



## **Investigating the Construct Validity of the MELAB Listening Test through the Rasch Analysis and Correlated Uniqueness Modeling**

**Christine Goh and S. Vahid Aryadoust**

National Institute of Education  
Nanyang Technological University, Singapore

**ABSTRACT** This article evaluates the construct validity of the Michigan English Language Assessment Battery (MELAB) listening test by investigating the underpinning structure of the test (or construct map), possible construct-underrepresentation and construct-irrelevant threats. Data for the study, from the administration of a form of the MELAB listening test to 916 international test takers, were provided by the English Language Institute of the University of Michigan. The researchers sought evidence of construct validity primarily through correlated uniqueness models (CUM) and the Rasch model. A five-factor CUM was fitted into the data but did not display acceptable measurement properties. The researchers then evaluated a three-traits<sup>1</sup> confirmatory factor analysis (CFA) that fitted the data sufficiently. This fitting model was further evaluated with parcel items, which supported the proposed CFA model. Accordingly, the underlying structure of the test was mapped out as three factors: ability to understand minimal context stimuli, short interactions, and long-stretch discourse. The researchers propose this model as the tentative construct map of this form of the test. To investigate construct-underrepresentation and construct-irrelevant threats, the Rasch model was used. This analysis showed that the test was relatively easy for the sample and the listening ability of several higher ability test takers were sufficiently tested by the items. This is interpreted to be a sign of test ceiling effects and minor construct-underrepresentation, although the researchers argue that the test is intended to distinguish among the students who have a minimum listening ability to enter a program from those who don't. The Rasch model provided support of the absence of construct-irrelevant threats by showing the adherence of data to unidimensionality and local independence, and good measurement properties of items. The final assessment of the observed results showed that the generated evidence supported the construct validity of the test.

---

<sup>1</sup> In this article, the terms (latent) trait, factor, and construct have been used interchangeably.

### Introduction

The Michigan English Language Assessment Battery (MELAB) was founded in 1985 to measure English language proficiency of nonnative English speaking applicants to American and Canadian universities and professional workers who need to produce a certificate of English proficiency or anyone interested in testing her/his English language proficiency (*MELAB Technical Manual*, 2003). The predecessor of the test, the Lado Test of Aural Comprehension (a multiple-choice listening test), evolved gradually into the current MELAB, which has been informed by ongoing research efforts supported by the University of Michigan. At present, the test is administered in 29 states in North America and encompasses three compulsory sections: (a) composition: 200–300 words (30 minutes); (b) listening comprehension: 50 questions (25 minutes); and (c) grammar, comprehension, vocabulary, reading: 100 questions (75 minutes). An optional speaking test that lasts about 15 minutes can also be taken by the test taker.

The listening component of MELAB has been researched in the past few years. The underlying factors of the test have been investigated by Eom (2008); Eom tested a model comprising language knowledge and comprehension and provided support for the hypothesized underpinning structure of the test. However, the methodological problem in the study was to covary error terms heavily without a theoretical support. Wagner (2004) also studied the factor structure of the listening subtests of MELAB and the Examination for the Certificate of Proficiency in English (ECPE); the MELAB study did not statistically separate the hypothetical underlying factors—the ability to understand explicitly and implicitly stated information—in the listening section successfully, indicating that this dichotomy can be an artifact.

The present study seeks to continue efforts of the validation of the MELAB listening test and to address some of the limitations in the earlier studies. The main questions in the study include (a) postulating and evaluating the construct map of the test, and (b) investigating construct representation and irrelevant factors or contaminants. According to Wilson (2005), a construct map is a modeled graphical representation of the underlying continuum of the construct that entails “a coherent and substantive definition for the content and the construct” (p. 26). The construct map, which “precipitates an idea or a concept,” is the representation of a unidimensional latent trait that we seek to measure. Every test measuring a construct has a construct map representing the components and structure of the construct. Towards this end, a latent variable model was used in the present study to help develop the construct map. For investigating construct representation and the presence of construct contaminants, we used the Rasch model (Bond, 2003; Haladyna & Downing, 2004). Needless to say that the latent variable model analysis used for the first question further informed us about the presence of construct irrelevant factors (Haladyna & Downing, 2004).

The present study begins with a review of selected listening comprehension literature and the technical manual of the MELAB test. Next, a listening model and a validation framework are proposed, and a content analysis of the test is conducted. Subsequently, models are generated to replicate the traits detected in the content analysis. This is followed by the Rasch investigation of item properties and the test. Results and findings from the analyses are then grounded in an evidence-based construct validity framework proposed by Chapelle (1994) because the framework concerns the “cornerstone of the definition” of validity, construct validity; other validation frameworks, such as Kane’s (1992, 2001) are useful but more general and need an extensive set of data and support of validity in a general sense, a concern which is out of the scope of this study.

### **The Listening Comprehension Construct in Assessment**

Scholarly literature on second language (L2) listening comprehension includes several conventional models and frameworks. Richards' (1983) taxonomy, Tatsuoaka and Buck's (1998) cognitive assessment through the rule-space technique, and Buck's (2001) model are attempts to explore the constituent structure of the skill and students' cognitive processes. Several researchers also explored the divisibility of listening comprehension from other language skills and reported controversial findings (Buck, 1992; Farhady, 1983; Oller & Hinofotis, 1980; Oller, 1978, 1979, 1983; Scholz, Hendricks, Spurling, Johnson, & Vandenburg, 1980). While these hypothetical models and taxonomies have deepened our understanding of the listening skill, there is a need to provide a clear and unifying definition of the skill.

Whereas listening comprehension was once assumed to be entirely a bottom-up process, later models posited that top-down processing takes place to understand implied messages. These perspectives on listening process have guided test developers and analysts in contemporary tests of L2 listening (Brindley, 1998; Buck, 1990; Rost, 1990; Tsui & Fullilove, 1998; Wagner, 2002, 2004). However, there is neither consensus over methods of testing listening skills nor an absolutely unified listening construct in terms of its definition (Dunkel, Henning, & Chaudron, 1993). For example, Glenn (1989) conducted a content analysis of 50 definitions of the listening construct and concluded that there was no universal agreement on the nature of this skill. Glenn further noted that this lack of agreement impeded research into listening assessment and even other areas where listening is involved, such as communication studies.

L2 researchers have used the two-level strategic comprehension model for discourse comprehension, which was originally proposed by Kintsch and van Dijk (1978) (also van Dijk & Kintsch, 1983) to define the listening construct (for example, Buck, 2002; Wagner, 2002; 2004). Kintsch and van Dijk's (1978) theory is a mix of Kintsch's research on psychology, which developed the concept of propositions, and van Dijk's studies on functional linguistics, which introduced macro-operators. According to this model, comprehenders have two types of strategy to comprehend discourse: local and global coherence strategies. Local strategies connect components within sentences or clauses throughout the text to make sense of the text at a sentential level. Global strategies generate the "macrostructure"; it helps comprehenders explore the theme, main ideas and their interrelations, and the entire discourse structure. These two strategies do not operate independently: when comprehenders process consecutive clauses, they use local strategies to process meaning of individual utterances; simultaneously, they also use global strategies to ensure a comprehensive interpretation of the textbase which is being generated.

Using global strategies in L2 listening is sometimes taken synonymous with top-down information processing (Nation & Newton, 2009; Shohamy & Inbar, 1991). Top-down processes help listeners make inferences and expectations about the text structure. They are different from bottom-up processes which depend on local strategies. Bottom-up processing helps deciphering the phonological stimulus, and involves rebuilding individual sounds into words and constructing clauses. Kintsch and van Dijk's (1978) model has been used in some L2 listening studies (Buck, 1993; Hansen & Jenson, 1994; Shohamy & Inbar, 1991). Shohamy and Inbar in particular emphasized that a competency-based approach to testing L2 listening should focus on top-down and bottom-up processing skills.

Similar to Kintsch and van Dijk's (1983) model, Buck (1991, 1992, 1994, 2001) offered a listening construct encompassing the ability to understand explicitly articulated

information and the ability to understand implicitly stated information; understanding explicit information is the ability to comprehend the verbal presentation of the message, and understanding implicit information is the ability of making inferences based on the world knowledge and schema. Buck (2001) refers to this model as a “default listening model,” stating that the model is general and flexible and can be expanded in various contexts. The validity of this model has been investigated via multivariate data analysis methods such as exploratory factor analysis (EFA) (Wagner, 2002, 2004) and confirmatory factor analysis (CFA) (Liao, 2007). Some researchers have argued that this model “delimits focus to the cognitive operation” of comprehension (Dunkel et al., 1993) and disassociates listening processes from higher, complex processes that concern cognition, such as synthesis and evaluation. However, Wagner’s (2004) factorial study, which was intended to show the discriminability (divisibility) of the ability to understand explicitly and implicitly articulated information, provided only limited evidence supportive of the two-factor model. Conversely, Liao (2007) reported that the variation in items of a listening test was accounted for by the two hypothesized latent traits. Liao also reported significantly high correlations between the two latent traits.

In another listening model proposed for the Test of English as a Foreign Language (TOEFL), Bejar, Douglas, Jamieson, Nissan, and Turner (2000) regarded L2 listening as a two-stage process: listening and response. In the listening stage, concurrent with hearing the aural message, pertinent situational knowledge (context role), linguistic knowledge (phonology, lexicon, morphology, and syntax), and background knowledge are activated to construct a set of simple statements or propositions; response takes two major forms: aural and written. According to Bejar et al., test developers should not base the test too heavily on either of these response types, because they can introduce construct irrelevant factors to the assessment of listening: if this stage overloads the mental processes of listeners, the measurement error will be overwhelming.

Some researchers tried to separate the listening construct from other language skills in an effort to demonstrate that listening is a separate trait (construct). Oller and Hinofotis (1980), Oller (1978, 1979, 1983), and Scholz, Hendricks, Spurling, Johnson, and Vandenburg (1980) used exploratory factor analysis (EFA) to isolate listening as a trait among other traits. However, EFA did not separate this trait. The researchers proposed that language proficiency is a unique and monolithic trait that cannot be partitioned. Interestingly, other researchers offered counterevidence and argued for the separability of language traits and listening (Buck, 1992; Farhady, 1983; Sawaki, Sticker, & Andreas, 2009).

This brief review shows that listening comprehension has different underlying processes. Wagner (2002) summarizes these processes as a general listening comprehension model comprising multiple major components: ability to understand details—indicative of bottom-up comprehension process—and large stretches of discourse (Buck, 2001; Richards, 1983), ability to comprehend major points or gist—recognized as top-down comprehension process—as stated by Richards (1983), ability to make inferences (Hildyard & Olson, 1978), and the ability to guess the meaning of unknown words from the context. We seek to investigate the operationalized MELAB listening construct in this study. We anticipate that we will identify some of these skills in the test.

### **Michigan English Language Assessment Battery Listening Test**

The listening section of the Michigan English Language Assessment Battery (MELAB) has three parts, consisting of a total of 50 multiple-choice items in the entire test.

Test instructions are delivered to test takers and they are asked to answer questions which are read to them after hearing the stimuli. Following this, candidates choose the most appropriate response among three printed alternatives in the test booklet.

According to the Michigan English Language Assessment Battery technical manual, referred to as *MELAB Technical Manual* hereafter (English Language Institute of the University of Michigan, 2003), there are four test forms (DD, EE, FF, and GG); DD and EE forms are fairly older than the other two forms and are now retired. While the DD and EE forms comprised emphasis item type, conversations, and extended talks, the FF and GG forms include minimal context questions, short conversations, and long radio interviews. Emphasis items are retired and not used in the new test forms. In minimal context items, the listener assumes the role of an interlocutor to provide an answer to a question, invitation, etc., or to select the best paraphrase of a short utterance they have heard. Conversations, long talks, and radio interviews have a more extended context compared with minimal context items.

The principal aim of the test is summarized as follows (English Language Institute of the University of Michigan, 2003):

The listening test of the MELAB is intended to assess the ability to comprehend spoken English. It attempts to determine the examinee's ability to understand the meaning of short utterances and of more extended discourse as spoken by university-educated, native speakers of standard American English. It requires that examinees activate their schemata to interpret the meaning of what they hear and to use various components of their schemata to interpret the meaning of what they hear and to use various components of their linguistic system to achieve meaning from the spoken discourse, and presumes the activation of various comprehension abilities such as prediction, exploitation of redundancy in the material, and the capacity to make inferences and draw conclusions while listening. The test does not attempt specifically incorporate a variety of English dialects or registers but focuses on general spoken American English—conversational as well as semi-planned and planned speech, e.g., lectures based on written notes and radio interviews with topic experts. (p. 34)

This paragraph is the principal resource identifying the types of listening comprehension abilities that the MELAB listening test is intended to measure. Based on the description of the listening test above, the competencies examined are summarized as follows:

1. Ability to use the individual's schemata to interpret meaning
2. Ability to use components of the individual's linguistic system (e.g., grammar, vocabulary, etc.) to construct their understanding
3. Ability to use a range of comprehension skills and strategies
4. Ability to make inferences and draw conclusions

Some of these competencies have been studied previously; as noted earlier, Wagner (2004) investigated the factor structure of long talks through exploratory factor analysis. In this study, he did not find strong evidence that this section of the test targets the ability of making inferences and understanding explicitly articulated information. Following this study and using a CFA model, Eom (2008) reported that language knowledge and comprehension are two underlying factors measured by the MELAB listening test. While the baseline latent

trait model in the study did not fit the data well, Eom allowed the error terms to covary heavily. This measure can improve the fit of the model, but it will also yield a less parsimonious model (lower degrees of freedom). A parsimonious model is less complex in terms of the relationship between indicators, error terms, and latent variables and is able to efficiently explain the underlying cognitive processes of the test. That is, the more paths we add to the model, the better the fit, but the lower the parsimony; good fit in un-parsimonious models does not always translate into a well-fitting model (Raykov & Marcoulides, 1999; Schumacker & Lomax, 2004). Because the modification does not appear to be directly informed by theory in Eom's study, the implications of the study are limited. In the present study, we seek to provide a less complex (more parsimonious) model which captures the complexity of the MELAB listening construct and approximates the cognitive processes of the test takers.

## **The Study**

### **Objectives of the Study**

The major objectives of the study are:

1. To determine the underpinning factor structure or the construct map of the MELAB listening test.
2. To determine construct-underrepresentation and construct-irrelevant threats to the construct validity of the listening test, if any.

### **Methodology**

#### *Participants*

A data set of the performance of 916 candidates who took the MELAB test was provided by the English Language Institute (ELI) of the University of Michigan. Although the participants in the test were from 78 countries, the ELI announced that the data are not from all countries where the test is administered. All test takers whose data were used in the current study took the same test. Of these, 425 were female and 427 male; the information on gender of 64 people was missing.

#### *Materials*

The ELI provided the test materials, including the scripts of the audio stimuli and 50 items which were of three types: (a) 15 minimal context items, (b) 20 short conversation items, and (c) 15 long radio interview items (three interviews). In minimal context items, test takers should choose the correct response to an invitation, offer, etc. The following example is from the MELAB information and registration bulletin (2009):

You hear: When's she going on vacation?  
 You read: a. last week      b. to England      c. tomorrow  
 The correct answer is c, *tomorrow*. (p. 8)

The *MELAB Technical Manual* refers to this item type as "minimal context items" (English Language Institute of the University of Michigan, 2003). The manual states that these questions measure different aspects of comprehension at the item and test levels; on the one hand, they assess the ability to comprehend the "conversational patterns" in daily spoken

English, and they test the ability to understand new information on the other. An element of predicting what the other interlocutors would say is fundamental to answering the items in minimal context tests.

The *MELAB Information and Registration Bulletin* (English Language Institute of the University of Michigan, 2009) explains that short conversation items evaluate the understanding of test takers of short conversations or talks. An example of short conversations follows:

You hear:

A: Let's go to the football game.

B: Yeah, that's a good idea. I don't want to (wanna) stay home.

You read:

a. They'll stay home.

b. They don't like football.

c. They'll go to a game. (p. 8)

This item assesses the comprehension of a longer stretch of discourse. Understanding the illocutionary forces of the items (e.g., requests, invitations, apologies, etc.) alongside the literal meaning in conversations is necessary to successfully answer these items (*MELAB Technical Manual*, 2003). In order to interpret the illocutionary meaning of these exchanges, the candidate will have to make inferences and draw conclusions where needed.

In the final part of the test are longer audio inputs. In this section, simulated radio talks and conversations are delivered to test takers. They are allowed to take notes. The presence of "graphic materials" serves to further contextualize the aural input. As a general observation, the printed options throughout the test are short, ranging from two to seven words each. This helps minimize the use of graphological knowledge and the effects of reading skills on listening. The *MELAB Technical Manual* states that grammatical, textual, lexical, functional, and sociolinguistic knowledge are the principal components of the comprehension items.

### *Analysis*

According to Kirsch and Guthrie (1980), the notion of validity is dependent upon "the congruence between the stated purpose of the test and what is being measured by the test" (p. 90). To investigate this congruence, Messick (1989) and other researchers suggest that researchers use such statistical methods as factor analysis; Borsboom et al. (2004) proposed latent trait models; and Wright and Stone (1999) recommended the Rasch model (also, see Bachman [2004] for a review of quantitative methods of validation). The present study seeks to use the following methods:

1. A particular confirmatory factor analysis modeling approach known as correlated uniqueness model for building a construct map
2. Rasch measurement for investigating construct underrepresentation and construct irrelevant threats

Before the results of the analysis are reported, we present the proposed listening models for MELAB, as well as explain the rationale for the use of the two methods of statistical analyses. In the results section, we will also describe the compensatory strategies employed and the results for each type of analysis before arriving at a conclusion for each research question.

### Proposed Listening Models for MELAB

Figure 1 illustrates a conceptualized listening construct in the MELAB listening test based on the Test Aims paragraph (see next paragraph). As stated previously, the ability to use local and global coherence strategies which help comprehend explicitly and implicitly stated information are two components of a listening trait. We use the explicit/implicit terminology because it is used in MELAB literature. The ability to make propositional and enabling inferences alongside understanding context-reduced stimuli, explicit aural input, and making close paraphrases are present in the three test models in Figure 1 (see the Materials section for a review of the test). A major objective of the MELAB listening test is to assess the ability to “make inferences and draw conclusions,” which we have divided into propositional and enabling inferences based on L2 literature (Hildyard & Olson, 1979). Likewise, we divided the ability to understand explicit information into close paraphrase and detailed information.

Figure 1 presents three listening models for the MELAB listening test. In each model: big circles represent latent traits; boxes represent observed variables or test items; error terms are displayed as small ellipses with arrows pointing to boxes; regression paths are indicated as one-headed or unidirectional arrows; and correlations are indicated as two-headed or bidirectional arrows. In each model, only ten items are displayed for reasons of space.

Model 1 will be modified if it does not fit the data satisfactorily. It is hypothesized that there are five separable traits underpinning the test in Model 1 and that five types of items measure test takers' trait levels: minimal context questions (MCQ), detailed or explicit information questions (DIQ), close paraphrase questions (CPQ), propositional inference questions (PIQ), and enabling inference questions (EIQ); all of these traits are correlated. This model is a correlated uniqueness model (CUM) with uncorrelated error terms because the error terms are *not* free, meaning that they are not covarying. If we free the error terms (covary them using double-headed arrows), Model 2 is generated.

Model 2 is proposed in the event that the first CFA model does not fit the data. This carries the implication that there is a method factor effect in the data. Model 2 is similar to Model 1 except that it allows error terms to correlate. Correlation is indicated by the double-headed arrows covarying (connecting) error terms. Some error terms do not correlate with others: methods in Model 2 comprise two types which are implied in the *MELAB Technical Manual*, i.e., short stimuli method and long stimuli method. Based on this definition, we consider items in section 1 and 2 the short stimuli method and items in section 3 representing the long stimuli method and covary their error terms. The justification for Model 2, with correlated traits and error terms (uniquenesses), is that the observed variance in data is assumed to be a joint function of traits and methods. Accordingly, we expect an increment in the fit of the model when we free (covary) the error terms if there is a method effect in the model. The error correlations, in turn, help us model the shared method variance which is unique to the measuring tool (see Bachman, 2004, p. 283–287).

While some studies show that listening comprehension is a general and nondivisible latent factor (Wagner, 2004), some studies have separated theory-informed factors (Buck, 1992, Hansen & Jenson, 1994). It is likely that there is a two-level latent trait, one level as a general listening trait and the other as listening components. This hypothesis is evaluated as a model competing with Model 1 and 2. Model 3, which is a second-order CFA model, has a major latent variable whose indicators are also latent; that is, there are “two layers of latent constructs” (Hair, Black, Babin, & Anderson, 2010, p. 815) in which the higher order latent variable—in our model, the listening construct—cause lower order latent variables—in our

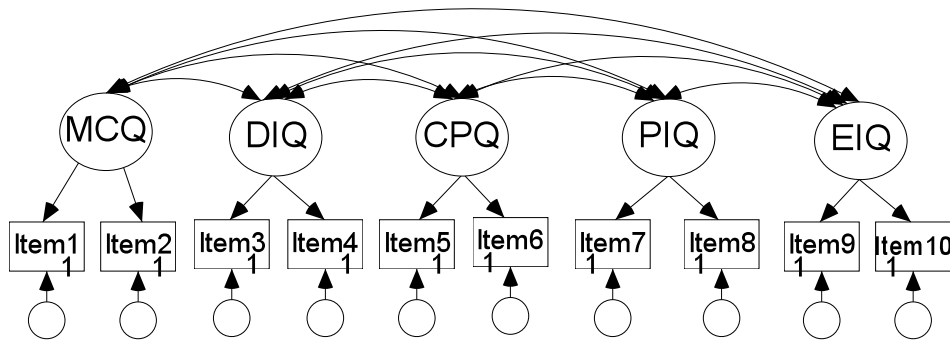
model, minimal context questions (MCQ), detailed or explicit information questions (DIQ), close paraphrase questions (CPQ), propositional inference questions (PIQ), and enabling inference questions (EIQ). Therefore, the distinctive feature of Model 3 is the presence of a higher order factor (a general listening ability) which is hypothesized to cause the proposed separate traits. If method effect is present, then Model 3 (on the next page) will not display good fit, indicating a CUM with covarying error terms is more suitable to explain the underpinning structure of the MELAB listening construct. Competing models may fit the data equally well, but the best one would be the most parsimonious and theory-informed model.

## **Statistical Analysis**

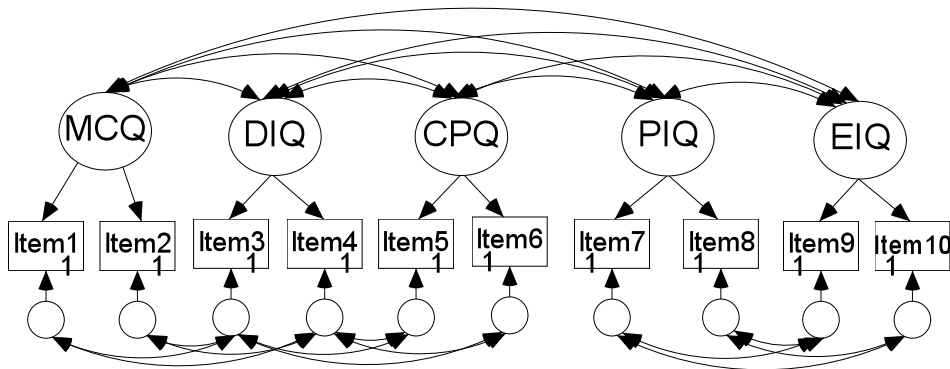
### **Confirmatory Factor Analysis**

The present study employs a confirmatory factor analysis approach (CFA) to build a construct map representing the constituent structure of the test. It provides evidence for the power and specification of the a priori factor model (Schmitt & Stults, 1986; Brown, 2006). Conway, Lievens, Scullen, and Lance (2004) classified CFA into four variants: (a) linear-additive CFA (Widaman, 1985, 1992), (b) hierarchical second-order factor (SOF) CFA (Marsh & Hocevar, 1988), (c) correlated uniqueness models (CUM) (Kenny, 1976; Marsh, 1989, Brown 2006), and (d) direct product (DP) of multiplicative trait-method effects (Campbell & O'Connell, 1967; Cudeck, 1988). CUM is used in the present study to solve the problem of multitrait-multimethod (MTMM) matrix. Proposed by Campbell and Fisk (1959) to examine construct validity, MTMM is a matrix assuming that each factor or trait is measured by several methods and the matrix is arranged in a way that traits are entailed in methods, i.e., the matrix is laid out by method blocks, each comprising at least three trait cells. The method leads the researcher to multiple evidence of construct validity, most notably the correlation between tests assessing the same trait (convergent validity) which should be high, and the correlation between tests assessing different traits (divergent validity) which should be low (Bachman, 2004). There are some problems in the analysis of MTMM, such as negative degrees of freedom, non-positive definite matrices (see Schumacker & Lomax, 2004), and that each trait in the matrix should be assessed by at least three methods. Correlated uniqueness modeling is proposed as an alternative less demanding approach to solve the MTMM matrix. To produce a CUM, researchers need to define at least two factors ( $F$ ) to be measured by three methods ( $M$ ). But a  $2F \times 2M$  model may fit when the factor loading indices of indicators (items) loading on the same latent variable are constrained to be equal (Brown, 2006, p. 220).

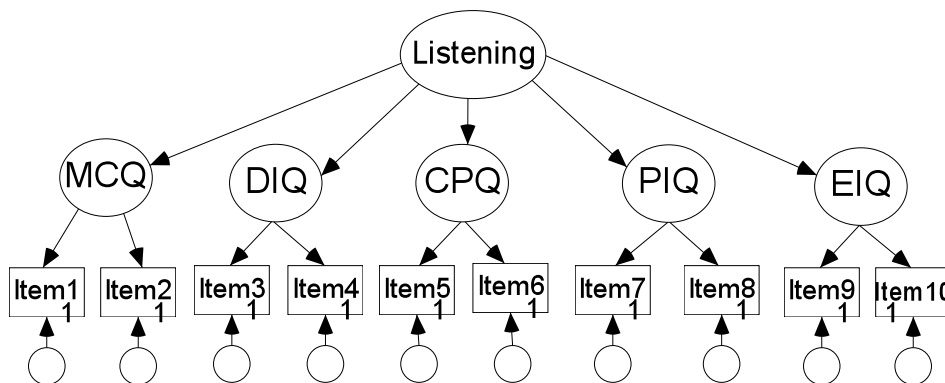
Like other latent trait models, CUM has faced criticisms in its literature. Whereas some researchers have recommended CUM for solving MTMM matrices (Brown, 2006; Lievens & Conway, 2001; Marsh, 1989), others argue that CUM estimates are biased (Lance, Noble, & Scullen, 2002). However, this bias has been shown to be trivial. Marsh and Baily (1991) used simulated data to explore the bias and concluded that the purported bias is not significant. The study by Tomás, Hontangas, and Oliver (2000) provided further evidence for Marsh and Baily's finding. In summary, in contexts where there are only two latent traits and two methods, CUM is a good way to solve the matrix (Brown, 2006).



Model 1 - Correlated Uniqueness Model, with Uncorrelated Uniquenesses



Model 2 - Correlated Uniqueness Model, with Correlated Uniquenesses



Model 3 - 2nd Order Model, with Uncorrelated Error Terms

Figure 1. Three Proposed Models for the MELAB Listening Test. Only ten items are displayed in this figure for reasons of space. Model 1 with uncorrelated error terms (uniquenesses) is included mainly as a baseline model. Model 2 with correlated error terms (uniquenesses) is the modified model. Model 3 is a second-order model with listening as the higher order trait.

(Legend: MCQ = minimal context questions. DIQ = detailed information questions (explicit). CPQ = close paraphrase questions. PIQ = propositional inference questions. EIQ = enabling inference questions.)

## **Rasch Measurement**

We use the Rasch model to investigate construct underrepresentation and construct threats in the present study. This model has two central features: (a) expected probabilities for persons and items and (b) fit indices to argue how persons and items fit the model. These features are valid when the local independence and unidimensionality criteria are established (McNamara, 1996). The Rasch model provides item and person measures based on the mathematical modeling of the data. The basic model from which we derive other models is:

$$\theta_{nil} = \frac{\exp(\beta_n - \delta_{il})}{1 + \exp(\beta_n - \delta_{il})}$$

where  $\theta_{nil}$  is person  $n$ 's probability of scoring 1 or answering item  $i$  correctly,  $\beta_n$  is the ability of person  $n$  on the entire test, and  $\delta_{il}$  is the difficulty level of item  $i$ . According to the model, the probability of success in answering an item is governed by person ability and item difficulty. The Rasch model can help investigate the construct validity of measuring tools by providing the opportunity for investigating construct-irrelevant variance (CIV) and construct underrepresentation (CUR) which are discussed in Messick (1989). The Rasch model is also suitable for assessing item bias, which is a major source of CIV across sub-samples. This analysis is a test of invariance which is also known as Differential Item Functioning (DIF). At the item level, DIF detects items which function significantly differently in different groups and flags them for further analysis and deletion (Bond & Fox, 2007; Wright & Stone, 1999). At the test level, it identifies the covariates that have contaminated the measurement thus introducing some construct-irrelevant variance to the measurement.

## **Results**

### **Descriptive Statistics and Reliability Analysis**

We examined the quantitative features of the data. Descriptive statistics summarize the features of the data in an understandable and concise way. We calculated mean, standard deviation (SD), skewness, and kurtosis using the Statistical Package for the Social Sciences (SPSS) computer program, Version 16 (see Table 1). The table presents items in three test sections; section one contains minimal context questions (items 1–15); section two entails short conversations (items 16–35); and section three contains simulated long radio interviews (items 36–50).

Normality of the observed data should hold in factor analysis studies. Univariate normality was investigated through the analysis of the skewness and kurtosis of data. Normal distributions have a skewness index of zero although a range of -2 to +2 is an acceptable span (Bachman, 2004). (Sometimes some random errors occur, which can change the value). Kurtosis is the degree of flatness or peakedness. Negative values indicate a fairly flat distribution and positive values indicate that the shape of the data has a high peak. Item 8 had skewness and kurtosis indices greater than |2|, indicating that it was slightly deviating from the properties of a normal distribution (Bachman, 2004). This item also had the highest mean but the lowest SD index. Other items did not display unusual skewness and kurtosis values, an indicator of the normality of distribution.

Next, using KR-21 formula, we investigated the reliability or internal consistency of the observed scores. Internal consistency indicates how much of the variation in observed

scores is attributable to error and how much to the true score. Respective KR-21 indices for sections 1, 2, and 3 are 0.65, 0.71, and 0.65. We also computed KR-21 indices for items according to the results of content analysis, as displayed in Table 2: detailed (explicit) information: 0.45 (6 items); close paraphrase 0.60 (13 items); propositional: 0.50 (9 items); enabling: 0.42 (7 items); minimal context 0.65 (15 items). Additionally, KR-21 for all items assessing explicitly said information (19 items) was 0.68 and for items assessing implicitly said information (16 items) was 0.64. According to Pallant (2007), low reliability indices are indicators of a small number of items, resulting in high measurement errors. Therefore, when the number of well-designed items in analysis increases and, measurement error drops, the reliability index tends to increase. The reliability index for the entire test was 0.85. The reliability index of 0.85 is very close to the average KR-21 index of 0.81 and closer to the average reliability index of 0.87 for candidates intending to further their education, as stated in the *MELAB Technical Manual* (2003).

Table 1. Descriptive Statistics for the MELAB Listening Test

Items	Mean	SD	Skewness	Kurtosis
V1	.48	.50	.08	-1.99
V2	.53	.50	-.10	-1.99
V3	.64	.48	-.60	-1.64
V4	.74	.44	-1.09	-.79
V5	.70	.46	-.86	-1.25
V6	.45	.50	.193	-1.96
V7	.59	.49	-.35	-1.88
V8	.87	.33	-2.22	2.93
V9	.80	.40	-1.50	.263
V10	.80	.40	-1.48	.212
V11	.73	.44	-1.06	-.86
V12	.49	.50	.048	-2.00
V13	.55	.50	-.22	-1.95
V14	.53	.50	-.12	-1.98
V15	.75	.43	-1.14	-.69
V16	.54	.499	-.158	-1.979
V17	.50	.500	-.017	-2.004
V18	.59	.491	-.383	-1.857
V19	.60	.489	-.429	-1.820
V20	.61	.487	-.462	-1.790
V21	.55	.498	-.207	-1.962
V22	.62	.486	-.481	-1.772
V23	.67	.469	-.742	-1.453
V24	.79	.406	-1.446	.090
V25	.68	.466	-.779	-1.396
V26	.40	.490	.406	-1.839
V27	.52	.500	-.066	-2.000

Items	Mean	SD	Skewness	Kurtosis
V28	.67	.470	-.736	-1.461
V29	.60	.491	-.401	-1.843
V30	.69	.463	-.823	-1.326
V31	.62	.485	-.515	-1.739
V32	.69	.464	-.806	-1.353
V33	.76	.429	-1.204	-.550
V34	.76	.430	-1.191	-.584
V35	.63	.483	-.544	-1.708
V36	.53	.500	-.105	-1.993
V37	.67	.469	-.747	-1.445
V38	.54	.499	-.149	-1.982
V39	.61	.487	-.472	-1.781
V40	.58	.494	-.314	-1.905
V41	.62	.485	-.510	-1.744
V42	.69	.461	-.839	-1.298
V43	.78	.412	-1.381	-.092
V44	.50	.500	.017	-2.004
V45	.84	.370	-1.820	1.316
V46	.77	.423	-1.268	-.392
V47	.65	.478	-.618	-1.622
V48	.67	.472	-.705	-1.507
V49	.47	.499	.123	-1.989
V50	.66	.474	-.674	-1.549

Note.  $n = 916$  in the sample.

The first section contains minimal context questions (items 1-15).

The second section contains short conversations (items 16-35).

The third section contains simulated long radio interviews (items 36-50).

### Content Analysis

From a competency-based viewpoint, the Test Aims paragraph in the *MELAB Technical Manual* assumes that the test measures different listening skills. We conducted a content analysis initially and categorized the items into five types. This analysis is informed by previous research as noted earlier (Buck, 2002; Hansen & Jensen, 1994; Shohamy & Inbar, 1991; Wagner, 2002) as well as the *MELAB Technical Manual*. These five categories are:

1. minimal context items
2. explicit items (close paraphrase)
3. explicit items (detailed information)
4. implicit items (propositional inferences)
5. implicit items (enabling inferences)

As demonstrated in Table 2, these five item types are classified into three major categories: according to the *MELAB Technical Manual*, minimal context items assess the ability to understand unexpected; according to Hildyard and Olson (1979), Hansen and Jensen (1994), and Wagner, (2002), the ability to understand detailed information and making close paraphrases is a general ability subsumed under the comprehension of explicitly said information; and the ability to make propositional and enabling inferences is subsumed under the ability to comprehend implicitly stated information (Hildyard & Olson, 1979). Therefore, three major skills—understanding the unexpected and assessing explicitly/implicitly said information—subsuming five item types are presented in Table 2.

A content analysis of items and texts was performed to map the items on the posited factor structure of minimal context, explicit information, close paraphrase, propositional inferencing and enabling inferencing. This stage in validation provides content-referenced evidence for the validity of the test. We performed three rounds of content analysis to increase the internal reliability of findings. Each researcher conducted a round of analysis separately. In the final phase, we discussed the item characteristics and the skills they assessed, based on the Test Aims paragraph. Any doubtful classification of items was further reviewed by both authors for a final decision on the classification of the items. Table 2 provides a summary of the findings.

Table 2. Results of the Content Analysis of the Items and Texts

Understanding the unexpected	Assessing explicitly said information		Assessing implicitly said information	
Minimal context	Detailed information	Close paraphrase	Propositional inferencing	Enabling inferencing
1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13 14, 15	40, 46, 47, 48, 49, 50	18, 21, 30, 35, 36, 37, 38, 39, 41, 42, 43, 44, 45	16, 22, 24, 25, 26, 27, 29, 31, 33	17, 19, 20, 23, 28, 32, 34

*Note.* This table presents three major skills—understanding the unexpected and assessing explicitly/implicitly said information—subsuming five item types—minimal context, explicit information, close paraphrase, propositional inferencing, and enabling inferencing.

Minimal context items in Table 2 (1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, and 15), assess the ability to comprehend minimally contextualized colloquial language especially the components of linguistic system like grammar, vocabulary, and idioms, and the ability to *respond* to the stimulus. Similarly, items which assess understanding detailed information (40, 46, 47, 48, 49, 50) and making close paraphrases (18, 21, 30, 35, 36, 37, 38, 39, 41, 42, 43, 44, and 45) require using the linguistic system alongside schema. Assessing the capacity to make inferences and drawing conclusions is a skill required in other items: propositional inferencing (16, 22, 24, 25, 26, 27, 29, 31, and 33) and enabling inferencing (17, 19, 20, 23,

28, 32, 34) (*MELAB Technical Manual*). The majority of items in section 2 belong to the detailed and close paraphrase categories, whereas long interviews in section 3 (in particular the last interview) assess understanding detailed information.

### Confirmatory Factor Analysis

To examine the fit of the postulated models into the data, we carried out a series of CFA to test the proposed models (as displayed in Figure 1), their fit indices, and parsimony. We used several different fit indices to investigate the fit of the proposed models: the root mean square error of approximation (RMSEA) is an index which displays how well a model fits a population. This index should be smaller than 0.08 and ideally smaller than 0.05 (Hair et al., 2010). To explore the precision of the RMSEA, the RMSEA 90% confidence interval is reported. The interval between the lower and higher bounds of this value should be as narrow as possible (Byrne, 2001). The chi-square ( $\chi^2$ ) index is a comparison between the correlation matrix implied and the correlation matrix produced. If they are significantly different, then the  $\chi^2$  value is significant. Normal  $\chi^2$  was also used, a ratio of sample discrepancy ( $\chi^2$ ) to the degree of freedom; better-fitting models generally have a ratio below 3. It should be noted that the  $\chi^2$  index is sensitive to the sample size: large or small samples may produce significant values. Therefore, other indices have been developed to further examine the fit of the model to the data (Miles & Shevlin, 2007; Steiger, 2007; Hair et al., 2010). We further used (a) CFI (Comparative Fit Index, an incremental index evaluating the fit of a model to data relative to a baseline model), (b) GFI (Goodness of Fit Index, an absolute fit index developed to solve the sensitivity of the chi-square index to the sample size), (c) NNFI (Non-normed fit index, also known as Tucker-Lewis Index, basically very similar to the CFI; used to compare the proposed model and the baseline model).

In the first attempt, we postulated a five-factor CUM with no correlation among its error terms, as illustrated in the first model in Figure 1. We used the PRELIS application to produce an asymptotic covariance matrix and a matrix of polychoric correlations for the ordinal data (Du Toit & Du Toit, 2001) because Pearson correlation matrix is not suitable for the factor analysis of dichotomous data (Uebersax, 2006). The underlying assumption in the matrix of polychoric correlations is that the underlying variable is continuous but the data is dichotomous and/or ordinal (Uebersax, 2006). Then, the LISREL software version 8.8 (Jöreskog & Sörbom, 2006) was used to construct and test the model (simplified as Model 1 in Figure 1). The five-factor model did not fit the data satisfactorily.

In a post hoc modification stage, we tried to isolate the test methods measuring separate factors in the test. As implied in the *MELAB Technical Manual* and noted above, the five factors identified are measured by two major test methods: items in section 1 and 2 of the test are the short stimuli method and items in section 3 represent the long stimuli method. Therefore, to generate Model 2 in Figure 1, short stimuli items should measure the same factors and long stimuli items should measure different factors. Yet, as Table 2 shows, there is no clear pattern of measurement in long and short items, indicating that the construction of Model 2 is impossible; therefore, a CUM solution with correlated error terms and factors is not possible.

Next, a second order CFA was performed to investigate whether a major listening factor can cause lower order factors. This model is simplified and displayed as Model 3 in Figure 1. Table 3 summarizes the properties of the CFA models.

Table 3. CFA Models to Confirm the Factor Structure of the MELAB Listening Test

Model	$\chi^2$	df	$\chi^2/df$	NNFI	CFI	GFI	RMSEA	RMSEA 90% confidence interval
CUM	1548.48**	1165	1.33	0.43	0.46	0.93	0.021	0.029 to 0.033
2 <sup>nd</sup> order model	1588.36*	1122	1.41	0.96	0.96	0.93	0.021	0.018 to 0.023
Three-factor CFA	1638.88*	1172	1.39	0.96	0.96	0.93	0.021	0.018 to 0.023
Parcel Items CFA	197.52	149	1.32	0.99	0.98	0.99	0.019	0.011 to 0.026
Constraint tenable	Non-sign.	—	< 3	.95	.95	.95	< 0.08	Narrow interval

*Note.*  $n = 585$  in the sample. \*\* $p < 0.001$ . \* $p < 0.01$ . NNFI: Non-Normed Fit Index. CFI: Comparative Fit Index. Goodness of Fit Index: GFI. RMSEA: Root Mean Square Error of Approximation.  $df$ : degree of freedom. In the CUM model, traits are correlated and error terms are uncorrelated.

According to Table 3, the first five-factor model (CUM) does not fit the data well (NNFI = 0.43; CFI = 0.43; GFI = 0.93; RMSEA = 0.021). This model has a significant  $\chi^2$  (1548.48), an acceptable normal  $\chi^2$  ( $\chi^2/df$ ) of 1.33; its CFI and NNFI are very low although the root mean square error of approximation (RMSEA) is acceptable. The summary of the item loading statistics is available from Appendix 1.

Table 4. Bivariate Correlations of Traits in the CUM

	Min context	Explicit	Close_Para	Enabling	Proposition
Min_context	1.00				
Explicit	1.04	1.00			
Close_Para	0.94	1.07	1.00		
Enabling	0.78	0.97	0.84	1.00	
Proposition	0.98	1.07	0.96	0.86	1.00

*Note.* The CUM model in this table has correlated error terms and correlated traits. Min = minimal. Para = paraphrase. Proposition = propositional inference.

As Table 4 presents, another problem with the CUM model (Model 1) is the emergence of unreasonably high correlation statistics among traits, which are greater than 1.00; the correlation matrix in this case is “non-positive definite”, indicating that “the determinant of matrix is zero or the inverse of the matrix [which is used to estimate the parameters] is not possible” (Schumacker & Lomax, p. 47). Therefore, the solution is not admissible and parameter estimations are not correct.

In Figure 1, Model 3 presents a simplified higher-order model with fewer items and Table 5 displays the fit indices ( $\chi^2 = 1588.36$ ; NNFI = 0.96; CFI = 0.96; GFI = 0.93; RMSEA = 0.021). This model has good fit indices but a significant  $\chi^2$  value. The observed problem in this model is the presence of extremely high loading indices of the lower order factors on the higher order factor (minimal context: 0.96; close paraphrase: 1.01; Explicit: 0.85; propositional inference: 1.11; enabling inference: 0.99). This also indicates that, like Model 1, the correlation matrix in this case is non-positive definite. Accordingly, we adopted a

compensatory strategy to fit a better CFA model into the data; because the models tested above did not fit the data, we opted for a model based on the structure of the MELAB listening test, which is further explained below.

### **First Compensatory Strategy**

We took another approach to redefine the basis of the CFA model. The models produced in Figure 1 depend on a more competency-based definition of listening, which is drawn from the Test Aims paragraph. A task-based theoretical framework highlights the tasks that candidates will encounter in real life situations, and also considers the theory of the construct (Bachman, 2002). According to the *MELAB Technical Manual*, the MELAB listening test targets three major tasks in three sections: understanding and responding to (a) the unexpected requests, invitations, offers, etc., (b) short conversation items, and (c) longer talks or radio interview items, which resembles the factor analysis stated in *MELAB Technical Manual* (English Language Institute of the University of Michigan, 2003, p. 46).

Based on this classification, we performed a CFA to investigate the fit of the three-factor model to the data. Results are demonstrated in Table 3; the three-factor model has similar properties as the 2<sup>nd</sup> order model ( $\chi^2=1638.88$ ; NNFI=0.96; CFI=0.96; GFI=0.93; RMSEA=0.021) but it does not display the problem of correlation coefficients greater than 1. As shown by two-headed arrows connecting the latent traits—MinimCon (minimal context items representing unexpected requests), ShortCon (short conversation items), and LongTalk (longer talks or radio interview items)—in Figure 2, the three-factor model has acceptable correlation coefficients among traits: the correlation indices do not exceed 1.00 and are greater than 0.70 (Hair et al., 2010); the model consisting of three factors (minimum contexts, short conversations and long talks) fits the data satisfactorily.

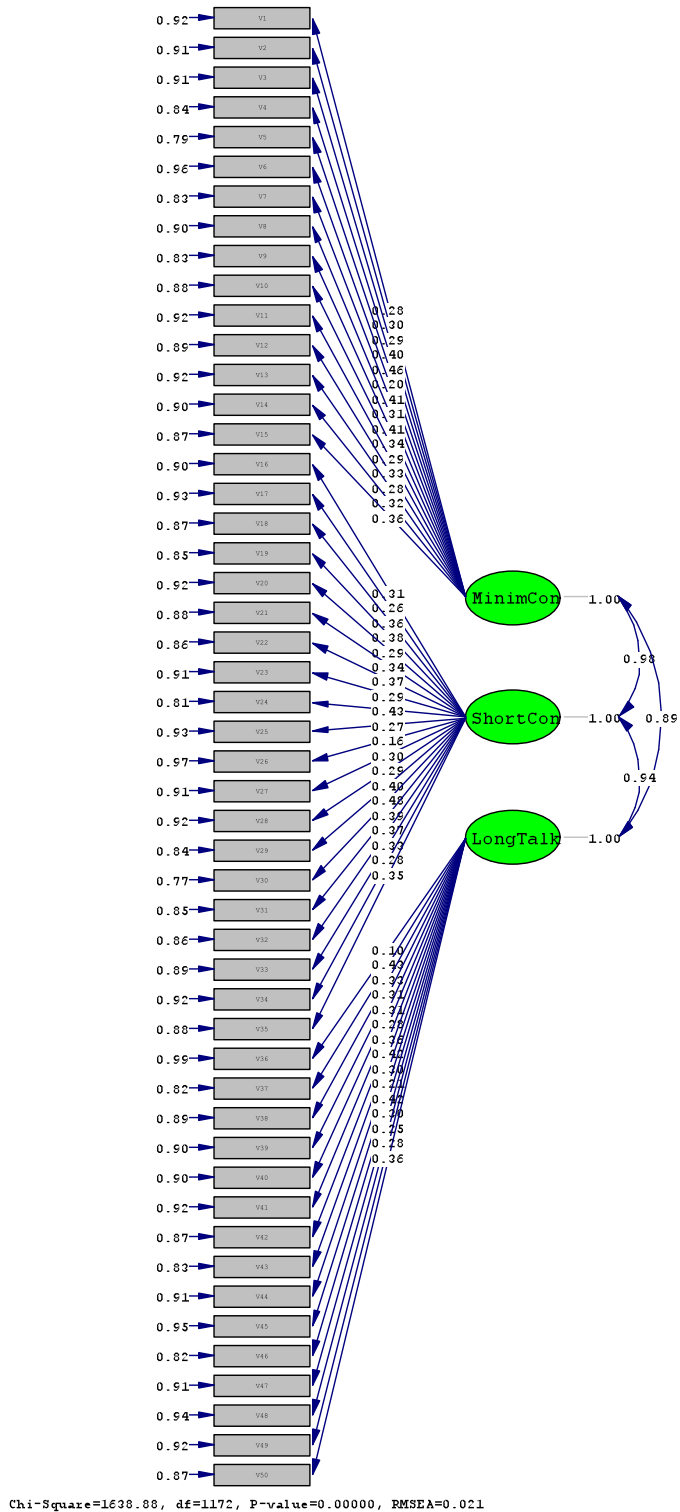


Figure 2. Three-factor Model of the MELAB Listening Test. Oval shapes represent latent traits and rectangles represent the measured variables or items.

### Second Compensatory Strategy

The second strategy is based around computing parcel scores or testlets. The model is presented in Figure 3. Following the *MELAB Technical Manual*, we constructed “short question odd-numbered items, short question even-numbered items, short conversational exchange odd-numbered items, short conversational exchange even-numbered items, three testlets for the three radio interview sets of items” (*MELAB Technical Manual*, 2003, p. 47); we summed up odd-numbered items and then even-numbered items, and the five items testing the comprehension of a long radio interview; so, we built seven aggregate (parcel) items for section 1, nine items for section 2, and three items for section 3.

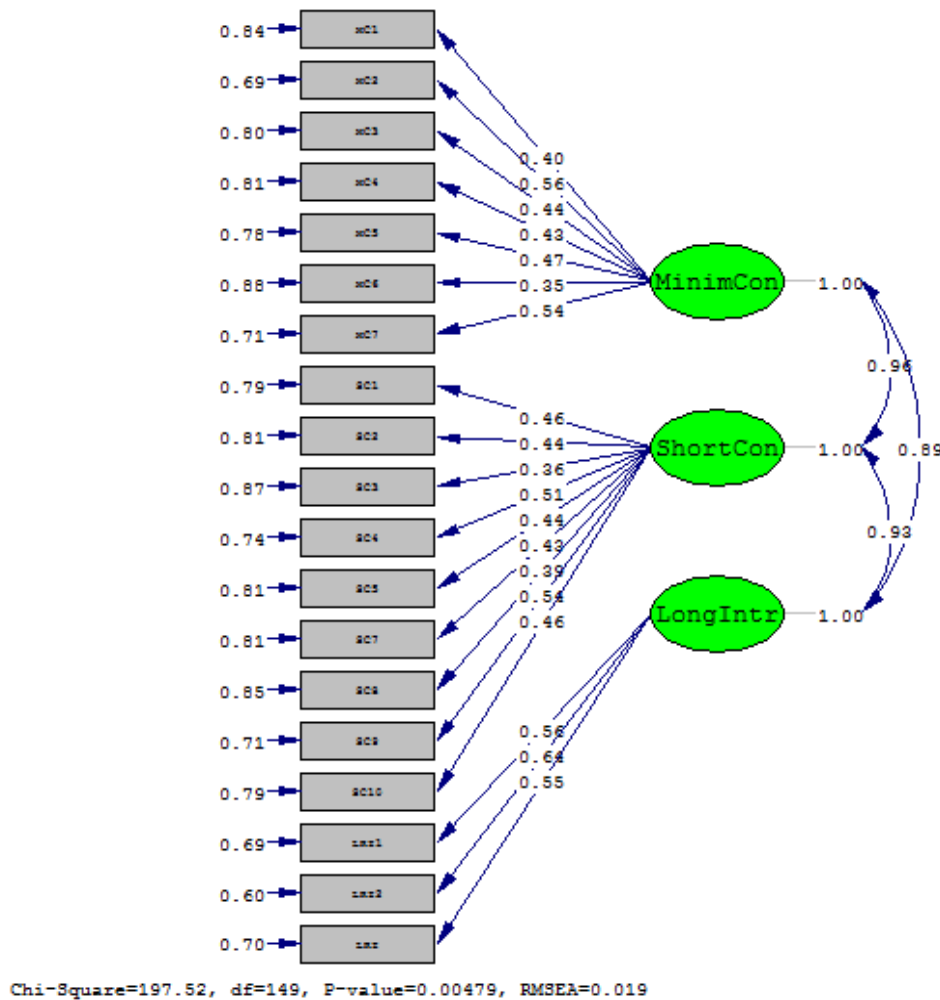


Figure 3. Three-factor Model Based on Parcel Scores. Oval shapes are latent traits and rectangles are the measured parcel variables.

Of all proposed models, the section-based model (Figure 3) with testlets (parcel scores) fit the data the best (NNFI=0.99; CFI=0.99; GFI=0.98; RMSEA=0.019). This model shows that the correlation coefficients of the proposed factors are sufficiently high. Resonant

with the *MELAB Technical Manual*, this observation testifies to the presence of a firm three-factor construct which underpins this version MELAB listening test.

### Rasch Analysis of the Test

We performed two Rasch-based analyses to investigate the item and person measurement properties. Initially, we calibrated the 50 items simultaneously (concurrent calibration) and checked the item fit statistics to see if these items can construct a scale. Person and item measures were generated in this analysis. In the second round, we conducted differential item functioning for gender.

*First Rasch analysis.* We used WINSTEPS package version 3.57 (Linacre, 2005) to fit the Rasch model into the data. Person and item reliability indices were 0.84 and 0.98, respectively. Separation indices were 2.30 and 7.30 for persons and items. The reliability index is evidence for the internal consistency of the person ability indices and item difficulty measures. Separation values are “the ratio of “true” variance to error variance” (Linacre, 2009, p. 462). This is another expression of reliability; ranges from 0 to infinity; and indicates the number of performance levels in the test or heterogeneity of people. Item reliability and separation index point to the ability of the measuring device to establish a similar item hierarchy along the variable in a similar sample from the same population; the item reliability of 0.98 indicates that the item estimates would be reproducible in a similar sample.

Next, Rasch item difficulty and person ability measures were computed. Figure 4 is an item-person map (or Wright map) which plots person ability against item difficulty. Items are laid out on the right side according to their difficulty measure and test takers on the left. The distribution of persons is consistent, making a curve-like shape which peaks around the mean. Person ability and item difficulty mean estimates were 0.68 and 0.00, respectively (in this analysis, the mean of items was anchored to 0.00; the person mean is 0.68 logits higher than the anchored item mean). This is an indicator that items were relatively easy for this sample of test takers. The *SD* indices for persons and items were 0.87 and 0.57, respectively. Figure 4 also demonstrates that some of the candidates with greater demonstrated ability (in red) did not get sufficient questions in the test that can further distinguish their ability levels (this observation is further examined in Figure 5). As will be discussed below, this inflates the standard error of measurement in the estimated ability measures.

To assess the fit of the Rasch model to the data, we examined infit mean-square (information-weighted mean-square statistic which is more sensitive to the unexpected behavior of items closer to persons’ measures) and outfit (unweighted mean-square sensitive to outliers). Mean-square (MNSQ) is computed as the chi-square value divided by the degree of freedom. MNSQ fit indices show useful, as opposed to perfect, fit of the data to the model. An infit MNSQ of, say, 1.2 means 1 unit of modeled information is observed and 0.2 units of unmodeled noise sneaks in. The t-test significance (ZSTD) is used to investigate the *perfect* fit of the data to the model (acceptable range:  $|2|$ ). In a sample size greater than 250, the infit ZSTD tends to exceed  $|2|$ . Therefore, Linacre (2003) recommended that researchers consider MNSQ indices in large samples to show that the Rasch model fit the data *usefully*. Another advantage of MNSQ over ZSTD is that as the sample size increases, the MNSQ power to find discrepancies in the data increases (Linacre, 2003). Bond and Fox (2007) considered 0.6—1.4 an acceptable infit MNSQ range (similar to Linacre’s (2003, 2009) recommendation of 0.5—1.5 for productive measurement).

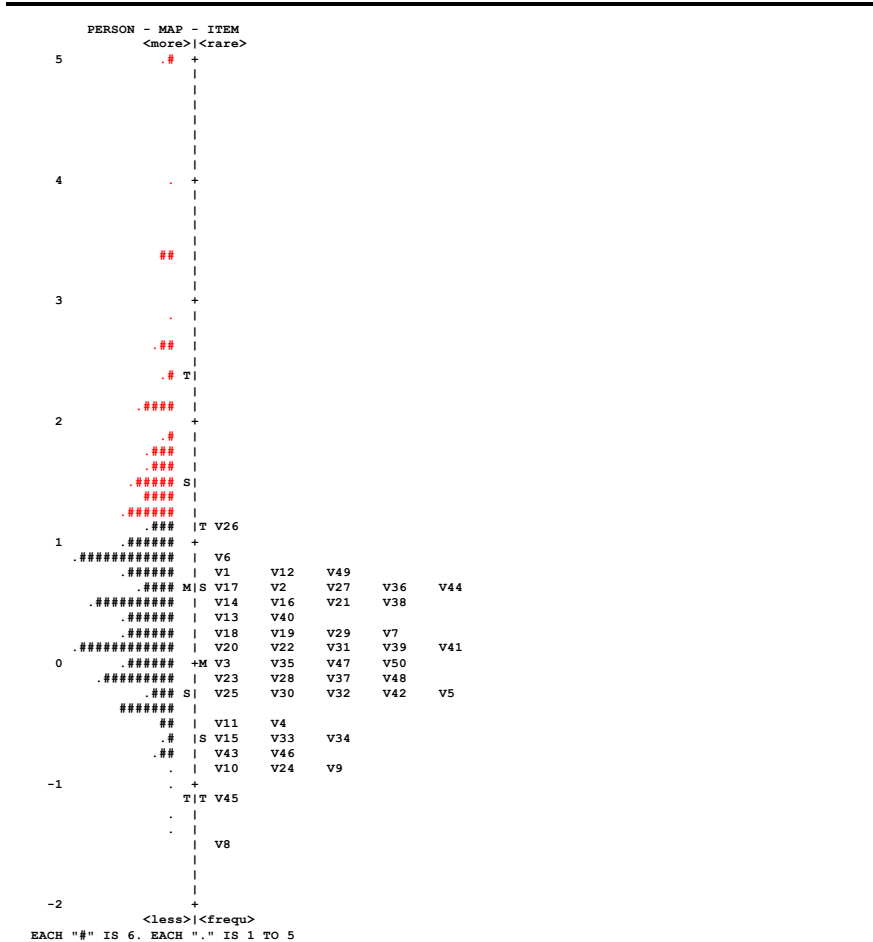


Figure 4. Rasch Analysis Performed on 50 Items. Each “#” sing represents seven persons in the sample.

Item fit statistics and difficulty measures are summarized in Table 5. The score column expresses the raw score assigned to each item according to students' performance and the measure is the converted raw score into logits (log-odds units). Standard errors (SE) indicate the imprecision of the item locations. The lower the SE, the higher the confidence in the location of item difficulty measures. Inflated SE indices are observed when there are not enough items to measure people's ability or when a test is administered to a small sample. According to Table 5, Infit and outfit MNSQ indices have an acceptable range (0.6—1.4). This is an important indicator of the lack of erratic responses and validity of scores. That is, as MNSQ indices show, there may be only few outliers (low ability people who unexpectedly answered a difficult item and high ability people who did not get an easy item right) that affect the Rasch model. Further, the mean scores of infit and outfit MNSQ statistics of 1.00 for items and .98 for people mirror the average fit of the items to the Rasch model's expectations.

Table 5. Item Measures and Fit Indices of all Items

Item	Score	Measure	SE	Infit MNSQ	Outfit MNSQ
1	432	.76	.07	1.05	1.04
2	474	.55	.07	1.03	1.04
3	582	-.01	.07	1.02	1.03
4	670	-.52	.08	0.93	.85
5	632	-.29	.08	0.90	.84
6	406	.90	.07	1.13	1.13
7	529	.27	.07	0.96	.93
8	790	-1.46	.10	0.96	.85
9	725	-.90	.09	0.91	.82
10	723	-.88	.09	0.96	.89
11	665	-.49	.08	1.01	1.02
12	439	.73	.07	1.02	1.02
13	500	.42	.07	1.05	1.04
14	479	.52	.07	1.02	1.01
15	677	-.57	.08	0.95	.91
16	486	.49	.07	1.02	1.00
17	454	.65	.07	1.06	1.05
18	536	.23	.07	0.98	.96
19	546	.18	.07	0.96	.91
20	553	.14	.07	1.03	1.05
21	497	.43	.07	1.00	.98
22	557	.12	.07	0.97	.91
23	609	-.16	.07	1.02	.97
24	718	-.85	.09	0.90	.76
25	616	-.20	.08	1.02	1.02
26	359	1.14	.07	1.16	1.22
27	465	.59	.07	1.03	1.06
28	608	-.16	.07	1.02	1.03
29	540	.21	.07	0.95	.95
30	624	-.25	.08	0.89	.79
31	564	.08	.07	0.96	.95
32	621	-.23	.08	0.96	.96
33	686	-.63	.08	0.96	1.08
34	684	-.61	.08	1.01	1.02
35	570	.05	.07	0.99	.96
36	474	.55	.07	1.19	1.26
37	610	-.17	.07	0.93	.85
38	484	.50	.07	1.02	1.01
39	555	.13	.07	1.01	1.00
40	521	.31	.07	1.03	1.00
41	563	.09	.07	1.04	1.09
42	627	-.26	.08	0.97	.98
43	710	-.79	.08	0.92	.78

Item	Score	Measure	SE	Infit MNSQ	Outfit MNSQ
44	446	.69	.07	1.05	1.11
45	758	-1.16	.09	1.01	1.08
46	695	-.69	.08	0.92	.84
47	585	-.03	.07	1.02	1.04
48	602	-.12	.07	1.07	1.15
49	422	.81	.07	1.07	1.08
50	596	-.09	.07	0.98	.94
Mean	573.3	.00	.08	1.00	.98

Note.  $n = 916$ .

MNSQ = Mean Square. SE = standard error of measurement.

To examine fit and person ability/item difficulty measures concurrently, Bond and Fox (2007) generated a bubble chart that plots measures and fit statistics. This analysis displays visually the relationship between ability/difficulty measures and the magnitude of measurement error. Figure 5 displays bubble charts of items' MNSQ statistics plotted against item difficulty (upper part) and person ability (lower part) measures; all item infit MNSQ statistics are closely distributed around the item fit mean (1.00), indicating good measurement properties of items. Figure 5 further shows that, as we expected, standard error (SE) of measurement is especially high in persons landed at the top of the hierarchy. The magnitude of SE is displayed as the size of the circles: the bigger the circle, the higher the SE. The reason for observing high SE indices for high-ability persons is that they did not get enough items corresponding to their ability level. Located at the top of the chart, these high-ability people answered all or the majority of items correctly; there is not enough information about their ability. For example, even if an individual with a high measured ability is most probably able to answer all items correctly, what type of item can inform us about the upper boundary of their ability? If these individuals answer a sufficient number of items, then we can collect more information about their ability as compared with when they do not receive sufficient number of items. As we move down the person bubble chart, the SE size decreases. This is due to the fact that lower ability people received enough items which corresponded to their ability; and their ability was therefore estimated with lower error.

There was no misfitting person in the sample. According to Pollitt and Hutchinson (1987), if person misfit does not exceed 2% of the data, then there is no significant erratic response pattern; we can opt for acceptable person performance, indicating that their performance accords with the expectations of the model.

To analyze possible patterns or structures in residuals, we performed a principal component analysis of residuals (PCAR). This analysis demonstrates “*contrasts* between opposing factors, not loading on one factor”, i.e., the contrast between positive and negative loading values (Linacre, 2009, p. 216). PCAR is a test of unidimensionality of the data set, a prerequisite to the Rasch model analysis; unidimensionality holds when test scores are not contaminated by any irrelevant factor and means that no datum affects the other one in the data set (Linacre, 2009). If no structure or pattern is observed in residuals, the variation in data which is not explained by the Rasch model is “random noise” (Linacre, 2009). It is expected that the correlation between the random noise of two items be ideally zero or very weak.

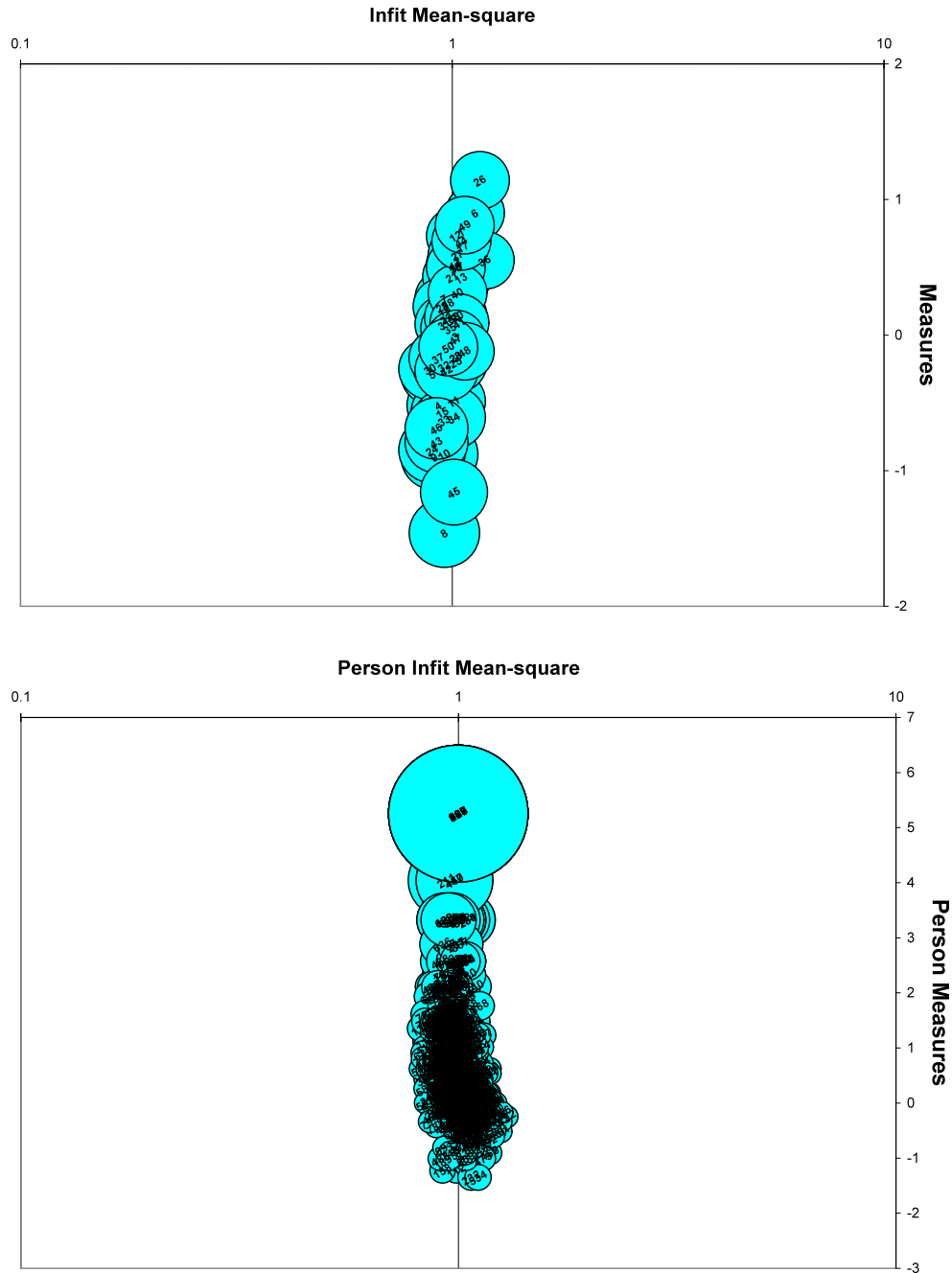


Figure 5. Bubble Chart Plotting Item and Person Infit MNSQ Statistics against Item and Person Measures. The size of bubble for items (upper chart) is consistent and small. This implies that the SE was small for items. Bubbles representing persons (lower chart) range from big to small, indicating that the ability of more proficient people has been estimated with greater amounts of SE.

According to our PCAR analysis, total variance explained by measures was 27.7%. This Rasch dimension is similar to the first factor in a principal component analysis of raw data, but it is based on linearized data (Wright, 1996). If the difference between the variance explained by Rasch dimension and the noise is considerably high, the unidimensionality of the test is supported.

Three weak factors were identified in residuals. The first one extracted 1.7 out of 50 Eigenvalue units, which is less than 5%; the strength of the Rasch dimension is almost 25 times this factor. Linacre (2009) argued that the smallest Eigenvalue regarded a structure or pattern in residuals is three; the observed value (1.7) does not reach Linacre's benchmark (1.7 > 3). This factor comprises two items, 17 and 25. But, as we observe in Table 5, their infit and outfit MNSQ statistics are only slightly deviating from 1, indicating that they are as predictable as the model expects. Factors two and three did not extract considerable Eigenvalue units. So, the observed factors are not "contradictory dimensions" (Linacre, 2009); this observation provides evidence for the unidimensionality of the test (Wright & Stone, 1999). Also, the analysis of correlation between item residuals showed there was no significant correlation between item residuals. This finding backs up adherence to the local independence of items.

*Second analysis: Testing for invariance.* As an additional step in understanding construct threats, we performed a uniform differential item functioning (DIF) analysis to examine any gender bias. According to Linacre (2009), for a DIF to be significant, two criteria should hold: "1. probability so small that it is unlikely that the DIF effect is merely a random accident 2. size so large that the DIF effect has a substantive impact on scores/measures on the test" (p. 148). The minimum noticeable DIF difference is 0.5 logits for items and the probability of observing DIF in items should be less than 0.05. Thus, a considerable DIF is not merely a function of the significance, but the difference should also have statistical substance.

DIF measure in Table 6 displays the difficulty of each item for a gender class; class 1 is male and class 2 female. For example, the local difficulty of item 1 for the Male Class is 0.88 logits and for the Female Class is 0.65 logits. A positive DIF contrast index indicates that the item is more difficult for the first group and a negative index shows the item is more difficult for the second group. As we observe in Table 6, item 1 is 0.23 logits more difficult for male candidates whereas item 2 is -0.18 logits more difficult for female candidates. The Welch t test expresses DIF significance as a two-sided Student's t-statistic. The null hypothesis is that the two DIF estimates are equal, considering measurement errors. The *p* value column shows the probability of the t with the degree of freedom (Linacre, 2009). Eight items have significant DIF t-tests ( $p < 0.05$ ). Items 6, 7, 21, 35, and 44 are more difficult for the Female Class (male candidates are more able on these items) and items 39 and 43 are more difficult for the Male Class (female candidates are more able on these items).

The observation of DIF needs to be further investigated to ascertain whether the observed DIF is a construct issue. If there is strong evidence that the DIF observed concerns some known construct issues, the item would most probably be retained in future administrations of the test. In the present analysis, because DIF is not balanced out in three items, the observed DIF, as we view it, attenuates the construct validity argument of the test. However, the effect of four DIF items is balanced out, which supports the construct validity of the test.

Table 6. Gender Differential Item Functioning in the MELAB Listening Test

DIF measure 1	DIF Measure 2	DIF Contrast	Welch <i>t</i>	<i>p</i>	Item
0.88	0.65	0.23	1.52	.129	1
0.48	0.65	-0.18	-1.18	.237	2
0.16	-0.11	0.27	1.77	.076	3
-0.56	-0.51	-0.05	-0.30	.764	4
-0.43	-0.16	-0.27	-1.72	.086	5
0.72	1.04	-0.31	-2.11	.035	6*
0.01	0.50	-0.49	-3.27	.001	7*
-1.38	-1.50	0.12	0.58	.560	8
-0.86	-1.00	0.13	0.73	.468	9
-0.83	-0.93	0.10	0.55	.580	10
-0.39	-0.66	0.28	1.68	.093	11
0.73	0.74	-0.01	-0.04	.966	12
0.35	0.48	-0.13	-0.85	.398	13
0.42	0.58	-0.16	-1.05	.294	14
-0.71	-0.49	-0.22	-1.32	.187	15
0.41	0.57	-0.16	-1.05	.293	16
0.74	0.55	0.19	1.27	.203	17
0.31	0.11	0.20	1.34	.179	18
0.07	0.31	-0.24	-1.58	.114	19
0.15	0.14	0.01	0.06	.955	20
0.28	0.64	-0.37	-2.46	.014	21*
0.01	0.20	-0.19	-1.24	.217	22
-0.10	-0.17	0.07	0.45	.654	23
-0.88	-0.81	-0.07	-0.41	.681	24
-0.18	-0.24	0.06	0.35	.724	25
1.21	1.08	0.13	0.85	.398	26
0.62	0.62	0.00	0.01	.991	27
-0.25	-0.08	-0.16	-1.03	.302	28
0.15	0.30	-0.15	-0.97	.331	29
-0.18	-0.34	0.15	0.97	.331	30
-0.03	0.16	-0.19	-1.25	.212	31
-0.20	-0.31	0.12	0.74	.461	32
-0.55	-0.68	0.13	0.76	.445	33
-0.58	-0.66	0.09	0.51	.609	34
-0.13	0.22	-0.35	-2.32	.020	35*
0.64	0.52	0.12	0.82	.414	36
-0.25	-0.10	-0.15	-0.96	.338	37
0.58	0.35	0.23	1.53	.125	38
0.40	-0.13	0.53	3.51	.001	39*
0.47	0.19	0.28	1.88	.060	40
0.22	0.02	0.20	1.33	.184	41
-0.23	-0.30	0.07	0.42	.675	42
-0.64	-0.98	0.34	1.96	.050	43*

DIF measure 1	DIF Measure 2	DIF Contrast	Welch <i>t</i>	<i>p</i>	Item
0.35	1.01	-0.66	-4.40	.000	44*
-1.05	-1.30	0.25	1.28	.200	45
-0.71	-0.68	-0.03	-0.18	.854	46
-0.04	0.04	-0.08	-0.51	.608	47
-0.18	-0.16	-0.03	-0.18	.856	48
1.04	0.55	0.48	3.23	.001	49*
0.05	-0.16	0.20	1.31	.191	50

*Note.* DIF measure 1 is the local difficulty of each items for male participants and DIF measure 2 is this index for female participants. The “\*” sing means that the item has a significant DIF.

### Weighing the Evidence for the Validity of MELAB

To sum up the findings in the current study, we use Chapelle’s (1994) table to display the evidence supporting or attenuating the validity of test scores’ interpretations. Table 7 demonstrates two groups of evidence. Evidence supporting construct validity consists of the results of the reliability analysis (cases above .70), content analysis which identified the factors and skills stated in the *MELAB Technical Manual*, CFA supporting the factor structure of the test, Rasch measures, fit, reliability, and PCAR which supported the absence of construct irrelevant factors, invariance analysis showing that the majority of items functioned similarly across gender subgroups. On the other hand, the reliability indices smaller than .70, the CUM and higher-order models, and DIF in three items attenuate the construct validity of the test.

Table 7. Evidence Supporting and Attenuating the Construct Validity Argument of the MELAB Listening Test

Evidence supporting construct validity	Evidence against construct validity
1) KR-21 analysis (above .70)	1) KR-21 analysis (below .70)
2) Content analysis	2) DIF not balanced out in three items
3) CFA	
4) Rasch measures in 50-item analysis	
5) Infit and outfit in 50-item analysis	
6) PCAR with 50 items	
7) Rasch reliability indices	
8) Invariance analysis: Four DIF are balanced out	

According to Table 7, the argument for construct validity of the MELAB test is supported by more evidence than it is attenuated by counterevidence.

## Discussion

### First Objective: Factor Structure of the Test

To examine the features of the hypothesized three- and five-factor models and the cause of variation in items, we performed a CUM and a second-order CFA analysis to build the construct map or factor structure of the test. The five-factor model had a competency-based approach towards the test, which was not supported. Because the proposed five-factor model has a supported theoretical underpinning in literature and *MELAB Technical Manual*, we argue that future research should address this area in other test forms.

If item correlations are erratic, factors may not be successfully separated, and the expected patterns will not be generated in factor analysis. In the five-factor CFA model we assumed that in answering some items test takers rely on their ability to comprehend explicitly stated information and in others the ability to comprehend implicitly stated information. However, as Wagner (2004) argued and as this study showed, separating these two skills may not produce optimum results or models in measurement. Even in Kinsch and van Dijk's (1983) model of comprehension, these two processes take place simultaneously. Therefore, we suggest that this dichotomy may be only an artefact.

The main hurdle to performing the CUM analysis was that the traits were not measured by two or more common methods. As a compensatory strategy, we posited a three-factor, task-based model according to the test sections. We thus "moved from a strictly confirmatory mode to an exploratory mode...to arrive at a model that would provide a reasonable explanation for the correlations among their variables" (Bachman, 2004, p. 285); we revised the CUM model and hypothesized that the underlying factors—ability to understand minimal context stimuli, short conversations, and longer radio interviews—are separable and cause the observed variation in data. This model had acceptably good fit and provided good support for the causality of test behavior by the three hypothesized latent traits: that all items loaded significantly onto the posited latent traits, as significant path coefficients showed, indicates that the variance in items is significantly accounted for by the latent traits, and latent traits are the cause of indicators (Hair et al., 2010). The analysis further showed that the correlation among these traits was significantly high.

The question of separability of listening traits has been dealt with in previous research. For example, Liao (2007) reported the correlation coefficient of 0.97 between explicit and implicit listening factors in the CFA study stating "these two factors are closely interrelated, but still not identical" (p. 60). Such a conclusion is in variation with the common school of thought: considerable correlation coefficients above .80 are indicative of significant similarity among the hypothesized factors and their inseparability (Hair et al., 2010). We argue that significantly high correlations of the three traits in turn can be evidence of the concurrent occurrence of the local and general comprehension strategies when test takers answer these items (van Dijk & Kintsch, 1983); the model fitting the data illustrates this relationship. The results show that comprehension is a complex and intertwined process and attempts to separate comprehension stages and skills may not be completely successful (see Bae and Bachman's [1998] study, where the separability of listening and reading traits, as two major and distinct skills, is not clearly established).

More recently, Borsboom, Mellenbergh, and Van Heerden (2004) redefined validity and argued that high correlations between different traits do not necessarily point to invalidity. According to Borsboom et al., the argument that significantly high correlations in CFA imply the presence of same traits leads the researcher into murky waters, pointing out that:

For instance, suppose one is measuring the presence of thunder. The readings will probably show a perfect correlation with the presence of lightening. The reason is that both are the results of an electrical discharge in the clouds. However, the presence of thunder and the presence of lightening are not the same thing under a different label. They are strongly related—one can be used to find out about the other—and there is a good basis for prediction, but *they are not the same*. (p. 1066; emphasis added)

By the same token, while the three hypothesized traits in the present study (the ability to understand minimal context stimuli, short conversations, and longer radio interviews) have caused the variation in the scores, as shown by significant regression weights, significantly high correlation coefficients among the traits (or factors) do not testify to the presence of identical traits. They are different, as lightening and thunder are, but also have high correlations, as lightening and thunder readings do. A more important observation is that the hypothesized traits were found to be causing a great amount of variation in scores.

That the arrows in Figure 2 and 3 move from the latent variable to items indicates that the variance in items is mainly caused by the trait, significant evidence of validity of the hypothesized trait (Borsboom et al., 2004). This observation carries an important implication: hypothetically, neither textual competence nor functional knowledge introduces measurable construct-irrelevant variance to measurement in minimal context items because they are principal component of the postulated trait; a reliable assumption would then be that minimal context items measure textual competence and functional knowledge. Yet, while they tap the intended construct, minimal context items belong to an older generation of listening items known as discrete-point items (Buck, 2001). The discrepancy between our findings, generally in favor of these items, and the mainstream literature, which highlights the reduction of context and communication as a shortcoming of these items, is worth further investigation: it seems that to answer the minimal context items, candidates use their prior knowledge and, more importantly, activate their textual competence, including vocabulary, syntax, and phonology (Bachman, 1990) and functional knowledge (*MELAB Technical Manual*, 2003). It is important to determine the extent to which candidates' inability to comprehend and respond correctly is caused by his/her inability to understand the meaning of phrases or certain lexical items and their lexico-grammatical relationships as apposed to the lack of a context. For example, Goh (2000) showed that some EFL learners can hear the words exactly and match them to sounds and words in their mental lexicon (recognition), but they may not be able to understand the prompt or stimulus. This question about minimal context items, we believe, should be further researched.

The second item type entails short conversations. From a competency-based viewpoint, these items are intended to measure the ability to comprehend explicitly and implicitly stated information. From a task-based viewpoint, items measure the ability to comprehend messages in short daily conversations if the interlocutor gets involved in such transactions. The former delineation assumes two dimensions for listening comprehension, whereas the latter definition hypothesizes a broader and less clearly partitioned construct.

When completing a task which is mainly based on listening comprehension skills, an interlocutor may use the ability to understand both explicitly and implicitly stated information, but they happen at the same time (Kintsch & van Dijk, 1978) and separating them is not a sound practice although some researchers, such as Shohamy and Inbar (1991), asserted that items measuring these two skills must be always present in any test of listening comprehension. We should not always expect having two clearly separate dimensions which function on their own. Because the content analysis supported Shohamy and Inbar's (1991) hypothesis, which resonates with a competency-based approach, but the trait was not divisible in the present study (see also Hansen and Jensen, 1994; Buck, 2001; Wagner, 2004), we tentatively propose that variations caused in the conversation items in the present study are attributable to a more general trait: the *ability to understand short context conversations* and the subskills have functioned to answer the items. Further research will be needed to elaborate on the connection between the general ability to understand short context conversations and its subskills.

The third section in the test includes the ability to comprehend longer interviews. The content analysis in our study showed that items tap two skills, understanding explicitly articulated information and the ability to make close paraphrases. Evidence was proposed that variation in the items is attributed to a general trait which we refer to as *the ability to comprehend lengthy pieces of discourse*, e.g., longer interviews and/or talks. This representation of the trait includes both task and competence features, which would be more properly construed as a task-based definition (Bachman, 2002).

In this light, we propose a tentative model with three correlated and relevant factors to explain the structure of this form of the MELAB listening test in this study (see Figure 6).

Figure 6 is a construct map or factor structure which displays traits and manifest variables (items). Now that such a map is proposed, future research can further evaluate its validity and reliability. In this model, there are three latent causal connections (double-headed arrows) that link the latent traits to manifest or measured variables. Therefore, three constructs presumably cause the responses and variation in them. For example, the presence of skills to understand minimal context stimuli, as a latent trait, is measured by items which operationalize this trait. These items target textual competence and textual knowledge, as is proposed in the construct map of the trait. On the whole, the results point to three clear factors in the test as defined by the sections in the test structure itself.

### **Second Objective: Construct Threats and Underrepresentation**

We performed two Rasch-based analyses to fulfill the second objective. The first analysis showed there was no misfit according to the fit MNSQ indices. When unidimensionality and local independence hold, the fact that all items fit the Rasch model supports "item function validity" (Wright & Stone, 1999). Item function validity (IFV) concerns the integrity of items and their functions: whether and how much their function agrees with or deviates from the expectations of the Rasch model. IFV assures the good measurement properties of the items in terms of their consistency with the model. In our study, easy items functioned according to the expectations of the model—in other words, high ability test takers answered easy items correctly but low ability test takers did not answer difficult items. This provides evidence for the absence of construct-irrelevant factors in items because erratic response patterns are the function of a trait other than the hypothesized one.

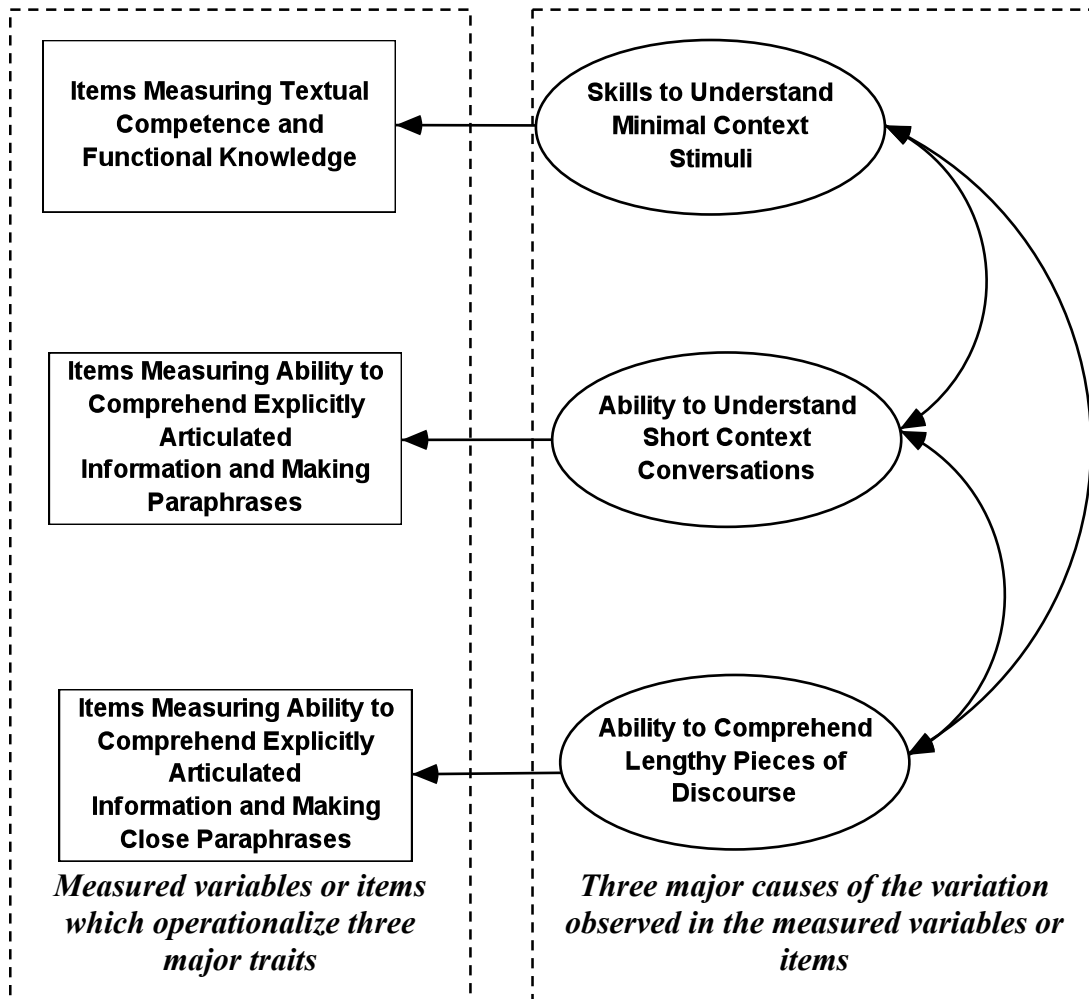


Figure 6. The Operationalized Listening Comprehension Model in the MELAB Listening Test. This model is a tentative construct map of the test which was analyzed in this study.

The principal component analysis of residuals (PCAR) showed that there was no substantive factor in residuals although two items loaded weakly onto the first identified factors in the residuals. The observation from PCAR lends support to the response validity (RV) of the test, which is determined from the observed differences between a response set and our expectations (residuals or random noise) (Wright & Stone, 1999). Large residuals are observed when lower ability persons answer a difficult item unexpectedly or when higher ability persons fail to answer easy items. That we found support for RV means that the Rasch dimension is dominant in the data and there is no conspicuous dimension beside it (Linacre, 2009). Therefore, both high and low ability students' performance resonates with the Rasch model expectations. One implication is that cheating, miskeying the data, fatigue, environmental factors, such as temperature, familiarity with personnel, and other facets (Bachman, 1990) did not contaminate the measurement. Taken as a whole, RV and IFV support the validity of the tests scores' meaning by providing counterevidence against construct irrelevant variance.

Another observation from Rasch measurement is that the item mean was slightly below the person mean index, and items were skewed downward on the item-person map. This indicates that the test has been relatively easy for the sample. According to the item-person map, there are few items suitable for measuring many persons with ability measures greater than 1 logit; the test displays a ceiling effect by not taking on a higher value. The consequence of the ceiling effect is that higher ability individuals are underestimated whereas lower or intermediate ability candidates tend to be overestimated. It may introduce some degree of construct underrepresentation in the MELAB test.

As a rule of thumb, adding more difficult items to the test may resolve the ceiling effect issue by revealing the true abilities of candidates in relation to these the targeted listening skills. This recommendation would be useful if the intent of the testing centre is to obtain detailed trait levels of candidates, particularly those with better overall listening abilities. However, if the aim (and we believe this may be the case) is to distinguish candidates who are able to perform satisfactorily based on a set of minimal requirements for entry to institutions of higher learning, then the MELAB listening test has sufficient validity to make this distinction.

We performed the second Rasch analysis to explore the invariance of the scores. The invariance or lack of DIF analysis helps generalizing the observed test results to expected scores (Aryadoust, 2009). However, observation of DIF in gender and other person factors is not always explicable. The inexplicability is either in terms of the item structure or what is known about the population. DIF may be observed in an item but other similar items may not display any DIF. Geranpayeh and Kunnan (2007) reported this “mysterious” DIF in a study of a Cambridge English exam. If we consider the content of such items to be the major cause, then we would expect to observe the same phenomenon across all similar items targeting the same trait. Analyzing the items that display DIF for gender in the present study, we found that DIF items did not lack any other feature that would have affected students’ performance, and neither did they possess an extra feature to affect performance. Also, DIF does not cause measurement problems if some items biasing a group are balanced out by another set of items biasing another group. In our study, four items are balanced out whereas three items are more difficult for females. The MELAB listening test has managed to keep some construct irrelevant factors at the minimum, such as the skill to read items and response options. This makes the task of interpreting DIF more complex because DIF is likely to be caused by a confounding variable.

## Conclusion

Validation does not always provide a definitive “yes” or “no” answer to validity inquiries (Chapelle, 1994). It is a dynamic process that never ends but develops as the science of measurement improves (Kane, 2004). The present study set out to determine the construct map of one form of the MELAB listening test, construct underrepresentation, and construct irrelevant threats; the validity of the MELAB listening test is supported by a considerable amount of evidence; multiple evidence from reliability and content analysis, CFA, and the Rasch model clearly support the construct validity argument although part of reliability analysis and DIF do not.

The study also showed the efficacy of CFA as a latent trait model in investigating the causality of test behavior and proposing construct maps underpinning a test. It also showed the efficiency of the Rasch model in investigating construct underrepresentation and irrelevant

factors. The main limitation of this study is that it only examined one form of the listening test. Although all forms of standardized tests are parallel, the claims in this study pertain to the test form that we examined. A replication of this study using other samples of participants and forms of the listening test will help deepen our understanding of some of the issues identified in this study.

As noted earlier, a future step in this area of research could be the examination of the influence of candidate's functional knowledge and textual competence on their test performance. This investigation helps us understand the validity of the minimal context items as a way of measuring listening comprehension. Further, although it was confirmed that the dichotomy of the comprehension of explicit or implicit information is an artefact, it is important to study further the effect of such test objectives on the difficulty of items in future research. Therefore, we propose two further validation inquiries:

- (a) What is the status of construct representation and construct irrelevant variance in other MELAB listening test versions?
- (b) How does the objective of the item (testing the comprehension of explicit or implicit information) affect item difficulty?

It is hoped that this study has provided some useful insights into the issues surrounding the examination of construct validity of the MELAB listening test.

## References

- Aryadoust, S.V. (2009). Mapping the Rasch-based measurement onto the argument-based validity framework. *Rasch Measurement Transactions*, 23(1), 1192–1193.
- Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.
- Bachman, L.F. (2002). Some reflections on task-based language performance assessment. *Language Testing* 19(4), 453–476.
- Bachman, L.F. (2004). *Statistical analyses for language assessment*. Cambridge: Cambridge University Press.
- 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 program. *Language Testing*, 15(3), 380–414.
- Bejar, I., Douglas, D., Jamieson, J., Nissan, S., & Turner, J. (2000). *TOEFL® 2000 listening framework: A working paper*. (TOEFL Research Report No. RM-00-07, TOEFL-MS-19). Princeton, NJ: Educational Testing Service.
- Bond, T.G. (2003). Validity and assessment: A Rasch measurement Perspective. *Methodologia de las Ciencias del Comportamiento*, 5(2), 179–194.
- Bond, T.G., & Fox, C.M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences*. London: Lawrence Erlbaum Associates.
- Bonk, W. J. (2000). Second language lexical knowledge and listening comprehension. *International Journal of Listening*, 14, 14–31.
- Borsboom, D., Mellenbergh, G.J., & Van Heerden, J. (2004). The concept of validity. *Psychological Review*, 111(4), 1061–1071.
- Brindley, G (1998). Assessing listening abilities. *Annual Review of Applied Linguistics*, 18(1), 171–191.

- Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York: The Guilford Press.
- Buck, G. (1990). *The testing of second language listening comprehension*. Unpublished PhD dissertation: University of Lancaster.
- Buck, G. (1991). The testing of second language listening comprehension: An introspective study. *Language Testing*, 8(1), 67–91.
- Buck, G. (1992). Listening comprehension: Construct validity and trait characteristics. *Language Testing*, 42(3), 313–357.
- Buck, G. (1994). The appropriacy of psychometric measurement models for testing second language listening comprehension. *Language Testing*, 11(2), 145–170.
- Buck, G. (2001). *Assessing listening*. UK: Cambridge University Press.
- Byrne, B. N. (2001). *Structural equation modeling with AMOS*. Rahwah, NJ: Lawrence Erlbaum Associates.
- Campbell, D. T., & Fiske, D. W. (1959) Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56(1), 81–105.
- Campbell, D. T., & O'Connell, E. J. (1967). Methods factors in multitrait-multimethod matrices: Multiplicative rather than additive? *Multivariate Behavioral Research*, 2(3), 409–426.
- Chapelle, C. (1994). Are C-tests valid measures for L2 vocabulary research? *Second Language Research*, 10(2), 157–187.
- Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (Eds.). (2008). Test score interpretation and use. In C.A., Chapelle, M.K., Enright, & J.M., Jamieson (Eds.), *Building a validity argument for the Test of English as a Foreign Language* (pp. 1–25). New York: Routledge.
- Conway, J. M, Lievens, F., Scullen, S. E., Lance, C.E., (2004). Bias in the correlated uniqueness model for MTMM data. *Structural Equation Modeling*, 11(4), 535–559.
- Cudeck, R. (1988). Multiplicative models and MTMM matrices. *Journal of Educational Statistics*, 13(2), 131–147.
- Dijk, T. A. van, & Kintsch, W. (1983). *Strategies of discourse comprehension*. New York: Academic Press Inc.
- Du Toit, M., & Du Toit, S. (2001). *Interactive LISREL: User's guide*. Lincolnwood, IL: Scientific Software International, Inc.
- Dunkel, P., Henning, G., & Chaudron, C. (1993). The assessment of a listening comprehension construct: A tentative model for test specification and development. *Modern Language Journal*, 77(2), 180–191.
- English Language Institute, University of Michigan. (2003). *Michigan English language assessment battery technical manual 2003*. Ann Arbor: English Language Institute, University of Michigan.
- English Language Institute, University of Michigan. (2009). *MELAB information and registration bulletin*. Ann Arbor: English Language Institute, University of Michigan.
- Eom, M. (2008). Underlying factors of MELAB listening construct. *Spain Fellow Working Papers in Second or Foreign Language Assessment*, 6(1), 77–94.
- Farhady, H. (1983). On the plausibility of the unitary language proficiency factor. In Oller, J.W., (Ed.), *Issues in language testing* (pp. 11–28), Rowley, MA: Newbury House.
- Geranpayeh, A., & Kunnan, A.J. (2007). Differential item functioning in terms of age in the certificate in advanced English examination. *LanguageAssessment Quarterly*, 4(2), 190–222.

- Glenn, E. (1989). A content analysis of fifty definitions in listening. *Journal of the International Listening Association*, 3(1), 21–31.
- Goh, C. (2000). A cognitive perspective on language learners' listening comprehension problems. *System*, 28(1), 55–75.
- Hair, J.F., Black, W.C., Babin, B.J., & Anderson, R.E. (2010). *Multivariate data analysis* (8<sup>th</sup> ed.). New Jersey: Pearson Educational Product.
- Haladyna, T.M., & Downing, S.M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement, Issues and Practices*, 23(1), 17–27.
- Hansen, C., & Jensen, C. (1994). Evaluating lecture comprehension. In J. Flowerdew (Ed.), *Academic listening: Research perspective* (pp. 241–268). Cambridge: Cambridge University Press.
- Hildyard, A., & Olson, D. (1978). Memory and inference in the comprehension of oral and written discourse. *Discourse Processes*, 1(1), 91–107.
- Jöreskog, K.G., & Sörbom, D. (2006). LISREL 8.8 for Windows [Computer software]. Lincolnwood, IL: Scientific Software International, Inc.
- Kane, M. (1992). An argument-based approach to validity. *Psychological Bulletin*, 112, 527–535.
- Kane, M. (2001). Current concerns in validity theory. *Journal of Educational Measurement*, 38(4), 319–342.
- Kane, M. (2002). Validating high-stakes testing programs. *Educational Measurement: Issues and Practice*, 21(1), 31–41.
- Kane, M. (2004). Certification testing as an illustration of argument-based validation. *Measurement: Interdisciplinary Research and Perspectives*, 2(3), 135–170.
- Kane, M. (2006). Validation. In R. L. Brennan (ed.), *Educational measurement* (4<sup>th</sup> ed.), (pp. 17–64). USA: American Council on Education, Praeger Series on Higher Education.
- Kenny, D. A. (1976). An empirical application of confirmatory factor analysis to the multitrait-multimethod matrix. *Journal of Experimental Social Psychology*, 65(4), 507–516.
- Kintsch, W., & van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review*, 85, 363–94.
- Kirsch, I.S., & Guthrie, J.T. (1980). Construct validity of functional reading tests. *Journal of Educational Measurement*, 17(2), 81–93.
- Lance, C. E., Noble, C. L., & Scullen, S. E. (2002). A critique of the correlated trait-correlated method and correlated uniqueness models for multitrait-multimethod data. *Psychological Methods*, 7(2), 228–244.
- Lievens, F., & Conway, J. M. (2001). Dimension and exercise variance in assessment center scores: A large-scale evaluation of multitrait-multimethod studies. *Journal of Applied Psychology*, 86, 1202–1222.
- Linacre, J. M. (2003). Size vs. significance: Standardized chi-square fit statistic. *Rasch Measurement Transactions*, 17(1), 918. Retrieved May 04, 2009, from <http://www.rasch.org/rmt/rmt171n.htm>.
- Linacre, J. M. (2005). WINSTEPS Rasch measurement [computer program]. Chicago: Winsteps.com.
- Linacre, J. M. (2009). *A user's guide to WINSTEPS*. Chicago: Winsteps.com.
- Liao, (2007). Investigating the construct validity of the grammar and vocabulary section and the listening section of the ECCE: Lexico-Grammatical ability as a predictor of L2

- listening ability. *Spain Fellow Working Papers in Second or Foreign Language Assessment*, 5, 37–78.
- Marsh, H.W. (1989). Confirmatory factor analysis of multitrait-multimethod data: Many problems and a few solutions. *Applied Psychological Measurement*, 13, 335–361.
- Marsh, H.W., & Hocevar, D. (1988). A new, more powerful approach to multitrait-multimethod analyses: Application of second-order confirmatory factor analysis. *Journal of Applied Psychology*, 73, 107–117.
- McNamara, T. (1996). *Measuring second language performance*. Longman: New York.
- Meccarty, F. (2000). Lexical and grammatical knowledge in reading and listening comprehension by foreign language learners of Spanish. *Applied Language Learning*, 11, 323–348.
- Messick, S. (1989). Validity. In R. L. Linn (Ed), *Educational measurement* (pp. 13–103). New York: ACE and Macmillan.
- Miles, J., & Shevlin, M. (2007). A time and a place for incremental fit indices. *Personality and Individual Differences*, 42(5), 869–874.
- Nation, I.S.P., & Newton, J. (2009). *Teaching ESL/EFL listening and speaking*. New York: Routledge.
- Oller, J. W., Jr. (1978). How important is language proficiency to IQ and other educational tests. In J.W. Oller, Jr. & K. Perkins (Eds.), *Language in education: Testing the test* (pp. 1–16). Rowley, Massachusetts: Newbery House Publishers.
- Oller, J. W., Jr. (1979). *Language tests at school*. London: Longman.
- Oller, J. W., Jr. (1983). Evidence for a general proficiency factor: An Expectancy grammar. In J.W. Oller, Jr. (Ed.), *Issues in language testing research* (pp. 3–10). Rowley, Massachusetts: Newbery House Publishers.
- Oller, J. W., Jr., & Hinofitis, F.B. (1980). Two mutually exclusive hypotheses about second language ability: Indivisible or partly divisible competence. In J. W. Oller, Jr., & K. Perkins (Eds.), *Research in language testing* (pp. 13–23). Rowley, Massachusetts: Newbery House Publishers.
- Pallant, J. (2007). *SPSS survival manual: a step by step guide to data analysis using SPSS for Windows*. New York: McGraw Hill/Open University Press.
- Pollitt, A., & Hutchinson, C. (1987). Calibrating graded assessments: Rasch partial credit analysis of performance in writing. *Language Testing*, 4(1), 72–92.
- Raykov, T., & Marcoulides, G., A. (1999). On desirability of parsimony in structural equation model selection. *Structural Equation Modeling: A Multidisciplinary Journal*, 6(3), 292–300.
- Richards, J. C. (1983). Listening comprehension: Approach, design, procedure. *TESOL Quarterly*, 17(2), 219–39.
- Rost, M. (1990). *Listening in language learning*. New York: Longman.
- Sawaki, Y., Sticker, L.J., & Andreas, H.O. (2009). Factor structure of the TOEFL Internet-based test. *Language Testing* 26(1), 5–30.
- Schmitt, N., & Stults, D. M. (1986). Methodology review: Analysis of multitrait-multimethod matrices. *Applied Psychological Measurement*, 10(1), 1–22.
- Scholz, G., Hendricks, D., Spurling, R., Johnson, M. & Vandenburg, L. (1980). Is language ability divisible or unitary? a factor analysis of 22 English language proficiency tests. In J.W. Oller, Jr. & K. Perkins (Eds.), *Research in language testing* (pp. 24–33). Rowley, Massachusetts: Newbery House.

- Schumacker, R.E., & Lomax, R.G. (2004). *A beginner's guide to structural equation modeling*. Mahwah, NJ: Lawrence Erlbaum.
- Shin, S. (2008). Examining the construct validity of a web-based academic listening test: An investigation of the effects of response formats. *Spaan Fellow Working Papers in Second or Foreign Language Assessment*, 6(1), 95–129.
- Shohamy, E., & Inbar, O. (1991). Construct validation of listening comprehension tests: The effect of text and question type. *Language Testing*, 8(1), 23–40.
- Steiger, J.H. (2007). Understanding the limitations of global fit assessment in structural equation modeling. *Personality and Individual Differences*, 42(5), 893–898.
- Thompson, B. (2004). *Exploratory and confirmatory factor analysis: Understanding concepts and applications*. Washington, DC: American Psychological Association.
- Tomás, J. M., Hontangas, P. M., & Oliver, A. (2000). Linear confirmatory factor models to evaluate multitrait-multimethod matrices: The effects of number of indicators and correlation among methods. *Multivariate Behavioral Research*, 35(4), 469–499.
- Tsui, A. B. M., & Fullilove, J. (1998). Bottom-up or top-down processing as a discriminator of L2 listening performance. *Applied Linguistics*, 19(4), 432–451.
- Uebersax, J. S. (2006). *The tetrachoric and polychoric correlation coefficients*. Retrieved April, 15, 2010, from the Statistical Methods for Rater Agreement web site at: <http://john-uebersax.com/stat/tetra.htm>.
- Wagner, A. (2002). Video listening tests: A pilot study. *Working Papers in TESOL and Applied Linguistics, Teachers College, Columbia University*, 2/1.
- Wagner, A. (2004). A construct validation study of the extended listening sections of the ECRE and MELAB. *Spaan Fellow Working Papers in Second or Foreign Language Assessment*, 2, 1–23.
- Widaman, K. F. (1985). Hierarchically nested covariance structures models for multitrait-multimethod data. *Applied Psychological Measurement*, 9(1), 1–26.
- Widaman, K. F. (1992). Multitrait-multimethod models in aging research. *Experimental Aging Research*, 18(2), 185–201.
- Widaman, K. F. (2002). To parcel or not to parcel: Exploring the question, weighing the merits. *Structural Equation Modelling*, 9(2), 151–173.
- Wilson, M. (2005). *Constructing measures: An item response modelling approach*. New York: Psychology Press.
- Wright, B. (1996). Local dependency, correlations, and principal components. *Rasch Measurement Transactions*, 10(3), 509–511.
- Wright, B., & Stone, M. (1979). *Best test design*. Chicago: MESA.
- Wright, B. D., & Stone, M. H. (1999). *Measurement essentials*. 2<sup>nd</sup> ed. Wilmington, Delaware: Wide Range, Inc.

### Acknowledgements

We gratefully thank the Spaan Fellowship program for supporting this research.

**Appendix 1**

Items	Minimal context	Explicit information	Close paraphrase	Propositional	Enabling	Error	R
V1	0.15					0.20	0.095
V2	0.15					0.20	0.099
V3	0.14					0.19	0.10
V4	0.17					0.14	0.18
V5	0.21					0.14	0.23
V6	0.097					0.21	0.042
V7	0.21					0.17	0.20
V8	0.11					0.078	0.13
V9	0.16					0.12	0.18
V10	0.14					0.11	0.14
V11	0.12					0.16	0.086
V12	0.17					0.19	0.13
V13	0.13					0.20	0.081
V14	0.16					0.20	0.12
V15	0.17					0.13	0.17
V16				0.15		0.20	0.098
V17					0.12	0.20	0.072
V18			0.17			0.19	0.14
V19					0.19	0.18	0.16
V20					0.14	0.19	0.097
V21			0.17			0.19	0.13
V22				0.17		0.19	0.13
V23					0.14	0.16	0.11
V24				0.17		0.11	0.20
V25				0.11		0.17	0.072
V26				0.066		0.20	0.021
V27				0.15		0.19	0.10
V28					0.14	0.18	0.099
V29				0.19		0.19	0.16
V30			0.23			0.14	0.26
V31				0.19		0.16	0.19
V32					0.18	0.16	0.17
V33				0.13		0.14	0.11
V34					0.12	0.16	0.089
V35				0.17		0.19	0.13
V36				0.045		0.22	0.0093
V37				0.19		0.16	0.19
V38				0.15		0.19	0.11
V39				0.14		0.19	0.098
V40		0.16				0.19	0.12
V41			0.13			0.19	0.079
V42			0.16			0.16	0.15
V43			0.17			0.12	0.18
V44			0.16			0.19	0.11
V45			0.075			0.11	0.047
V46		0.20				0.12	0.25
V47		0.16				0.18	0.12
V48		0.14				0.17	0.10
V49		0.15				0.20	0.11
V50		0.18				0.17	0.17