International Journal of Scientific Research in Education, MARCH 2016, Vol. 9(1), 7-19.

International Journal of Scientific Research in Education (IJSRE)
ISSN: 1117-3259

Establishing the Assessment Model for English Language Continuous Writing Component

Molefhe O. Mogapi[i]
Department of Educational Foundations
University of Botswana
Gaborone, Botswana
Molefhe.mogapi@mopipi.ub.bw

Abstract

Validation of assessment instruments has often been confined to ensuring that items in the test are a representative sample of the domain. However, structural validation of psychological instruments through multivariate statistical procedures can be used to enhance the construct validity of psychological scales by providing an empirical model of both the dimensionality of the construct itself and fidelity of the scale developed to measure the construct. The current research is an attempt to establish the assessment model of the English Language continuous writing component. The study uses exploratory factor analysis with orthogonal rotation techniques to map-out the dimensional structure the construct. Specifically, final examination composition and letter writing scores were submitted to a principal component analysis extraction method with Varimax rotation technique to arrive at a more parsimonious and theoretically meaningful language proficiency model. Items with significant loadings on the recovered components were subsequently used to name each component in the scale.

Keywords: Structural validity, Principal component analysis, Varimax rotation, Eigenvalue, Psychometric uniqueness.

Reference to this paper should be made as follows:

Mogapi, M. O. (2016). Establishing the Assessment Model for English Language Continuous Writing Component. *International Journal of Scientific Research in Education*, *9*(1), 7-19. Retrieved [DATE] from http://www.ijsre.com

INTRODUCTION / BACKGROUND

Validation of achievement tests has typically been concerned with issues of content validity. The major objective of content validation of tests is to ensure that items used in the test are a representative sample of the domain. However, construct validation of tests and scales is much more imperatives as it is concerned with the extent to which a scale actually conform to the structure of a construct of interest. The underlying dimension or structure of a psychological construct may be derived from a theory or gleaned from experience. A good example in this case is the construct of intelligence. Psychologists have been fascinated with human intelligence and the best way to measure it. Different theories about the structure of the intelligence construct have been offered and one of the theories suggests that intelligence construct is two dimensional(Spearman, 1904). Therefore, any measurement instrument that purports to measure intelligence with a set of items should conform to the two factor structure. Likewise, achievement instruments designed to assess performance in a given area should exhibit internal structure that conforms to the domain structure.

Messick (1996) defines validity as "…an overall evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores or other modes of assessment' (p. 6). The author proposed a unified concept of validity that is characterized by six elements to be considered when assessing a measurement scale. According to Messick (1989) quoted in Messick (1996):

> In particular, six distinguishable aspects of construct validity are highlighted as a means of addressing central issues implicit in the notion of validity as a unified concept. These are content, substantive, structural, generalizability, external, and consequential aspects of construct validity. In effect, these six aspects function as general validity criteria or standards for all educational and psychological measurement (p. 12).

Specifically, structural validity demands that the internal structure of a measurement instrument be consistent with the theoretical model of the construct of interest. Currently, English Language subject matter at primary school level is assessed by means of two dimensions; namely Comprehension and Language Use. Comprehension assess the ability of the candidate to extract relevant information from a variety of sources while Language Use dimension focuses on correct application of language devices such as conjunctions, adjectives and the ability to present ideas in a coherent and logical manner. Composition and letter writing falls exclusively under Language Use dimension. Each modality (i.e., Composition and Letter) is assessed by means of 10 criteria as indicated in Table 1.

Table 1: Composition and Letter writing assessment criteria

Composition Writing Criteria	Letter Writing Criteria
Introduction	Correct Address
Descriptive solution	Key Sentence
Sequence of events	Consistency in Content
Realistic composition	Different Adjectives
Correct spelling	Correct Conjunctions
Appropriate Adjectives	Sentence Coherent
Defined Sentence	Varied sentence length
Correct punctuation	Reasons for the job
Assistance required	Consistent Register
Feeling for assistance	Consistent Tense

The categorization of all the 20 criteria under one dimension implies a unidimensional structure for continuous writing. The unidimensional model has not been verified empirically and there is no theoretical justification of its use. The major research objective for the current study, therefore, is to establish the goodness-of-fit of the two factor model for continuous writing scores. The covariance matrix from PSLE composition and letter writing scores were factor analyzed to map out patterns of relationships amongst the items. Multiple model fit criteria were used to extract reliable factors that best represents the language proficiency construct.

LITERATURE REVIEW

The need to consult relevant literature sources on language assessment stems from the controversy surrounding the number of reliable factors that represent English language proficiency construct. According to Oller (1976), English language ability is represented by only one unitary factor; in other words, performance in English language is unidimensional. However, subsequent research work (e.g., Sasaki, 1999) has largely disconfirmed the unitary hypothesis in favour of the multi-componential structure. The general consensus amongst language researchers is that language proficiency is best represented by a high-order secondary factor and several correlated first order factors (Stricker, Rock & Lee, 2005). The controversy relating to the unitary hypothesis claim and multi-componential structure viewpoint can only be resolved by frequent and sustained research work to determine which of the two arguments is supported by the data. One institution that has carried out extensive and intensive research work on the dimensionality of language proficiency construct is the Educational Testing Service (ETS) in the USA. ETS administers a Test of English as a Foreign Language (TOEFL) to candidates all over the world. As a result, ETS has managed to accumulate a wealth of knowledge on language

assessment through its numerous research studies and publications. A few of the relevant studies are reviewed below to shed light on the language dimensionality issue.

TOEFL was developed in 1963 by the National Council on the Testing of English as a Foreign Language (Sawaki, & Orange, 2008). It is specifically designed to test the English language proficiency of non-native speakers of the language applying for admission to institutions in the United States. TOEFL has developed over the years from a paper-based test to an internet-based format (i.e., TOEFL iBT). Extensive research work has been done to assess the relevance and appropriateness of TOEFL test to different populations across the globe. For example, from 1977 to 2005, nearly 100 research and technical reports have been published (ETS, 2008). Several of the TOEFL studies focus on establishing the dimensionality of the test and the invariance of its factors across different subgroups. Some of the key research studies are enumerated below.

Stricker, Rock and Lee (2005) conducted a study aimed at establishing the factor structure of the LanguEdge test and its invariance across various TOEFL populations around the world. The test has four sections; namely, Listening, Reading. Speaking, and Writing Sections. The LanguEdge test was administered to a total of 472 candidates recruited from both domestic and international TOEFL testing centers. The sample was made up of mostly Arabic (N=160), Chinese (N=225), and Spanish (N=114). Maximum likelihood estimation procedures were applied to analyse the covariance matrix for the normalized scores of each subgroup using LISREL version 8.53 (Joreskog & Sorbom, 1996b). The researchers tested the composite hypotheses of the latent factor structure of TOEFL language test and the invariance of the final solution across the three subgroups. Initially four competing models were tested.

The first model tested was a four factor model which proposes a unitary language proficiency trait. In this model, the four language modalities (Reading, Listening, Speaking and Writing) load on a single language ability factor. The second model tested hypothesized existence of two factors defined by a separate factor for Speaking and a second factor created by a fusion of Reading, Writing and Listening. The third model proposed three correlated factors representing Listing and Speaking and a combination of Teaching and Writing components. The fourth model had four distinct factors that corresponded to the four sections of the test(e.g., Reading, Listening, Speaking and Writing). Items in each of the modalities were expected to have significant loadings on their respective sections. Several goodness-of-fit indices were employed to identify the model that has a superior fit to the data

The analysis converged on only two models; the one factor model and two factor model. The fit indexes for the individual subgroups and the overall analysis were also deemed to be satisfactory for the two models. Subsequent detailed comparison of the two competing models indicated a superior fit of the two factor model. Once the researchers established the baseline model; secondary analysis of the data was carried out to determine the extent of the measurement non-invariance of the LanguEdge test. Hierarchically ordered nested models were tested about the factor invariance specifically relating to the invariance of the number of factors; invariance of the factor loading; the invariance of the error variances; and the invariance of the factor intercorrelation. The analysis showed that only the factor correlations were deemed to be invariant across groups. Therefore, the researchers were able to establish that the LanguEdge test is essentially invariant across groups and can best be represented by two factors; these factors were identified as Speaking and a fusion of Reading, Writing, and Listening. However, the two factor solution was at variance with earlier research studied by Carrol (1983) and Bachman et al

(1995) and Kunnan (1995); these studies extracted three correlate factors with Listening defined as a distich factor.

Another TOEFL instrument that has been thoroughly researched is the TOEFL Internet Based or TOEFL iBT (ETS, 2005). The test was developed in 2005 and consist of five section; Reading, Writing, Speaking and Listening. Candidates who write this test receive a total of five scores; a separate score for each of the five language modalities and a global language ability score (Sawaki, et al., 2008). The internet based test has a compulsory Speaking section thus making its design to be different from the LanguEdge test. Therefore, it became necessary to investigate the new test's internal structure as well as generate evidence to support the five score reporting policy. Sawaki, Stricker and Orange (2008) conducted a confirmatory factor analysis study to model the test's internal structure. Separate analyses were done for Reading, Listening and a combination of Writing and Listening. Writing and Listening were combined due to limited number of items in these two sections (Sawaki, et al., 2008). The researchers adopted the multitrait-multimethod factor analytic approach to test for the presence of trait and method effects in the correlation matrix. The traits identified were listed as Basic Comprehension, Inferencing, and Reading to learn. On the other hand, the items in the sections were grouped to define item set factors. Four different models that differed according to the number of path coefficients created in each model were sequentially tested. The models were as follows:

- Model A: Correlated traits and correlated item set model. Within this model the three traits and item sets were allowed to correlate among themselves but not between the two factors.
- Model B: Correlated traits and Uncorrelated item set model. This model limited the degree of correlation of the item sets and as such the model is nested within Model A.
- Model C: Correlated trait model. This model imposes maxim restriction on the data as it essentially eliminates item set correlation.
- Model D: Correlated trait and correlated uniqueness model. The model is different from the other models above in the sense that variance within each item is partitioned into specific and error variance to estimate residual variance. However, due to its complexity, the model was not considered as a likely candidate for representing the factor structure of the test but was included to evaluate the appropriateness of other models (p. 16).

The researchers used three criteria to identify the best fitting model to the data. The criteria were: (1) the extent to which the solution was proper, (b) Substantive interpretation, (c) goodness of fit for the model. Also the magnitude of inter-factor correlation cut of point of .90 or more was also used to identify high inter-factor dependency. The pre-determined fit criteria suggested that none of the four models exhibited satisfactory fit to the data. ModelA factor loadings were un-interpretable while the remaining three models had high inter-factor loadings exceeding .90. High inter-factor correlation indicates lack of psychometric uniqueness of the factors (Bagozzi & Yi, 1992). The results therefore point to the likelihood that the three traits (e.g., Basic Comprehension, Inferencing, and Reading to Learn) are evidence of a single trait. As a result, the inter-factor correlations were set at 1 essentially creating one factor upon which all items loaded. Further analysis confirmed the superiority of the single trait model and it was subsequently adapted as a parsimonious representation of the trait structure of the reading section. This suggested a unidimensional nature of reading ability (p. 27).

A similar analysis was done for the Listening section and another one for combined Speaking and Writing sections. In the case of listening, the model specified three traits as being Basic Understanding, Pragmatic Understanding, and Connecting Information. The results showed lack of psychometric distinctiveness amongst the three traits due to significant inter-factor correlations that was well above the pre-determine cut off level of .90. A single trait approach was then followed to establish the best fitting model. The analysis indicated that the single trait model and the single trait uniqueness model had equivalent fit to the data. For this reason, the single trait model was adopted on the principle of parsimony.

The analysis of the Speaking and Writing sections was not as thorough as the previous ones for Reading and Listening. This was mainly due to the number of items available for analysis. However, the analysis converged on a correlated trait model that combined Speaking and Writing traits. A chi-square difference test suggested a significantly better fir for a two factor model.

To establish the best fitting model for the entire TOEFL iBT test, the four preceding models were combined. As a result, four nested model were tested at the global level; these are the Bifactor model, the Correlated trait model, the Single trait model, and the Higher-order model. The main difference between the Bifactor model and the High-order model was that the former allowed for direct influence of the second-order factors on the measured variables whereas the latter specified presence of a common underlying dimension that influenced the individual items only through the first order factors (p. 53). Several goodness of fit indices were used to evaluate model fit (e.g., NNFI, CFI, RMSEA, GFI, ECVI). The fit indices converged on two models; the Correlated trait model and the Higher-order model. The fit statistics indicated that the two models had an equivalent fit. The high-order model was considered to be a reasonable representation of the factor structure of the entire TOEFL iBT test, thus, supporting claims that language ability was essentially multi componential in nature.

Another very informative study on TOEFL assessment instruments was done by Shin (2000). This particular study was conducted to provide evidence that will shed some light on the relationship between language proficiency and the structure of language tests. The issue to be addressed was whether the structure of language tests is invariant across different ability levels. Previous research that sought to understand the relationship led to the emergence of two schools of thought. The first school of thought proposes that as language proficiency improves, the factor structure also becomes more differentiated. A study by Swinton and Powers (1980) for example, 'pointed to greatest amount of factor differentiation for the group with the highest proficiency and the least amount of factor differentiation for language group with the lowest proficiency (p. 33). The observed positive relationship was also confirmed by Ginther and Stevens (1980) in their multi-group structural equation modelling research. On the other hand, some research studies have provided evidence showing an inverse relationship between language proficiency and factor differentiation. The studies show that as learners become more proficient in language use, the factor structure of the test becomes less differentiated. Studies by Kunnan (1992) and Oltman, Stricker, and Barrows (1988) demonstrated that the factor loadings for the low ability groups become more salient.

The primary goal of Shin (2000) study was, therefore, to investigate the relationship between language proficiency and the structure TOEFL. The sample of the study comprised 779 candidates who participated in the Cambridge TOEFL Comparability Study (Bachman, 1995). The sample was divided into three groups; low, intermediate, and high language ability groups. The candidates wrote TOEFL test and the scores were used to generate a covariance matrix of

correlation coefficients. A principal axis factoring modeling techniques were used to extract initial components. Amongst the several competing models, the high-order model was selected as best fitting model because it was able to account for the high correlation amongst the three factors identified. Also, the researcher noted that the selected model structure replicated previous research findings demonstrating that .language proficiency consists of one general factor and several distinct abilities (e.g., Fouly xx, 1990).

Since the analysis indicated the suitability of the second order model, the model was further analyzed to test its invariance across the three language proficiency groups. The model was subsequently tested for measurement in variance and structural invariance. Measurement invariance holds when the factor loadings of a measure are found to be equal across groups (p.). The results largely indicated that the TOEFL test exhibited measurement and structural invariance; however, the pronunciation and fluency subscales were an exception to the general observation. The evidence lead the researcher to conclude that final model supported neither of the hypotheses of increasing factor differentiation nor that of decreasing factor differentiation as a function of increasing examinee proficiency (p. 53). The results also supported the policy of using a single score for all language groups and attest to the high construct validity of the TOEFL test.

Recently, a study was also undertaken by Gu, Turkan, & Gomez, (2015) to investigate the dimensionality of the Test of English for Teaching or TEFT. The main objective of the study was to build a validity argument in support of the test's internal structure and the relevance of the score reporting practice. Specifically, the researcher noted that 'With regard to score reporting, the most prominent information on the score report is the total score scale and the associated band and band descriptors' (Gu et al., p. 3).The researchers sampled 1,307 participants from a group of students who wrote one form of TEFT administered during a pilot study in 2012. TEFT is designed in such a way that it measures language proficiency in two broad areas. Firstly, the test items are categorized into four primary language skill area; namely, Reading, Writing, Speaking, and Listening. Secondly, the test is organized into three content areas based on the functional use of language in a classroom setting. The functional skill areas are Managing the classroom, Understanding and Communication lesson content, and Providing feedback(Gu et al, 2015).. A candidate receives a separate score for language skill area and another score for the content component. The overall score is obtained by combining the two component scores.

According to Gu et al. (2015), "A multitrait-multimethod confirmatory factor analysis approach was taken to examine the influence of both skill and content on test performance' (p. 3). The analysis first focused on the establishment of a plausible baseline model for skill and content dimensions separately. The best fitting models for the separate dimensions were then combined. The decision to test separate models was motivated by the score reporting policy (i.e., reporting separate score for skill and content area dimensions).

In terms of the skill dimension, three competing models were tested; (a) A correlated four factor model, (b) A higher-order model, and (c) A Bifactor model. The analysis indicated an adequate fit for the four factor skill model and the higher-order model. A parallel analysis was done for the content section; three models were also hypothesized. However, all the three content models produced unsatisfactory fit indices and were subsequently considered to be inadmissible. This scenario compelled the researchers to create a new model where the Managing the classroom and Understanding and Communication lesson content factors were combined to form a new factor. This essentially created a two factor content model.

To test an overall model for the entire test, the best fitting skill model and best fitting content models were combined to create four models (i.e., . High-Order Skill Relationships and Two Content Factors, Four Skill Factors and Two Content Factors, High-Order Skill Relationships and Three Content Factors, Four Skill Factors and Three Content Factors). The analysis showed that only the Higher-Order model converged. The researchers noted that 'Generally, the results support the current score reporting practice, that is, to report a total scaled score along with score information on skills and language use in specific content areas' (p. 9).Also, the results supported findings by other researchers suggesting that language proficiency is hierarchical and multi-componential in nature (Fouly, Bachman & Cziko, 1990; Bachman, Davidson, Ryan & Choi, 1995; Sawaki, Stricker & Orange, 2008; Sawaki et al., 2009). Therefore, the literature sources reviewed indicates an apparent consensus amongst language experts that language proficiency is best represented by a multi-componential hierarchical model.

METHODS

The study follows the quantitative approach as correlation coefficients for PSLE English Language continuous writing component scores are used in the analysis. The correlation matrix of the scores is submitted for analysis using CFA procedures to identify reliable dimensions.

Sampling procedures

There are ten educational districts in Botswana; each district is divided into regions. The South Central District has four educational regions with a total of about 30 public primary schools. The current study derived the sample from the four regions within South Central District. Simple random sampling procedures were applied at the level of schools; as a result, 22 schools were sampled and all of the composition and letter scripts in each of the sampled schools were included for analysis making a total sample of 1800.

Instrument

The candidates wrote PSLE English Language composition and letter examination in 2003. An analytic marking scheme comprising twenty criteria was used to determine the language proficiency of each candidate. Though composition and letter modalities are graded separately, the two marks are combined to produce a score for continuous writing.

Analysis

The suitability of the data for factor analysis was assessed by means of Kaiser-Meyer-Olkin Measure of Sampling Adequacy (KMO) and Bartlett's Test of Sphericity. Both tests yielded favourable results as indicated in Table 2.

Table 2: KMO and Bartlett's Test of Sphericity

KMO and Bartlett's Test		
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.899
Bartlett's Test of Sphericity	Approx. Chi-Square	8111.610
	df	190
	Sig.	.000

The methodological literature on factor analysis strongly recommends use of multiple goodness-of-fit indices because each criterion has its own advantages and disadvantages(Henson, Capraro & Capraro, 2004; Hu & Bentler, 1999). For the present data, two factor extraction techniques were used; these are the Kaiser criterion (K1 rule) and the Scree Plot. The rationale of the K1 rule is that a factor or dimension with substantive meaning should account for variance in the data that exceeds the variance explained by a single item. Since the maximum variance that an item can explain is 1, the K1 rule leads to the retention of dimensions with eigenvalues greater than 1 (Costello & Osborne, 2005). Table 3 shows that four factors satisfy the KI rule and their cumulative variance explained is 48.684%. It should be noted that the fifth factor missed the cut of point by a very small margin.

Table 3: Total variance explained

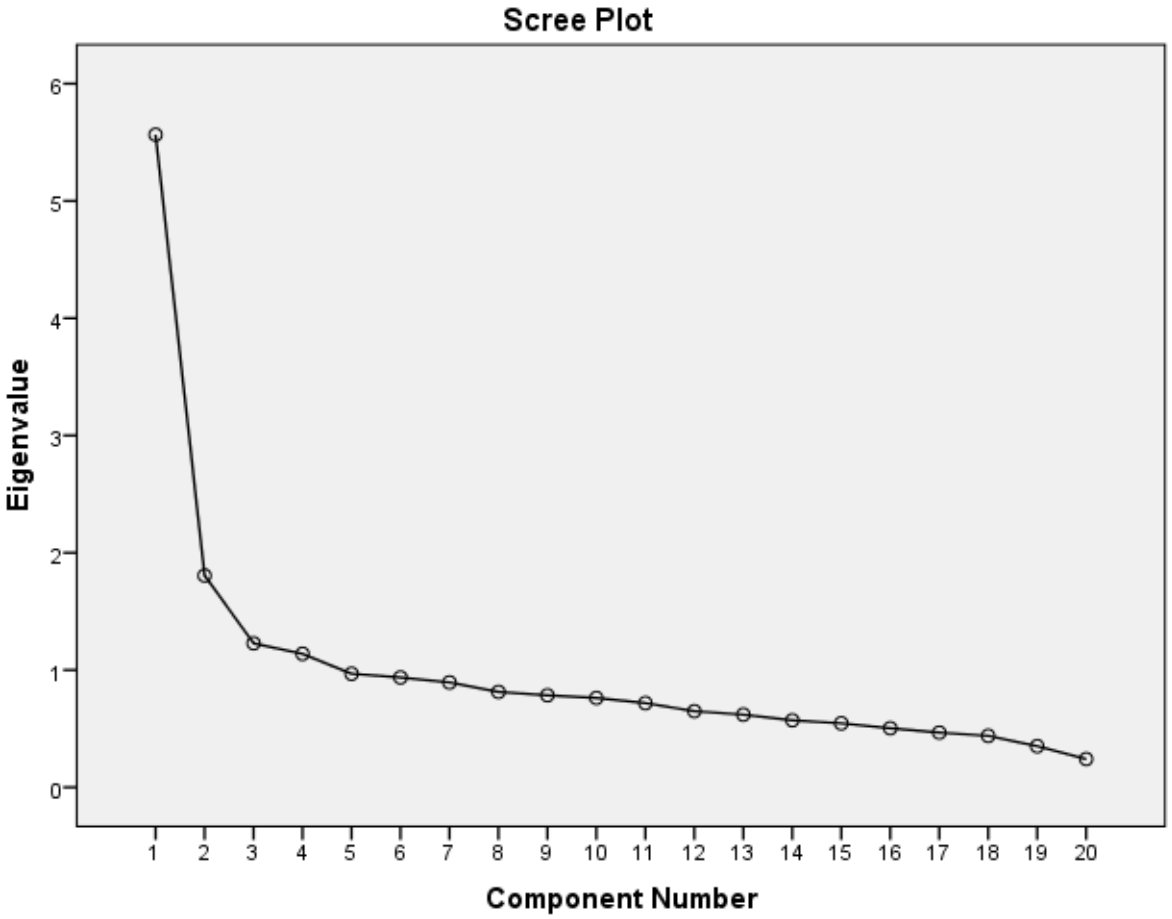
Componen	Initial Eigenvalues			Extraction Sums of Squared Loadings		
t	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	5.565	27.826	27.826	5.565	27.826	27.826
2	1.806	9.028	36.854	1.806	9.028	36.854
3	1.228	6.139	42.993	1.228	6.139	42.993
4	1.138	5.691	48.684	1.138	5.691	48.684
5	.968	4.841	53.525			

.

Extraction Method: Principal Component Analysis. Rotation method: Varimax

The four dimensional structure represented by the KI rule is corroborated by the Scree plot (Figure 1). The Scree test arranges dimensions by order of significance with the factor that explains the larger variance listed first. An examination of the plot shows that a bend or 'elbow' occurs after the fourth factor. Factors that occur after the bend lie in an almost horizontal position. However, it should be noted that the point where the lines changes form vertical to horizontal is not so pronounced. This could mean that Dimension4 is not well represented in the data or Dimension 5, though excluded on the basis of the K1 rule, has some theoretical value. Generally, the K1 rule and Scree plot suggest a four dimensional solution as a more parsimonious representation of the language proficiency construct.

Figure 1: The Scree Plot



Factor Loadings

The rotated component matrix (Table 4) below shows the factor loading of the 20 measured variables used to assess language proficiency. Varimax rotation was used to identify items that have significant loading on one of the four dimensions extracted. This ensures that each item has a higher loading on only one factor to avoid cross loading items and at the same time bring about simple structure. For example item 1 (Introduction) has a higher loading of .715 on Dimension 2 an insignificant loading of .096 and .065 on Dimension 1 and 2 respectively.

Table 4: Variable loading for the rotated component matrix

	Component			
	1	2	3	4
1. Introduction	.096	**.715**	.065	.030
2. Descriptive solution	.122	.374	**.581**	-.014
3. Sequence of events	.107	.021	**.615**	-.112
4. Realistic composition	.253	**.428**	.292	-.070
5. Correct Spelling	.016	.115	**.613**	.204
6. Appropriate Adjective	.189	**.666**	.305	.034
7. Defined Sentence	.057	.204	**.531**	.111
8. Correct punctuation	.149	**.584**	.034	.115
9. Assistance required	.238	**.525**	.226	-.040
10. Feeling for assistance	.175	**.693**	.085	-.039
11. Correct address	.121	-.091	.254	**.633**
12. Key Sentence	**.632**	.158	.151	-.081
13. Consistency in content	**.783**	.277	.054	.117
14. Different adjectives	**.672**	.063	.183	.116
15. Correct conjunctions	.375	.208	-.165	**.464**
16. Sentences coherent	**.781**	.262	.073	.089
17. Varied Sentence length	**.488**	-.062	.180	-.542
18. Reasons for the job	**.745**	.212	.089	.168
19. Consistent Register	.374	-.002	.130	**.430**
20. Consistent Tense	**.667**	.294	-.058	.055

Extraction Method: Principal Component Analysis.

Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 7 iterations.

Naming of the four dimensions

Since the analysis converged on four dimensions, the next logical stage is to find an appropriate name for each of the dimensions. Naming of a dimension is determined by content of items that have significant loading on the dimension under consideration. Dimension 1 was named 'Logical Development of Ideas' because most of the items that load on this dimension dealt with the ability of the candidate to put ideas in a logical and coherent manner. The Dimension 2 was named 'Communication of Feelings' as the majority of items associated with the dimension required the candidate to show their feelings and/ or emotions. Dimension 3 and Dimension 4 were labelled as 'Correct Use of Language Devices' and Appropriate Register'. Dimension 3 deals with correct use of adjectives and conjunctions while Dimension 4 examines the ability to write an address and salutation correctly when writing a letter.

DISCUSSION

The original 20 metric variables used to measure language proficiency were submitted to PCA with Varimax rotation. The factor analysis applied successfully generated four dimensions which were subsequently named. The components extracted had substantive meaning as the minimum number of items loading on a dimension was three. The four factor solution is in support of the multi-componential nature of language previously established other language researcher (). Most importantly, the four factor solution is contrary to the current PSLE assessment practice that regards continuous writing as being unidimensional. Categorizing all items relating to continuous writing under the Language Use dimension does not seem to be a true reflection the continuous writing construct. The current results identify a four factor model as a more parsimonious structure for continuous writing. Consequently, candidates should be assessed by means of four separate scores; there should be a score for Logical Development of Ideas, Communication of Feelings, Correct Use of Language Devices and Appropriate Register. However, more research work needs to be done to establish the substantive and theoretical relevance of Dimension 4 and Dimension 5. The cu-off point between these two dimensions is currently ambiguous.

CONCLUSION

Research on the dimensionality of language proficiency strongly points to the multidimensional nature of the language proficiency construct. The extraction of four factor model that account for 48.684% of the variance in the matrix provides additional evidence for the four multi-componential hypothesis. Therefore, score reporting practice that utilizes four separate scores for each of the identified dimensions would serve as a better diagnostic tool than the current unidimensional approach.

REFERENCES

Bagozzi, R. P., & Yi, Y. (1988).On the evaluation of structural equation models. Journal of the Academy of Marketing Science, *16*, 74-94.

Carrol, J. B. (1983). Psychometric theory and language testing. In J. W. Oller, Jr. (Ed.), *Issues in language testing research* (pp. 80-107). Rowley, MA: Newbury House.

Costello, Anna B. & Jason Osborne (2005). Best practices in exploratory factor analysis: four recommendations for getting the most from your analysis. *Practical Assessment Research & Evaluation, 10*(7). Available online: http://pareonline.net/getvn.asp?v=10&n=7

Fouly, K. A., Bachman, L. F. & Cziko, G. A. (1990). The divisibility of language competence: A confirmatory approach. *Language Learning. 40,* 1-21

Gu, L., Turkan, S., & Gomez, P. G. (2015). Examining the internal structure of the test of English for Teaching (TEFT). New Jersey: Educational Testing Service.

Henson, R. K., Capraro, R. M., & Capraro, M. M. (2004). Reporting practice and use of exploratory factor analysis in educational research journals: Errors and explanations. *Research in Schools, 11*(2), 61-72.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional versus new alternatives. *Structural Equation Modeling, 6 (1), 1-55.*

Joreskog, K. G., & Sorbom, D. (1996b). PRELIS 2: User's reference guide [Computer software manual] Chicago: Scientific Software.

Kunnan, A. J. (1995). Test taker characteristics and test performance: A structural modeling approach. Cambridge, England: Cambridge University Press.
Messick, S. (1996). Validity and washback in language testing. Princeton, New Jersey: Educational Testing Service.
Oller, J. W., Jr. (1976). Evidence of a general language proficiency factor: An expectancy grammar. *Die Neuren Sprachen*, *76*, 165-174.
Oltman, P. K., Stricker, L. J., & Barrows, T. (1988).Native Language, English proficiency, and the structure of the TOEFL. TOEFL research report 27. Princeton, NJ: Educational Testing Service.
Sasaki, M. (1999). Second language proficiency, foreign language aptitude, and intelligence: Quantitative and qualitative analyses. New York: Lang.
Sawaki, Y., Stricker, L., & Orange, A. (2008). Facto structure of the TOEFL Internet Based Test (iBT): Exploration in the field trial sample. New Jersey: Educational Testing Service.
Shin, S. (2005). Did they take the same test? Examinee language proficiency and the structure of language tests. Language Testing, *22*(1), 33-57.
Spearman, C. (1904). General intelligence, objectively determined and measured. *American Journal of Psychology*, *15*, 201-293.
Stricker, J. L., Rock, A. D., & Lee, Y. (2005). Factor structure of the LanguEdge test scores across groups. Monograph Series. New Jersey: Educational Testing Service.
Swinton, S. S., & Powers, D. F. (1980). Factor analysis of the TOEFL. Research Report 6. New Jersey: Educational Testing Service. Retrieved from: http://www.

---

[i] Mogapi, M. O. is a lecturer in the Department of Educational Foundations, University of Botswana Gaborone, Botswana. He can be via email at: Molefhe.mogapi@mopipi.ub.bw