A Construct Validation of the Malaysian University English Test and the English Placement Test: Two High-stakes English Language Proficiency Tests

NOOR LIDE ABU KASSIM
AINOL ZUBAIRI
NURAIHAN MAT DAUD
International Islamic University Malaysia

Abstract: *Malaysian students are normally required to show a certain level of English language proficiency before they are offered a place by a local institution of higher learning. One of the examinations conducted to ascertain their level of proficiency is the Malaysian University English Test (MUET). Apart from MUET some universities also conduct their own test for the same reason. This paper discusses a study on the comparability of MUET with the English Placement Test (EPT) conducted by the International Islamic University Malaysia (IIUM). This study was conducted to see whether the two tests could be used interchangeably for admission and placement purposes. The two tests were administered on second year pre-university students of IIUM. Analyses show that though significant, the correlation coefficients of the two tests were not high enough for them to be used interchangeably.*

INTRODUCTION

Many leading universities require their intending students to show a certain level of proficiency particularly in the language of instruction before they are admitted into the institutions. TOEFL and IELTS scores are two of the common yardsticks where English language proficiency is concerned. Such tests, however, are not easily accessible to candidates in some countries, and they could also be costly by local standards. To overcome this problem, countries such as Malaysia offer their own version of the test. The Malaysian University English Test (MUET) is conducted by the Malaysian Examinations Council, and it is a requirement for all students intending to pursue their studies in Malaysian public universities. Apart from the MUET, some Malaysian universities – public and private – also administer their own English language

proficiency tests which are tailored to their requirements. One such example is the English Placement Test (EPT) which is conducted by the International Islamic University Malaysia (IIUM).

In the context of the IIUM, incoming students are required to sit for both tests. Since it is not economical for students to take two tests of English language proficiency, a logical step would be to investigate the comparability of the two tests. This project, therefore, was undertaken to investigate the comparability of the MUET and the EPT in terms of the construct measured. Empirical evidence gained from this study could then be used to decide whether it would be best to adopt both or whether one was adequate for placement decisions. This study therefore investigated the construct validity of the two tests. The approach used involved the quantitative investigation of patterns of relationships of examinee performance among the different subtests of the MUET and the EPT based on test scores. Specifically, the study aimed to answer the question: Do the different subtests of the MUET and EPT measure the same underlying language ability?

Construct Validity

Construct validity is perceived as one of the three most common types of validity evidence, the other two being content and criterion validity (Alderson, Clapham & Wall, 1995; Henning, 1987). Historically, the notion of construct validity grew out of efforts by the American Psychological Association in the 1950's to develop standards for adequate psychological testing (Bachman, 1990; Gray, 1997). Construct essentially refers to the psychological domain, which is a theoretical conceptualization about an aspect of human behaviour (American Psychological Association, 1985). In a language testing context, construct is viewed as "...an underlying ability or trait that is hypothesized in a theory of language" (Hughes, 1989, p. 26). Commonly acknowledged constructs of language ability are: reading and listening comprehension and writing and speaking abilities.

In the traditional view of validity, construct validation studies are concerned with the evidence to justify the usefulness of test interpretation, i.e. the relationship between test performance and what it intends to measure. Therefore, evidence from construct validation studies is about how well test scores can be interpreted in terms of some psychological quality (Gronlund, 1981; Hughes, 1989). In recent years, the status of the multiple definitions and types of validity has been challenged by a more unified perception of validity. This more

unified conceptualisation of validity is that construct validity is the central or fundamental principle that subsumes all the other previously recognized aspects of validity (Cumming, 1995).

The most influential and widely cited model which considers this conception of construct validity is proposed by Messick (1989a, 1989b). He argues that although validity has been traditionally perceived in terms of its functional worthiness (how well a test does its intended job), the approach to validity has neglected crucial forms of validity evidence – the social consequences and value implication of test interpretation and use (Messick, 1989b).

Messick (1989a, 1989b) therefore, proposes a unified framework for classifying validity which essentially perceives construct validity to be central to all aspects of validity, and considers the value implications and social consequences. The model consists of two interconnected facets, which are 'source of justification of the testing' and 'the function or outcome of a test'. The source of justification is based on appraisal of evidence or consequence, while the function or outcome of testing on test interpretation or test use. The four dimensions form a 'progressive matrix' (see Table 1) in which construct validity can be determined and systematically appraised.

Table 1: *Facets of Validity (Messick, 1989a)*

	Function or outcome of testing	
Sources of justification of the testing	**Test interpretation**	**Test Use**
Evidential Basis	Construct validity	Construct validity + Relevance/Utility
Consequential Basis	Constrcut validity + Value implication	Construct validity + Relevance/Utility + Social consequences

One of the reasons for reconceptualizing construct validity is that it places the test construct in the centre of focus, and therefore requires testers to consider what is known about language knowledge, and how these can be operationalised (Alderson & Banarjee, 2002).

That means that conceptualizing the language construct is also central to testing as it establishes the ways to elicit people's language and language use in attempts to assess people's ability.

Comparability Studies and Construct Validity

There has been continuous interest in comparing students' performances on different tests, particularly in influential test of proficiency in English as foreign language (Alderson & Banarjee, 2002). Examples of large scale comparability studies that have been undertaken are the Cambridge-TOEFL study (Bachman, Kunnan, Vinniarajan & Lynch, 1988; Kunnan, 1995; Bachman, Davidson, Ryan & Choi, 1995; Kunnan, 1995) and the ELTS validation studies that compared the English Proficiency Test Battery (EPTB), the English Language Battery (ELBA) and ELTS (Criper & Davies, 1988). Such comparability studies are valuable in that they help in giving a clearer picture of the different language constructs. For example, the Cambridge-TOEFL comparability study revealed problems related to lack of parallelism in some of the sub-tests which later led to significant improvements in those tests.

METHOD

The Tests

The MUET is an influential test of proficiency in English as a foreign language in the Malaysian context. It is a national standardized English test developed to assess and measure the English language proficiency and achievement of pre-university students. It is administered by the Malaysian Examinations Council twice a year to test the English language ability of those who intend to pursue first-degree courses in local public institutions of higher learning. Students with a Higher Certificate of Education (equivalent to 'A' levels), Diploma or Matriculation certificate who plan to undertake undergraduate studies at a local university are required to sit for the MUET. The MUET consists of four sub-tests, each assessing one area of language ability – reading, listening, speaking and writing. An overview of the test is given in Appendix 1.

The EPT is a tailor-made English language test administered to incoming students at the International Islamic University to ascertain the level of their English language ability, and to place them in the appropriate language support courses. The test is not only meant for

Malaysians, but also international students who come from around a hundred different countries. The test is given to those students who are unable to produce, upon entry, results of the TOEFL or other international tests of English proficiency.

The first part of the EPT focuses on the assessment of students' grammatical competence and performance whereas the second consists of a battery of skills-based tests covering reading, listening, speaking and writing skills. These tests aim at assessing students' competency and performance in the four language skills. A summary of the EPT is given in Appendix 2.

The observable similarities of both tests are the utility of the test scores and the language constructs measured. In summary, both tests are:

- based on the view that language ability is partially divisible as advocated by Bachman (1990);
- used to ascertain language ability level of candidates entering the tertiary levels of education in Malaysia;
- used for selection and/or placement;
- intended to assess student's ability in four main language constructs: listening, speaking, reading and writing;
- designed to test different aspects of the language separately but the scores are reported as a composite.

Subjects

The subjects for this study were 1,556 students (545 males and 1011 females) studying at the IIUM Matriculation Centre. These students were enrolled in the following programmes: Medicine, Laws, Engineering, Economics, Information Technology, Science, Applied Science, Applied Health, Social Science and Islamic Studies. The EPT, followed by the MUET were administered several months before the end of their matriculation programme. Administration of the EPT was conducted by the English Language Department, IIUM Matriculation Centre, while the MUET was conducted by the Malaysian Examinations Council-appointed examiners.

Data Analysis

Two types of statistical procedures were conducted. The first involved the bivariate correlation procedure. The second was an exploratory factor analysis using the Principal Component Analysis procedure. Both

analyses were conducted using SPSS Windows 12.1. The bivariate correlation procedure was conducted to examine the inter-relationships among the subtests in this study. The exploratory factor analysis was carried out to examine the convergent and divergent validity of the two tests. The results from this analysis served as evidence of construct validity of the two tests and also their comparability.

RESULTS

Before conducting the factor analysis, a few tests were carried out to examine whether assumptions underlying factor analysis were met. The first was a correlational analysis of the EPT and MUET subtests. This was to investigate the degree of collinearity among the subtests (variables). The second was the Bartlett's test of sphericity and the third, the Kaiser-Meyer-Olkin measure of sampling adequacy. These tests were carried out to investigate the factorability of the intercorrelation matrix.

Intercorrelation Matrix of Subsets

The intercorrelation matrix indicates a significant correlation among the subtests. The result of the analysis is given in Table 2.

Table 2: *Intercorrelation Matrix of the MUET and EPT subtests*

	Listening EPT	Reading EPT	Writing EPT	Listening MUET	Speaking MUET	Reading MUET
Speaking EPT	.33**	.32**	.28*	.24**	.34**	.35**
Listening EPT		.56**	.37**	.42**	.32**	.55**
Reading EPT			.42**	.47**	.25**	.58**
Writing EPT				.24**	.22**	.35**
Listening MUET					.23**	.46**
Speaking MUET						.36**
Reading MUET						

*$p < .01$ (2-tailed)

**$p < .05$ (2-tailed)

The intercorrelation matrix indicates that for the MUET, the correlation between writing and speaking was not statistically significant. In the EPT, the correlation between the two was the weakest ($r = 0.28$). The strongest correlation was observed between reading and listening for both tests (MUET: $r = 0.46$, EPT: $r = 0.56$). This reflects a common pattern between the tests.

Bartlett's Test of Sphericity and the Kaiser-Meyer-Olkin Measure of Sampling Adequacy

To determine the factorability of an intercorrelation matrix, two tests were carried out: the Bartlett's Test of Sphericity and the Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy. The first test is an indicator of the strengths of the relationship among variables and was used to test the null hypothesis that the variables in the correlation matrix were uncorrelated. This hypothesis was made based on the Principal Component Analysis. The results are presented in Table 3.

Table 3: *Bartlett's Test of sphericity and the Kaiser-Meyer-Olkin measure of sampling adequacy*

Kaiser-Meyer-Olkin measure of Sampling Adequacy		.86
Bartlett's test of Sphericity	Approx. Chi-Square	707.55
	df	28
	Sig.	.01

As the observed significance level is less than 0.01, it is therefore appropriate to reject the null hypothesis. The second is an index for comparing the magnitudes of the observed correlation coefficients to the magnitudes of the partial correlation coefficients. As the KMO indicates a 'meritorious' (Kaiser, in Pett, Lackey & Sullivan, 2003) value at .86, the factors extracted from a factor analysis of the variables (subtests) would account for quite a substantial amount of variance.

Factor Analysis

The purpose of this analysis was to explore and verify patterns of correlation coefficients among the subtests of the MUET and the EPT. In this way, convergent (where similar tests load together) and divergent (tests of different traits load separately) validity could be ascertained. In this analysis, the Principal Component Analysis with Varimax Rotation (orthogonal rotation) was utilized. In the initial analysis eigenvalue of 1.00 or higher was used to determine the number of factors to be extracted, and from the eigenvalue two factors were extracted. Table 4 shows the results of the 2-Factor Analysis after Varimax Rotation:

Table 4: *Results of the 2-Factor Analysis after Varimax Rotation*

Tests	Factor 1	Factor 2	*h*
Listening EPT	**.75**		.61
Listening MUET	**.67**		.46
Reading EPT	**.68**	***(.41)***	.63
Reading MUET	**.73**	*(.30)*	.63
Writing EPT	*(.39)*	**.58**	*.49*
Wrinting MUET		**.89**	*.79*
Speaking EPT	**.65**		*.34*
Speaking MUET	**.55**		*.44*
Proportion of Variance	*.36*	*.19*	*.55*

The eigenvalue for the first factor was 3.367 whereas the eigenvalue for the second factor was 1.01. The total variance accounted for by this 2-Factor Model is 54.70% with the first factor accounting for 36.10% and the second factor accounting for 18.50% of the variance in these tests/subtests. The factor loadings (Stapleton, 1997) on the first factor ranged from .55 to .75 indicating a moderate to high correlation between the subtests and the factors. The communalities (squared loadings) ranged from .34 to .79 indicating that the 2-Factor Model accounted for 34.00% to 78.50% of the variance among the subtests, with an average of 54.70 % of the variance accounted for.

The 2-Factor analysis also shows that the Listening, Reading and Speaking subtests of both the EPT and the MUET loaded on the first factor whereas the Writing subtests loaded on Factor 2. The Reading EPT, Reading MUET and Writing EPT subtests cross loaded on two factors. This meant that both factors accounted for some amount of variance in these subtests.

Gorsuch's (1983) suggestion on accounting for at least 70% of the total variance, and also the suggestion to increase the number of factors until all nontrivial variance is accounted for were taken into consideration; hence, the possibilities of other factor models were explored. Two other factor models were then generated, a 3-Factor and a 4-Factor analysis. The results of the 3-Factor analysis are given in Table 5.

Table 5: *Results of the 3-Factor Analysis after Varimax Rotation*

Tests	Factor 1	Factor 2	Factor 3	*h*
Listening EPT	**.73**			.63
Listening MUET	**.79**			.63
Reading EPT	**.77**			.70
Reading MUET	**.73**			.65
Writing EPT	*(.34)*		**.57**	*.50*
Wrinting MUET			**.90**	*.80*
Speaking EPT		**.73**		*.63*
Speaking MUET		**.83**		*.73*
Proportion of Variance	*.31*	*.18*	*.17*	*.66*

In this 3-Factor model, the total amount of variance accounted for was 65.90% with Factor 1 accounting for 30.80%, Factor 2 accounting for 18.40% and Factor 3 accounting for 16.80% of the total variance. The communalities ranged from .50 to .80, which means that this analysis accounted for 50.20% to 80.40% of the variance in each of the eight subtests. This meant that there was a moderate to high correlation between the subtests and the factors. This solution also added a non-trivial amount of variance as the 65.90% of the variance accounted for in this model was 11.20% more than the 54.70%

accounted for in the two-factor model. Furthermore, the second and third factor accounted for 18.40% and 16.80% of the variance respectively, and at least two tests loaded above 0.30 on each factor.

The results of the 3-Factor analysis generally supported both convergent and divergent validity to a certain degree. The EPT and the MUET writing tests and speaking tests loaded on Factor 3 and 2 respectively, thus indicating convergent validity and divergent validity. However, the reading and listening tests did not demonstrate the same pattern as they all loaded on the same factor. A cross loading of the Writing EPT (Factor 1 & Factor 3) was also noted.

In the 4-Factor analysis (see Table 6), 74.70% of the total variance of all the eight subtests were accounted for with Factor 1 accounting for 29.20%, Factor 2 accounting for 17.80%, Factor 3 accounting for 14.50% and Factor 4 accounting for 13.10% of the variance. The main difference between this analysis and the 3-Factor analysis was that the writing subtests loaded on different factors – the EPT on Factor 3 and the MUET on Factor 4. The Listening EPT and the Reading EPT cross loaded on Factors 1 and 3.

Table 6: *Results of the Four-Factor Analysis after Varimax Rotation*

Tests	Factor 1	Factor 2	Factor 3	Factor 4	*h*
Listening EPT	**.68**		*(.36)*		.66
Listening MUET	**.83**				.70
Reading EPT	**.74**		*(.36)*		.71
Reading MUET	**.73**				.67
Writing EPT			**.91**		*.91*
Wrinting MUET				**.96**	*.96*
Speaking EPT		**.74**			*.64*
Speaking MUET		**.83**			*.74*
Proportion of Variance	*.29*	*.18*	*.15*	*.13*	*.75*

This 4-Factor analysis was carried out based on the theory that "language proficiency consists of distinct ability" (Bachman et al., 1988, p. 155) and that the subtests testing the four language skills would load

on separate factors (divergent validity). However, it is apparent that divergent validity was only partially supported. The reading and listening subtests still loaded on the same factor, indicating that they were highly intercorrelated and working as one variable, and hence measuring the same latent trait.

DISCUSSION

The analyses show that there were significant correlations between most of the eight sub-sets of the two tests. The correlation coefficients of the two tests, however, were not high enough for the two tests to be interchangeable. All of the sub-tests of the EPT were found to be significantly correlated, with the highest between the reading and listening components and the lowest between speaking and writing. On the other hand, for the MUET sub-tests, lower correlation coefficients were found. As in the EPT, the highest correlation was between reading and listening, but no significant correlation was observed between speaking and writing. Overall, there is sufficient evidence to support the notion that different aspects of language proficiency are inter-related, as attested in the current literature.

When the patterns of student performance in the eight sub-sets of the four different language constructs were subjected to further analyses (exploring the patterns of performances through factor analysis), there was evidence of convergent and divergent validity. The EPT and MUET speaking subtests loaded on one factor while the writing subtests loaded on another. However, it is important to highlight that the EPT writing subtest cross-loaded on Factor 1 in the 2- and 3-factor models along with the reading sub-tests. This suggests that the skills tapped in the EPT writing sub-test were closely related to those in the reading sub-tests. The most likely explanation for the finding is that the reading passage for summary writing in the EPT writing sub-test was based on the reading passage found in the EPT reading sub-test.

From the 2-factor, 3-factor and 4-factor analyses, it was observed the reading and listening sub-tests of the MUET and EPT loaded on the same factor (Factor 1). This strongly suggests that these two sub-tests were measuring the same underlying constructs. To understand why the listening and reading subtests loaded on the same factor, a content analysis of the two sub-sets was conducted. It was discovered that the tasks in both the reading and listening sub-sets of

MUET and EPT tapped similar types of abilities. First, the ability to perform the listening comprehension task to a certain extent depended on the ability to process written information (reading ability), as most of the listening tasks involved some amount of reading. Second, both the EPT and the MUET listening and reading subtests were of the same test format: multiple-choice questions.

These results suggest that there is no clear evidence of divergent and convergent validity in support of the two subtests as the characteristics of the listening comprehension construct tapped in the MUET and EPT were similar to those tapped in the reading comprehension construct. Interestingly enough, other empirical research has also come to similar conclusions (Buck, 1992; Bae & Bachman, Buck, 1992; 1998; Freedle & Kostin, 1994; 1999). This suggests that some revisions are necessary, particularly to the listening subtests, to ensure that the constructs of interest are clearly defined to measure what they are intended to measure.

CONCLUSION

Based on the analyses, it is concluded that although the MUET and the EPT appear to measure the same constructs, they are not interchangeable. The intercorrelation matrix shows that the degree of correlation among the relevant subskills was not sufficient for them to be truly comparable and interchangeable. The study also found that for both tests, the reading and listening subtests loaded on the same factor. The developers of the two tests may need to look into these constructs to improve the tests.

REFERENCES

Alderson, J. C., & Banerjee, J. (2002). Language testing and assessment (Part 2). *Language Teaching, 35*, 79- 113.

Alderson, J. C., Clapham, C. & Wall, D. (1995). *Language Test Construction and Evaluation.* Cambridge: Cambridge University Press.

American Psychological Association. (1985). *Standards for Educational and Psychological Testing.* Washington D.C.: Author.

Bachman, L. F. (1990). *Fundamental Considerations in Language Testing.* Oxford: Oxford University Press.

Bachman, L. F., Davidson, F., Ryan, K. & Choi, I. C. (1995). *An Investigation into the Comparability of Two Tests of English as a Foreign Language.* Cambridge: Cambridge University Press.

Bachman, L. F., Kunnan, A., Vinniarajan, S., & Lynch, B. (1988). Task and ability analysis as a basis for examining content and construct comparability in two EFL proficiency test batteries. *Language Testing, 5*(2), 128-159.

Bae, J. & Bachman, L. F. (1998). A latent variable approach to listening and reading: testing factorial invariance across two groups of children in the Korean/ English two-way immersion irogram. *Language Testing, 15*(3), 380-414.

Buck, G. (1992). Listening comprehension: Construct validity and trait characteristics. *Language Learning, 42*(3), 313-357.

Criper, C. & Davies, A. (1988). *English Language Testing Services ELTS Validation Project Report, Research Report 1.* British Council Local Examination Syndicate/University of Edinburgh.

Cumming, A. (1996). The concept of validation in language testing, in Cumming, A. & Berwick, R. (eds.), *Validation in Language Testing.* Clevedon: USA: Multilingual Matters.

Freedle, R,. & Kostin, I. (1994). Can multiple choice reading comprehension test be construct valid? *Psychological Science, 5*, 107-110.

Freedle, R,. & Kostin, I. (1999). Does text matter in multiple-choice test of comprehension? The case for the construct validity of TOEFL minitalks. *Language Testing, 6*(1), 2-32.

Gorsuch, R.L.(1983). Factor analysis (2nd ed.). Hillsdale, N.J.: Lawrence Erlbaum.

Gray, T. B. (1997). Controversies regarding the nature of score validity: still crazy after all these years. *Paper presented at the Annual Meeting of The Southwest Educational Research Association, Austin, January 1997,* available at http://ericae.net/ft/tamu/valid.htm

Gronlund, N. E. (1981). *Measurement and Evaluation in Teaching.* New York.: MacMillan.

Henning, G. (1987). *A Guide to Language Testing- Development, Evaluation, Research.* London: Newbury House.

Hughes, A. (1989). *Testing for Language Teachers.* Cambridge: Cambridge University. Press.

Kunnan, A. J. (1995). *Test Taker Characteristics and Test Performance: A structural modelling approach. (Studies in Language Testing Series, vol 2).* Cambridge: University of Cambridge Local Examinations Syndicate/Cambridge University Press.

Messick, S. (1989a). Meaning and values in test validation: the science and ethics of assessment. *Educational Researcher, 18*(2), 5-11.

Messick, S. (1989b). Validity, in Linn R.L (ed.), *Educational Measurement.* New York: McMillan, pp 13-103.

Pett, M.A., Lackey, N.R., & Sullivan, J.J. (2003). Making Sense of Factor Analysis. The Use of Factor Analysis for Instrument Development in Health Care Research. Thousand Oaks, CA: Sage.

Stapleton, C.D. (1997). Basic concepts and procedures of confirmatory factor analysis. Paper presented at the annual meeting of the Southwest Educational Research Association, Austin, January 1997. Retrieved November 1, 2004 from http://ericae.net/ft/tamu/Cfa.htm

	PAPER 1 LISTENING	PAPER 2 SPEAKING	PAPER 3 READING COMPREHENSION	PAPER 4 WRITING
TIME	½ hour	-	2 hours	1 ½ hours
SECTION / QUESTION	3 sections 15questions (5 questions per section)	2 tasks Task A – individual presentation Task B – group discussion	50 Questions • CLOZE (MCQ) – 15 items • Non-linear text (graph) – 3 items • Information organization – 4 items • 4 reading comprehension passages (Skimming, scanning, reference, vocabulary, main idea, inference, etc)	2 Questions • Summary writing (100 words) • Essay writing (250 words)
	Each recording is played twice	Task A: • 2 minutes for preparation • 2 minutes for presentation Task B: • 2 minutes for preparation • 10 minutes for group discussion	• All multiple choice questions • Objectively marked	• Subjectively marked • Single rater • Holistic Scoring • Scoring descriptors (task fulfillment and language)

APPENDIX 1

	LISTENING	SPEAKING	READING COMPREHENSION	WRITING
TIME	1 hour	10 minutes oral interview	1 ½ hours	2 hours
SECTION / QUESTION	5 sections: • Fill in the blanks with stressed words • Form filling • Filling in missing information • • Note-taking (based on a lecture)	5 sections: • Ice-breaking • Short talk (candidates are given 5 minutes to prepare) • Question time • Extended conversation • Leave-taking	• 50 questions based on three thematically related reading passages • Question type – MCQ and summary close	3 sections: • Data Interpretation (10 marks) • Essay writing (25 marks) • Summary writing (15 marks)
		Holistic scoring based on band descriptors	Objective scoring and one-word answer	Based on the same theme as reading. • Essay based on reading theme Scoring (analytic scale) • Summary based on one of the reading passages. Scoring (content and language/ organization)

APPENDIX 2