle, 1994; Wigglesworth, 1993).

Though Rasch analysis has made important contributions to instrument/ scale validation, this empirical study turned to Confirmatory Factor Analysis (CFA) as it is the most appropriate tool to examine (1) the correspondence of the factorial structure of the MUET essay writing rating scale to the theoretical construct of ESL writing, and (2) for testing model-data fit and invariance. All confirmatory and invariance analyses were conducted using AMOS 4.0 (Arbuckle, & Wothke, 1999). These techniques helped provide empirical evidence for the factorial validation of the rating scale in

question. This research draws attention to the construct validity of rating scales to ensure that the interpretation of test scores is consistent with the theoretical views of the ability or construct tested.

METHODOLOGY

Data and Data Collection Method

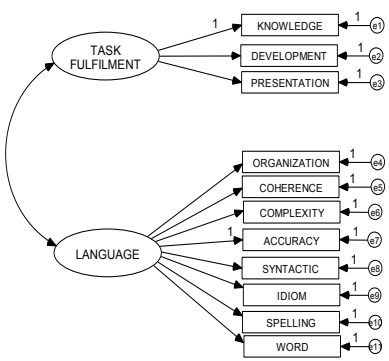
Data for analysis were sourced from a writing test of a university placement examination. From a total of 400 essays that were randomly selected, 392 were found suitable, and rated by an experienced ESL writing instructor. The other 8 essays had to be discarded as they were incomplete and therefore would not give an accurate representation of actual writing ability. The essays were scored using a 7-point scale which was developed based on the construct definition of ESL writing, embodied in the MUET essay writing rating scale. At the initial stage of the rating process, several essays were randomly selected and rated by another experienced writing instructor. Ratings from the two instructors were discussed, together with the researcher, to ensure scoring consistency and accurate interpretation of the underlying construct. Final ratings for all the essays were later entered onto the SPSS data editor and prepared for data analysis. Following that, the data were randomly split into two sets to allow for testing of factorial invariance.

Data Analysis Procedure

A measurement model was developed based on the criteria description of the MUET essay writing rating scale, and tested using CFA. Factorial validity of the model was established by testing the fit of the measurement model to the data. The following goodness-of-fit indices were used in determining model-data fit: the chi-square statistic (c^2), Bentler's comparative fit index (CFI), the chi-square to degree of freedom ratio (c^2/df), the Tucker Lewis Index (TLI), and the root mean square of approximation (RMSEA) (see Kline, 1998 for discussion of fit statistics). CFI and TLI values of > .90 indicate good fit; values close to 1.0 are therefore desirable. CFI and TLI may sometimes exceed 1.0 (Kline 1998). RMSEA values of < .05 are indicative of good fit. Values ³ 0.05 and £ 0.08 demonstrate acceptable fit; whereas, values ³ 0.08 and £ 0.10 indicate mediocre fit.

As the hypothesized measurement model produced unsatisfactory model-data fit, it was re-specified and re-tested. Re-specification of the model was guided by theoretical considerations and relevant fit indices. Factorial invariance, on the other hand, was examined through a series of more restrictive models using a multiple group analysis (see Kline, 1998; Hair et al., 1998). Fit of the invariance models was determined primarily using the change in chi square per change in degrees of freedom between the models. Change in CFI was also used to determine differences in model fit in the invariance analyses as it has been shown to be appropriate (see Cheung & Rensvold, 2002). A change of 0.01 in CFI is indicative of a substantial difference in models.

Model Specification



Labels and Descriptions of Latent and Observed Variables:

Latent Variable: *Task Fulfilment*
Observed variables/Indicators

- Knowledge/ Understanding of Topic (KNOWLEDGE)
- Development of Ideas (DEVELOPMENT)
- Presentation of Ideas (PRESENTATION)

Latent Variable: *Language*
Observed variables/Indicators

- Organization of Ideas/ Paragraphing (ORGANIZATION)
- Linking of Ideas & Coherence (COHERENCE)
- Complexity of Sentence Structures (COMPLEXITY)
- Language Accuracy (ACCURACY)
- Effective Sentence Structures (SYNTACTIC)
- Appropriacy of Vocabulary (IDIOM & WORD)
- Spelling (SPELLING)

Figure 1. The 2-Factor Measurement Model of the MUET Essay Writing Rating Scale.

The MUET essay writing rating scale measures two main components or constructs of ESL writing: task fulfillment and language. Task fulfillment includes aspects related to knowledge or understanding of topic, development of ideas, and presentation of ideas. Language, on the other hand, assesses the appropriate use of the English language. It includes language accuracy, spelling, effective use of sentence structure, complexity of sentence structures, appropriacy of vocabulary & idioms, linking of ideas & coherence, organization of ideas and paragraphing. Based on this construct definition of ESL writing, a 2-factor measurement model was constructed (Figure 1).

Model Identification

Before a model can be tested, it must be ensured that it does not have any identification problems (Hair et al., 1998; Tabachnick & Fidell, 2001). For a model to be identified, *"a unique numerical solution for each of the parameters in the model"* must be achieved (Tabachnick & Fidell, 2001, p. 691). The constructed model was over identified with 43 degrees of freedom.

Assumptions

Evaluation of assumptions for CFA was carried out using SPSS for Windows version 12.0.

- *Assumptions of Normality and Linearity*

  Two of the observed variables had skewness of below -1 whereas three variables had kurtosis greater than 1 but below 2. To assess linearity, randomly selected pairs of variables were examined using scatterplots. They were found to be linearly related to varying degrees. Multivariate normality was also met with a critical ratio of 1.445. The Mahalanobis Distance also indicated absence of outlying values.

- *Multicollinearity and Singularity*

  Inter-item correlations of the variables indicate that most of the variables had correlations of .8 and below. Only a few variables had inter-item correlations of above .8. Generally SEM programs abort and provide warning messages if the covariance matrix is nonsingular. As convergence was achieved in all the analyses conducted, the covariance matrix was assumed to be nonsingular and free from multicollinearity.

Model Estimation

Maximum likelihood estimation (MLE) was employed to estimate the hypothesized and re-specified models. The maximum likelihood estimation method (MLE) was used for three reasons. First, MLE "*makes estimates based on maximizing the probability (likelihood)*

that the observed covariances are drawn from a population assumed to be the same as that reflected in the coefficient estimates. That is, MLE picks estimates which have the greatest chance of reproducing the observed data" (Garson, 2002, p. 6). Second, it is the most commonly used method, and thirdly, it can be used with sample sizes of between 100 and 200 (Hair et al., 1998).

RESULTS

Two-Factor Structure Model Not Supported

The CFA result indicates that the 2-factor structure of the MUET writing rating scale lacks empirical support. The large chi square value (χ^2 = 169.309) and chi square to df ratio (χ^2/df = 3.937), along with the root mean square approximation value of above 0.10 (RMSEA = .127) indicates poor model-data fit. TLI of just above .90 provides further evidence of poor fit (Figure 2). The model was, therefore, re-specified and re-tested.

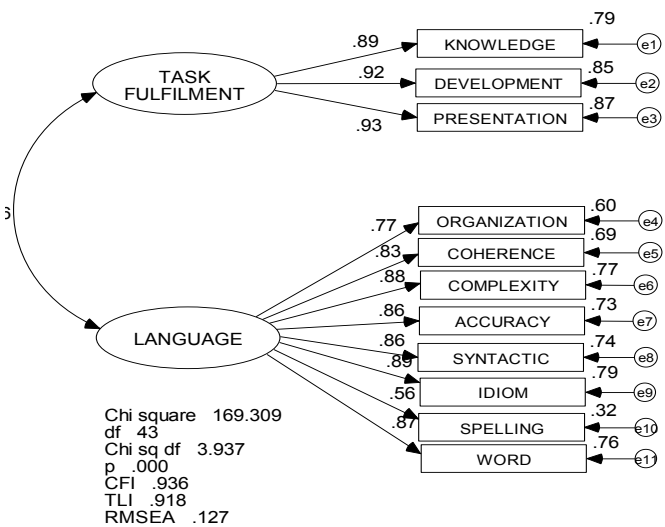


Figure 2. Standardized Parameter Estimates and Goodness-of-Fit Statistics of the 2-Factor Structure Model

Re-specified Model A

Literature on ESL writing has identified two distinct but interrelated constructs: writing expertise and language. Writing expertise has been argued and shown to encompass aspects related to content and organization, whereas second language proficiency relates to linguistic elements (Cumming, 1990). In the MUET essay writing scale, organization and coherence are postulated as sub-constructs of language, not task fulfilment, which pertains to content-related aspects of writing proficiency. Therefore, in this re-specified model (Model A), the observed variables, *ORGANIZATION* and *COHERENCE* (which has been theorized and empirically shown to be related to writing expertise) were placed under *'Task Fulfilment'*, together with elements related to content. The re-specified model yielded better goodness-of-fit indices: the chi square value is smaller ($\chi^2 = 105.631$); the chi square to df ratio improved considerably (χ^2/ df = 2.457), values for CFI and TLI were above the threshold value of .90. Despite the improvements in certain model-data fit indices, the root mean square approximation value showed mediocre fit (RMSEA = .090).

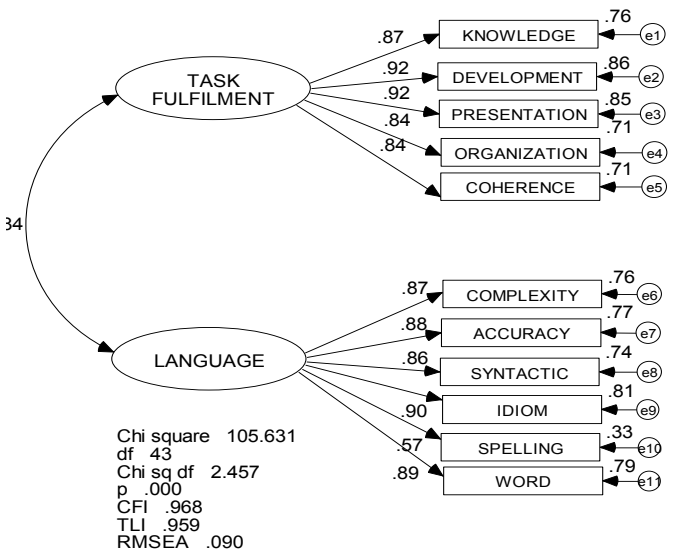


Figure 3. Standardized Parameter Estimates and Goodness-of-Fit Indices of the Re-Specified Model (Model A)

Re-specified Model B

The model was further modified based on literature on ESL writing and findings on construct validation of rating scales used in the assessment of ESL writing. The second re-specified model (Model B) comprised three factors: *Task Fulfilment* (aspects related to content), *Text Organization* (aspects related to organization of content), and *Language* (aspects to do with linguistic elements). The re-specified Model B was over identified with 41 degrees of freedom.

The CFA results indicated a much better model-data fit. The chi square to df ratio fell below 2.0 (χ^2/df = 1.813), values for CFI and TLI were above .95 and close to 1. The root mean square approximation value was below .05 (RMSEA = .046) indicative of good fit. With the exception of *SPELLING*, all factor loadings (regression weights) for this 3-factor structure model were above .70 supporting the validity of the indicators as sub-constructs of the three dimensions. Consistent with the literature on ESL writing, aspects related to writing expertise (content and organization) were found to be very highly correlated (.93) while their correlations with language were of smaller magnitude.

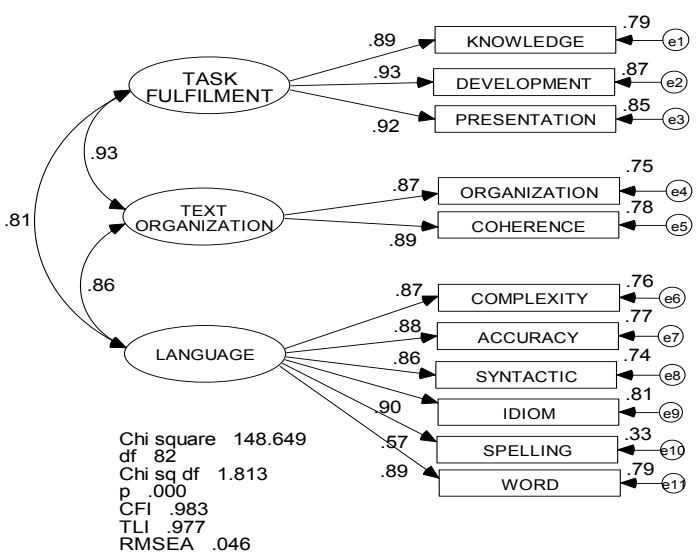


Figure 4. Standardized Parameter Estimates and Goodness-of-Fit Indices of the 3-Factor Model (Model B)

Factorial or Measurement Invariance

Tests of factorial invariance are generally conducted using a multiple group analysis which involves comparing nested models (Jöreskog, 1971). The goodness-of-fit indices and chi-square statistics of increasingly restrictive factor models (where more constraints are subsequently imposed on model parameters) are compared against a baseline model. If the restricted models produce better fit to the data and yield a non-significant chi square change, it can be concluded that the more restrictive models fits better and demonstrate invariance. If not, the baseline model is better and invariance is not supported.

In this study, three levels of invariance (each more restrictive) were tested: configural, metric, and factor variance/covariance. Configural invariance signifies that the same factor structure is applicable to groups tested. Metric invariance means that a particular indicator or sub-construct has the same scaling unit across groups. This is necessary for substantive comparisons between groups on the latent construct or variable. Factor variances, on the other hand, determines whether groups are using the same range of the construct continuum, and factor covariances test whether the same associations between the latent constructs/variables are supported across groups.

Configural invariance was first tested in the baseline model via multiple group analysis. No equality constraints were imposed, and parameters were estimated separately for the two population groups. The results of the multiple group analysis showed acceptable fit; thus configural invariance was deemed supported. To test for metric invariance, a constrained (restricted) model was tested where equality constraints were imposed on the factor loadings of the two groups (Figure 5). The fit indices for this model demonstrate adequate model-data fit but a significant change in chi square, $^{\Delta}\chi^{2}_{.05,\,8} > 15.507$ (Table 1), indicating that metric invariance may not hold.

As full measurement or metric invariance is rarely found in empirical research (Romhild, 2008), partial metric invariance of the 3-factor model was examined. In this subsequent analysis, a new model is run but without equality constraints. Critical ratios for differences in parameter estimates were examined. Z-scores that exceed the critical value of z for $p < .05$ ($\pm$ 1.96) indicate that parameters are

significantly different from each other. From Table 2, it is evident that all parameters, with the exception of PRESENTATION, were not significantly different across the baseline and constrained model. Partial measurement invariance of the 3-factor structure is therefore maintained.

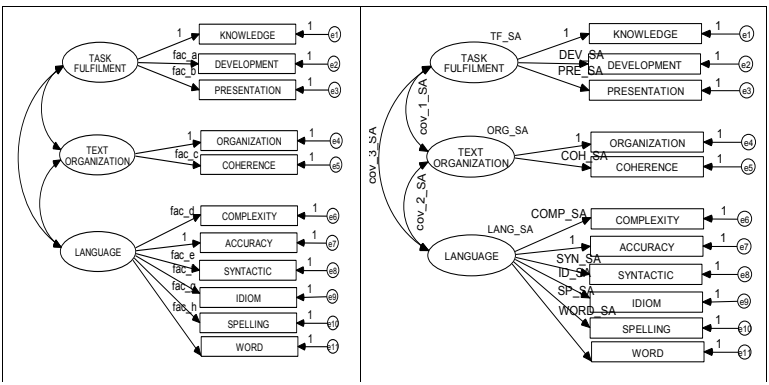


Figure 5. The Constrained Model with Equality Constraints Imposed on Factor Loadings

Table 1

Comparison of Baseline and Constrained Models to Determine Metric Invariance

	χ^2	df	χ^2/df	CFI	TLI	RMSEA	*P*
Unrestricted Model	148.649	82	1.813	.983	.977	.046	.000
Constrained Model (i.e., Metric Model)	168.942	90	1.87	.980	.975	.048	.000
$^{\Delta}\chi^2_{p,df}$ (χ^2 Difference)	20.293* (C.R. 15.507)	8					

* $p < 0.05$

Table 2

Critical Ratios for Differences between Parameters across Models

Parameter/factor loading	Critical Ratios
KNOWLEDGE	-
DEVELOPMENT	0.324
PRESENTATION	-2.998*
ORGANIZATION	-
COHERENCE	-1.698
COMPLEXITY	0.907
ACCURACY	-
SYNTACTIC	0.304
IDIOM	1.379
SPELLING	-0.581
WORD	0.449

Given the adequate fit of the metric model, and the non-significant change in CFI (CFI < 0.01), the metric model was re-tested with the factor loading for *PRESENTATION* unconstrained. A factor variance/ covariance model, which is more restrictive model, was also tested. For this analysis, factor loadings (except *PRESENTATION*) and factor variances/ covariances were set equal between the two groups (Figure 6). The results of analysis showed that the factor variance/ covariance model produced the best fit to the data (Table 3) with no significant $^{\Delta}\chi^2$ and CFI.

Table 3

χ^2 Difference Test for the Restrictive Models (with PRESENTATION Unconstrained)

Model	χ^2(df)	χ^2/df	CFI	TLI	RMSEA	*p*	D $c^2_{p, df}$
Baseline Model (Unrestricted)	148.649(82)	1.813	.983	.977	.046	.000	6.209 (D $c^2_{.05, 7}$ = 14.067)
Metric Invariance Model	154.858(89)	1.877	.983	.979	.044	.000	
Factor variance/ covariance Model	164.689(95)	1.734	.982	.979	.044	.000	16.04 (D $c^2_{.05, 13}$ = 22.362)

DISCUSSION

This study sought to provide empirical evidence on the factorial validity of an essay writing rating scale used for the assessment of writing ability in a national-level standardized test. Findings of this study suggest that the rating scale in question has some shortcomings in terms of its factorial structure. Not only is the 2-factor structure model underlying the MUET essay writing rating scale not supported by empirical data, it is also incongruent with the theoretical definition of ESL writing construct. On the contrary, the factorial validity of the 3-factor structure model is well-supported as it shows good model-data fit. All of the factor loadings of the individual indicators/observed variables except 'SPELLING' exceed the threshold value, thus explaining more than 50% of the variance in the observed variables. The use of these indicators of ESL writing therefore is justified. The multiple group CFA establishes the factorial validity of the measurement model in each group. It suggests that the constructs (latent variables) are the same in each group and allows for the substantive theoretical comparisons between groups.

This study has provided strong empirical evidence for the improvement of the rating scale used for the assessment of the MUET essay writing component. However, further validation should be carried out across different ESL populations to garner more evidence for its factorial invariance. To date, similar research involving writing assessment is still lacking and is mainly conducted by large language testing agencies, such as, the Educational Testing Service (ETS). As construct validity is considered to be the overarching validity, research of this kind must be given due attention. This is to ensure that the interpretation of test scores is consistent with the theoretical views of the ability tested (Bachman, 1990).

Finally, in the ESL context, writing is a very challenging task for ESL learners due to the complexity of factors and tasks involved in writing. For instance, writing requires both cognitive ability and skilled manipulation of linguistic structures to produce concise and coherent discourse. In other words, the ESL writer must have the necessary cognitive ability to formulate his or her thoughts and ideas logically, accurately and coherently alongside his or her linguistic ability that shapes all thoughts through selecting appropriate lexical items, correct morphological forms and appropriate syntactical

structures (see Noor Lide Abu Kassim, 2001). Hence, ESL teachers should take into consideration students' development in the three areas (i.e. Task fulfilment, Task organization and Language) in their teaching of writing and its assessment. Furthermore, ESL instructional material designers should ascertain that materials used are appropriate for the development of ESL learners' writing in the three areas.

CONCLUSION

This study sought to validate the factorial validity and invariance of the Malaysian University English Test (MUET) essay writing rating scale in terms of its correspondence to empirical data and the theoretical construct of ESL writing. The findings support a 3-factor structure model consisting of *Task Fulfilment*, *Text Organization* and *Language*. This contradicts the current MUET writing scale that combines *Text Organization* and *Language* in one sub-scale. The findings suggest that language and text organization are two separate constructs in ESL writing; and hence, should be separately assessed.

The findings provide further support for the notion that ESL learners' development of writing skills may vary across different sub-skills or may develop unevenly across sub-skills. Therefore, scoring of the different sub-skills should be done separately. In addition, this finding suggests that the teaching of L2 writing skills should provide sufficient emphasis on the two sub-skills – *Text Organization* and *Language use* – separately so that learners are given fair learning experience for both skills in writing. Additionally, this study also demonstrates the importance of rigorous empirical validation of rating scales used in high stakes assessment.

MUET is a high stakes test used for entry requirement into Malaysian universities; thus, accuracy in the measurement of test takers' proficiency for all components is critical. Although it is practical, in terms of scoring, to use rating scales that combine several dimensions, it is important to ensure that the dimensions fit together, and therefore will provide accurate measures, especially when there is clear evidence of uneven development in the different dimensions of ESL writing. It is recommended that further validation be carried out across different ESL populations to provide more empirical evidence.

REFERENCES

Arbuckle, J. L., & Wothke, W. (1999). *AMOS 4.0 user's guide*. Chicago: Smallwaters Corporation.

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

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences*. London: Routledge.

Brown, A. (1995). The effect of rater variables in the development of an occupation-specific language performance test. *Language Testing*, *12*(1), 1-15.

Brown, J. D., & Bailey, K. M. (1984). A categorical instrument for scoring second language writing skills. *Language Learning*, *34*(4), 21-38.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 9*(2), 233-255.

Cumming, A. (1990). Expertise in evaluating second language compositions. *Language Testing*, *7*(1), 31-51.

Cumming, A. & Mellow, D. (1997). An investigation into the validity of written indicators of second language proficiency. In A. Cumming & R. Berwick (Eds.), *Validation in language testing*. Clevedon: Multilingual Matters.

Curtis, D. D., & Boman, P. (2007). X-Ray your data with Rasch. *International Education Journal, 8*(2), 249-259.

Garson, D. (2002). *Structural equation modeling*. Retrieved form https://bit.ly/2LVcpCD.

Diederich, P.B., French, J. W., & Carlton, S.T. (1961). *Factors in the judgment of writing quality*. Princeton, NJ: Educational Testing Service.

Hair, J. F., Anderson, R. E, Tatham, R. L., & Black, W. C. (1998). *Multivariate data analysis* (5th ed.). NJ: Prentice Hall.

Hamp–Lyons, L., & Henning, G. (1991). Communicative writing profiles: An investigation of the transferability of a multiple–trait scoring instrument across ESL writing assessment contexts. *Language Learning*, *41*(3), 337-373.

Hoyle, R. H., & Smith, G. T. (1994). Formulating clinical research hypotheses as structural equation models: A conceptual overview. *Journal of Consulting and Clinical Psychology, 62*(3), 429-440.

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

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika*, *36*(4), 409–426.

Kimberlin, C. L., & Winterstein, A. G. (2008). Validity and reliability of measurement instruments used in research. *American Journal of Health-System Pharmacy*, *65*(23), 2276-2284.

Kline, R.B., (1998). *Principles and practice of structural equation modeling*. New York: Guildford Press.

Kobayashi, H., & Rinnert, C. (1996). Factors affecting composition evaluation in an EFL context: Cultural rhetorical pattern and readers' background. *Language learning*, *46*(3), 397-433.

Kondo-Brown, K. (2002). A FACETS analysis of rater bias in measuring Japanese second language writing performance. *Language Testing, 19*(1), 3-31.

Krapels, A. R. (1990). An overview of second language writing process research. In B. Kroll (Ed.), *Second language writing: Research insights for the classroom,* (pp. 37–56). Cambridge: Cambridge University Press.

Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing, 12*(1), 55-71.

Malaysian Examinations Council (2006, 2011, 2015). *Malaysian university English test (MUET): Regulations, test specifications, test format and sample questions*. Retrieved from http://bit.ly/2pM4gYa .

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

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist, 35*(11), 1012-1027.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*(9), 741-749.

Noor Lide Abu Kassim (2001). *Matching instructional materials to students' needs: An evaluation of an ESL writing course* (Unpublished master thesis). Universiti Sains Malaysia.

Roid, G.H. (1994). Patterns of writing skills derived from cluster analysis of direct-writing Assessments. *Applied Measurement in Education*, *7*(2), 159-170.

Romhild, A. (2008). Investigating the invariance of the ECPE factor structure across different proficiency levels. *Spaan Fellow Working Papers in Second or Foreign Language Assessment,* 6, 29-55.

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

Schoonen, R., Vergeer, M., & Eiting, M. (1997). The assessment of writing ability: Expert readers versus lay readers. *Language Testing*, *14*(2), 157-184.

Tabachnick, B. G. & Fidell, L. S. (2001*). Using multivariate statistics*. Needham Heights, MA: Allyn & Bacon.

Weigle, S. C. (1994). Effects of training on raters of ESL compositions. *Language Testing*, *11*(2), 197-223.

Wigglesworth, G. (1993). Exploring bias analysis as a tool for improving rater consistency in assessing oral interaction. *Language Testing*, *10*(3), 305-319.