Examining the notion of listening sub-skill divisibility and its implications for second language listening

Christine Chuen Meng Goh and Vahid Aryadoust

International Journal of Listening, 29(3), 109-133

Taylor & Francis (Routledge)

Copyright © 2014 Taylor & Francis

This is an Accepted Manuscript of an article published by Taylor & Francis in International Journal of Listening on 30/09/2014, available online: http://www.tandfonline.com/doi/full/10.1080/10904018.2014.936119

Notice: Changes introduced as a result of publishing processes such as copy-editing and formatting may not be reflected in this document. For a definitive version of this work, please refer to the published source.

Examining the Notion of Listening Sub-Skill Divisibility and its Implications for Second Language Listening

Abstract

The testing and teaching of listening has been partially guided by the notion of sub-skills, or a set of listening abilities that are needed for achieving successful comprehension and utilisation of the information from listening texts. Although this notion came about mainly through applications of theoretical perspectives from psychology and communication studies, the actual divisibility of the sub-skills has rarely been examined. This paper reports an attempt to do so by using data from the answers of 916 test takers of a retired version of the Michigan English Language Assessment Battery (MELAB) listening test. First, an iterative content analysis of items was carried out, identifying five key sub-skills. Next, the discriminability of sub-skills was examined through confirmatory factor analysis (CFA). Five independent measurement models representing the sub-skills were evaluated. The overall CFA model comprising the measurement models showed excessively high correlations between factors. Further tests through CFA resolved the inadmissible correlations, though the high correlations persisted. Finally, we made 23 aggregate-level items which were used in a higher-order model, which induced best fit indices and resolved the inadmissible estimates. The results show that the sub-skills in the test were empirically divisible, hence lending support to scholarly attempts in discussing components in the listening construct for the purpose of teaching and assessment.

Keywords: confirmatory factor analysis; listening sub-skills; second language listening; sub-skill divisibility;
The teaching and testing of listening has been guided in part by the notion of sub-skills, or a set of listening abilities that are needed for achieving successful comprehension and utilisation of the information from listening texts. An underlying assumption of this approach is that by distinguishing the listening components and operationalizing them in a language curriculum or a language test, English learners will be able to build up a repertoire of skills that enable effective comprehension while test takers’ ability levels can be accurately reflected by their test results. The ability to use certain sub-skills or a combination of these skills would therefore distinguish one learner or test taker from another (Buck & Tatsuoka, 1998). In spite of this received practice of focusing on listening sub-skills, the psychological reality of some of these individual skills and the way they combine to bring about comprehension has yet to be ascertained empirically. This paper attempts to offer some insights from a study that examined whether different abilities considered to be important for academic listening can be empirically separated based on test-takers’ performance in an international standardised listening test. It begins with a review of the introduction of sub-skills into second language teaching, and some findings and discussions that have contributed to our current understanding of the issue of sub-skill divisibility. Next, it describes our study which investigated the discriminability of different parts of the test and discusses the results in relation to expressed test aims. It discusses the implications of the results for teaching and assessing listening.

**Sub-skills and Second Language Listening**

The notion of sub-skills was introduced in various discussions about second language (L2) listening skills during the dawn of the communicative language teaching era. Listening and other language skills were defined as complex skill operations comprising a number of
EXAMINING THE NOTION OF LISTENING SUB-SKILL DIVISIBILITY
AND ITS IMPLICATIONS FOR SECOND LANGUAGE LISTENING

constituent elements (i.e., sub-skills) which enable listeners to achieve comprehension. Some well-known frameworks for sub-skills include:

- Carroll’s (1972) two-stage taxonomy of parsing spoken language which comprised the ability to comprehend linguistic information and relating it to the wider discourse;
- Oakeshott-Taylor’s (1977) model of listening macro- and micro-comprehension attributes;
- Aitken’s (1978) categories of major listening skills for a) comprehending vocabulary and syntax, b) identifying prosodic patterns, c) identifying attitudes, views, and intentions of interlocutors, and rhetorical clues, and d) making inferences; and
- Richard’s (1983) theoretical taxonomy of top-down and bottom-up sub-skills organised around the context of conversational and academic listening.

The sub-skill approach to assessing L2 listening comprehension recognizes the interwoven relationships between listening elements and seeks to create coherence in the development of sub-skills. Richard (1983) proposed a list of 18 academic listening “micro-skills” to describe simple and complex skills such as “[the] ability to identify purpose and scope of lecture [and] identify relationships among units within discourse” (Richards, 1983, pp. 229-239). Similarly, Flowerdew (1994, pp. 11-12) argued that academic listening has distinct features and challenges, as compared with conversational listening, such as “[the] amount of implied meaning or number of indirect speech acts” and “ability to concentrate on and understand long stretches of talk”.

Munby’s (1978) list of 250 sub-skills for each of the four language skills arguably is the lengthiest list ever proposed. For listening sub-skills, he proposed a complex model including recognising the role of intonation and discourse markers and selecting key points, resulting in the precise scope of the term sub-skill becoming relatively undefined. Other authors have also
suggested lists of listening phases, for example, Rixon’s (1981) five-stage framework which stresses the importance of predicting contents, sample discourse, and checking the hypotheses formed during listening.

Listening researchers have applied a variety of latent trait models to support the validity of speculative taxonomies. Drawing on Buck’s (2001) default listening construct, Wagner (2004) posited that measurable components of L2 listening would be correlated and include the ability to comprehend explicitly and implicitly articulated information. Although similar models have been verified in reading studies, Wagner’s exploratory factor analysis (EFA) did not support the hypothesis, a finding which led to the assumption that listening sub-skills were not discriminable, that is, listening comprehension was not empirically multidivisible. He attributed this finding to the simultaneous operation of the hypothesized sub-skills.

Liao (2007) investigated the validity of Wagner’s model (2004). Contrary to Wagner’s findings, her research provided evidence favouring the statistical divisibility of items measuring the ability to understand explicit and implicit information, although some items did not load on the pre-specified factors. However, confirmatory factor analysis (CFA) showed that the correlation coefficient of these factors was .97, indicating that the factors were hardly discriminable.

Soon after Liao’s (2007) research, Eom (2008, p. 81) proposed that the MELAB listening test would rest upon a two-factorial structure consisting of “linguistic knowledge” and “comprehension” as measured by 14 subcomponents. As indicated by their correlation index (.85), these factors proved to be separable yet highly related. In the same year, Shin (2008) researched the underlying structure of a Web-based listening test and showed that a higher-
order multitrait-multimethod CFA model where aggregate-level test items—variables that are made on the basis of the combination of independent test items—could best capture the complexity of the underlying structure of the listening test. The traits measured by the test constructed a hierarchy and comprised the abilities to identify overarching main ideas, major ideas, and supports. Apart from the methodological differences between Shin’s and previous research, Shin used polytomous constructed response test items¹ which reliably measured the academic listening construct. Shin’s finding is in line with Bodie et al.’s (2011, p. 38) study of the Watson–Barker Listening Test (WBLT), a test of listening comprehension, who argued that dichotomous scales might not “fully reflect listening ability”.

More recently, cognitive diagnostic assessment (CDA) has been applied to listening comprehension. For example, Sawaki, Kim, and Gentile (2009, pp. 200-201) reported that the underlying test structure could be explained by three dominant sub-skills: understanding “general and specific information”, “text structure and speaker intention”, and “connecting ideas”.

The aforementioned studies suggest that research into L2 and academic listening has produced mixed results with regard to the divisibility of sub-skills (see also Field, 2008; Flowerdew & Miller, 2006; Goh, 2010; Rost, 2002; Tsui & Fullilove, 1998; Vandergrift 2007). However, the supporting studies (e.g., Shin, 2008) indicate that listening comprehension sub-skills would generally include one or more of following sets of sub-skills, which also feature in many recent discussions of the nature of learner listening:

(a) Understanding lexicogrammatical features (e.g., Shin, 2008);

(b) Understanding explicitly stated information (e.g., Field, 2008);

¹ Unlike dichotomous test items which have only two possible results (e.g., 0 and 1), polytomous items have more than two possible results (e.g., 0, 1, and 2).
(c) Understanding paraphrases (e.g., Wagner, 2004);
(d) Identifying intentions, attitudes, and rhetorical clues (e.g., Vandergrift 2007);
(e) Making inferences (e.g., Tsui & Fullilove, 1998); and
(f) Drawing conclusions (e.g., Liao, 2007; Sawaki et al., 2009).

Proponents of listening sub-skills nevertheless cautioned against any attempt at assessing individual sub-skills, particularly for decoding (e.g., identifying vowels or recognizing vocabulary) because of the view that these skills or processes function concurrently to achieve an outcome and therefore cannot be evaluated separately (e.g., Dunkel, Henning, & Chaudron 1993). For example, tests that assess the ability to identify prosodic features cannot be said to assess comprehension, because the perception of prosodic patterns is important only to the extent that it assists the listener to use more global skills such as understanding the surface meaning (Dunkel et al., 1993).

In effect, the verification of the divisibility of listening comprehension would confer important benefits on listening assessment and teaching (Hansen & Jensen, 1994; Shohamy & Inbar, 1991). For example, it would attest to the genuineness of the speculative taxonomies that have been used in both testing and teaching. Such evidence can be utilized to construct fair (diagnostic) assessment systems which furnish fine-grained information concerning weaknesses and strengths of test takers. An efficient (diagnostic) assessment system—which would operate for the benefit of test takers—would provide both subscores and overall scores on the basis of test takers’ performance on each bundle of items which tap a specific sub-skill (Kunnan & Jang, 2008;). Such a system demands evidence of divisibility of sub-skills, because supporting evidence which can provide subscores to assist test takers would make for a more reliable practice (Sawaki, Quinlan, & Lee, 2013).
Learners also benefit from a sub-skill approach to listening in L2 pedagogy and (diagnostic) assessment. For example, learners who receive helpful skill-based instructions on preparing for academic lectures would perform better at the comprehension of lectures and tutorials. The focus on sub-skills in academic listening is a significant part in the process of developing L2 listening syllabi and materials. For example, Espeseth’s (2004) and Sanabria’s (2004) course books of English for Academic Proficiency (EAP) listening focus on improving students’ ability to listen for details, summarize the message, take notes, and understand implicit messages. These materials expose students to relatively extensive oral texts to improve their attention span and concentration, alongside intensive materials to strengthen other sub-skills such as “whole utterance recognition” (Lynch, 2010, p. 82).

In light of the importance of listening sub-skills ascribed to listening pedagogy, particularly in academic listening, and the paucity of empirical research in examining the listening construct, the present study was an attempt to investigate the issue of divisibility through an analysis of data from an academic listening test. It aimed to answer the following questions:

(a) Are there empirically divisible listening sub-skills underlying L2 listening test performance?

(b) If so, what kind of relations do they have?

The Study

The Test

A retired version of the Michigan English Language Assessment Battery (MELAB) listening test was used in this study. The test has three sections, with items tapping various sub-skills that could potentially form part of the test takers ability to understand English spoken in academic contexts in a comprehensive manner. The sections are a) 15 minimal-context items,
b) 20 short conversation items, and c) 15 radio interview items. The format is entirely multiple-choice, and the audio stimuli are played only once (see the CaMLA website).

In Section 1, the candidate hears a question or a statement and chooses the best answer in the test booklet according to the illocutionary force of the utterance, including understanding “the unexpected” statements or questions that would be posed to them in short daily exchanges where they are expected to instantaneously provide a response, and vocabulary that had been misunderstood by language users in real life (Johnson 2003, p. 36). In Section 2, the candidate hears short conversations and chooses an answer option that most accurately reflects the intent or outcome of the conversation. In Section 3, the candidate hears relatively long simulated radio broadcast excerpts, mini lectures, or interviews. She may take notes and refer to visual aids such as graphs and tables to help them answer comprehension questions based on the texts. Each section intends to tap into various listening sub-skills, which are discussed below (see Item Analysis subsection).

Data Source

The data were provided by the English Language Institute at the University of Michigan (ELI-UM), now CaMLA, comprised a retired version of the MELAB, and graded answers from 916 test takers, 425 female, 427 male, and 64 missing, who took the test from 76 countries. The participant-to-item ratio (916/50 = 18.32) was sufficient for CFA (Marsh & Hocevar, 1988). Information about test structure, aims and focus was obtained from the MELAB manual and other related documents.

---

2 Examples of test items for the current design of the MELAB test are available at http://www.cambridgemichigan.org/.
**Item analysis**

Item analysis was chiefly guided by the relevant literature and Johnson’s (2003) *MELAB Technical Bulletin*. We adopted the listening sub-skills that Johnson describes in the bulletin and, where necessary, relied on the findings of the literature survey, which was previously discussed. Johnson (2003, p. 34) stated that

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.

Johnson (2003, p. 34) further argued that, in addition to understanding surface meaning, the test demands activating a variety of listening comprehension sub-skills including “the capacity to make inferences and draw conclusions”. Based on this, we carried out an iterative and exploratory content analysis of individual test items in multiple meetings, discussing extensively our individual results before reaching a consensus on key ones. This resulted in a preliminary list of listening sub-skills, as follows:

(a) understanding vocabulary,

(b) understanding grammatical structures,

(c) understanding and responding to minimal context stimuli,

(d) understanding explicitly articulated information,

(e) making inferences,

(f) understanding propositional paraphrases,

(g) understanding the attitude of speaker(s), and

(h) drawing conclusions.
This initial set of sub-skills reflects those within the listening construct which was discussed earlier; however, the list was revised after multiple rounds of discussion because the sub-skills had considerable commonalities in definition and function (e.g., inference-making and understanding implicit information); because they would not result directly in comprehension although they would facilitate it (e.g., understanding vocabulary and grammatical structures); and because none of the test items seemed to assess some sub-skills that emerged from the literature (e.g., the ability to draw conclusions).

Only effective and measurable comprehension sub-skills were retained. For example, we decided that perception and morpheme recognition sub-skills would facilitate comprehension but were different from comprehension, as we made a decision to posit models that elicited major or complex sub-skills (see Dunkel et al., 1993, for a discussion).

In addition, every effort was made to postulate the “conceptually distinguishable” sub-skills that can be identified in “a minimal number of items” and agree with the item design descriptions (Kim & Jang, 2009, p. 839). Haberman and von Davier (2007) had cautioned against positing many excessively fine-grained sub-skills, as this would likely contribute to the invalidity of the model and that there would also be a trade-off between the number of sub-skills and convergence of a model (see Table 1).
Examining the Notion of Listening Sub-Skill Divisibility and Its Implications for Second Language Listening

Table 1
Description of the Listening Sub-skills Identified in the MELAB Listening Test

<table>
<thead>
<tr>
<th>Skills</th>
<th>Description</th>
<th>Items</th>
<th>No. of items</th>
</tr>
</thead>
<tbody>
<tr>
<td>Understanding and responding to the unexpected statements and/or questions (minimal context items)</td>
<td>Ability to understand illocutionary forces (e.g., invitations, offers, suggestions, and so forth) and find an appropriate response among the item options.</td>
<td>1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15</td>
<td>15</td>
</tr>
<tr>
<td>Understanding details and explicit information</td>
<td>Ability to comprehend individual lexical items, utterances and supporting information.</td>
<td>40, 46, 47, 48, 49, 50</td>
<td>6</td>
</tr>
<tr>
<td>Making close paraphrases</td>
<td>Ability to paraphrase the key phrases or use synonyms to answer the item.</td>
<td>18, 21, 30, 35, 36, 37, 38, 39, 41, 42, 43, 44, 45</td>
<td>13</td>
</tr>
<tr>
<td>Making propositional inferences</td>
<td>Ability to understand what might logically follow in a text.</td>
<td>16, 22, 24, 25, 26, 27, 29, 31, 33</td>
<td>9</td>
</tr>
<tr>
<td>Making enabling inferences</td>
<td>Ability to make inferences based on the causal relationships implied in the text.</td>
<td>17, 19, 20, 23, 28, 32, 34</td>
<td>7</td>
</tr>
</tbody>
</table>

Providing inter-coder reliability information is an important research requirement. A commonly adhered-to methodology is to propose guidelines for item coding and train coders to carry out the coding. The agreement of coders is subsequently sought and reported. It has been argued that, although useful, this method sometimes yields low reliability statistics in the first and/or second rounds due to the biases of coders; reliability improves and coder biases are “neutralized” only when raters convene, discuss with the rest of the group, and refine their performance (Compton, Love, & Sell, 2012, p. 351). Whereas there is no inaccuracy inherent in this method, it seems that the initial rounds of coding are sometimes inefficient, as they basically resemble individual training sessions for coders. To overcome these limitations, we took a directed qualitative content analysis approach where we started with the findings of our literature survey “as guidance for initial codes” (Hsieh & Shannon, 2005, p. 1277). This approach has been adapted in Buck and Tatsuoka’s (1998) listening

---

3 In the studies where coding rehearsal is conducted, if the targeted reliability is not obtained, an external independent coder will be required to make the final decision.
study with relative success. In short, we convened for multiple meetings and discussed the coding scheme and test items in detail. We concurred on the coding of the majority of test items (100% agreement) but modified the coding of others several times across meetings till we reached a full agreement. Approximately two weeks from the final round of joint coding, one of the researchers re-examined the finalized codes to ascertain the trustworthiness of the codes over time.

Here we provide further details about the coding process using items similar to the ones in the paper. The actual items are secure. According to Johnson (2003), the first sub-skill, to which an entire test section has been devoted, is the ability to understand minimal context stimuli, that is, the ability to comprehend and respond to statements (e.g., requests or invitations) or questions that are asked unexpectedly by people around us, for example, on a campus or at an office. (Interested readers are referred to CaMLA website.)

In trying to understand an utterance in the minimal context item presented in this item, a candidate needs to recognise and understand the word “When’s” in order to select ‘c’ as it is the correct answer. In addition, the candidate also needs to be able to parse the unit “When’s she going” grammatically to know that the speaker is referring to an event in the future. The listener is also expected to understand the illocutionary force of this inquiry and, in response, provide the date of her vacation. Although grammatical and vocabulary knowledge alongside understanding the illocutionary forces of the stimuli can facilitate and lead up to comprehension, they were not regarded as independent sub-skills (Dunkel et al., 1993).

Another sub-skill is the ability to understand details or explicit information. The following example engages this sub-skill:
The tea plant can grow to a very tall tree if it is left untrimmed. The plant grows in tropical regions; it cannot tolerate the cold climate of mountainous areas or the regions that receive an extremely low amount of precipitation ....

Q: Where does the tea plant grow?

a. mountainous regions
b. warm regions
c. tropical regions *

The short text in which the answer is located states that “The plant grows in tropical regions”. We identified five items that would engage this sub-skill. However, the response is not always explicit and the reader needs to understand paraphrases of the oral language to arrive at the answer. For example, understanding the text above could be evaluated through the following item:

Q: Where does the tea plant grow?

a. mountainous regions
b. cold regions
c. warm and humid regions *

The correct option is c but the text does not include the “exact wording” of the answer. To arrive at the answer, the successful listener is supposed to know that “tropical regions” in this text refers to “warm and humid regions”. We found that 13 items elicit the ability to understand paraphrases.

We also found that 16 items engage inference-making sub-skills. Two related inferences emerge from literature: propositional and enabling (Hildyard & Olson, 1978). Propositional
inferences are typically instances of syllogism or logical appeal. They are logical arguments without explicitly stated conclusions about which listeners should make inferences. We adapt a listening test item from Wagner (2004) which tests the ability to make propositional inferences:

There are probably more skunks alive today than there were a thousand years ago. This is mostly because human development usually involves clearing the land of tree cover. This is fortunate for skunks, because they like to live in open areas.

Q: According to the speaker, skunks have_____.

a. been raised by human as pets
b. evolved in the last thousand years
c. benefitted from the presence of humans*
d. almost gone extinct because of human development (Wagner, 2004, p. 10)

This item is an instance of syllogism, because the listener hears two sentences based on which she should draw an inference: She hears “human development usually involves clearing the land of tree cover” followed by “they like to live in open areas.” These two premises follow that skunks benefit from human’s deforestation activities, though the conclusion is not stated explicitly. Propositional inferences do not rely on the world knowledge; they are implicit conclusions of explicitly articulated sentences or ideas (Hildyard & Olson, 1978). By contrast, enabling inferences require that listeners connect the ideas or statements by using their world knowledge and schemata. Hildyard and Olson (1978) stated that “Enabling inferences are inferences that must be draw to make discourse coherent and, therefore, comprehensible” (p. 95). Wagner (2004) gives the following example:
A skunk’s odor works as a very good defence against predators. Skunks spray their musk at predators, causing a really bad smell.

Q. Why is the skunk’s smell a good defence against predators?

a. The predators can smell the skunks.
b. Skunks use their odor to blend in to their environment.
c. Skunks do not taste very good to predators because their odor is so bad.
d. The predators know that if they attack the skunks, they will end up smelling very bad.* (Wagner, 2004, p. 10)

There are two important pieces of information in the text, which must be understood to draw the inferences: "Skunks spray their musk at predators” and “causing a really bad smell”. Successful listeners realize that because spraying the musk at predators would irritate them, they avoid approaching the skunks.

**Modelling the Sub-skills**

A five-factor listening model comprising the following sub-skills was constructed:

(a) Understanding and responding to the unexpected statements and/or questions (minimal context items);

(b) Understanding details and explicit information;

(c) Making propositional inferences;

(d) Making enabling inferences; and

(e) Drawing conclusions.

Figure 1 shows the model accommodating five divisible sub-skills underlying the test. The boxes represent test items and the big ovals represent factors (latent variables or sub-skills) for listening. Each item has an error of measurement which is expressed as a small circle. Arrows going from latent variables to items represent prediction relationships (or statistical
causality): the observed variance in the item is mainly accounted for by the posited latent variables. Regression paths are indicated as unidirectional arrows (e.g., the unidirectional path connecting Minimal context to Item 1) and suggest that the latent variables cause variations in examinees performance on items; and correlations are indicted as bidirectional arrows (e.g., correlations of Minimal context and Propositional inferences).

**Factor Analysis**

Factor analysis (FA) is basically a variable simplification technique to identify factors or underlying constructs of tests or questionnaires and to reduce items into smaller groups of variables. It is performed by exploring the intercorrelation patterns between test items. It is supposed that the variance of test items is due to the variation in latent (unobservable) traits represented by factors (Schumacker & Lomax, 2010). Numerous researchers have successfully proposed factors representing cognitive and metacognitive strategies (e.g., Phakiti, 2008; Vandergrift, Goh, Mareschal, & Tafaghodatari, 2006) and different language abilities (Bae & Bachman, 1993; Bodie, 2011; Bodie et al., 2011; Oller, 1983; Oller & Hinofitis, 1980).

Confirmatory factor analysis (CFA) models are developed based on *preconceived* theories, (as opposed to exploratory FA models which discover the underpinning factor structure of data sets where there is no preconceived theory [Fabrigar, Wegener, MacCallum, & Strahan, 1999]). CFA allows the analyst to verify whether the postulated relations between items and latent variables exist (Schumacker & Lomax, 2010). In the present study, we specified the model by drawing the diagrams (e.g., Figure 1) and evaluated the significant features of the model such as fit statistics and the magnitude of correlation and regression coefficients.
Figure 1. The proposed models for the MELAB listening test.
We used Kline’s (1998) method for testing CFA because of its comprehensiveness when compared to other methods. It includes a two-stage CFA analysis where the measurement model is tested prior to the full structural model. The measurement model comprises a latent variable with strictly causal relationships to test items. For example, a measurement model in Figure 1 includes Minimal context latent variable, 15 items, and their error terms; it is hypothesized that Minimal context latent variable would account for the performance of test takers on those items (Schumacker & Lomax, 2010). The full structural model includes all relationships among the measurement models, which are primarily causal, although they can also be correlational (bidirectional arrows in Figure 1). If the structural model includes correlational relationships, then the entire analysis is called CFA (Jöreskog & Sörbom, 2001).

CFA was performed on the LISREL statistical computer package, Version 8.8 (Jöreskog & Sörbom, 2006). Measurement models were initially developed, followed by the CFA model for all the latent variables. This was followed by a calculation of a range of fit statistics for each of the models we developed.

To evaluate the adequacy of the model, we examined a number of features of the proposed model: fit of the model expressed as fit indices, regression statistics which show how much of the variance observed in the performance data is explained by (or caused by) the latent variables, and the correlation coefficients of the latent variables, which should fall between .30 and .80 (Schumacker & Lomax, 2010). Following Schumacker and Lomax, we used the following fit indices to report the fit of the model:

a) Chi-square test (\( \chi^2 \)), an index that shows the difference between the observed and implied covariance matrices. For larger samples, it tends to be significant; therefore,
other fit statistics have been developed that impose a penalty for the sample size and number of variables in the model.

b) $\frac{\chi^2}{df}$ (or normed $\chi^2$), the ratio of $\chi^2$ to the degree of freedom ($df$). This ratio should be small (preferably below 3) in well-fitting models.

c) RMSEA (Root Mean Square Error of Approximation), a measure that corrects for the tendency of the chi-square test to be significant in large samples. Lower RMSEA indices are desirable.

d) NNFI (Non-Normed Fit Index) or Tucker-Lewis Index, to compare the posited and the baseline models. It is possible for the NNFI to be greater than unity, but it is usually set at unity. Indices greater than .90 are desirable.

e) CFI (Comparative Fit Index), an index to test the fit of a postulated model relative to a baseline model; indices greater than .90 are desirable.

First, we evaluated five independent measurement models including minimal context, explicit information, close paraphrasing, enabling inference, and propositional inferencing latent variables. Overall, 14 items were deleted due to their low regression coefficients (< 0.30), leaving 36 items for CFA modelling (Hair et al., 2010). The 36-item CFA model was tested: the results indicated reasonable fit but the emergence of a few “offending estimates”, which in this study includes extremely high correlation indices some of which are greater than unity, made the 36-item model solutions non-admissible.

Large CFA factor correlations can be due to negligible item cross-loadings (i.e., there might not be exact zero loadings) or the dependence of some sub-skills on others (Schumacker & Lomax, 2010). Small cross-loadings can be eliminated if we aggregate (combine) related items into parcels and run an aggregate-level CFA analysis (see Little, Cunningham, Shahar,
Widaman, 2002). Aggregation converts the dichotomous into polytomous items whose joint statistical information is the same as their combination into a polytomous item at every point on the latent variable (local independence assumed) (Huynh & Mayer, 2003). As a general rule, the more categories in a polytomous item, the more is the total summed information in the item across the latent variable (Bond & Fox, 2007). Following Johnson’s (2003) FA study of MELAB, we combined odd-numbered items and then even-numbered items tapping the five sub-skills, making 23 aggregate-level items. Given the large correlations among factors, we further added a higher-order general listening factor on which the five factors were statistically regressed (Chen, West, & Sousa, 2006; Mulaik & Quartetti, 1997).

To examine the higher-order CFA model with aggregate or parcel items, we incorporated prediction relationships among the factors (i.e., single-headed arrows running from one lower-order factor to another one). That is, we postulated that the ability to understand explicit information would predict inferencing as well as understanding paraphrases. We further discuss parcel items and prediction relationships in the CFA model below.

**Parcel items**

The use of parcel or bundle items has become widespread in second language and educational assessment research (e.g., Aryadoust, 2013; Goh & Hu, 2013; Kunnan, 1998; Purpura, 1998; Phakiti, 2007, 2008; L. Zhang, Aryadoust, & L. J. Zhang, in press). Bandalos & Finney (2001) showed that approximately 20% of the articles published in a number of measurement journals had employed parcel items in CFA studies. Indeed, parcelling is an efficient technique when the test items are scored dichotomously, as dichotomous scales can inflate the chi-square index in CFA models (Bandalos, 2002) and/or result in inadmissible regression
or correlation coefficients (Schumacker & Lomax, 2010). Parcelling can also improve the reliability and fit of the CFA models (Little et al., 2002).

However, in discussions of parcelling, one controversial issue has been the number of parameters estimated. On the one hand, advocates of parcelling argue that using parcel items results in model parsimony (i.e., fewer parameter to estimate) and accordingly stable models (Bagozzi & Edwards, 1998). On the other hand, Hall, Snell, & Singer Foust (1999) contend that parcel-item CFA would not be as robust as CFA of individual items. In the present study, we advocate the former view which is in line with the common practice in language assessment literature. We agree with Bandalos (2002, p. 100) who argued that the application of parcel items to dichotomous items can “ameliorate the effects of coarsely categorized and nonnormally distributed item-level data on model fit.”

It is also worth noting that Pearson correlation matrix is not suitable for the factor analysis of dichotomous data (Uebersax, 2006). Therefore, the CFA models were all drawn from polychoric correlation matrices.

**Prediction relationships in CFA**

Researchers attempting to establish the empirical multidivisibility of second language sub-skills have used CFA and cognitive diagnostic models (CDMs) (Aryadoust et al., 2012; Lee & Sawaki, 2009; Sawaki et al., 2009). Traditionally, in both approaches, the statistical correlations of sub-skills—as represented by dimensions in CDMs and factors in CFA—are estimated. However, new research by Sawaki et al. (2013) and Aryadoust, Goh, and Lee (2012) shows that the relations between second language sub-skills are more complex than commonly postulated correlations. Sawaki et al. (2013, p. 88) found that, contrary to their initial presumption, some of the posited sub-skills would predict test takers’ performance on other sub-skills in a sourced-based writing test. They reported that the CFA model
incorporating the prediction relationships was “the most plausible” model among the tested CFA models. Sawaki et al. argued that if a parsimonious CFA model incorporating correlations among factors does not plausibly fit the second language test performance data, it might indicate that some of the important relationships among the factors are missing. The researcher then should consider incorporating the prediction relationships among factors to capture the complexity of the data.

Results

Preliminary Statistics and Reliability

SPSS for Windows (Version 16, SPSS Inc., Chicago, IL) was used to examine the mean (or facility indices), standard deviation, skewness, and kurtosis coefficients of the data set. Most items possessed skewness and kurtosis coefficients in the range between -2 and +2, an indicator of univariate normality (Schumacker & Lomax, 2010). The composite Cronbach’s alpha reliability coefficient was .84, which fell within the reliability range from .74 to .84 as reported in the test manual (Johnson, 2003). High Cronbach’s alpha values indicate the internal consistency of the data in the sense that items are interrelated.

Confirmatory Factor Analysis

Initially, we manufactured five independent measurement models which were the constituents of the five-factor CFA model, and then the five-factor CFA model accommodating all latent variables and their indicators (Kline, 1998). Following Hair et al., 2010, we deleted 14 items (items 2, 3, 6, 16, 17, 23, 26, 28, 29, 31, 36, 40, 44, and 45) out of 50 through iterative analysis of measurement models because their regression coefficients were below .30 and therefore their contribution to measuring the latent trait was fairly low.
The fit statistics of the measurement models and their modified forms are presented in Table 2. For example, the modified Minimal context model where Items 2, 3, and 6 were deleted fitted the data reasonably well ($\chi^2 = 48.94$, NNFI = 0.95, CFI = 0.96, RMSEA = 0.029). Next, the measurement models were incorporated into the CFA model which comprised 36 test items and five latent variables and fitted the data well ($\chi^2 = 795.70$, NNFI = 0.97, CFI = 0.97, RMSEA = 0.020). One correlation coefficient greater than 1.00 nevertheless emerged, thus rendering the model basically inadmissible (see Table 3). This is caused when the variance-covariance matrix, based on which CFA parameters are estimated, is non-positive definite (NPD), which is an “offending estimate” (see Brown, 2006, pp. 187). Thus, while the model fits the data, it is not admissible.

Next, we performed a higher-order aggregate-level CFA analysis (Widaman, 2002) by combining odd-numbered items and then even-numbered items and making 23 aggregate-level or parcel items, whose descriptive statistics and correlations are presented in Appendix A and B, respectively. As Figure 2 shows, the general listening predicts and lower-order latent variables, as opposed to the previously evaluated CFA models where the relationships were correlational. In other words, the lower-order factor are all sub-components of a general listening factor and the prediction relationships extends to the lower-order factors; for example, understanding explicitly stated information can predict understanding paraphrases. This model further postulates a prediction relationship between lower-order latent variables (Hair et al., 2010).
EXAMINING THE NOTION OF LISTENING SUB-SKILL DIVISIBILITY
AND ITS IMPLICATIONS FOR SECOND LANGUAGE LISTENING

Table 2
CFA Models of the MELAB Listening Test

<table>
<thead>
<tr>
<th>Model</th>
<th>$\chi^2$</th>
<th>df</th>
<th>$\chi^2/df$</th>
<th>NNFI</th>
<th>CFI</th>
<th>RMSEA</th>
<th>RMSEA 90% confidence interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minimal context model</td>
<td>99.46*</td>
<td>65</td>
<td>1.54</td>
<td>0.95</td>
<td>0.96</td>
<td>0.023</td>
<td>0.013 – 0.032</td>
</tr>
<tr>
<td>Modified minimal context $^a$</td>
<td>48.94*</td>
<td>27</td>
<td>1.81</td>
<td>0.95</td>
<td>0.96</td>
<td>0.029</td>
<td>0.014 – 0.042</td>
</tr>
<tr>
<td>Explicit information</td>
<td>5.07</td>
<td>9</td>
<td>0.56</td>
<td>1.03</td>
<td>1.00</td>
<td>0.000</td>
<td>0.000 – 0.023</td>
</tr>
<tr>
<td>Modified explicit information $^b$</td>
<td>0.70</td>
<td>5</td>
<td>0.14</td>
<td>1.06</td>
<td>1.00</td>
<td>0.000</td>
<td>0.000 – 0.023</td>
</tr>
<tr>
<td>Close paraphrase</td>
<td>99.46*</td>
<td>65</td>
<td>1.53</td>
<td>0.95</td>
<td>0.96</td>
<td>0.023</td>
<td>0.013 – 0.032</td>
</tr>
<tr>
<td>Modified close paraphrase $^c$</td>
<td>47.24</td>
<td>27</td>
<td>1.74</td>
<td>0.95</td>
<td>0.96</td>
<td>0.029</td>
<td>0.014 – 0.042</td>
</tr>
<tr>
<td>Propositional inference</td>
<td>36.61*</td>
<td>37</td>
<td>0.99</td>
<td>0.96</td>
<td>0.96</td>
<td>0.020</td>
<td>0.000 – 0.034</td>
</tr>
<tr>
<td>Modified propositional inference $^d$</td>
<td>4.50</td>
<td>5</td>
<td>0.90</td>
<td>1.01</td>
<td>1.00</td>
<td>0.000</td>
<td>0.000 – 0.043</td>
</tr>
<tr>
<td>Enabling inference</td>
<td>11.90</td>
<td>14</td>
<td>0.85</td>
<td>1.02</td>
<td>1.00</td>
<td>0.000</td>
<td>0.000 – 0.027</td>
</tr>
<tr>
<td>Modified enabling inference $^e$</td>
<td>0.58</td>
<td>2</td>
<td>0.29</td>
<td>1.06</td>
<td>1.00</td>
<td>0.000</td>
<td>0.000 – 0.045</td>
</tr>
<tr>
<td>Five-factor 36-item CFA</td>
<td>631.88*</td>
<td>454</td>
<td>1.40</td>
<td>0.97</td>
<td>0.97</td>
<td>0.021</td>
<td>0.017 – 0.024</td>
</tr>
<tr>
<td>Higher-order 25-item model</td>
<td>255.37</td>
<td>221</td>
<td>1.15</td>
<td>1.00</td>
<td>1.00</td>
<td>0.013</td>
<td>0.013 – 0.019</td>
</tr>
</tbody>
</table>

Note. $n = 916$. $df =$ degree of freedom. NNFI = Non-Normed Fit Index. CFI = Comparative Fit Index. RMSEA = Root Mean Square Error of Approximation.

$^a$ Deleted items = 2, 3, 6. $^b$ Deleted items = 40. $^c$ Deleted items = 36, 44, 45. $^d$ Deleted items = 16, 26, 29, 31. $^e$ Deleted items = 17, 23, 28.

Good fit is indicated by NNFI and CFI $\geq 0.95$, RMSEA $\leq 0.05$, $\chi^2/df < 3$, and non-significant $\chi^2$.

* $p < .01$.

Table 3
Correlation Coefficients of the Latent Variables of the 36- and 30-Item CFA Models

<table>
<thead>
<tr>
<th></th>
<th>Minimal context</th>
<th>Enabling inference</th>
<th>Propositional inference</th>
<th>Close paraphrase</th>
<th>Explicit information</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minimal context</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Enabling inference</td>
<td>0.92* (0.94)</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Propositional inference</td>
<td>0.98 (0.99)</td>
<td>0.92 (0.91)</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Close paraphrase</td>
<td>0.97 (0.98)</td>
<td>0.94 (0.89)</td>
<td>1.04* (0.95)</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Explicit information</td>
<td>0.74 (0.70)</td>
<td>0.77 (0.73)</td>
<td>0.88 (0.81)</td>
<td>0.87 (0.91)</td>
<td>1</td>
</tr>
</tbody>
</table>

Note. * non-admissible correlation index.

$^a$ Correlation coefficients of the latent variables in the 36-item five-factor model.

$^b$ Correlation coefficients of the latent variables in the 30-item five-factor model.
### Figure 2

Path model of the higher-order aggregate-level 25-item model. ($\chi^2 = 255.37$, NNFI = 1.00, CFI = 1.00, RMSEA = 0.013).
As expected, the higher-order aggregate-level 23-item model fitted the data extremely well ($\chi^2 = 255.37$, NNFI = 1.00, CFI = 1.00, RMSEA = 0.013) and resolved the inadmissibility of correlation coefficients. As displayed in Figure 2, coefficients of the regression paths between the higher-order listening factor and lower-order factors are significantly high (> 0.001). For example, for a one-unit increase in the higher-order latent variable *listening*, we expect to see 0.97 increase in the magnitude of the lower-order *minimal context* variable. Similarly, if the magnitude of *explicit information variable* increases for one unit, we expect to see a small but significant increase in *enabling inferences* variable (i.e., 0.18).

**Discussion**

This study investigated the divisibility of listening sub-skills in a retired version of the MELAB listening test through an examination of mixed evidence from modelling listening as a five-factor model. We found evidence for the divisibility of the five sub-skills posited earlier:

(a) Understanding and responding to the unexpected statements and/or questions (minimal context items)

(b) Understanding details and explicit information

(c) Making propositional inferences

(d) Making enabling inferences

(e) Drawing conclusions

We discuss our findings in more detail here. A notable finding in the item-level CFA modelling was the presence of high correlations among factors representing sub-skills ($p < 0.001$), making the models inadmissible because discriminant validity has been violated (Brown 2006; Hair et al., 2010). After the data were further tested through the higher-order
aggregate-level CFA model where five lower-order factors were statistically regressed on a general listening factor, the results supported distinctions among factors. This observation resonates with Bodie et al. (2011), Eom (2008), Hansen and Jensen (1994), Sawaki et al. (2009), Shin (2008), and Shohamy and Inbar (1991), and partly with Liao (2007) who suggested or found that listening sub-skills are divisible, but contradicts the findings of Wagner (2004). In addition, we found that the factor structure of the test is much more complicated than what a simple EFA or CFA could represent. That is, all factors were attributed to a higher-order factor, and the inference making and understanding paraphrase factors were partly predicted by the ability to understand explicit information. We suggest that divisibility was not observed in previous EFA research by Oller (1983), Oller & Hinofitis (1980), and Wagner (2004) due to the statistical devices employed which were not able to represent otherwise known interconnections of factors which represent sub-skills.

An important reason for this interconnection could be that listening performance is attributable to the presence of a general listening ability. Therefore, the identified sub-skills might not operate in isolation but in unison to facilitate the achieving of a listening comprehension goal. For example, although a group of items measure the ability to make inferences and understand paraphrases, to achieve this type of comprehension test takers must initially understand the explicitly stated information. This relationship is represented by regression paths, indicating that while parts of the performance on inferences and paraphrase items can be explained by the ability to understand explicit information, these sub-skills are indeed separate and contribute to the general listening ability (Aryadoust, Goh, & Lee, 2011).

Relatively, to make inferences or draw conclusions from short conversations, the candidate also has to use similar sub-skills to interpret the outcome of the conversation. For example,
the questions in the minimal context section would require candidates to concurrently activate their linguistic competence, including vocabulary, syntax, and phonology. The single utterance prompts in the version of the test we examined (see Item analysis section) contained what could be new or unfamiliar lexical items including phrasal verbs and idioms. Choosing the correct answer depended heavily on the candidate’s hearing, simultaneous recognising and understanding these vocabulary items before they were able to apply their functional knowledge to interpret the illocutionary force of the message. Thus, performance in this section appears to rely on fundamental decoding skills of sounds discrimination and making sound-script connection and even vocabulary in some instances when unfamiliar lexical items are included in the prompts. These skills are already in use when test takers answer test items tapping other sub-skills. For example, the inclusion of slightly less common words and idioms requires listeners to activate top-down processes to assist meaning building, which seems to combine two uses of context: compensatory use to deal with unknown words and use to build higher-level meaning.

Another example is the test items such as:

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.* (Fleurquin el al., 2009).

The test taker needs to recognise the units of relevant information (RIs), “Let’s”, “Yeah”, and “good”. A difference (with minimal context items) here is that a simple context has been built up and the candidate may use the clues from B’s second utterance “I don’t want to (wanna)
stay home” to monitor and evaluate their understanding, or as a strategy for choosing the right option. From this we can see the redundancy and overlaps of the sub-skills needed for answering different types of questions which are aimed at testing different sub-skill, such as understanding details and drawing conclusions. This kind of model raises questions about how connected to each other sub-skills really can be; and the test formats that aim to tap different aspects of a global listening skill actually load on to the different yet connected set of central processes. This has implications for the on-going effort in the field to postulate a model of listening.

This hypothesis alludes to the theoretical propositions made by Bechtel and Abrahamsen (1991) about human cognition, which propose parallel processing of language through a spreading activation of interconnected or associative neural networks in the brain (Bechtel & Abrahamsen, 1991). An effective L1 and L2 listener is someone who is able to carry out parallel processing through the activation of different skills and items of knowledge (Hulstijn, 2003; Vandergrift, 2007).

It seems that sub-skills in the present L2 context functioned in an interactive and interdependent manner in listening events particularly those that are interactional in nature. This relationship between different levels of listening sub-skills is appropriately illustrated in Wolvin and Coakley’s (1996) metaphor of a tree where the roots are discriminative listening and the trunk represents general comprehension, while the branches refer to different types of higher-level listening, such as critical listening, which is contingent upon successful discriminative listening and general overall surface level comprehension (see Imhof & Janusik, 2006).
The second reason why the item-level CFA failed to support the divisibility of listening sub-skills might have to do with the nature of the scale and/or the statistical model used. A close examination of the relevant literature shows that the application of CFA to dichotomously scored tests of L1 or L2 listening has resulted in mixed findings. For example, Bodie et al. (2011) investigated the structure of the Watson–Barker Listening Test (WBLT)—a dichotomously scored L1 listening test—and found that the test structure did not emerge as multi-factor, contradicting the theoretical structure of the test but consistent with Wagner (2004). They discuss the implications for constructing listening tests, as follows:

Perhaps dichotomous scoring does not fully reflect listening ability, with the valid use of dichotomous scoring likely dependent on context […] Research across the academic landscape suggests that deriving meaning from conversation is more complex than picking out a single, correct meaning […] Consequently, right–wrong scoring may misrepresent the multitude of meanings that may be viable alternatives. Indeed, a person who is able to generate multiple alternative meanings from a given utterance may be a more proficient or competent listener. (Bodie et al., 2011, p. 38)

It follows that either L2 listening tests would discriminate high from low ability test takers better if they use polytomous scales, or that with binary tests researchers should use modelling approaches other than item-level CFA. Evidence favouring the former postulation is that aggregate-level items in the higher-order model established the divisibility and interrelations of the sub-skills, which is consistent with Shin’s (2008) findings. It would prove worthwhile to carry out further investigation in L2 listening field with tests which are scored on polytomous scales or apply other modelling approaches for dichotomously scored tests. Relatedly, it would be crucial to further investigate the applicability of CFA for the
context of L2 listening. We speculate that sub-skill divisibility might be established by the statistical tools that would suit dichotomously scored tests (see Aryadoust & Goh, 2013).

Conclusion

The issue of the sub-skills and the divisibility of the listening construct into separate ones will continue to capture the attention of researchers while curriculum developers and teachers will continue to be concerned with the practical challenges of helping learners improve their language skills in order to communicate better socially, pass examinations, and do well academically. There are benefits in focusing on some aspects of the listening ability which are crucial to successful comprehension be they at the decoding level or a higher processing level (Field, 2008). For example, tasks that bring the separate sub-skills together so that they can be used in conjunction with each other and the issue of whether they are separable or not ceases to be of central importance in pedagogical terms (Sawaki et al., 2013). Alternatively, it may be useful to focus separately on the development of different kinds of processes according to the kinds of listening context within an academic environment. The development of listening for academic contexts is in many ways similar to general listening development although distinctions can be made in the types of discourse that students participate in. Thus, there is rightly a great deal of emphasis on helping students construct meaning of what they hear through extended discourse such as lectures and seminar presentations.

As pointed out by Brown & Yule (1983), utterances during informal social interaction tend to contain grammar that is “loose”; they are typically short and often strung together through the use of relevant coordinators such as ‘and’ to achieve cohesion. In contrast, the utterances in lectures and talks which are of a transactional nature are longer and the grammar is “tighter”,

31
containing many features of the syntax of written language. Learners would certainly benefit from learning how to handle the two types of discourse in an academic context.

The findings from this study suggested that the sub-skills that the test aimed to assess through separate sections of the instrument were empirically divisible, should the right modelling approach be used. This lends support to the view that it is possible to distinguish sub-skills from one another, hence supporting the language assessment and teaching experts who have proposed theoretical taxonomies in an attempt to deconstruct listening into manageable parts for testing and teaching purposes.

It is worth noting that sub-skills function in an interactive and interdependent manner in listening events particularly those that are interactional in nature such as when someone is participating in or listening to a conversation. Because of this, it may be worthwhile that test developers focus on processes that are important for different listening contexts and types of discourses encountered in an academic context, as language processing may be constrained or assisted by the discourse structure of texts and the way meaning is packaged in utterances (see Bodie, 2013). The key skills that are needed for listening to a short conversation and those for listening to a lecture may not be the same even though some low-level skills such as decoding are fundamental to both. It would still be advisable to separate the sub-skills to develop tasks in assessment and pedagogy, and ensure that sub-skills are integrated in extended listening tasks such as those that focus on comprehension.

Finally, evidence favouring divisibility of sub-skills is useful as it may be used to generate assessment systems where both subscores and composite scores are provided to test takers. Subscores disaggregate the performance of test takers on the basis of the sub-skills tapped by
each set of items. Test takers and teachers will benefit from such score reports, as it highlights the weaknesses and strengths of test takers and helps the educational institutes develop the kind of educational programs that would best meet the language needs of the test takers.

Limitations and Further Research
An obvious limitation of our study is that we used data from only one version of the test. Future research could repeat the same modelling procedures or other statistical methods over a larger number of tests (specifically the tests that possess polytomous scales). Researchers would need to constantly refine the methods for gathering evidence of divisibility of sub-skills not only through a statistical quantitative approach but also through rigorous qualitative methods wherever possible.

In this study, we examined the test takers’ performance with reference to the attributes of test items and the sub-skills that they would engage. As one of the reviewers suggested, metacognition, first and second language vocabulary, (implicit) grammar, and working memory are extremely important factors in predicting listening comprehension (see, for example, Goh & Hu, 2013). Future research can adapt independent measures for these factors and examine their predictive power in listening test performance of low- and high-ability test takers. Although this approach has been used in reading studies (e.g., D. Zhang, 2012; L. Zhang et al., in press), it has scarcely been researched in listening comprehension studies and is one line of research that has great potential.
Acknowledgement

Authors would like to thank the Spaan Fellowship Program of Cambridge-Michigan Language Assessment Institute (CaMLA) for their support of the project.

References


EXAMINING THE NOTION OF LISTENING SUB-SKILL DIVISIBILITY
AND ITS IMPLICATIONS FOR SECOND LANGUAGE LISTENING


Rost, M. (2002). Teaching and researching listening. New York: Longman.


### Appendix A

Descriptive statistics of the aggregate-level test items.

<table>
<thead>
<tr>
<th>Item</th>
<th>Mean</th>
<th>Standard deviation</th>
<th>Skewness</th>
<th>Kurtosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>i1i3</td>
<td>1.12</td>
<td>0.727</td>
<td>-0.195</td>
<td>-1.094</td>
</tr>
<tr>
<td>i2i4</td>
<td>1.26</td>
<td>0.698</td>
<td>-0.419</td>
<td>-0.907</td>
</tr>
<tr>
<td>i6i8</td>
<td>1.32</td>
<td>0.602</td>
<td>-0.286</td>
<td>-0.645</td>
</tr>
<tr>
<td>i5i7</td>
<td>1.28</td>
<td>0.738</td>
<td>-0.508</td>
<td>-1.022</td>
</tr>
<tr>
<td>i9i11</td>
<td>1.53</td>
<td>0.637</td>
<td>-1.044</td>
<td>-0.019</td>
</tr>
<tr>
<td>i10i12i14</td>
<td>1.81</td>
<td>0.884</td>
<td>-0.186</td>
<td>-0.833</td>
</tr>
<tr>
<td>i13i15</td>
<td>1.30</td>
<td>0.683</td>
<td>-0.467</td>
<td>-0.823</td>
</tr>
<tr>
<td>i18i30</td>
<td>1.28</td>
<td>0.712</td>
<td>-0.474</td>
<td>-0.937</td>
</tr>
<tr>
<td>i21i35</td>
<td>1.18</td>
<td>0.746</td>
<td>-0.310</td>
<td>-1.158</td>
</tr>
<tr>
<td>i36i38</td>
<td>1.06</td>
<td>0.724</td>
<td>-0.096</td>
<td>-1.090</td>
</tr>
<tr>
<td>i37i39</td>
<td>1.28</td>
<td>0.706</td>
<td>-0.477</td>
<td>-0.910</td>
</tr>
<tr>
<td>i41i43i45</td>
<td>2.24</td>
<td>0.803</td>
<td>-0.835</td>
<td>0.065</td>
</tr>
<tr>
<td>i42i44</td>
<td>1.18</td>
<td>0.682</td>
<td>-0.257</td>
<td>-0.866</td>
</tr>
<tr>
<td>i40i46</td>
<td>1.34</td>
<td>0.679</td>
<td>-0.555</td>
<td>-0.761</td>
</tr>
<tr>
<td>i47i49</td>
<td>1.11</td>
<td>0.726</td>
<td>-0.182</td>
<td>-1.091</td>
</tr>
<tr>
<td>i48i50</td>
<td>1.32</td>
<td>0.705</td>
<td>-0.555</td>
<td>-0.857</td>
</tr>
<tr>
<td>i16i22</td>
<td>1.15</td>
<td>0.744</td>
<td>-0.260</td>
<td>-1.161</td>
</tr>
<tr>
<td>i24i26</td>
<td>1.19</td>
<td>0.674</td>
<td>-0.254</td>
<td>-0.829</td>
</tr>
<tr>
<td>i25i27</td>
<td>1.19</td>
<td>0.711</td>
<td>-0.306</td>
<td>-0.998</td>
</tr>
<tr>
<td>i29i31i33</td>
<td>1.98</td>
<td>0.890</td>
<td>-0.455</td>
<td>-0.671</td>
</tr>
<tr>
<td>i17i19i23</td>
<td>1.78</td>
<td>0.911</td>
<td>-0.226</td>
<td>-0.818</td>
</tr>
<tr>
<td>i20i28</td>
<td>1.28</td>
<td>0.701</td>
<td>-0.461</td>
<td>-0.900</td>
</tr>
<tr>
<td>i32i34</td>
<td>1.44</td>
<td>0.667</td>
<td>-0.792</td>
<td>-0.496</td>
</tr>
</tbody>
</table>

*Note: n = 916.*

*a* Each aggregate-level item is made by combining two or three test items. For example, i1i3 was made by combining Items 1 and 3.
### Appendix B

Correlation of the aggregate-level items.

|    | 1  | 2  | 3  | 4  | 5  | 6  | 7  | 8  | 9  | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 |
|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
| 1  | 1  |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| 2  | .20| .08|    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| 3  | .18| .19| .01|    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| 4  | .19| .23| .18| 1  |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| 5  | .15| .18| .17| .24|    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| 6  | .23| .22| .21| .29| .25|    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| 7  | .16| .20| .14| .25| .17| .25|    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| 8  | .18| .27| .18| .34| .28| .31| .18| 1  |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| 1  | .13| .13| .08| .05| .10| .11| .08| .17| .09| 1  |    |    |    |    |    |    |    |    |    |    |    |
| 0  | .41| .70| .80| .20| .20| .20| .20| .25| .25| .18| 1  |    |    |    |    |    |    |    |    |    |    |
| 1  | .19| .24| .23| .17| .20| .19| .21| .20| .25| .25| .18| 1  |    |    |    |    |    |    |    |    |    |
| 2  | .22| .18| .14| .29| .18| .22| .18| .26| .16| .17| .22| 1  |    |    |    |    |    |    |    |    |    |
| 4  | .16| .20| .10| .19| .17| .23| .22| .26| .21| .13| .23| .20| .21| 1  |    |    |    |    |    |    |    |
| 5  | .15| .13| .11| .19| .08| .16| .15| .20| .13| .14| .22| .21| .13| .23| 1  |    |    |    |    |    |    |
| 6  | .18| .14| .04| .19| .11| .18| .18| .18| .16| .10| .17| .19| .20| .24| .21| 1  |    |    |    |    |    |
| 8  | .11| .16| .12| .23| .19| .16| .16| .18| .17| .10| .17| .14| .17| .19| .12| .15| .16| 1  |    |    |
| 9  | .14| .15| .14| .15| .17| .22| .18| .18| .18| .16| .25| .17| .14| .18| .11| .19| .17| .16| 1  |    |
| 10 | .17| .18| .16| .21| .10| .17| .18| .17| .18| .16| .25| .17| .14| .18| .11| .19| .17| .16| .17| 1  |
| 11 | .13| .14| .15| .16| .17| .22| .18| .18| .18| .16| .25| .17| .14| .18| .11| .19| .17| .16| .17| .16| 1  |
| 12 | .18| .31| .17| .31| .26| .29| .25| .36| .29| .15| .27| .26| .27| .26| .20| .23| .22| .16| .22| .16| .22| .1 |
| 13 | .21| .27| .15| .26| .28| .20| .21| .27| .25| .22| .23| .20| .15| .27| .17| .16| .25| .18| .18| .23| .29| 1  |
| 14 | .14| .15| .06| .21| .17| .20| .14| .20| .17| .07| .18| .21| .17| .21| .15| .13| .17| .15| .13| .25| .21| 1  |

Note: *p < .05. **p < .01.