Validation and Invariance of Factor Structure of the ECPE and MELAB across Gender

Shudong Wang
Harcourt Assessment Inc.

The purpose of this study is twofold: (1) to validate the internal structure of the Examination for the Certificate of Proficiency in English (ECPE) and the Michigan English Language Assessment Battery (MELAB), and (2) to examine the invariance of the factor structures of both the MELAB and the ECPE across gender. For both the MELAB and the ECPE, a one-factor, or one-dimensional model was postulated and tested. The results for both tests support one-factor models. The study results also show that the internal structure of the MELAB and the ECPE are equivalent across male and female examinees, which implies that the two tests are fair across gender groups. This study supports the claim that the total score of the MELAB measures "proficiency in English as a second language for academic study" (English Language Institute, 2003) and the claim that the total score of the ECPE measures English language proficiency for admission to North American colleges and universities.

The construct underlying a test is a theoretical representation of the underlying trait, concept, attribute, process, or structures that the test is designed to measure (Cronbach, 1971; Messick, 1989). Factorial validity (Guilford, 1946), or the investigation of the factor structure underlying a test, can be a valuable component of validity evidence (Messick, 1995). Validity, according to the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 1999), is the most important consideration in test development and evaluation. Fairness is also required by the *Standards*; "Regardless of the purpose of testing, fairness requires that all examinees be given a comparable opportunity to demonstrate their standing on the construct(s) that test is intended to measure" (p. 74). In seeking evidence of test fairness, the researcher should address whether the test measures the same construct in all relevant subgroups of the populations. Fairness is closely related to the factor structure validity of the test. Factorial structure analysis can be used not only to evaluate the dimensionality of an exam, but also to provide evidence of fairness. Similarity of factor structure across gender groups, for example, suggests that the test measures the same construct(s) for males and females. Different factor structures could imply that different constructs are being measured for the two groups. If evidence of differential factor structures is found, further investigation is needed. Differential factor structures for subgroups of examinees per se cannot tell which group's scores are more valid, nor can they explain why group differences occur. They can only serve as a flag to identify where psychological constructs may be structured differently over different subpopulations.

One of the goals of construct validation of test scores is to capture the important aspects of the internal construct. Several studies (Jiao, 2004; Saito, 2003; Wagner, 2004) have been conducted to examine the internal construct validity or dimensionality of the Examination for the Certificate of Proficiency in English (ECPE) and of the Michigan English Language Assessment Battery (MELAB). However, these studies only focused on partial sections of the tests, such as the Cloze section (Jiao, 2004; Saito, 2003), the Listening section (English Language Institute, 1994; 2003; Wagner, 2004), and the Grammar/Cloze/Vocabulary/Reading (GCVR) section (English Language Institute, 1994; 2003; Jiao, 2004). Because both the ECPE and the MELAB report the total/average/final test (scale) scores as major evidence for their uses and interpretations (in fact, the ECPE is awarded only to those who obtain passing scores on all five sections), it is very important to gather internal structure validity evidence to support the claim that the total score of MELAB really measures the "proficiency in English as a second language for academic study" (English Language Institute, 2003) and the claim that the total score of ECPE measures English language proficiency for admission to North American colleges and universities. Despite substantial investment in test development and the establishment of content validity of both the ECPE and the MELAB, there is surprisingly little published research describing factorial or internal construct validity of the whole tests.

Previous studies (English Language Institute, 1994; 2003; Saito, 2003; Wagner, 2004) did report factor analysis results of the Listening and GCVR sections in the MELAB and of Listening and Cloze sections in the ECPE. However, the analysis was done at either item level or subtest level using testlet or component scores. This study makes full use of information from multiple subtests and examines English language proficiency construct validity taking these subtests as a whole, and thus offers a new perspective to evaluate construct validity. Furthermore, this study tests the degree of construct equivalence across gender groups. The purposes of this study are first to validate the internal structure of the ECPE and the MELAB, and then to examine the invariance of factor structure of both the MELAB and the ECPE across gender.

Method

Sample and Instrument

MELAB

The MELAB data used in this study are from 216 examinees who took one particular combination of Listening and GCVR test forms, referred to here as Form X and Form Y. The testing for both forms took place from April 4, 2003, to June 6, 2004. There are 19 possible Composition scores, ranging from 1 to 19 for analysis purposes in this study. In addition to a Composition item (essay), there are 150 multiple-choice (MC) items for measuring other language skills. MC items 1 through 50 are 3-option Listening items. Among these Listening items, there are 10 short question items, 16 short statement items, 8 emphasis items, 7 lecture comprehension items (from one lecture), and 9 conversation comprehension items (from one conversation), administered in sequence. The remaining 100 4-option MC items measure Grammar, Cloze, Vocabulary, and Reading (GCVR). Among the 100 items, there are 30 Grammar items, 20 Cloze items (from one passage), 30 Vocabulary items (the first 14 are synonym type, the next 16 are completion type), and 20 Reading comprehension items (5 each from four passages).

ECPE

The ECPE data were collected from 2011 examinees during the 2005 administration (data with missing values were deleted). The data were from test centers mostly located in North and South America, while the largest group of examinees, tested in Greece, was not included in this study. There is one speaking item with a rating scale from 1 to 4. Among the total 150 MC items, 50 are Listening items and 100 are GCVR items. Included in the Listening items are 14 short conversation items, 21 short question items, and 15 radio interview items. Included in the 100 GCVR items are 30 Grammar items, 20 Cloze items, 30 Vocabulary items, and 20 Reading comprehension items.

MELAB and ECPE Data Analysis

To investigate the factor structure of the MELAB and the ECPE and the equivalence of the factor structure for each test across gender groups, a series of analyses were conducted, as follows.

First, descriptive statistics, internal consistency, and intercorrelations of raw scores of subtests/tests were used to provide general information about the test scores.

Second, a series of exploratory factor analyses (EFA) using classical factor analysis procedures was conducted for the internal structural validity study. For the EFA of the MELAB, the potential models include the measurement models that use subtests (Writing, Listening, Grammar, Cloze, Vocabulary, Reading, and Speaking) and sub-subtests (Writing, short question, short statement, emphasis, lecture, Grammar, Cloze, synonym completion, reading 1–4, and Speaking) as observed variables. For the EFA of the ECPE, the potential models include the measurement models that use subtests (Speaking, Listening, Listening Interview, Grammar, Cloze, Vocabulary, and Reading) and sub-subtests (Speaking, short conversations, short questions, listening radio interview 1–3, Grammar, Cloze, Vocabulary, and reading comprehension passages 1–4) as observed variables.

Third, after identifying a potential model that best explains the data in terms of theory and model fit, a confirmatory factor analysis (CFA) using structural equation modeling (SEM) was used to test the invariance of the factorial model. For the purpose of cross-validation, subjects were randomly split into two samples to form a calibration and a validation sample (Byrne, 2001). One of the purposes for using a cross-validation strategy is to assess the reliability of model fit. Having chosen a SEM model that is best for a particular sample of examinees, it is not proper to automatically assume that this SEM model can be reliably applied to other samples of the same population. However, the model that fits the data using the calibration sample can be further validated by using another sample from the same population. In order to evaluate the adequacy of the factor models to fully account for the relationships among observed variables, a series of SEMs with the maximum likelihood estimation was conducted on the calibration sample. Once model fit for each calibration sample was determined, the invariance of the model structure for the validation samples was investigated across gender. All tests of model invariance begin with a global test of the equality of covariance structures across groups (Joreskog, 1971). The data for all groups were analyzed simultaneously to obtain efficient estimates (Bentler, 1995). Then, a series of nested constraints was equally applied to the same parameters across gender groups in order to detect the configuration and factor pattern difference across gender groups. The constraints used include, from weaker to stronger: (1) model structure, (2) model structure and factor loadings, and (3) model structure, factor loadings, and unique variance.

Evaluation of Model Fit

Changes in goodness-of-fit statistics have been examined to detect differences in structure parameters. Several well-known goodness-of-fit indices were used to evaluate model fit: the chi-square χ^2, the comparative fit index (CFI), the unadjusted goodness-of-fit indices (GFI), the normal fit index (NFI), the Tucker-Lewis Index (TLI), the root mean square error of approximation (RMSEA) and the standardized root mean square error residual (SRMR).

Goodness-of-fit (GOF) indices provide "rules of thumb" for the recommended cutoff values to evaluate data-model fit. Hu and Bentler (1999) recommend using combinations of GOF indices to obtain a robust evaluation of model fit. The criterion values they list for a model with good fit are CFI > 0.95, TLI > 0.95, RMSEA < 0.06, and SRMR < 0.08 for assessing fit in structural equation modeling. Hu and Bentler offer cautions about the use of GOF indices, and current practice seems to have incorporated their new guidelines without sufficient attention to the limitations noted by Hu and Bentler. Moreover, some researchers (Beauducel & Wittmann, 2005; Fan & Sivo, 2005; Marsh, Hau, & Wen, 2004; Yuan, 2005) believe that these cutoff values are too rigorous and the results by Hu and Bentler may have limited generalizability to the levels of misspecification experienced in typical practice. In general practice, a "good enough" or "rough guideline" approach is that for absolute fit indices and incremental fit indices (such as CFI, GFI, NFI, and TLI), cutoff values should be above 0.90 (0.90 benchmark) and for fit indices based on residuals matrix (such as RMSEA and SRMR), values below 0.10 or 0.05 are usually considered adequate.

For the group comparisons with increased constraints, the χ^2 value provides the basis of comparison with the previously fitted model. A non-significant difference in χ^2 values between nested models reveals that all equality constraints hold across the groups. Therefore, the measurement model remains invariant across groups as the constraints are increased. Sample size must be taken into account, however, in interpreting a significant χ^2. A significant χ^2 does not necessarily indicate a departure from invariance when the sample size is large. All analyses were conducted using AMOS 4.0 (Arbuckle & Wothke, 1999) and SAS. All models were identified by fixing the one factor variance at 1.0.

Results

Summary Descriptive Statistics

Tables 1 summarizes the n-counts, median, minimum, maximum, range, and the first four moments describing the distributions of subtest and test raw scores for the MELAB by group (total, male, and female groups). The four moments are: mean, standard deviation, skewness, and kurtosis. Table 2 provides the same information for the ECPE test. There are unequal n-counts across gender for both the MELAB and the ECPE; for the MELAB, female examinees have slightly higher mean test scores than male examinees, while for the ECPE, the mean test score of male examinees is slightly higher than that of female examinees. For both tests, female examinees have less variation of test scores than male examinees.

Reliability of Subtests and Test Scores

Internal consistency coefficients were computed for the subtests and the total test scores for both the MELAB and the EPCE, and are shown in Table 3. The coefficient alpha can be considered as the mean of all possible split-half coefficients. All reliability coefficients of subtests and test scores range from moderate (0.85) to high (0.95).

Linear Correlations among Subtest Scores

It is expected that all subtest scores within each test would show some degree of correlation to one another, based on the assumption that the subtests measure general language proficiency. On the other hand, since each subtest measures different skills, it would be expected that the intercorrelations of subtests would not be very high. Pearson's correlation coefficient was used to analyze the relationship between subtest scores. Table 4 reports the intercorrelations among the subtests of the MELAB, and Table 5 summarizes the intercorrelations among the subtests of ECPE. For the MELAB, the correlations between Composition scores and the rest of the subtest scores are very low due to the restriction of the scale range for the Composition score.

Exploratory Factor Analysis

Exploratory factor analysis without rotation (orthogonal solution) was used to extract the language proficiency factor underlying both MELAB and ECPE test items. Figures 1 and 2 show the scree plots of eigenvalues for the MELAB and ECPE, respectively, based on subtest scores. A similar pattern was observed for both tests. In each plot there was one large break in the data following factor 1 and then the plots flatten out beginning with factor 2. This indicates only factor 1 was dominant and accounted for meaningful variances and only this factor should be retained. The eigenvalues from the EFA for both the MELAB and the ECPE are given in Tables 6 and 7, respectively. For the MELAB, the first factor had an eigenvalue of 3.54 and accounted for approximately 60% of the common variances. For the ECPE, the first factor had eigenvalue of 2.20 and accounted for more than 90% of the common variances. Hattie (1985) suggests using the difference of eigenvalues between the first factor and the second factor divided by the difference of eigenvalues between the second factor and the third to evaluate unidimensionality. If the ratio is large (usually larger than 3), the first factor is relatively strong. Both MELAB and ECPE EFA results show that the ratio high: 5.69 for the MELAB and 37.72 for the ECPE. Lord (1980) argues that a rough procedure for determining unidimensionality was the ratio of first to second eigenvalues and inspection as to whether the second eigenvalue is not much larger than any of the others. Based on both criteria, the results in Tables 6 and 7 support the statement that there is only one meaningful factor as a dominant factor in both the MELAB and the ECPE data.

Table 1.Descriptive Statistics of MELAB Total Test and Subtest Scores for All, Female, and Male Students

Sample	Test/SubTest	N	Mean	Std Dev	Median	Minimum	Maximum	Range	Skewness	Kurtosis
All	Composition	216	11.11	2.68	11.00	4.00	19.00	15.00	0.47	0.36
	Listening	216	32.31	8.32	33.00	5.00	49.00	44.00	-0.24	-0.16
	Grammar	216	16.28	6.36	16.00	3.00	30.00	27.00	0.22	-0.64
	Cloze	216	10.35	3.79	10.00	1.00	20.00	19.00	0.00	-0.48
	Vocabulary	216	18.22	6.54	19.00	2.00	30.00	28.00	-0.12	-0.84
	Reading	216	10.92	4.00	10.00	2.00	20.00	18.00	0.12	-0.49
	Total Test	216	88.07	24.45	86.00	28.00	147.00	119.00	0.19	-0.39
Female	Composition	147	10.95	2.78	11.00	4.00	19.00	15.00	0.42	0.06
	Listening	147	32.71	7.81	33.00	12.00	49.00	37.00	-0.29	-0.29
	Grammar	147	16.45	6.35	16.00	4.00	30.00	26.00	0.32	-0.66
	Cloze	147	10.61	3.71	10.00	1.00	20.00	19.00	-0.05	-0.36
	Vocabulary	147	18.36	6.35	19.00	5.00	30.00	25.00	-0.08	-0.82
	Reading	147	11.27	3.91	11.00	2.00	20.00	18.00	0.07	-0.47
	Total Test	147	89.40	23.39	86.00	30.00	142.00	112.00	0.27	-0.42
Male	Composition	69	11.46	2.45	11.00	6.00	19.00	13.00	0.78	1.38
	Listening	69	31.45	9.31	30.00	5.00	49.00	44.00	-0.09	-0.09
	Grammar	69	15.91	6.43	16.00	3.00	29.00	26.00	0.03	-0.62
	Cloze	69	9.78	3.93	10.00	2.00	19.00	17.00	0.12	-0.60
	Vocabulary	69	17.93	6.97	19.00	2.00	30.00	28.00	-0.16	-0.92
	Reading	69	10.16	4.12	10.00	2.00	20.00	18.00	0.28	-0.40
	Total Test	69	85.23	26.53	85.00	28.00	147.00	119.00	0.16	-0.42

Table 2. Descriptive Statistics of EPCE Total Test and Subtest Scores for All, Female, and Male Students

Sample	Test/SubTest	N	Mean	Std Dev	Median	Minimum	Maximum	Range	Skewness	Kurtosis
All	Speaking	2011	3.20	0.62	3.00	1.00	4.00	3.00	-0.25	-0.16
	Listening	2011	39.23	6.85	40.00	14.00	50.00	36.00	-0.77	0.19
	Grammar	2011	21.80	4.61	22.00	7.00	30.00	23.00	-0.40	-0.32
	Cloze	2011	12.56	3.77	13.00	0.00	20.00	20.00	-0.32	-0.45
	Vocabulary	2011	17.45	4.35	17.00	5.00	30.00	25.00	0.27	-0.09
	Reading	2011	15.47	3.32	16.00	1.00	20.00	19.00	-1.08	1.07
	Total Test	2011	97.14	15.50	98.00	42.00	132.00	90.00	-0.39	-0.03
Female	Speaking	1179	3.23	0.61	3.00	1.00	4.00	3.00	-0.23	-0.19
	Listening	1179	39.30	6.67	40.00	16.00	50.00	34.00	-0.74	0.08
	Grammar	1179	21.91	4.50	22.00	8.00	30.00	22.00	-0.40	-0.23
	Cloze	1179	12.17	3.74	12.00	0.00	20.00	20.00	-0.27	-0.44
	Vocabulary	1179	17.14	4.31	17.00	5.00	30.00	25.00	0.28	0.00
	Reading	1179	15.28	3.28	16.00	3.00	20.00	17.00	-1.00	0.76
	Total Test	1179	96.86	15.30	98.00	42.00	132.00	90.00	-0.40	0.11
Male	Speaking	832	3.15	0.64	3.00	1.00	4.00	3.00	-0.25	-0.15
	Listening	832	39.13	7.09	40.00	14.00	50.00	36.00	-0.80	0.28
	Grammar	832	21.65	4.75	22.00	7.00	30.00	23.00	-0.38	-0.44
	Cloze	832	13.12	3.74	14.00	1.00	20.00	19.00	-0.41	-0.42
	Vocabulary	832	17.88	4.37	18.00	5.00	30.00	25.00	0.25	-0.20
	Reading	832	15.74	3.36	17.00	1.00	20.00	19.00	-1.21	1.59
	Total Test	832	97.55	15.78	99.00	46.00	132.00	86.00	-0.38	-0.20

Table 3. Internal Consistency of MELAB and ECPE Tests and Subtests

Subtest/Test	Coefficient Alpha	
	MELAB	ECPE
Listening	.87	.85
GCVR	.94	.90
Total Test	.95	.92

Table 4. Intercorrelations of Raw Score of MELAB Subtests for Total Sample

Subtest	CO	L	G	C	V	R
Composition (CO)	1.00					
Listening (L)	0.17	1.00				
Grammar (G)	0.17	0.67	1.00			
Cloze (C)	0.16	0.62	0.69	1.00		
Vocabulary (V)	0.14	0.52	0.74	0.65	1.00	
Reading (R)	0.20	0.58	0.60	0.70	0.58	1.00

Table 5. Intercorrelations of Raw Score of ECPE Subtests for Total Sample

Subtests	S	L	G	C	V	R
Speaking (S)	1.00					
Listening (L)	0.37	1.00				
Grammar (G)	0.43	0.61	1.00			
Cloze (C)	0.29	0.52	0.62	1.00		
Vocabulary (V)	0.29	0.39	0.58	0.53	1.00	
Reading (R)	0.20	0.53	0.46	0.51	0.38	1.00

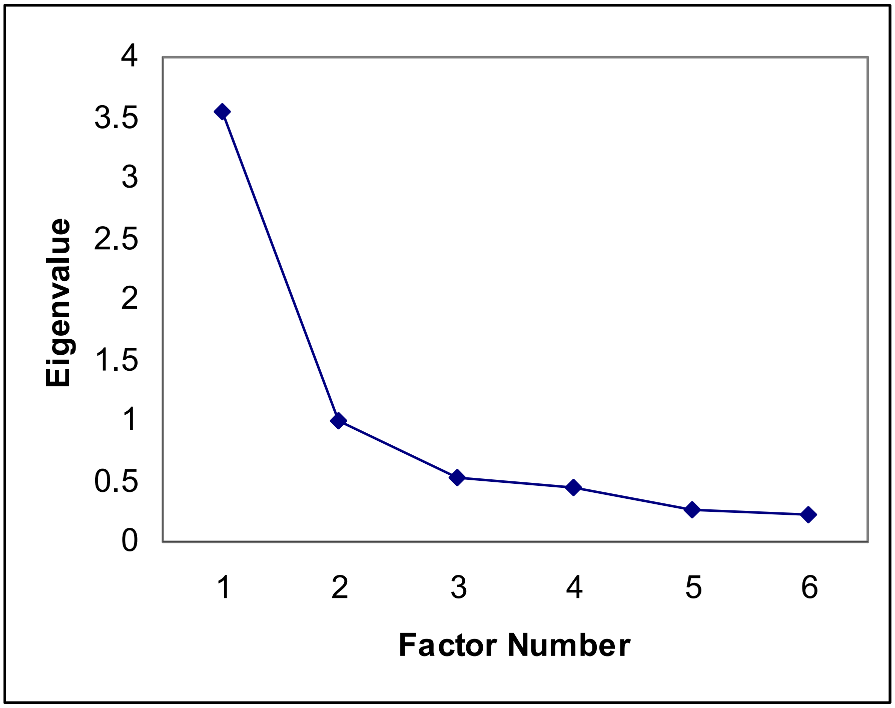

Figure 1. MELAB Factor Scree Plot

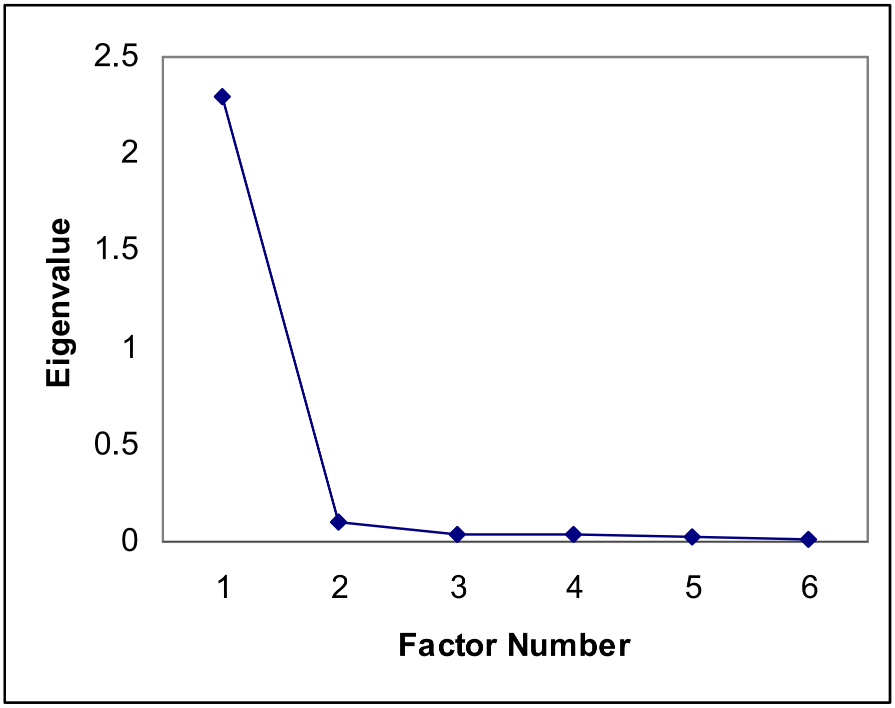

Figure 2. ECPE Factor Scree Plot

Table 6. Eigenvalues and Common Variance Explained by the Factors of MELAB Test

Factor	Eigenvalues	Difference*	% of Variance	Cumulative %
1	3.54	2.55	59.05	59.05
2	0.99	0.46	16.50	75.55
3	0.53	0.08	8.83	84.38
4	0.45	0.20	7.56	91.94
5	0.26	0.03	4.30	96.24
6	0.23		3.76	100.00

*Ratio of difference of Eigenvalues: (E1-E2)/(E2-E3) = 5.69.

Table 7. Eigenvalues and Common Variance Explained by the Factors of ECPE Test

Factor	Eigenvalues	Difference*	% of Variance	Cumulative %
1	2.30	2.20	91.68	91.68
2	0.10	0.06	4.04	95.73
3	0.04	0.00	1.61	97.34
4	0.04	0.02	1.44	98.78
5	0.02	0.01	0.83	99.60
6	0.01		0.40	1.00

*Ratio of difference of Eigenvalues: (E1-E2)/(E2-E3) = 37.72.

Confirmatory Factor Analysis

Evaluation of Model Fit

First, for the purpose of validating the factorial structure of the test, a CFA model was investigated. Second, for the purpose of cross-validation, subjects were randomly split into two groups to form a base calibration sample and a validation sample. Figures 3 and 4 present the one-factor linear models tested using AMOS for the MELAB and the ECPE across original and validation samples. The model-fit statistics for different samples are summarized in Tables 8 (MELAB) and 9 (ECPE).

For the full and cross-validation samples, the majority of values satisfy the Hu and Bentler criteria for the four fit statistics CFI, GFI, NFI, and TLI. All values satisfy the 0.90 benchmark criteria except the value for the ECPE base validation sample. All SRMR index values show that the data fit the model, while all RMSA values show that model fit is not good. All Chi-squares statistics are significant. Based on all model fit indices, the MELAB and the ECPE models fit quite well and are quite comparable for the base calibration and validation samples.

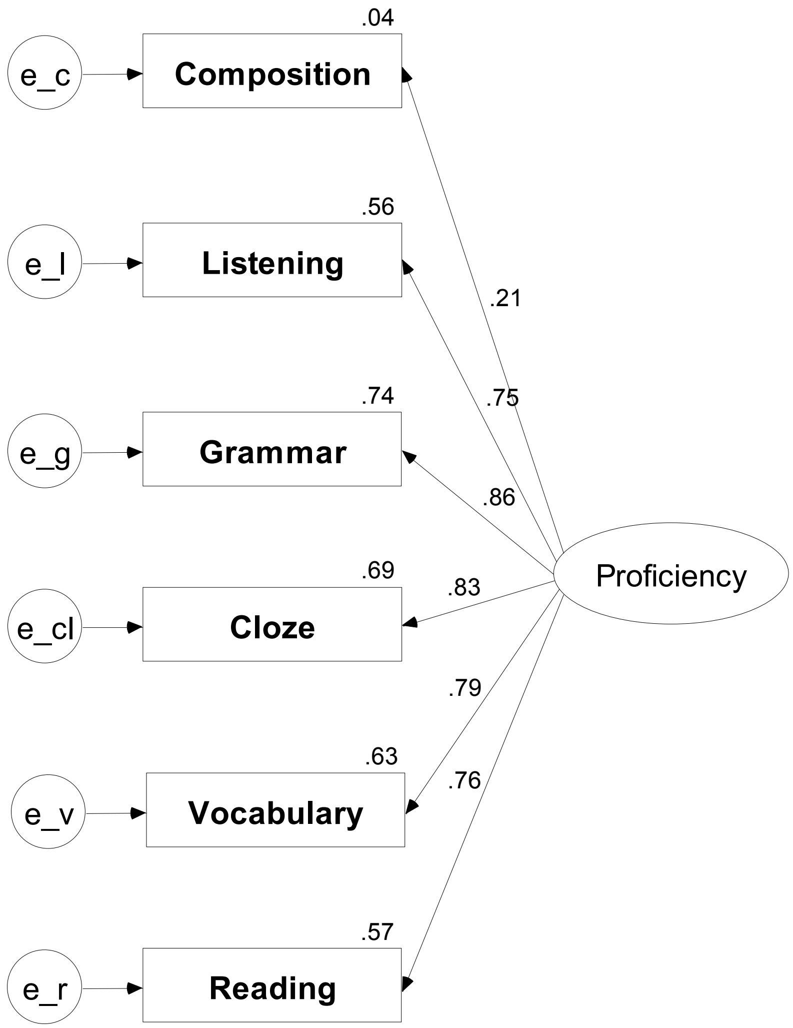

Figure 3. Structure of MELAB Tested with Full Sample

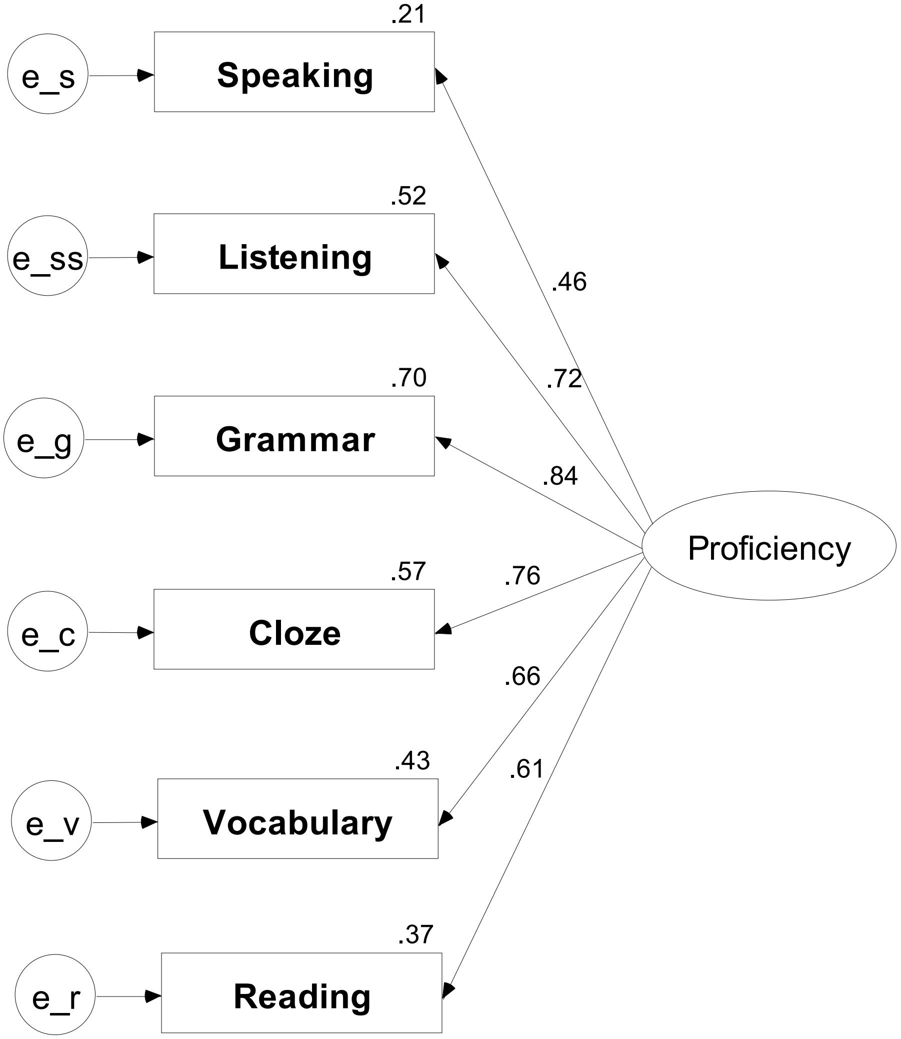

Figure 4. Structure of ECPE Tested with Full Sample

Table 8. Summary of Fit Indices of One-Factor Model of MELAB for Full and Cross-validation Samples

Sample	N	df	χ2	CFI	GFI	NFI	TLI	RMSEA	SRMR
All	216	9	31.97	.96	.95	.95	.94	.11	.03
Base Calibration	108	9	15.38	.98	.95	.95	.97	.08	.05
Validation	108	9	25.82	.93	.91	.96	.91	.11	.05

Table 9. Summary of Fit Indices of One-Factor Model of ECPE for Full and Cross-validation Samples

Sample	N	df	χ2	CFI	GFI	NFI	TLI	RMSEA	SRMR
Full	2011	9	250.11	.95	.96	.94	.91	.12	.04
Base Calibration	1006	9	160.51	.93	.95	.93	.88	.13	.05
Validation	1005	9	102.91	.96	.97	.95	.93	.10	.04

Test of Factorial Structure Equivalence across Gender Samples

The goodness-of-fit indices for a series of nested tests of different degrees of equivalence of the factorial structure across gender under a one-factor model are presented in Tables 10 and 11, for the MELAB and the ECPE, respectively. The specified parameters for each condition were constrained to be equal for both genders. The equivalence of the factor loading and the variance of three factor models (parallel, τ-equivalent, and congeneric) were tested by placing different constraints (equal loading or variance) on two compared models. Two tests are said to be psychometrically parallel if they share an equal amount of factor loading and the specific variance. If two tests have the same factor loading, but different variance, they are τ-equivalent. Congeneric tests have the similar factor loading and variance, but not necessarily to the same degree (Byrne, Shavelson, & Muthén, 1989; Jöreskog & Sörbom, 1979; Loehlin, 2004; Lord, 1957). Some fit values satisfy the Hu and Bentler criteria for the four fit statistics, CFI, GFI, NFI, and TLI, and some do not. All values satisfy .90 benchmark criteria. Both RMSEA and SRMR indices show that the data fit the model based on the 0.10 and 0.05 criteria. All χ^2 differences between nested models are not statistically significant. To select alternative models among the three models tested, a statistically non-significant difference in χ^2 suggests that stronger models are correct. The parallel model showed the best fit to the data, which demonstrates that models for male and female students have structure, factor loading, and variance equivalence.

Table 10. Test of Factorial Equivalence of One-Factor Model for MELAB across Gender

Sample	N	df	χ2	CFI	GFI	NFI	TLI	RMSEA	SRMR
Congeneric (I)	216	19	54.32	.95	.93	.92	.91	.09	.07
Tau-equivalent (II)	216	24	56.82	.95	.93	.92	.94	.08	.07
Parallel (III)	216	30	58.87	.95	.92	.91	.96	.07	.07

The levels of model constraints that were constrained to be equal across gender are:
I. Model structure and latent variable variance.
II. Model structure, latent variable variance, and factor loading.
III. Model structure, latent variable variance, factor loading, and unique variance.

Table 11. Test of Factorial Equivalence of One-Factor Model for ECPE across Gender

Sample	N	df	χ2	CFI	GFI	NFI	TLI	RMSEA	SRMR
Congeneric (I)	2011	19	248.00	.95	.96	.94	.92	.08	.04
Tau-equivalent (II)	2011	24	248.79	.95	.96	.94	.94	.07	.04
Parallel (III)	2011	30	266.78	.95	.96	.94	.95	.06	.04

The levels of model constraints that were constrained to be equal across gender are:
I. Model structure and latent variable variance.
II. Model structure, latent variable variance, and factor loading.
III. Model structure, latent variable variance, factor loading, and unique variance.

Summary

This study examined the internal construct of the MELAB and ECPE tests. For the MELAB, although the speaking section data were not available at the time of this study, the results of overall internal structure are informative and provide insights into the construct validity of test. And, in spite of the missing writing section data, the results also show a clear picture of the internal structure of the ECPE. The one-factor, or one-dimensional model postulated and tested here supports the claim that the total score of MELAB really measures the "proficiency in English as a second language for academic study" (English Language Institute, 2003) and also supports the claim that the total score of the ECPE measures English language proficiency for admission to North American colleges and universities. The study results also show that the internal structure of the MELAB and the ECPE are equivalent across male and female examinees, which implies that the two tests are fair across gender groups. In summary, this study underscores the importance of empirical validation of language tests and provides evidence supporting the validity and fairness of the widely used MELAB and ECPE language exams. It carries the validation process beyond the content-related evidence that often serves as the sole documented support of validity for language exams.

References

American Educational Research Association, American Psychological Association & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: AERA, APA, NCME.

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

Beauducel, A., & Wittmann, W. (2005). Simulation study on fit indices in confirmatory factor analysis based on data with slightly distorted simple structure. *Structural Equation Modeling*, *12*(*1*), 41–75.

Bentler, P. M. (1995). *EQS: Structural equations program manual*. Encino, CA: Multivariate Software, Inc.

Byrne, B. M. (2001). *Structural equation modeling with AMOS*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Bryne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for equivalence of factor covariance and mean structures: The issues of partial measurement invariance. *Psychological Bulletin*, *105*(*3*), 456–466.

Cronbach, L. (1971). Validity. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443–597). Washington DC: American Council on Education.

English Language Institute, University of Michigan. (1994). *The Michigan English language assessment battery technical manual: 1994*. Ann Arbor, MI: English Language Institute, University of Michigan.

English Language Institute, University of Michigan. (2003). *The Michigan English language assessment battery technical manua:l 2003*. Ann Arbor, MI: English Language Institute, University of Michigan.

Fan, X., & Sivo, S. A. (2005). Sensitivity of fit indices to misspecified structural or measurement model components: Rationale of two-index strategy revisited. *Structural Equation Modeling*, *12*(*3*), 343–367.

Guilford, J. P. (1946). New standards for test evaluation*.* *Educational and Psychological Measurement*, *6*, 427–439.

Hattie, J. (1985). Methodology review: Assessing unidimesionality of tests and items. *Applied Psychological Measurement*, *9*, 139–164.

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

Jiao, H. (2004). Evaluating the dimensionality of the Michigan English language assessment battery. *Spaan Fellow Working Papers in Second or Foreign Language Assessment*, *2*, 27–52. Ann Arbor, MI: English Language Institute, University of Michigan.

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

Jöreskog, K. G., & Sörbom, D. (1979). *Advances in factor analysis and structural equation models*. Cambridge, MA: Abt Books.

Lord, F. M. (1957). A significance test for the hypothesis that two variables measure same trait except for error of measurement. *Psychometrika*, *22*, 20.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Marsh, H.W., Kit-Tai Hau and Z. Wen (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling*, *11*(*3*), 320–341.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: American Council on Education, Macmillan.

Messick, S. (1995). Standards-based score interpretation: Establishing valid grounds for valid inferences. In *Proceedings of the joint conference on standard setting for large-scale assessments of the National Assessment Governing Board (NAGB) and the National Center for Education Statistics (NCES)* (Vol. 2, pp. 291–305). Washington, DC: National Assessment Governing Board and National Center for Education Statistics.

Saito, Y. (2003). Investigating the construct validity of the cloze section in the Examination for the Certificate of Proficiency in English. *Spaan Fellow Working Papers in Second or Foreign Language Assessment*, *1*, 39–82. Ann Arbor, MI: University of Michigan English Language Institute.

Wagner, E. (2004). A construct validation study of the extended listening sections of the ECPE and MELAB. *Spaan Fellow Working Papers in Second or Foreign Language Assessment*, *2*, 1–25. Ann Arbor, MI: University of Michigan English Language Institute.

Yuan, K.H. (2005). Fit indices versus test statistics. *Multivariate Behavioral Research*, *40*(*1*), 115–148.