Spaan Fellow Working Papers in Second or Foreign Language Assessment
Copyright © 2008
Volume 6: 77–94

English Language Institute
University of Michigan
www.lsa.umich.edu/eli/research/spaan

Underlying Factors of MELAB Listening Constructs

Minhee Eom
University of Texas–Pan American

ABSTRACT This study examined underlying factors of the listening construct of the Michigan English Language Assessment Battery (MELAB). Confirmatory factor analysis tested the hypothesis that the listening construct had two underlying factors: language knowledge and comprehension. The analysis of the input facet of MELAB resulted in 14 listening ability variables. A total of 2,133 test takers' listening scores were used. The initial model showed several high standard residuals and significant model chi-square values suggesting model respecification. In the respecified model, the measurement errors were allowed to covary while the factor associations remained unchanged. The respecified model showed good fit of both parameters and overall model. Thus the second model of listening construct was accepted as evidence to support the hypothesis. The two-factor model supported the complex nature of listening. The two factors were part of various knowledge sources in L2 listening, and other sources could account for the measurement error covariances.

Listening constructs refer to the ability measured by a given listening test and are different from genetic listening abilities due to the test method effect. It is important to understand what is tested in order to argue for the validity of meaningful inference about listening abilities of test takers.

This study investigates the underlying factors of the listening construct of the Michigan English Language Assessment Battery (MELAB). The underlying factors represent the dimensions of listening abilities and are grounded on theoretical aspects of listening abilities. In order to conceptualize the factors, this study considers language theories, comprehension processing, and test method formats. Language theories provide a theoretical basis for defining the components of the listening construct. Comprehension processing explains a cognitive complexity of the listening construct. Test formats are structural characteristics of a given test that place restrictions in defining the construct.

Background Review

Second language listening is a complex process involving such sources as linguistic knowledge, contextual knowledge, general world knowledge, and co-text knowledge (Buck, 2001). Among these sources, linguistic knowledge is most in need for second language learners. Bachman's model of communicative language ability defines language knowledge with respect to organizational and functional knowledge (Bachman, 1990; Bachman & Palmer, 1996). Organizational knowledge governs the linguistic knowledge such as syntactic, phonological, lexical, and textual knowledge. Functional knowledge refers to social aspects of language use, such as illocutionary and sociocultural knowledge.

When listening abilities are related to Bachman's model, they have both linguistic and functional aspects. Phonological knowledge plays a core role in linguistic knowledge of listening because listeners have to comprehend the aural input. The functional aspect of listening is also important because listeners have to understand the intent of speakers.

Numerous researchers proposed listening taxonomies (Brown & Yule, 1983; O'Malley, Chamot, & Kupper, 1989; Dunkel, Henning, & Chaudron, 1993; Mendelsohn, 1994; Richards, 1983; Rost, 1994). These taxonomies describe various listening activities in different contexts. While agreeing that listening comprehension is a complex, multidimensional process, they lack an agreement on the components of listening.

For general listening skills, Richards (1983) proposed micro- and macroskills of listening based on an analysis of a variety of sources. It was suggested that the microskills were required for conversational listening while the macroskills were relevant to academic listening. Rost (1994) provided an extensive list of listening skills for the curriculum design of listening classes. However, theoretical taxonomies lacked empirical evidence (Buck, 1995).

In testing, Powers (1985) found that listening skills were considered important to academic success across disciplines. Powers conducted surveys of the faculty members, students, and admissions officers at universities and found various listening activities important to academic success. These listening activities were related to identification of various ideas, information retention, inference, and vocabulary.

Empirical studies in listening suggested various factors affecting listening in testing contexts. Some research studies investigated the textual and item characteristics that were relevant to item difficulty (Freedle & Kostin, 1998; Nissan, DeVincenzi & Tang, 1995; Kostin, 2004). Nissan et al. (1995) studied the stimulus-related and item-related features of TOEFL dialogue items that contributed to item difficulty. The study found five variables that have a significant impact on item difficulty: word frequency, utterance pattern, negative in stimulus, explicit/implicit information, and role of speakers. It also found that combinations of variables had stronger impact on the item difficulty index than any individual variable.

Kostin (2004) replicated the study of Nissan et al. with additional variables. The aim of the study was to provide practical information for the test developers to create easier or more difficult items in TOEFL dialogue tests. The author grouped the contributing factors into macrolevels of classifications: word, sentence, discourse and task-processing level factors. The multiple-regression analysis found some variables significant predictors of item difficulty. Those variables were related to vocabulary, idioms, negations, syntactic structures, and text contents.

Similarly, Freedle and Kostin (1996) conducted a study to predict listening item difficulty for TOEFL minitalk items. Based on the fact that the scores of TOEFL listening were highly correlated with its reading scores, the variables used in the reading study were reused in this study. They included negations, referentials, rhetorical organizers, fronted structures, serial position effects, lexical overlaps, and vocabulary. A set of new variables was introduced in the listening study: emphatic text words, pauses, the redundancy of information, lexical overlap, and topical differences. Using a multiple-regression analysis, the study found significant predictors in characteristics of texts, types of inference, and text/item overlaps.

Other studies supported that item difficulty is affected by item types. Jensen, Hansen, Green, and Akey (1996) showed that item difficulty was manageable in item writing. They investigated the effect of characteristics of texts and items on the item difficulty of a listening test for academic purposes. The study found none of the text-related predictors significant. The authors attributed this nonsignificant finding to the "leveling effect" in item writing. That is, considerations in item writing such as pacing questions appropriately and asking verbatim responses may have leveled out the effect of textual features on item difficulty. Item characteristics may override the difficulty created by textual features, and item characteristics may be the primary source of item difficulty. The authors concluded that the texts did not have a strong influence on the scores as long as the texts were not too technical. In addition, by controlling item types, test writers can use a variety of texts.

Thomson (1993) examined two types of item tasks used in a Russian language test: recall of structurally important information, which was relevant to the main context, and recall of incidental information, which was not relevant to the main context. The recall of incidental information appeared more difficult, but the result was not statistically significant. On the other hand, Shohamy and Inbar (1991) found that less-skilled learners performed better mostly on questions that required identifying details and facts while high-level learners performed well on the questions that required synthesizing information, drawing conclusions, and making inferences.

Some studies had a global view on listening abilities measured by listening tests. Buck and Tatsuoka (1998) and Buck, Tatsuoka, Kostin, and Pelphs (1997), for example, incorporated theoretical taxonomies and empirical data to define listening attributes. Using rule spacing methodology, they investigated the listening attributes measured by the test. The listening attributes believed to underlie the performance on a test were identified in terms of task identification, context, information processing, and response construction, and each item was coded according to which attributes it measured. While the prime attributes indicated independent knowledge states, the interaction attribute indicated co-occurrence of attribute. The prime listening attributes they found significant were related to types of information, information loads, and uses of contextual information, stress patterns, speech rates, inference, and background knowledge, among others (Buck & Tatsuoka, 1998).

Aim of the Study

This study proposes and tests the hypothesis: the listening abilities measured by the MELAB consist of two factors associated with language knowledge and comprehension.

Methods

Data

The English Language Institute at the University of Michigan has provided the data for the present research. The data set has 2,133 test takers' scores of the MELAB listening section. The listening section has 50 items in three parts. The first part has 15 questions, each of which is given after a short utterance of one speaker. The second part has 20 questions, each of which is given after a short dialogue of two speakers. Finally, the third part has 15 questions. Four to five questions are asked after a lengthy monologue passage on academic topics. The data have 0 and 1 values for all the items of each test taker. Test-takers' backgrounds are not provided.

Variables

This study uses 14 listening abilities as variables (see Table 1). This set of 14 abilities is derived from the analysis of the test input. The input facet of the test is characterized by its format and its language (Bachman, 1990; Bachman & Palmer, 1996). The input format includes channel, form, language, length, type, degree of speededness, and vehicle.

The input text length and types of linguistic knowledge are major criteria in identifying the MELAB listening abilities. The listening section of the MELAB shares the same features of the format except for the text length and text types (monologue and dialogue). Parts 1 and 2 have short input texts that have one or a few sentences, whereas Part 3 has long input texts that consist of a few discourse paragraphs. Thus, the input text length is a variation that can affect test takers' performance.

Linguistic knowledge is another criterion. Language knowledge of input is described in terms of organizational characteristics, including grammatical and textual aspects, and pragmatic characteristics including functional and sociolinguistic aspects (Bachman, 1990). The listening abilities of the MELAB include both organization and pragmatic knowledge. The descriptions of listening abilities and the corresponding items are reported in Table 1.

Data Analysis: Confirmatory Factor Analysis

In order to test the stated hypothesis, this study uses confirmatory factor analysis (CFA). Confirmatory factor analysis allows researchers to construct models in advance (Bollen, 1989). Researchers can determine the number of latent variables and have prior hypotheses about the effect of the latent variables on the observed variables in terms of directionality and magnitudes.

This study follows a model-generating situation that allows model respecifications (Joreskog & Sorbom, 1993). The researcher has a tentative initial model and if the initial model does not fit the data well, it is modified and tested again on the same data set. The respecification of each model is theoretically or statistically determined.

Confirmatory factor analysis consists of five stages: model specification, model identification, model estimation, model evaluation, and model respecification (Bollen, 1989; Kline, 2005; Kunnan, 1998). When the first model does not have good fit, then the cycle of five stages starts again with a respecified model until a theoretically and statistically good fit model is found.

Table 1. Variables

Var.	Descriptions of Listening Abilities	No. items	Item Number	Mean	SD
x_1	Decoding various verb tenses	4	1, 7, 9, 32	.640	.280
x_2	Decoding prepositional verbs	4	3, 5, 8, 28	.702	.259
x_3	Decoding key vocabulary	4	4, 11, 23, 33	.727	.275
x_4	Decoding grammatical lexicon	4	6, 12, 14, 15	.746	.270
x_5	Decoding idiomatic expression	3	2, 24, 26,	.669	.303
x_6	Comprehending illocutionary inference stated by a speaker	4	13, 22, 30, 35	.830	.301
x_7	Comprehending conversational inference	3	16, 34, 31	.716	.266
x_8	Processing key information stated by a speaker	3	10, 17, 20	.624	.321
x_9	Processing key information in conversation	2	18, 19	.771	.263
x_{10}	Decoding auxiliary negatives	4	21, 25, 27, 29	.697	.270
x_{11}	Comprehending text-based inference	5	38, 39, 42, 45, 46	.697	.355
x_{12}	Comprehending stated specific details	4	36, 43, 47, 48	.612	.263
x_{13}	Comprehending stated details with explanation or repeated	4	40, 44, 49, 50	.656	.270
x_{14}	Comprehending and recalling the stated details in the beginning of discourses	2	37, 41	.649	.287

Results

All 2,133 test takers' MELAB listening scores are analyzed. The descriptive statistics show that the mean score of the listening test is 34.37 out of 50 with a standard deviation of 9.12. The highest score is 50 while the lowest is 8. The reliability coefficient (Cronbach's Alpha) is 0.90. The CFA analysis uses a covariance matrix of the observed variables, and it is reported in Table 2.

Defining Factors: Language Knowledge and Comprehension

This study conceptualizes two underlying factors of listening: language knowledge and comprehension. In language testing, language knowledge is a major component of the construct to be measured, and knowledge that varies from context to context may contribute to measurement errors (Buck, 2001). Language knowledge includes grammatical, phonological, and lexical knowledge, along with pragmatic knowledge. Comprehension is a cognitive process critical to all the modalities of language use. Second language listening comprehension has three stages (Rost, 2005). The decoding phase recognizes lexical items

Table 2. Covariance Matrix

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}
x_1	1.231													
x_2	0.582	1.143												
x_3	0.637	0.577	1.400											
x_4	0.844	0.781	0.944	1.978										
x_5	0.547	0.492	0.633	0.744	1.117									
x_6	0.633	0.524	0.635	0.814	0.539	1.141								
x_7	0.611	0.542	0.623	0.815	0.506	0.587	1.054							
x_8	0.484	0.425	0.570	0.683	0.442	0.437	0.470	1.098						
x_9	0.697	0.611	0.653	0.912	0.615	0.682	0.677	0.561	1.313					
x_{10}	0.624	0.573	0.643	0.876	0.565	0.602	0.613	0.501	0.691	1.501				
x_{11}	0.588	0.484	0.616	0.777	0.528	0.612	0.598	0.428	0.693	0.592	1.410			
x_{12}	0.399	0.383	0.385	0.492	0.344	0.396	0.409	0.255	0.421	0.375	0.537	1.027		
x_{13}	0.587	0.507	0.634	0.765	0.483	0.640	0.573	0.347	0.658	0.599	0.762	0.599	1.735	
x_{14}	0.368	0.344	0.376	0.449	0.342	0.396	0.375	0.223	0.453	0.413	0.503	0.327	0.515	0.885

and parses propositions. The comprehension phase connects input to relevant knowledge sources. The final phase involves interpretation of the listener in respect to response options.

Language knowledge and comprehension factors of the MELAB listening test are associated with the input component of test method facets. The main classification of test methods are setting, rubric, input, expected response, and relationship between input and response (Bachman 1990; Bachman & Palmer 1996). The input format is described in terms of its channel (aural, visual), form (language, nonlanguage, both), language (native, target, both), length, type (item, prompt), degree of speededness, and vehicle (live, reproduced, both). Among these, the input length varies across the parts of the test while other factors are consistent for the entire listening test.

The difference in the input texts seemed to tap into different aspects of listening abilities. The questions with short texts have listeners use linguistic aspects of listening, whereas the questions with long texts have them use comprehension aspects of listening. The hypothesized latent factors of listening could be interpreted in terms of input text length. However, the text length does not provide us with meaningful understanding of the construct. The latent factors are conceptualized with respect to language theories.

Notations

The factor analysis model uses various symbols and notations presented in Table 3. The listening ability types are observed variables represented by *x*. The language knowledge factor and comprehension factor are latent variables represented by KSI (ξ). The lambda (λ, Λ_x) refers to factor loading coefficients that are the regression coefficients of latent variables (ξ) on the observed variables (x). The DELTA (δ) refers to the measurement errors representing the unique factor of variables *x* uncorrelated with latent variables and with the other variables' DELTA. Variances of latent variables are represented by PHI (ϕ) and free to vary or covary.

Table 3. Symbols and Notations used in Model Specifications

	Symbol	Name	Definition
Variables	x	Ex	Observed variables of latent variables (ξ)
	ξ,	Ksi	Latent variables
Coefficients	λ, Λ_x	Lambda	Coefficients relating observed endogenous variables (x) to latent variables (ξ)
	δ, Δ	Delta	Measurement errors for observed endogenous variables (x)
Covariance	ϕ, Φ	Phi	Covariance of latent variables (ξ)
	θ_δ, Θ_δ	Theta-Delta	Covariance of measurement errors

Initial Model

The hypothesis postulates that listening constructs measured by the MELAB listening test consist of two latent factors. LAG represents the first factor (ξ_1) associated with language knowledge, while COM indicates the second factor (ξ_2) associated with comprehension. The observed variables x are the fourteen listening ability types that this research proposes.

Model Specification

The model specifications identify a total of 31 parameters to be estimated in three matrices: factor loading (Λ_x), covariance of latent variables ($\boldsymbol{\Phi}$), and covariance of measurement errors ($\boldsymbol{\Theta}_\delta$). The parameters to be estimated are called free parameters.

The initial model has the variables from x_1 to x_{10} factored onto LAG (ξ_1), and those from x_{11} to x_{14} onto COM (ξ_2). All variables (x) contain measurement errors (δ), which are uncorrelated with each other and with the latent variables. The linear equation of the measurement model, $x = \Lambda_x\xi + \delta$, shows the relationship between the observed variables and the latent variables.

The measurement errors of the observed variables (δ) in this initial model specification are uncorrelated with each other (COV (δ_i, δ_j) = 0, for all i and j) and uncorrelated with the latent variables (ξ) (COV (ξ_i, δ_j) = 0, for all i and j). Those errors are random and have the expected value equal to zero (E(δ_j) = 0, for all j). Φ (PHI) represents the covariance of latent variables (ξ). The matrix below shows the three covariance coefficients of latent variables to be estimated.

$$\boldsymbol{\Phi} = \begin{pmatrix} \phi_{11} & \\ \phi_{12} & \phi_{22} \end{pmatrix}$$

Model Identifications

Model identifications examine whether a unique solution exists for the parameter coefficients specified in the previous section. Unidentified models allow an infinite number of values for the parameters, which would produce the same covariance matrix. The estimations of not-identified models would result in indeterminacy, arbitrary estimates of the parameters, and meaningless interpretations. Kline (2005: 169–170) states two necessary conditions that any confirmatory factor analysis model has to meet in order to be identified: first, the number of free parameters is less than or equal to the number of observations (i.e., $df_M \geq 0$), and second, every latent variable including the measurement errors and the factors have to have a scale.

The model specified in this study satisfies the first condition. The number of observations equals $v(v+1)$, where v is the number of observed variables. This study has 14*15 = 210 observations, and the initial model has 31 free parameters to be identified. As the number of free parameters is less than the number of observation, it meets the first necessary condition and the model can be identified.

Similar to the first condition, Bollen (1989) suggests the *t*-rule as a necessary condition for model identification. That is, the number of free parameters (t) must be less than or equal to the number of unique elements in the covariance matrix of observed variables x, $t \leq ½(q)(q+1)$. The t refers to the number of free parameters in the residual covariance matrix, $\Theta\delta$. With the number of observed variables equal to q, there are $½(q)(q+1)$ known-to-be-identified elements. In this study, t is 14, less than 105 (½*14*15). Thus, the *t*-rule is satisfied, and the model is identifiable.

The second necessary condition in Kline (2005) regards scaling the latent variable. For scaling latent variables, this study uses a unit variance identification (UVI) constraint (Kline, 2005). The unit variance identification constraint fixes the factor variance to 1.0 and standardizes the factor.

$$\mathbf{\Phi} = \begin{pmatrix} 1.0 & \\ \phi_{12} & 1.0 \end{pmatrix}$$

In this case, all the factor loadings (λ) are free parameters. This method of scaling is simple but only applicable to exogenous factors. The model of this study has only exogenous factors, so the UVI constraint method is applicable.

Fixing the phi $\phi 11$ and $\phi 22$ to 1.00 decreases the total number of parameters by two. Therefore, scaling the latent variables by UVI method results in a decrease of the total number of parameter estimates from 31 to 29.

In sum, the model identification tests confirm that the specified model is identifiable. All the parameter estimations are reported in the following section.

Model Estimations

The parameters are estimated with LISREL software (see Table 4). The number of input variables is 14, which is equal to the number of the observed variables (X-variables). The number of latent variables represented by KSI (ξ) is two, and the number of observations is 2,133.

Table 4. Maximum Likelihood Parameter Estimates of the Initial Model

	Factor loadings (LAMBDA-X)			Measurement Errors	
	LAG (ξ_1)	COM (ξ_2)	t-value	(THETA-DELTA)	
x_1	0.79	-	37.09	VAR(δ_1)	0.60
x_2	0.71	-	33.29	VAR(δ_2)	0.64
x_3	0.83	-	35.89	VAR(δ_3)	0.72
x_4	1.07	-	40.38	VAR(δ_4)	0.83
x_5	0.70	-	33.45	VAR(δ_5)	0.63
x_6	0.77	-	37.63	VAR(δ_6)	0.54
x_7	0.77	-	39.22	VAR(δ_7)	0.47
x_8	0.61	-	28.60	VAR(δ_8)	0.72
x_9	0.87	-	40.26	VAR(δ_9)	0.56
x_{10}	0.80	-	32.73	VAR(δ_{10})	0.86
x_{11}	-	0.89	36.81	VAR(δ_{11})	0.63
x_{12}	-	0.61	27.99	VAR(δ_{12})	0.66
x_{13}	-	0.89	32.66	VAR(δ_{13})	0.93
x_{14}	-	0.57	28.12	VAR(δ_{14})	0.56

Factor variances and covariance (PHI)	
ϕ_{11}	1.00
ϕ_{21}	0.83
ϕ_{22}	1.00

The identified model is evaluated for the specific parameters as well as for the overall model fit.

Model Evaluation

In model evaluation, following Joreskog and Sorbom's (1993) suggestions, first the parameter estimates are examined for unreasonable values or anomalies. The R-square is evaluated as an indicator of the strength of the linear relationship between latent and observed variables. Finally, overall fit of the model is evaluated with various goodness-of-fit indices.

Parameter Evaluation

The parameter evaluation includes *t*-tests on the meaningful associations between the latent factors and the observed variables and the examination of standardized residuals.

Significant Tests

The null hypothesis of *t*-tests is that the factor loading is equal to zero in the population. A nonsignificant factor loading indicates that the involved variable is poor in measuring the underlying factor and is possibly reassigned or dropped (Hatcher, 1994). The *t*-test values show that the associations between the observed variables and the latent variables in the initial model are reasonable. The *t*-values for the factor loadings in Table 4 show that all the factor loadings are significant at $p < 0.001$.

Table 5. Standardized Residuals

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}
x_1	-													
x_2	1.71	-												
x_3	-1.61	-0.53	-											
x_4	-0.50	1.79	4.00	-										
x_5	-0.81	-0.21	4.05	-0.43	-									
x_6	1.70	-1.90	-0.38	-1.08	-0.23	-								
x_7	0.26	0.14	-0.92	-0.38	**-2.89**	-0.51	-							
x_8	-0.26	-0.58	**4.35**	1.75	0.91	**-2.98**	0.05	-						
x_9	0.49	-0.31	**-5.60**	-1.57	0.47	0.88	1.14	**2.20**	-					
x_{10}	-0.75	0.61	-1.13	1.30	0.38	-1.15	0.14	0.71	-0.30	-				
x_{11}	0.10	**-2.40**	0.34	-0.75	0.72	**2.93**	2.53	-1.49	**3.57**	0.15	-			
x_{12}	-0.26	1.64	**-2.12**	**-2.87**	-0.74	0.29	1.59	**-3.47**	-1.40	-1.68	-0.12	-		
x_{13}	-0.32	-1.07	0.87	-1.59	-2.17	**3.73**	0.13	**-5.64**	0.51	0.16	**-2.74**	**4.02**	-	
x_{14}	-0.54	0.73	-0.98	-3.47	0.79	**2.34**	1.09	**-4.47**	**3.14**	**2.18**	0.15	-1.55	0.63	-

Standardized Residuals

The residual matrix is a measure of differences between the sample covariance matrix and the implied covariance matrix. Standard residuals are residuals divided by their standard errors. Ideally, all residuals are to be near zero for a "good" model. Large standard residuals help locate the reasons for poor fit of the model (Joreskog & Sorbom, 2001).

The standardized residuals of the initial model show many values not close to zero (see Table 5). When a covariance matrix is used, the elements in the residual matrix are not standardized in a meaningful way. In general, the element values exceeding 2.00 should be considered large. Several residuals marked in bold have substantially high values, above 2.00. These are problematic and the second model is respecified to reduce these values.

Squared Multiple Correlations (SMC)

The squared multiple correlations indicate the proportion of variance in the observed variables accounted for by the latent factor variables. The initial model shows medium sizes of SMC that range from 0.34 for x_8 to 0.58 for x_4 (see Table 6). The first latent factor, F1, accounts for 34% of the total variance of variable x_8 and 58% of the total variance of x_4.

Table 6. Squared Multiple Correlations for X - Variables

SMC of the Initial Model													
x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}
0.51	0.44	0.49	0.58	0.44	0.52	0.56	0.34	0.58	0.42	0.56	0.36	0.46	0.36

Overall Model Fit

All measures of overall model fit are based on the covariance of the sample (S) and the covariance of the population (Σ). The fundamental assumption of model fit is that the population covariance matrix of observed variables is equal to the covariance matrix written as a function of the parameters of the model, called implied covariance matrix ($\Sigma(\theta)$) (Bollen, 1989). The fit indices estimate the closeness of the sample covariance matrix to the population covariance matrix as function of model parameters.

There are numerous fit indices, making it difficult to select which particular ones to use and which value to report. Kline (2005) suggests four fit indices as a minimum: (1) the model chi-square, (2) the root mean square error of approximation (RMSEA), (3) the comparative fit index (CFI), and (4) the standardized root mean square residual (SRMR) (see Table 7).

Chi-Square (χ2 M) Test

Chi-square is a badness-of-fit measure. A χ^2_M –test assumes the null hypothesis (H_0) that the researcher's model is a perfect fit to the population. Thus, researchers do not desire to reject the null hypothesis. The current study yields $p = 0.000$ for the minimum fit function χ^2 and the null is rejected. As a result, the initial model is found to be not a good fit to the population.

Though the χ^2_M –test concludes that the initial model is not a good fit, it is important to consider other indices of model evaluation because of the limitation in χ^2 model tests. All of the indices based on χ^2 assume multivariate normality of the endogenous variables. In the case of nonnormality, true models will be rejected too often. Also the model χ^2_M is sensitive to sample size. If the sample size is large, the value of χ^2_M tends to reject the model (Kline, 2005).

Table 7. Goodness of Fit Statistics for the Initial Model

Indices	Value	Evaluation
Degrees of Freedom	76	
Minimum Fit Function Chi-Square	269.41	
P-value of chi-square (χ^2_M)	P = 0.000	Not good fit
Root Mean Square Error of Approximation (RMSEA)	0.035	Good fit
90 Percent Confidence Interval for RMSEA	(0.030 ; 0.039)	Good fit
Comparative Fit Index (CFI)	0.99	Good fit
Incremental Fit Index (IFI)	0.99	Good fit
Relative Fit Index (RFI)	0.99	Good fit
Root Mean Square Residual (RMR)	0.027	Good fit
Standardized RMR	0.021	Good fit

Root Mean Square Error of Approximation

The Root Mean Square Error of Approximation (RMSEA) measures the error of approximation, which is concerned with the lack of fit of the model to the population covariance matrix (Kline 2005). Unlike χ^2 tests, the RMSEA does not assume a null hypothesis of a perfect fit of the researcher's model in the population, but the degree of falseness. The RMSEA is a "badness-of-fit" index, in which the higher the value is, the worse the fit is. RMSEA = 0.00 indicates the best fit. In general, a RMESA value smaller than 0.05 represents close approximate fit, while a value between 0.05 and 0.08 indicates a reasonable error of approximation. The RMSEA over 0.10 suggests poor fit.

The RMSEA of the initial model is 0.035 indicating a reasonably good fit of the model in the population.

Confidence interval for RMSEA

The 90% confidence interval for RMSEA reflects the degree of uncertainty associated with the RMSEA. The low bound of the interval is a cut off value for a good fit. That is, if the low bound of RMSEA is smaller than 0.05, the good fit null hypothesis cannot be rejected. Thus, the model has a close approximate fit in the population. On the other hand, the upper boundary is a cutoff value of poor fit. That is, if the upper value of the interval exceeds a cutoff value for a poor fit, such as 0.10, the poor fit null hypothesis cannot be rejected. Thus, an upper bound value larger than 0.10 indicates poor fit.

For this study the boundaries of the 90% confidence interval are at 0.030 and 0.039. The lower boundary is smaller than 0.05, and the upper bound is smaller than 0.10. The null hypothesis of poor fit is rejected indicating it is not a poor fit. Thus, the 90 % Confidence Interval for RMSEA index concludes this model is a good fit in the population.

Comparative Fit Index

Comparative Fit Index (CFI), Relative Fit Index (RFI), and Incremental Fit Index (IFI) assess the relative improvement in fit of the model compared with a baseline model or a null model. In general, values greater than roughly 0.90 indicate reasonably good fit of the researcher's model. In this study, the model shows high comparative fit indices; CFI = 0.99, IFI = 0.99, and RFI = 0.99, and thus indicate a good fit of the model.

Root Mean Square Residuals

The Root Mean Square Residual (RMR) is a measure of the mean absolute value of the covariance residuals. RMR = 0 indicates a perfect model fit, and the higher the value is, the worse the fit is. The Standard Root Mean Square Residual (SRMR) transforms the covariance matrices (unstandardized) into correlation matrices (standardized) and measures the mean absolute correlation residual. Values of the SRMR less than 0.10 are considered reasonable fit. The RMS and SRMS of the initial model are 0.0027 and 0.0021, respectively, supporting a good fit of the initial model.

In summary, the indices of overall model fit demonstrate a mixed outcome. The *t*-tests of factor loadings show reasonable associations between the latent factors and the observed variables, but the standard residuals show several high values suggesting the necessity of model respecification. In overall model fit evaluation, the *p* value of the model chi-square is far lower than the cut off value indicating bad fit of the model. However, all other indices including indicate a reasonable fit.

As a result, it is necessary to respecify the model. The current specifications regarding the association between the latent variables and the observed variables are kept. However, the several measurement errors are set to covary to reduce the unusually high standard residuals.

Respecified Model

The respecified model has some measurement error set free to allow them to covary. The covariance of measurement errors reduces the large standard residuals in the initial model. This change is theoretically and statistically appropriate. In language theories, listening is considered as a complex cognitive process engaging several sources of knowledge (Bachman, 1990; Bachman & Palmer 1996; Buck, 2001). Linguistic knowledge is one of many factors affecting the listening abilities of test takers. As this study examines only two factors, it is reasonable to postulate that other factors treated as errors might be correlated. Statistically, the allowance of error covariance reduces the chi-square indices and produces a better model fit.

The following 15 measurement errors are set to covary in the respecified model: COV(x_4, x_3), COV(x_5, x_3), COV(x_7, x_5), COV(x_8, x_6) , COV(x_9, x_3), COV(x_8, x_3), COV(x_{11}, x_9), COV(x_{11}, x_2), COV(x_{12}, x_8), COV(x_{13}, x_8), COV(x_{13}, x_6), COV(x_{13}, x_{12}), COV(x_{14}, x_4), COV(x_{14}, x_8), and COV(x_{14}, x_9).

Identification of the Respecified Model

First, the model is again examined for a unique solution for the parameter coefficients. The respecified model has 44 free parameters to identify. The t-rule for model identification (with *q* observed variables, $t \leq ½(q)(q+1)$) shows that the respecified model is identifiable. The number of free parameters (t) is 45, less than 105 (½*14*15) known-to-be-identified elements. Thus, the t-rule is satisfied, and the respecified model is uniquely identifiable.

Estimations of the Respecified Model

LISREL is used to analyze the respecified parameter estimates. See Table 8 for the values and Figure 1 for the path diagram of the resulting estimates. The estimations use the unit variance identification (UVI), where the variances of latent variables are set to 1.00. The variance of latent variables is 0.85, an increase of 0.02 from the initial model.

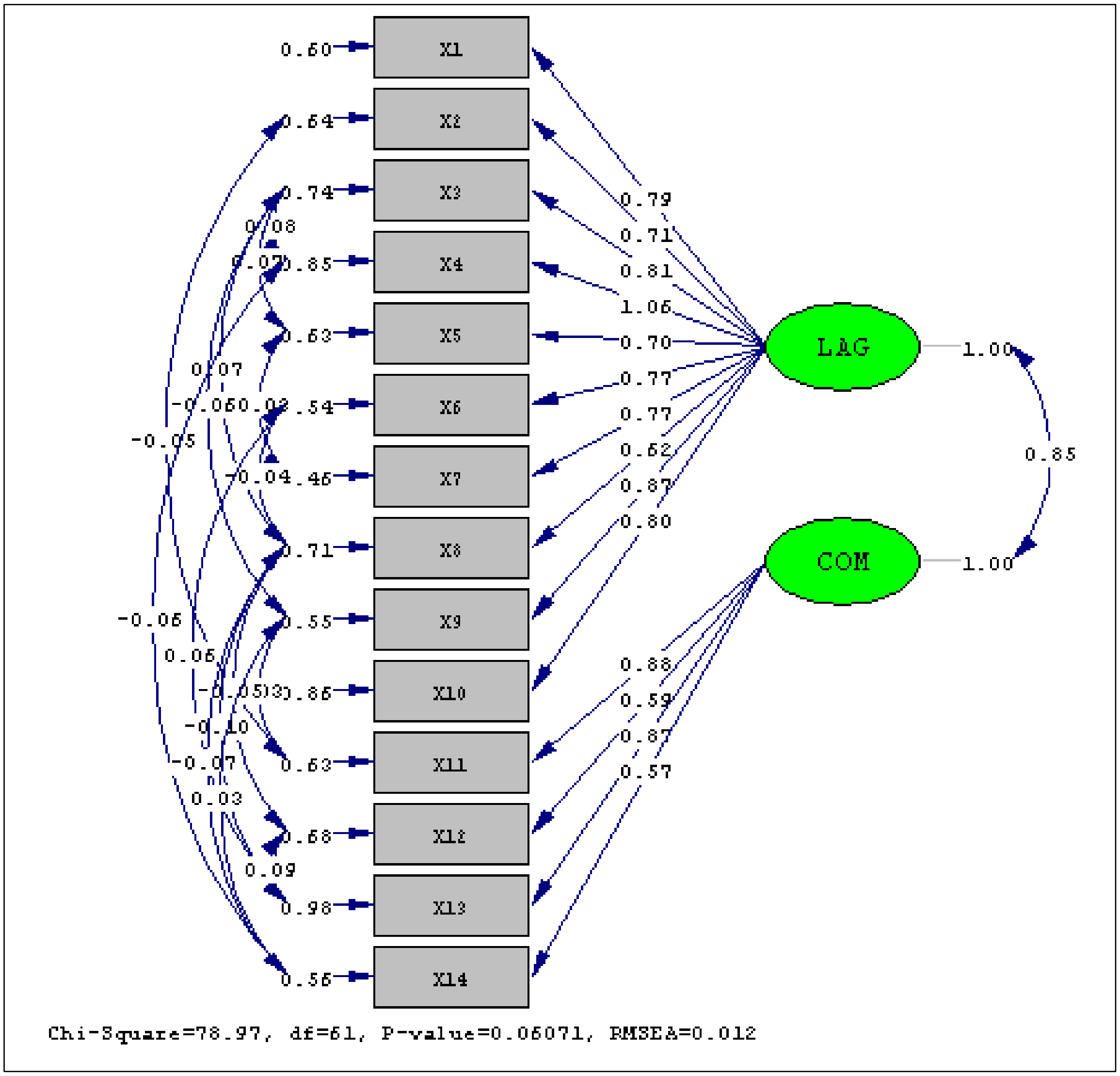

Figure 1. Path Diagrams of the Respecified Model Estimates

Table 8. Estimates of the Respecified Model

	Factor loadings (LAMBDA-X)			Measurement Errors	
	F1(ξ_1)	F2	t-value	(THETA-DELTA)	
x_1	0.79	-	37.09	VAR(δ_1)	0.60
x_2	0.71	-	33.41	VAR(δ_2)	0.64
x_3	0.81	-	34.28	VAR(δ_3)	0.74
x_4	1.06	-	39.86	VAR(δ_4)	0.85
x_5	0.77	-	37.63	VAR(δ_5)	0.63
x_6	0.77	-	37.63	VAR(δ_6)	0.54
x_7	0.77	-	39.40	VAR(δ_7)	0.46
x_8	0.62	-	28.66	VAR(δ_8)	0.71
x_9	0.87	-	40.37	VAR(δ_9)	0.55
x_{10}	0.80	-	32.74	VAR(δ_{10})	0.86
x_{11}	-	0.89	35.69	VAR(δ_{11})	0.61
x_{12}	-	0.59	26.49	VAR(δ_{12})	0.68
x_{13}	-	0.88	29.48	VAR(δ_{13})	0.95
x_{14}	-	0.57	28.07	VAR(δ_{14})	0.56

Covariance of Measurement Errors

Error covariance		t-value	Error covariance		t-value
COV(x_4, x_3)	0.08	3.82	COV(x_{12}, x_8)	-0.04	-2.71
COV(x_5, x_3)	0.07	3.99	COV(x_{13}, x_8)	-0.10	-5.17
COV(x_7, x_5)	-0.03	-2.40	COV(x_{13}, x_6)	0.06	3.14
COV(x_8, x_6)	0.07	3.91	COV(x_{13}, x_{12})	0.08	3.34
COV(x_9, x_3)	-0.06	-3.75	COV(x_{14}, x_4)	-0.05	-3.14
COV(x_8, x_3)	0.07	3.91	COV(x_{14}, x_8)	-0.07	-4.36
COV(x_{11}, x_9)	0.04	2.20	COV(x_{14}, x_9)	0.03	2.12
COV(x_{11}, x_2)	-0.05	-3.19			

Factor variances and covariance (PHI)	
ϕ_{11}	1.00
ϕ_{21}	0.85
ϕ_{22}	1.00

Evaluation of the Respecified Model

The factor loading parameters of the respecified model show large effect sizes for both latent factors. The standardized path coefficients with absolute values less than 0.10 may indicate a “small” effect, values around 0.30 a “typical” or “medium” effect; and with absolute values greater than 0.50 a “large” effect (Cohen, 1988 cited in Klein, 2005).

The parameter evaluations of the respecified model include both the significant tests of factor loadings and the standardized residuals. The t-values (see Table 8) are all higher than 3.291, indicating significant association between the observed variables and the latent factor variables. The standardized residual matrix improved, with only three values larger than 2.00, and the largest absolute value being 3.24.

The overall model fit improves significantly in terms of the chi-square test (see Table 9). Other indices indicate a good fit of the respecified model. Based on the evaluation of the respecified model, it is accepted as evidence to support the hypothesis.

Table 9. Goodness of Fit Statistics of the Respecified Model

Indices	Value	Evaluation
Degrees of Freedom	61	
Minimum Fit Function Chi-Square	79.40	
P-value of chi-square (χ^2_M)	P = 0.051	Good fit
Root Mean Square Error of Approximation (RMSEA)	0.012	Good fit
90 Percent Confidence Interval for RMSEA	(0.0 ; 0.019)	Good fit
Comparative Fit Index (CFI)	1.00	Good fit
Incremental Fit Index (IFI)	1.00	Good fit
Relative Fit Index (RFI)	1.00	Good fit
Root Mean Square Residual (RMR)	0.014	Good fit
Standardized RMR	0.011	Good fit

Discussion

Second language listening is a complex cognitive process involving various knowledge sources. The various knowledge sources required for listening can explain high variances and the medium level of squared multiple correlations of the variables. The significant tests show that all factor loading coefficients are significant at the critical level of p = 0.01 as their p values exceeded 3.29. However, the variances of each variable are also significantly high.

The squared multiple correlations (SMC) of variables, which indicate the variance proportions explained by the factors, range in the medium levels. The complex nature of listening constructs may account for the medium degrees of SMCs. The language knowledge and comprehension factors are two of several knowledge sources engaged in listening. The effects of other sources, such as background knowledge, are not included in this study and may contribute to the variances observed in the variables. In addition, other test method features may have not been represented by the two factors. Those features may also contribute to the values of variances and the medium size of SMCs.

The complexity of listening constructs may account for the covariance of measurement errors in the respecified model. In model respecification, the factor associations have not changed, but the measurement errors are specified to covary. The covariance of measurement errors represents that the unique variances of the variables are not unique but correlated with the unique variance of other variables. The covariance of measurement errors indicates the possibility of common factors other than language knowledge and comprehension factors between the variables of covarying measurement error. Considering the complex nature of listening, it is theoretically plausible to assume factors other than the two factors that are specified in this model of listening constructs exist.

The above explanations about the covariance of measurement errors are largely speculative. However, the presence of error covariance is theoretically grounded in the fact that listening is a complicated process involving more factors than the two factors highlighted in this study. This calls for further research to investigate other aspects of listening constructs.

References

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

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

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.

Brown, G., & Yule, G. (1983). *Teaching the spoken language*. Cambridge, UK: Cambridge University Press.

Buck, G. (2001). *Assessing listening*. Cambridge, UK: Cambridge University Press.

Buck, G. (1995). How to become a good listening teacher. In D. J. Mendelsohn and J. Rubin (Eds.). *A guide for the teaching of second language listening* (pp. 113–131). San Diego: Dominie Press.

Buck, G., & Tatsuoka, K. (1998). Application of the rule-space procedure to language testing: Examining attributes of a free response listening test. *Language Testing*, *15*(2), 119–157.

Buck, G., Tatsuoka, K., Kostin, I., & Phelps, M. (1997). The sub-skills of listening: Rule-space analysis of a multiple-choice test of second language listening comprehension. In A. Huhta, V. Kohonen, L. Kurki-Suonio, & S. Luoma (Eds.), *Current developments and alternatives in language assessment: Proceedings of LTRC 96* (pp. 589-624). Tampere: University of Jyvaskyla.

Dunkel, P., Henning, G., & Chaudron, C. (1993). The assessment of an L2 listening comprehension construct: A tentative model for test specification and development. *Modern Language Journal*, *77*(2), 180–191.

Freedle, R., & Kostin, I. (1993). The prediction of TOEFL reading item difficulty: Implications for construct validity. *Language Testing, 10*(2), 133–170.

Freedle, R., & Kostin, I. (1996). *The prediction of TOEFL listening comprehension item difficulty for mini-talk passages: Implications for construct validity*. (TOEFL Research Report No. 96–29). Princeton, NJ: Educational Testing Service.

Freedle, R., & Kostin, I. (1999). Does the text matter in a multiple-choice test of comprehension? The case for the construct validity of TOEFL's minitalks. *Language Testing*, *16*(1), 2–32.

Hatcher, L. (1994). *A step by step approach to using the SAS system for factor analysis and structural equation modeling*. Cary, NC: SAS Institute Inc.

Jensen, C., Hansen, C., Green, S. B., & Akey, T. (1997) An investigation of item difficulty incorporating the structure of listening tests: a hierarchical linear modeling analysis. In A. Huhta, V. Kohonen, L. Kurki-Suonio, & S. Luoma (Eds.), *Current developments and alternatives in language assessment: Proceedings of LTRC 96* (pp. 151-164). Tampere: University of Jyvaskyla.

Joreskog, K. G., & Sorbom, D. (1993). *Lisrel 8 user's reference guide.* Chicago: Scientific Software International.

Kline, R. B. (2005). *Principles and practice of structural equation modeling*. New York: Guilford Press.

Kostin, I. (2004). *Exploring item characteristics that are related to difficulty of TOEFL dialogue items*. (Research Report No. 79). Princeton, NJ: Educational Testing Service.

Kunnan, A. J. (1998). An introduction to structure equation modeling for language assessment research. *Language Testing*, *15*(3), 295–332.

Mendelsohn, D. J. (1994). *Learning to listen: A strategy based approach for the second language learner*. San Diego: Dominie Press.

Nissan, S., DeVincenzi, F., & Tang, L. (1996). *An analysis of factors affecting the difficulty of dialogue items in TOEFL listening comprehension*. (TOEFL Research Report No. 51). Princeton, NJ: Educational Testing Service.

O'Malley, J. M., Chamot, A. U., & Kupper, L. (1989). Listening comprehension strategies in second language acquisition. *Applied Linguistics, 10*(4), 418–437.

Powers, C. (1985). *A Survey of Academic Demands Related to Listening Skills*. (TOEFL Research Report No. 20). Princeton, NJ: Educational Testing Service.

Richards, J. (1983). Listening comprehension-approach, design, procedure. *TESOL Quarterly*, *17*(2), 219–240.

Rost, M. (1994). *Introducing listening*. Harmondsworth: Penguin.

Rost, M. (2005). L2 listening. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 503-528). Mahwah, NJ: Erlbaum Associates.

Shohamy, E., & Inbar, O. (1991). Validation of listening comprehension tests: The effect of text and question type. *Language Testing*, *8*(1), 23–40.

Thompson, I. (1995). Assessment of second/foreign language listening comprehension. In D. J. Mendelsohn and J. Rubin (Eds.). *A guide for the teaching of second language listening* (pp. 30–58). San Diego: Dominie Press.