Psychological Test and Assessment Modeling, Volume 54, 2012 (3), 227-256

Developing and validating an Academic Listening Questionnaire

Vahid Aryadoust[1]*, Christine C. M. Goh*[2] *& Lee Ong Kim*[2]

Abstract

This article reports on the development and administration of the Academic Listening Self-rating Questionnaire (ALSA). The ALSA was developed on the basis of a proposed model of academic listening comprising six related components. The researchers operationalized the model, subjected items to iterative rounds of content analysis, and administered the finalized questionnaire to international ESL (English as a second language) students in Malaysian and Australian universities. Structural equation modeling and rating scale modeling of data provided content-related, substantive, and structural validity evidence for the instrument. The researchers explain the utility of the questionnaire for educational and assessment purposes.

Key words: academic listening, language testing, Rating Scale Model, structural equation modeling

[1] *Correspondence concerning this article should be addressed to:* Vahid Aryadoust, Centre for English Language Communication, 10 Architecture Drive, Singapore 117511, National University of Singapore; email: elcsva@nus.edu.sg, vahidaryadoust@gmail.com

[2] National Institute of Education, Nanyang Technological University

Self-rating, a process by which students systematically appraise their own skills and abilities, has attracted significant attention among researchers as an effective tool in language and educational training and assessment (Adams & King, 1995; Brantmeier, 2005, 2006; Cameron, 1990; Heilenmann, 1990; Jafarpour, 1991; Little, 2005; Mowl & Pain, 1995; Orsmond, Merry, & Reiling, 1997; Rivers, 2001; Rolfe, 1990; Ross, 1998; Shore, Shore, & Thornton III, 1992; Stefani, 1994; Sullivan & Hall, 1997). Researchers report it to be a rigorous method of improving language learners' awareness of their own weaknesses and strengths (Ekbatani, 2000), a useful supplement to teacher evaluations (Nunan, 1988), and a way to help teachers understand language learners' self-perceptions, which can direct their teaching.

Educational assessment institutions have begun to use self-rating as an important process in language learning. Little (2005, p. 321) reported that the policies of the Common European Framework of Reference for Languages (CEFR) and the European Language Portfolio (ELP) have shifted toward a more learner-centered learning and assessment paradigm which provides independence to language learners and generates an educational context in which learners "take full account" of their own assessment. Little advocated establishing self-rating procedures that "bring the learning process into a closer and more productive relation to tests and examinations than has traditionally been the case" (Little, 2005, p. 324), and using these procedures in high-stakes assessment.

While numerous research studies have examined the principle of self-rating, analysis of its efficacy in academic contexts (e.g., universities and colleges) has been limited. The present study focuses on listening comprehension assessment, an area where self-rating is largely unexplored. Researchers have generated a few self-rating instruments for listening tests (see, for example, Sawaki & Nissan, 2009; Ford & Wolvin, 1992, 1993; Ford, Wolvin, & Sungeun, 2000), but their efficacy, and the utility of self-rating in listening comprehension generally, has not been examined.

We present an English academic listening self-rating questionnaire. This questionnaire is designed to help students in supplementary and academic English language courses improve their academic listening performance and become more autonomous and aware of their proficiency level. The questionnaire is based on an exploratory model of integrated academic listening macroskills (see Figure 1). Since multiple studies have reported weak or even negative correlations between student scores on the listening section of the International English Language Testing System (IELTS^{TM}) (a popular English proficiency test for university admissions) and their subsequent grade point averages (see Aryadoust, 2011a), the model we propose is also intended to help accurately represent and assess academic listening. The study provides content-referenced, substantive, and structural validity evidence for the questionnaire.

Self-rating

A number of empirical studies have found self-rating to be an effective pedagogical tool. Granville and Dison (2005) suggest that students can benefit from a teaching approach that incorporates self-reporting, and Butler and Lee (2006) find that self-rating improves

students' learning and self-confidence in English classes, although these effects can vary depending on the instructional context. Dragemark (2006) reported that, after time and practice, adult English language learners who participated in a self-rating program became better aware of their language skills, and that the program motivated students to reflect on their language learning by transferring some evaluation responsibilities from teachers to them. Dragemark (2006) also argues that self-rating can be useful in "virtual" educational contexts: since teachers are not physically present to evaluate and provide feedback on students' learning, creating independent methods of evaluation becomes highly important, and student-led assessment is among the most attractive of these methods.

The utility of self-rating depends largely on students' degree of motivation, experience, and comfort in evaluating their own performance. Oscarson (1984, 1999) showed that self-rating of language skills is often reliable, and has strong correlations with objective measurement methods, if students are instructed in its objectives and advantages. In Dragemark's (2006) study, students evaluated their performance more accurately toward the end of the semester, as they became familiar with the process. Bachman and Palmer (1989) conducted a confirmatory factor analysis study of self-reports of language proficiency, and found that second language (L2) speakers can reliably reflect on their communicative skills, given conditions similar to those articulated by Oscarson and Dragemark. Finally, Patri (2002) argues that less independent learners tend to assess their language proficiency less accurately: For example, many high-ability learners of English in Matsuno's (2009) study underestimated their writing proficiency, most likely due to "the tendency of many Japanese to display a degree of modesty" (p. 94), a finding that suggests that cultural background may have significant effects on self-rating.

Interestingly, however, the validity of self-rating does not appear to vary with the quality of instruction given to students. In a study of Dutch learners, van Dieten (2000) found that instructional procedures did not influence the accuracy of participants' self-reflection. Van Dieten describes this finding as "alarming" since proper instruction should be expected to help students appraise their own language ability.

Academic listening comprehension

To function usefully, self-rating requires a reliable and valid instrument. Here we discuss the theoretical background and underlying model of academic listening underlying the self-rating questionnaire we present. This model incorporates three complementary inputs: a general model of listening comprehension, the structure of academic lectures, and students' English language ability levels (Flowerdew & Miller, 1992; Powers, 1986; King, 1994).

The listening that enables most learning in university lectures, tutorials, and seminars is academic listening, a form of listening substantially different from ordinary conversational listening (Benson, 1989). Academic lectures require the listener to distinguish relevant information and draw on background knowledge of the topic to a much greater

extent than ordinary conversation, and involve comparatively little turn-taking and few indirect speech acts (Flowerdew, 1994).

Richards (1983), one of the first scholars to formalize the distinction between general and academic listening, proposed a list of academic listening micro-skills, some of which include: the ability to identify a lecture's purpose and scope; to identify relationships among units within discourse (such as major and supporting ideas, generalizations, and examples); and to infer relationships such as cause, effect, and conclusion (Richards, 1983). Drawing on Richards's taxonomy, Weir (1990) developed a list of micro-structures in academic language comprehension, and proposed a listening assessment method that assessed both extensive and intensive listening skills. Weir further proposed that these skills be evaluated by different item formats, such as multiple choice and open-ended questions.

Powers's (1986) survey of 144 university lecturers identified nine academic listening micro-skills thought to be especially important to learning, some drawn from general listening theory (such as understanding vocabulary and identifying major points and themes), some addressing lecture structure (such as inferring relationships between information), and some relating to specific student skills (such as note taking and retrieving information from notes). More recently, Jordan (1997, p. 180) described a similar taxonomy of academic listening micro-skills, some of which follow:

a) ability to identify purpose and scope of lecture

b) ability to identify topic of lecture and follow topic development

c) ability to identify relationships among units within discourse

Although academic listening is sometimes treated as a single global latent trait, a number of research studies suggest that it is actually multidivisible, composed of a number of separate but interrelated subskills (Buck, 2001; Goh & Aryadoust, 2010; Wagner, 2004; see also Imhof & Janusik, 2006). The taxonomies mentioned above appear to support the latter view.

Structure and style of academic discourse

Academic lectures are based on underlying discourse structures, which may vary across disciplines, and knowledge of which facilitates student learning and understanding (Dudley-Evan, 1994). Researchers have attempted to classify academic lectures according to their discourse structures: Dudley-Evan (1994), Olsen and Huckin (1990), and Strodt-Lopenz (1991) reported different macro-structure frameworks for plant biology, engineering, and social science lectures, respectively. Despite these differences, Young (1994) introduced a general framework for analyzing lecture structure, arguing that "there is consistency of codal choice across disciplines in terms of macro-structure, and between native and non-native speakers' discourse in this registerial variety—university spoken discourse" (p. 174). Similarly, Hansen (1994) presented a model of academic

discourse that divides lectures into topics based on sentential topic identification, which helps identify the markers that indicate a shift in topic.

Lectures also involve specific "lexico-grammatical" cues that differentiate them from other auditory activities (Flowerdew, 1994), and which can either facilitate or hinder listening comprehension. For example, Flowerdew and Miller (1992) argue that rapid delivery, new vocabulary and definitions, and distractions in lengthy lectures are likely to impede the comprehension of lectures in L2 students. Similarly, because understanding lectures requires an extensive vocabulary knowledge base (Kelly, 1991; Olsen & Huckin, 1990), low-proficiency L2 learners may rely too heavily on linguistic decoding, compromising comprehension of the aural message (Flowerdew & Miller, 1992; Goh, 2005; Rost, 1994).

Researchers have also investigated the "subsidiary" or metapragmatic discourse present in lectures (Coulthard & Montgomery, 1981). Metapragmatic cues facilitate comprehension and learning by pointing directly to shifts in, examples of, and support for main ideas. A potential difficulty in formal lectures is that they lack such features (Coulthard & Montgomery, 1981). Researchers have found that lecturers can facilitate comprehension by including these elements: by using stories and personal anecdotes (Strodt-Lopenz, 1991); by creating an atmosphere of cooperation, friendship, and sense of belonging to a group (Rounds, 1987); and by providing "discourse markers," contextualizing clues as to how discourse is to be interpreted (Eslami & Eslami-Raseck, 2007). Conversely, Flowerdew (1992, 1994) claims that the personal attitudes conveyed in lectures can adversely affect their comprehensibility.

Relatedly, the formality of academic discourse can affect its comprehension. Dudley-Evan (1994) divided lectures into four classes according to their content and discourse properties: (a) "reading style," delivered in a formal language similar to textbook language; (b) "conversational style," which presents information in the typical manner of spoken language; (c) "rhetorical style," in which the lecturer performs like an actor, changing tone, intonations, and body language; and (d) Fredrick's (1986) "participatory lecture" style, a discussion between the lecturer and students. Dudley-Evan found that moving toward less formal, more conversational lecturing styles facilitates learning and comprehension.

Student's language proficiency

Student performance in academic listening is most obviously affected by their own ability level. Since students' internal psychological processes cannot be gauged directly, much research on academic listening comprehension examines students' written notes as a concrete record of those processes. Note-taking itself is an important facet of academic comprehension: It "facilitates encoding or the impression of information in the memory" (Olmos & Lusung-Oyzon, 2008, p. 71), and it engages and thereby improves students' ability to memorize discourse (Williams & Eggert, 2002; as cited in Olmos & Lusung-Oyzon, 2008, p. 71). Partly for these reasons, effective note taking is strongly correlated with academic performance (Chaudron, Loschky, & Cook, 1994; Kiewra, 1985; King,

1994), and Dunkel and Davis (1994) report that the quantity (measured by word count and idea units) and quality of student notes is predicted by the students' proficiency, by their degree of fluency in English, and by the extent to which the lecture makes explicit use of "rhetorical cuing" to help situate them.

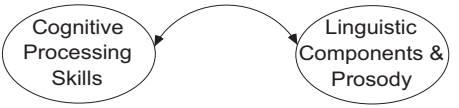
According to Olsen and Huckin (1990), students' note-taking strategies appear to reveal two major systems of recognizing lecture intent: "information-driven," employed by students who intend simply to identify and learn facts; and "point-driven," which is more hierarchical in that students attempt to distinguish major points from supporting ideas. However, Tauroza and Allison (1994) criticize Olsen and Huckin's (1990) study as being limited by small sample size and ambiguity in key terms, and propose five "idea units" as a constituent structure of lectures: topic, introduction, problem, solution, and evaluation (see also Chaudron, Loschky, and Cook, 1994, for another taxonomy).

Using their model, Tauroza and Allison (1994) investigated 50 students' recalls of a lecture. They found that most students (70%) successfully differentiated the topic from the introduction, and that 60%, 76%, and 74% could identify three key details of the solution, but that a much smaller percentage could accurately identify the problem and the solution itself (26% and 18%, respectively), and very few (6%) the evaluation. The evaluation was the "the only section where the subjects reported the opposite of what the lecturer said" (p. 43). Students' success in comprehending the evaluation and in distinguishing other sections were not correlated, although there was some evidence that higher-ability students (i.e., proficient in English) were more successful at evaluation, perhaps indicating the difficulty and complexity of evaluation skills (see Rost, 1994, for a study of note-taking and how it can mirror comprehension). According to Tauroza and Allison, the students who wholly misunderstood the evaluation section may have performed a "local interpretation," first described in Brown and Yule (1983): novice comprehenders attempt to understand stimuli based on the current context, while proficient listeners "refer to a more global context in order to reach a coherent interpretation consistent with textual evidence" (Tauroza & Allison, 1994, p. 45). Tauroza and Allison conclude that overreliance on the local interpretation of a message can cause students to misunderstand an ongoing lecture.

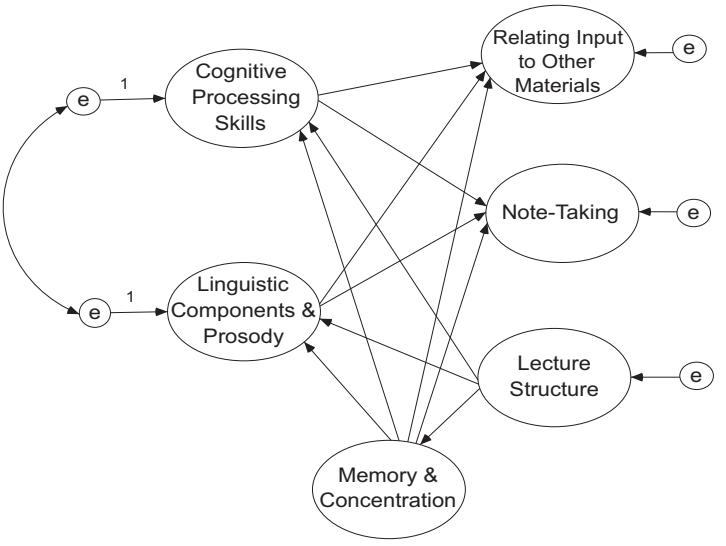
Modeling academic listening

The foregoing brief review suggests that success in L2 academic listening is determined by various dimensions of student language ability, as well as by the content, structure, and style of the academic discourse. To these considerations might be added a number of environmental factors, such as the presence of distractions. Researchers have not yet attempted a model designed to capture the complexity of academic listening, comprising a number of separate yet interrelated and interdependent components. We have made every effort to posit a model which is easy to explain, testable (see Bodie & Fitch-Hauser, 2010; Janusik, 2009), elegant, and accurately crafted (see Bodie, 2009, for details). Figure 1 presents this model.

The model consists of two major sections: the first (at top in Figure 1) is a general listening model comprising cognitive processing skills (CPSs) and linguistic components and prosody (LCP), and the second (at bottom in figure 1) represents the multidivisible components of an academic listening model. The proposed path model at the bottom incorporates various subskills that likely affect or are affected by CPSs and LCP. The "e" circles indicate error terms associated with endogenous (or dependent latent) variables, the bidirectional arrows indicate correlations between latent variables, and the unidirectional arrows indicate causal relationships.



Relationship between the Cognitive Processing Skills and Linguistic Components & Prosody



Relationship between academic listening skills and external factors contributing to success in academic listening

Note: The model posits two multidivisible components of listening ability at the top of the figure: Cognitive Processing Skills (CPSs) and Linguistic Components and Prosody (LCP). The proposed path model at the bottom incorporates various subskills that hypothetically affect and are affected by CPSs and LCP. The "e" circles indicate error terms associated with endogenous (or dependent latent) variables. The bidirectional arrows indicate correlations between latent variables and unidirectional arrows indicate cause-effect relationships. This model is taken as the baseline model. Its fit to the data might suggest that modification should be applied to the model.

Figure 1:
Illustration of the model of academic listening

We used this model to construct the Academic Listening Self-rating Questionnaire (ALSA), to allow L2 learners to evaluate their own academic listening performance. We seek to provide preliminary evidence of the validity of our model of academic listening and of the ALSA through structural equation modeling (SEM), and by evaluating various types of content-related, structural, and substantive validity evidence.

Methodology

Sample 1

We first administered the ALSA in an Australian university to thirty (30) international university students aged between 18 and 51 (M = 27.33; SD = 7.5). Participants studied a wide range of disciplines, were nonnative speakers of English, and signed consent forms prior to participating in the study.

Sample 2

In the second phase of the study, we administered the survey to one hundred and nineteen (119) international university students aged between 18 and 39 (M = 27.27; SD = 3.75) from the accessible population of six universities in Malaysia. The participants' first languages were Persian, Turkish, Arabic, Punjabi, and Nepali. Eight participants (6.7%) were pursuing undergraduate degrees, 86 (71.1%) master's degrees, and 27 (22.4%) PhD degrees. Informed consent was obtained from all participants before the study was conducted.

Since participants' familiarity with the concepts of the questionnaire was important, we endeavored to select students with the most similar areas of study possible. Also, most participants reported having taken preparation courses for language proficiency tests (IELTS or TOEFL), which had likely made them aware of their linguistic shortcomings. Table 1 provides an overview of participants' distribution in the target tertiary institutes of education.

For their English Language proficiency Test for university admissions, 196 (87.6%) participants reported having taken the IELTS test, nine (7.4%) TOEFL in its Internet-Based Test (iBT) administration, four (3.3%) the Paper-Based (PB) TOEFL, and one (0.8%) the Computer-Based (CB) TOEFL. The descriptive statistics of their IELTS scores are presented in Table 2. TOEFL scores are not reported because of the high incidence of missing data in the TOEFL dataset.

Table 1:
Distribution of the Consent Sample in Malaysian Universities

	Frequency	Percent
University A	34	28.57
University B	49	41.18
University C	22	18.49
University D	4	3.36
University E	4	3.36
University F	4	3.36
Missing	2	1.68
Total	119	100

Table 2:
Descriptive Statistics of IELTS test Participants

	Listening	Reading	Writing	Speaking	Total
Mean	6.21	6.10	6.07	6.59	6.31
Standard deviation	0.87	0.82	0.70	0.76	0.68
Skewness	0.37	-0.14	0.45	-0.02	0.22
Kurtosis	-0.32	0.45	-0.34	-0.43	-0.73
Minimum	4.50	3.50	5.00	5.00	5.00
Maximum	8.50	8.50	8.00	8.50	7.50

$n = 106$.

Material

The model of academic listening we propose posits two major factors in general listening: cognitive processing skills (CPSs) and linguistic components and prosody (LCP). Other variables that likely associate with academic listening performance include note-taking (NT), lecture structure (LS), relating input to other materials (RIOM), and memory and concentration (MC).

We developed a pool of 62 items designed to tap all hypothesized subskills. In iterative steps, we scrutinized item contents, clarified or omitted double-barreled items, deleted problem and redundant items, and attempted to solve items' potential linguistic problems in order to generate an item pool that usefully represented the concepts (Brace, 2008; Gillham, 2006; Willis, 2005). We kept 47 items in the final questionnaire, which we arranged on an alternating gray and white background.

Although the common assumption is that negatively worded items can help researchers detect response sets, we avoided using them because various exploratory and confirmatory factor analyses and RSM have demonstrated that data based on them can be biased and less reliable (Wolfe & Smith, 2007a). We chose a four-response Likert scale for the items: *poor (1), satisfactory (2), good (3),* and *excellent (4)*, to avoid a "neutral point," a likely place for noncooperative respondents to hide (see Wolfe and Smith, 2007a, b, for a discussion).

Two graduate students were contracted to administer the questionnaire at the target academic institutions in Malaysia and Australia. Consenting students received the questionnaire with a cover page explaining the aims of the research and assuring them that their participation in the study was voluntary, that they could refuse or discontinue participation or skip any item which they would not like to answer, and that no personally identifying information would be disclosed. Participants also received the contact information of one of the researchers in case they had questions.

Data analysis

Descriptive statistics. Using *SPSS* computer program, Version 16 (SPSS Inc., 2007), we computed descriptive statistics of all questionnaire items in the Australian and Malaysian data separately, including mean, standard deviation, skewness, and kurtosis values. Skewness and kurtosis values are two measures of univariate normality, and are expected to fall between -2 and +2 (Bachman, 2004) in a normal distribution.

Technical features of items. As a preliminary analysis of the item contents, we investigated the congruence of responses with our expectations (Wolfe & Smith, 2007a, b) by fitting the Rating Scale model (RSM) (Andrich, 1978) to the data. The RSM describes their psychometric properties and provides difficulty estimations for all items, including those with few observations in each response category (Linacre, 2000). The RSM is expressed as Equation 1:

$$P_{xi} = \frac{\sum_{j=o}^{x}\left(\theta - \left(\lambda_i - \delta_i\right)\right)}{\sum_{x=o}^{mi} e\left[\sum_{j=o}^{x}\left(\theta - \left(\lambda_i - \delta_i\right)\right)\right]} \quad (1)$$

where δ_i is the threshold between categories, λ_i is the item location parameter, and θ is the person trait level (Bond & Fox, 2007). We used the RSM in lieu of exploratory factor analysis (EFA), because EFA results are often obscured by "ordinality" of variables and high correlation among factors (Schumacker & Linacre, 1996, p. 470; Smith, 1996),

because Rasch modeling (and RSM as an expansion of it) is not biased by missing data, unlike the factor solution (Bond, 1994, Linacre, 1998; Wright, 1994a, b).

Since the academic listening model assumes six divisible components, we performed six individual RSM analyses, examining three major statistical indices in each analysis: item difficulty and person ability measures, the fit of the data to the model, and point-measure correlations.

We used *WINSTEPS* computer program, Version 3.70 (Linacre, 2010a) to compute infit and outfit MNSQ statistics. Infit is an information-weighted index that is sensitive to the erratic patterns of inliers, which usually indicate important confounding trends. Outfit is outlier-sensitive, and helps to find, for example, lucky guesses and mistakes due to fatigue or carelessness in tests (Linacre, 2002). In polytomous data, MNSQ values greater than 1.4 are said to underfit, and values smaller than 0.6 to overfit (Bond & Fox, 2007; see also Aryadoust, Goh, & Lee, 2011). MNSQ indices can be standardized by applying Wilson-Hilferty transformation, which represents their statistical significance or *p*-value. These standardized fit indices, expressed as infit and outfit *ZSTD*, are ideally expected to be zero, though the range between -2 and +2 is commonly accepted in small datasets (Linacre, 2010b).

Point-measure correlations (PMCs), another measure of the technical features of items, express the correlation between participants' responses on individual items and their overall measured latent trait levels (Wolfe & Smith, 2007b). PMCs are more accurately interpreted if they are compared with RSM expected values (Linacre, 2008). Observed PMC values that are much higher than expected indicate overly predictable response patterns, and observed values much lower than expected indicate unmodeled variation or noise in the data.

The RSM also produces reliability and separation indices for both persons and items. High person reliability indices suggest that the instrument discriminates well among respondents (Bond & Fox, 2007), though low reliability does not necessarily indicate a problem with the data as it may indicate higher homogeneity amongst respondents. Separation, the ratio of "true" variance to error variance, ranges from zero to infinity and indicates the number of distinct trait levels perceptible in the data (Linacre, 2010b).

Content-related validity evidence. To collect content-related validity evidence, we sought expert judgments and student feedback on the first draft of the questionnaire. Two university lecturers and two PhD students in applied linguistics provided feedback on the questionnaire items' clarity and possible cultural and linguistic bias. Two master's degree students, one in applied linguistics and one in education, provided further written and oral comments.

We carried out a pilot study with our first sample group of respondents, thirty English learners studying in Australia. As a result of this study and our review, we rewrote a few items, as well as the Likert scale, to be easier to understand.

Substantive validity evidence. This aspect of validity concerns the relationship between the questionnaire content and the observed responses of participants. Following Wolfe and Smith (2007b), we sought substantive evidence by examining the cognitive represen-

tation of the instrument (presented in Figure 1), the quality of the rating scale, person fit, and item endorsability. We used Linacre's (2004) exposition of RSM results that support an instrument's substantive validity argument, as follows: (a) each rating category must be selected by at least 10 respondents; (b) the rating scale categories (or category probability curves) should be clearly separated into hill-shaped structures; (c) response category difficulty measures should increase monotonically, and difficulty measure thresholds should increase monotonically by at least 1.1 log-odd units (logits), given the four-point scale used in the questionnaire; (d) the MNSQ statistics should not depart too markedly from unity; and (e) model expectations should resemble observed performance.

Structural validity evidence. We sought structural validity evidence by examining the internal relationships between individual subskills and their corresponding questionnaire items through confirmatory factor analysis (CFA), and the external relationships between subskills through structural equation modeling (SEM). CFA and SEM have proven useful in language assessment and modeling (Aryadoust, 2012; Bodie, Worthington, & Fitch-Hauser, 2011; Kunnan, 1994; Bae & Bachman, 1998). In CFA, the researcher proceeds according to a theory-derived model that both hypothesizes latent trait variables and specifies manifest (observable) variables (Hair, Black, Babin, & Anderson, 2010). CFA is one of the applications of SEM, a multivariate statistical technique that tests both the causal and correlational relationships among both latent and manifest variables. SEM models are graphically illustrated as path models or flowcharts displaying these relationships.

We performed two-stage analysis, testing each measurement model individually through CFA prior to testing the full structural model through SEM. Each measurement model comprised a latent trait (a subskill) with strictly causal relationships to item responses, the manifest variables in this study (Jöreskog & Sörbom, 2001). The SEM model included all causal and correlational relationships among the measurement models. We used *AMOS* computer program, Version 16, to perform the modeling. We used the Maximum Likelihood (ML) method of parameter estimation, and employed multiple fit criteria to evaluate the fit of the postulated measurement models and the full SEM model:

a) Chi-square test (χ^2): An index representing the difference between the observed and implied covariance or correlation matrices.

b) Normed χ^2 (χ^2/df): The ratio of χ^2 to the degrees of freedom (*df*). This ratio is small in well-fitting models (preferably below 3).

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

d) Two incremental indices: Non-Normed Fit Indices (NNFI) and Comparative Fit Indices (CFI). Both types of indices compare the postulated model to a baseline model that assumes that measures are not correlated. NNFI indices penalize increases in the number of model parameters, and can be greater than unity, but are usually set at unity. We chose indices of 0.90 or above as indicators of satisfactory fit.

e) Two model-parsimony fit indices: Consistent Akaike Information Criterion (CAIC) and Akaike Information Criterion (AIC) indices. CAIC penalizes sample size and parameter increases, while AIC merely adjusts for parameters (Bozdogan, 1987).

Results

Preliminary analysis of data: Descriptive statistics

Table 3 presents descriptive statistics for all questionnaire items in the Australian and Malaysian data separately. Except three items (1, 9, and 17) in the Australian data, both datasets satisfy skewness and kurtosis criteria, denoting approximation to statistical normality.

Table 3:
Descriptive Statistics of Items in Malaysian and Australian Samples

	Malaysian data				Australian data			
Items	**Mean**	**SD**	**Skewness**	**Kurtosis**	**Mean**	**SD**	**Skewness**	**Kurtosis**
1	3.46	0.606	-0.883	0.997	3.43	0.727	-1.477	2.910
2	3.03	0.740	-0.303	-0.416	2.73	0.739	-0.067	-0.178
3	2.96	0.784	-0.152	-0.848	2.58	0.732	0.265	-0.249
4	3.46	0.592	-0.589	-0.578	3.30	0.836	-1.014	0.393
5	3.16	0.734	-0.270	-1.098	3.26	0.739	-1.028	1.635
6	3.28	0.651	-0.371	-0.708	3.40	0.674	-0.693	-0.517
7	3.21	0.685	-0.304	-0.854	3.13	0.681	-0.170	-0.715
8	3.24	0.710	-0.541	-0.328	3.17	0.710	-0.263	-0.894
9	3.33	0.687	-0.695	0.000	3.43	0.727	-1.477	2.910
10	2.90	0.768	0.060	-1.037	3.10	0.803	-0.188	-1.406
11	3.00	0.701	-0.159	-0.496	2.86	0.730	0.214	-1.019
12	3.31	0.658	-0.439	-0.721	3.00	0.643	0.000	-0.364
13	3.48	0.593	-0.682	-0.481	3.16	0.647	-0.166	-0.502
14	3.40	0.665	-0.851	0.380	3.26	0.739	-1.028	1.635
15	3.13	0.773	-0.453	-0.549	3.10	0.758	-0.172	-1.187
16	3.07	0.811	-0.427	-0.622	2.86	0.819	0.259	-1.457
17	3.30	0.681	-0.631	-0.005	3.33	0.660	-1.251	3.827
18	3.15	0.744	-0.390	-0.694	3.06	0.907	-0.731	-.124
19	2.69	0.739	0.057	-0.433	2.46	0.681	0.478	0.072
20	2.83	0.813	0.030	-0.932	2.66	0.922	0.461	-1.214
21	3.21	0.709	-0.620	0.224	3.16	0.592	-0.040	-0.082
22	2.50	0.970	0.117	-0.961	2.57	0.878	-0.410	-0.410

23	3.06	0.806	-0.611	-0.034	2.83	0.592	0.040	-0.082
24	3.00	0.744	-0.248	-0.491	2.89	0.685	0.138	-0.721
25	3.19	0.692	-0.588	0.403	2.93	0.583	-0.003	0.229
26	3.27	0.700	-0.444	-0.882	3.03	0.614	-0.016	-0.092
27	3.28	0.721	-0.617	-0.347	3.13	0.681	-0.170	-0.715
28	3.27	0.670	-0.383	-0.779	3.10	0.758	-0.680	0.655
29	3.02	0.782	-0.258	-0.767	2.73	0.827	-0.231	-0.300
30	3.03	0.815	-0.436	-0.482	2.80	0.714	0.316	-0.911
31	3.07	0.828	-0.498	-0.504	2.76	0.773	-0.037	-0.403
32	2.88	0.886	-0.353	-0.655	2.46	0.899	-0.198	-0.668
33	3.03	0.795	-0.363	-0.570	2.80	0.924	-0.415	-0.501
34	3.00	0.831	-0.546	-0.216	2.56	0.935	-0.071	-0.753
35	3.15	0.683	-0.209	-0.842	3.00	0.787	-0.907	1.287
36	3.20	0.682	-0.285	-0.839	2.76	0.568	-0.013	-0.168
37	3.29	0.690	-0.625	-0.095	3.03	0.718	-0.647	1.085
38	3.45	0.696	-1.061	0.441	3.26	0.691	-0.409	-0.770
39	3.11	0.757	-0.434	-0.434	2.83	0.791	-0.132	-0.444
40	3.00	0.721	-0.273	-0.277	2.56	0.897	0.093	-0.674
41	3.32	0.687	-0.683	-0.007	3.06	0.827	-0.520	-0.300
42	3.24	0.710	-0.530	-0.329	2.80	0.805	-0.034	-0.606
43	3.25	0.689	-0.372	-0.858	2.93	0.739	-0.440	0.388
44	2.91	0.762	-0.437	0.051	2.30	0.876	0.007	-0.714
45	3.33	0.781	-0.987	0.373	3.13	0.730	-0.783	1.248
46	3.07	0.746	-0.246	-0.791	2.90	0.758	0.172	-1.187
47	3.37	0.674	-0.786	0.225	3.16	0.647	-0.166	-0.502

Note. Australian sample = 30. Malaysian sample = 119.

Rating Scale Model (RSM)

We first fit the RSM to the Australian sample. The results from this analysis helped identify problem items for revision before administration to the larger Malaysian sample. Table 4 presents item endorsability measures, item fit statistics, and expected and observed PMC indices for the Australian sample.

For example, Item 1 was highly endorsed by respondents (measure = -1.42): as expected, most respondents perceived their ability to understand "*isolated words and short phrases in spoken English, such as numbers and commonplace names*[3]" to be high. Item 40 was

[3] This item was revised after the first administration of the questionnaire (to the Australian students). The revised questionnaire is presented in Appendix A.

Table 4: Item Statistical Features of the Australian Data

Item	Measure	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	PMC	PMC Expected
*1	-1.42	1.34	1.22	1.84	1.7	0.65	0.71
4	-0.74	1.31	1.05	1.18	0.6	0.76	0.74
5	-0.58	0.8	-0.62	0.7	-0.91	0.81	0.74
6	-1.25	1.03	0.2	1.13	0.45	0.71	0.72
11	1.17	1.04	0.25	1.08	0.39	0.71	0.76
13	-0.1	0.79	-0.69	0.81	-0.58	0.77	0.76
14	-0.58	1.03	0.21	1.02	0.17	0.75	0.74
17	-0.91	0.64	-1.34	0.58	-1.26	0.81	0.73
26	0.49	0.66	-1.38	0.68	-1.27	0.78	0.76
37	0.49	0.98	0.00	0.99	0.05	0.73	0.76
*38	-0.58	1.75	2.18	2.51	3.35	0.53	0.74
*40	2.33	1.39	1.48	1.39	1.36	0.74	0.77
41	0.35	0.83	-0.57	0.78	-0.8	0.84	0.76
42	1.43	0.74	-1.21	0.73	-1.19	0.83	0.76
47	-0.1	0.51	-2.01	0.56	-1.67	0.84	0.76
7	-0.65	0.76	-0.94	0.83	-0.58	0.68	0.61
8	-0.79	1.08	0.4	1.17	0.69	0.56	0.61
9	-1.71	1.04	0.22	0.88	-0.26	0.68	0.56
16	0.18	0.81	-0.73	0.82	-0.71	0.79	0.63
19	1.3	0.61	-1.8	0.64	-1.59	0.72	0.65
20	0.75	1.38	1.44	1.4	1.53	0.66	0.64
21	-0.75	0.68	-1.29	0.65	-1.4	0.65	0.61
*22	1.1	2.54	4.39	2.62	4.58	0.1	0.65
25	-0.02	0.56	-2.00	0.6	-1.78	0.67	0.63
34	1.02	1.01	0.11	1.01	0.14	0.78	0.64
35	-0.22	0.74	-1.03	0.68	-1.35	0.79	0.62
36	0.47	0.6	-1.8	0.6	-1.82	0.63	0.64
43	-0.02	1.06	0.31	1.15	0.65	0.57	0.63
45	-0.65	1.15	0.62	1.18	0.74	0.55	0.61
*15	-1.09	1.4	1.26	1.62	1.6	0.74	0.82
29	0.72	0.73	-1.26	0.7	-1.26	0.86	0.83
33	0.42	1.00	0.07	1.01	0.12	0.86	0.83
46	-0.05	0.85	-0.55	0.83	-0.56	0.84	0.82
18	-0.78	0.91	-0.27	0.82	-0.59	0.83	0.70
28	-0.9	1.04	0.23	1.05	0.27	0.69	0.70
30	0.15	0.7	-1.22	0.69	-1.29	0.78	0.73
31	0.26	0.48	-2.46	0.5	-2.34	0.86	0.73
*32	1.23	1.58	2.06	1.64	2.21	0.62	0.76
39	0.04	1.32	1.22	1.3	1.15	0.61	0.73
12	-0.23	1.22	0.86	1.27	0.91	0.65	0.73
23	0.62	0.93	-0.19	0.97	0.01	0.69	0.72
24	0.54	0.91	-0.24	0.92	-0.16	0.78	0.71
27	-0.93	0.92	-0.22	0.86	-0.38	0.79	0.73
2	-0.75	0.92	-0.22	0.94	-0.11	0.74	0.73
3	-0.22	0.7	-1.19	0.74	-0.9	0.81	0.75
*44	0.96	1.32	1.17	1.23	0.85	0.74	0.79

Note. This table reports the results of the application of the Rasch Rating Scale model (RSM) to the Australian data (n = 30). The RSM was applied to individual subscales, which are separated from others by a horizontal line in the table. Measure is the endorsability of the item, which is analogous to item difficulty in a test: highly endorsed items are analogous to easy items and lowly endorsed items to difficult items. Problem items are indicated by a "*" sign. MNSQ = Mean square; PMC = Point measure correlation; ZSTD = standardized Z scores.

the least endorsable: most respondents reported having difficulty modifying their "*understanding of the lecture if it is incorrect.*" Fit estimates of a number of items fall outside the range between 0.6 and 1.4, indicating unpredictability (noise) in the data, as well as instances of overfitting. Given the small sample size, we decided that these erratic fit statistics were indicative of potential problems rather than decisive[4], though misfitting items were subjected to scrutiny and reworded.

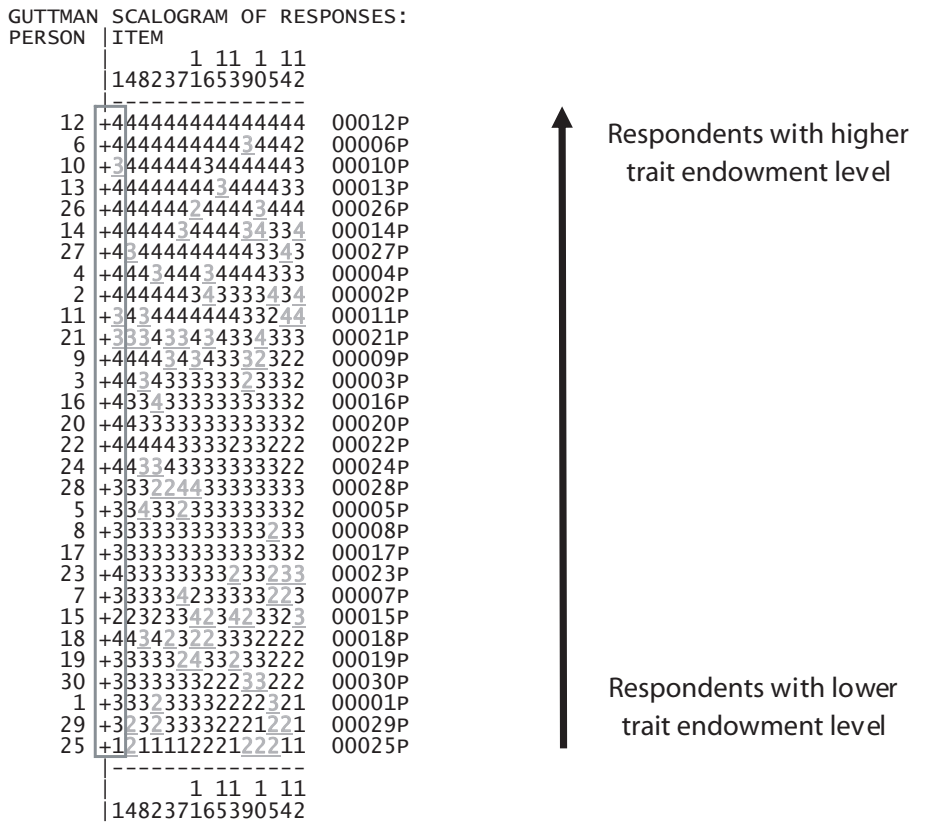


Figure 2:
Scalogram orders items from low to high measures vertically, and persons from high to low trait endowment levels horizontally. Unexpected responses are displayed in grey

[4] We investigated the residuals of persons for these given items to see who amongst the respondents has caused the large fit statistics. Few respondents were identified. We argue that sometimes it may be believable that some persons are genuinely better or poorer in some aspects of the test construct and cause the response patterns to be statistically "unexpected" but are possible human properties. It does get back to human judgment and if this is the case, the items are actually not having any problems at all.

We investigated the order of persons and items on the Scalogram, which orders items from low to high endorsability measures vertically, and persons from high to low trait endowment levels horizontally. The top left corner matches high trait level respondents with the most endorsable items, so a high number of "Excellent" (4) responses is expected, shifting to a high number of "Poor" (1) or "Satisfactory" (2) responses in the bottom right corner, which matches low trait level respondents with the least endorsable items. Outfit statistics are sensitive to perturbations and erratic responses toward the edges of the Scalogram, and infit statistics to perturbations in the middle zone. Most of our expectations were satisfied, but some unexpected responses were identified and are underlined and displayed in grey.

Rating Scale Model

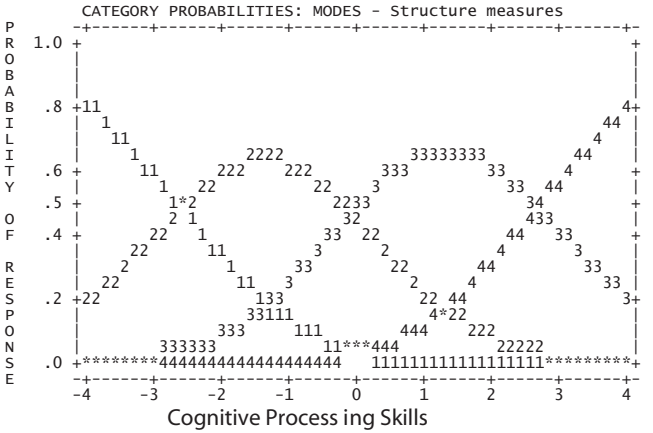
We analyzed each subscale using RSM modeling. Table 5 reports Rasch reliability and separation indices for both items and persons. (For example, the LCP subscale's item reliability and separation indices in the Malaysia dataset are .94 and 3.39, respectively, and its person reliability and separation indices are .86 and 2.49, indicating approximately 3 identifiable item endorsability strata and 2.5 person trait level strata.) The Rasch reliability of both datasets is comparable, although they differ in size.

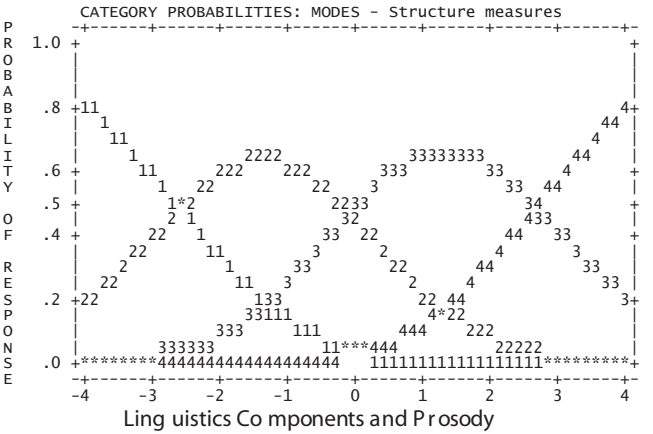
Table 5:
Item and Person Reliability and Separation statistics in the Rating Scale Model

Country	Subskills	Item reliability	Item separation	Person reliability	Person separation
Malaysia	Cognitive processing skills (CPSs)	.88	2.68	.86	2.49
	Linguistics component and prosody (LCP)	.94	3.91	.86	2.47
	Note-taking (NT)	.52	1.32	.69	1.48
	Lecture structure (LS)	.82	2.13	.76	1.78
	Relating input to other materials (RIOM)	.86	2.47	.64	1.33
	Memory and concentration (MC)	.51	1.01	.60	1.23
Australia	Cognitive processing skills (CPSs)	.83	2.22	.92	3.38
	Linguistics component and prosody (LCP)	.83	2.20	.84	2.40
	Note-taking (NT)	.61	1.26	.82	2.20
	Lecture structure (LS)	.74	1.68	.78	1.86
	Relating input to other materials (RIOM)	.54	1.08	.59	1.20
	Memory and concentration (MC)	.71	1.57	.40	0.82

Analysis of rating scale categories

We investigated the quality of rating scale categories in each component. Figure 3 shows the CPSs and LCP probability categories. The horizontal axis in each figure expresses the endorsability of rating scales in logits, and the vertical axis the probability of choosing each rating scale. For example, we would expect a person with trait level of -1.20 logits (horizontal axis) to choose category 2 on a CPSs question (probability = .65).





Note: For space reasons, we do not present other rating scales here. The horizontal axis is the endorsability of rating scales. The rating scale categories display hill-shape structures. The difficulty measure of each response category and their thresholds seems to increase monotonically.

Figure 3:
Illustration of the rating scale categories in the CPSs and LCP

The difficulty measures of all response categories and their thresholds increase monotonically in all subskills. For example, the difficulty measures of categories 1 through 4 in CPSs were -0.37, 0.28, 1.43, and 3.26, respectively, with thresholds (where adjacent category probabilities intersect) at -2.56, -0.67, and +2.67. In both CPSs and LCP, each rating category was selected at least by 10 respondents, although category 1 in NT, MC, and RIOM was selected by only 6, 7, and 4 people, respectively. If these results were reconfirmed in a larger sample, this might justify collapsing categories 1 and 2 for these subskills, but because they had acceptable fit indices, we did not collapse them in the present study.

Analysis of the Malaysian data: CFA and SEM

After adjusting problem items found in the pilot study, we administered the revised questionnaire to the Malaysian students. RSM analysis on this sample demonstrated that revised items functioned satisfactorily (see Appendix B for complete proofs). Table 6 presents fit statistics of the two-stage CFA and SEM modeling.

Table 6 presents the fit and parsimony indices of seven measurement models and a path model. All posited models fit the data satisfactorily, although the memory and concentration (MC) model did not yield χ^2 statistics likely[2] due to the small number of relevant observations. The CPSs model fits the data quite well, and its high correlation (.98) with the LPC model indicates that these two latent traits are orthogonal, or highly related. Appendix C, Figure C1 presents a path diagram for each model.

After finalizing the measurement models, we investigated the interrelationships between latent traits, presented as the academic listening model in Figure 1. The model did not converge, so we took a compensatory strategy: we generated aggregate scores for each latent trait by adding up the items in each subscale. The score obtained from aggregation represents respondents' self-rating of their endowment of that latent trait.

Appendix C, Figure C2 presents a path model exploring the relationship between aggregate scores. This path model's use of observed scores and their relationships in multiple regression analysis differentiates it from Figure 1, which relies on measuring latent variables. Since the measurement models (Table 6) fit satisfactorily, we anticipated that the fit of the path model in Figure C2 and the SEM model in Figure 1 would be very similar. Table 1 presents the statistical features of Figure C2 as Path 1. This model's fit was close to the constraints tenable, but the regression coefficients of paths from MC to CPS, RIOM, and NT were statistically insignificant.

We modified Path 1 by deleting statistically insignificant paths. The modified model, Path 2 (Figure C3), fits the data well with more parsimony, has significant paths regression coefficients, and closely resembles our initial proposed academic listening model.

Table 6:
Fit Statistics of the CFA and SEM Models

Model	χ^2	*P*	*df*	χ^2/df	NNFI	CFI	CAIC	AIC	RMSEA	RMSEA 90% confidence interval
CPSs	77.94	0.814	90	0.866	1.006	1.000	294.38	137.94	0.000	0.000 – 0.050
LCP	84.16	0.270	77	1.093	0.997	0.997	286.17	140.16	0.014	0.000 – 0.030
CPS-LCP	342.94	0.596	530	0.980	1.001	1.000	746.96	454.94	0.000	0.000 – 0.015
LS	11.15	0.265	9	1.240	0.996	0.997	97.73	35.15	0.022	0.000 – 0.058
NT	4.082	0.130	2	2.041	0.984	0.995	61.79	20.08	0.046	0.000 – 0.110
RIOM	3.898	0.142	2	1.949	0.984	0.995	61.61	19.89	0.044	0.000 – 0.108
MC	U	NA	0	NA	U	1.000	43.28	12.00	0.435	0.394 – 0.479
Path 1	10.04	0.018*	3	3.347	0.924	0.989	U	58.04	0.140	0.051 – 0.240
Path 2	12.714	0.048*	6	2.119	0.964	0.990	U	57.71	0.097	0.009 – 0.171

Note. Fit indices: AIC = Akaike Information Criterion; CAIC = Consistent Akaike Information Criterion; CFI = Comparative Fit Index; *df* = degree of freedom; GFI = Goodness of Fit Index; NA = Not applied; NNFI = Non-Normed Fit Index; RMSEA = Root Mean Square Error of Approximation; U = Unknown.

Models: CPS = Cognitive processing skills; LCP = Linguistics component and prosody; NT = Note-taking; LS = Lecture structure; RIOM = Relating input to other materials; MC = Memory and concentration.

Good model fit is indicated by: non-significant chi-squared (χ^2); $\chi^2/df < 3.00$; NNFI > 0.90; CFI > 0.90; relatively small CAIC and AIC; and RMSEA < 0.08 with a small confidence interval.

* $p < 0.05$.

Discussion

We set out to investigate content-related, substantive, and structural evidence of validity of the ALSA questionnaire. On the whole, we found supportive evidence of the validity of the instrument, although the criterion-related validity of the instrument needs to be investigated in the future.

Content-related validity evidence

The model of academic listening we propose resulted from an extensive examination of relevant literature, including existing taxonomies of academic listening. The model sorts the major subskills that appear to work together in academic listening, and considers causal and correlational relationship among these subskills.

After initially developing the ALSA based on the model, the researchers, along with experts and students, conducted iterative content analysis to improve the content representation of items. The results of our subsequent analysis of the items' technical features supported our findings from this stage: all hypothesized subskills had moderate to high Rasch person and item reliability indices, indicating successful discrimination among participants; and had similar observed and expected item difficulty, fit, and point-measure correlation measures, indicating similarity between theoretical expectations and observations. We also observed that revising items after the pilot study improved the fit of the data to the RSM.

Substantive validity evidence

Applying Linacre's (2004) guidelines to our RSM findings, we find that, except for rating category 1 in subskills NT, MC, and RIOM, each rating category in each subskill was selected by more than 10 respondents, indicating that rating categories adequately tapped trait endowment levels of participants. If the observed deficiencies in rating category 1 may be due to the features of the students in the sample (i.e., they believe their trait level to be beyond a "1" in these subskills), we would anticipate this issue to resolve itself in a larger sample. If, however, these deficiencies stem from problems with the rating categories themselves, they would persist in a larger sample, implying that the scale should be fine-tuned.

As displayed (for CPSs and LCP) in Figure 3, all rating scale categories for all subskills displayed hill-shaped structures and separations. This shape, coupled with the categories' fit statistics, small departures from unity, and monotonically increasing incremental difficulty measures and thresholds, indicate that the rating scale categories were well-ordered – that is, that a lower category always corresponded to a lower trait endowment level. Finally, fit statistics and PMCs indicated coherence between model expectations and observed performance. In general, this analysis supports the substantive validity of both the ALSA and its underlying model.

Structural validity evidence

We sought structural validity evidence for the model by examining the internal relationships between hypothesized latent traits and items (through CFA) and the external relationships between latent traits (through SEM). Various fit statistics showed that the measurement models (CFA) fitted the data sufficiently, and that most regression coefficients were greater than .50, indicating that latent trait variables caused most observed variance.

In SEM analysis, the model did not converge, likely due to the sample size; this assessment could be tested by administering the ALSA to a larger sample. In a compensatory strategy, we used aggregate-level scores to examine the fit of the model. The Path 1 model, which included all hypothesized causal and correlational relationships displayed in Figure 1, did not fit well and the paths going from NT to MC to RIOM did not have significant regression coefficients. We modified the model by deleting the insignificant relationships, which yielded the Path 2 model.

The Path 2 model demonstrated that higher and lower order listening skills, operationalized as cognitive processing skills (CPSs) and linguistic component and prosody (LCP), respectively, predict two major components of the academic listening model: relating input to other materials (RIOM) and note-taking (NT). The model further shows that lecture structure (LS) is likely to influence CPA and LCP. Both these findings are consistent with previous research on academic listening. We further find that LS is likely to exert a significant impact on memory and concentration (MC): this conclusion seems to be unique to this study, and implies that certain lecture structures might be less successful in maintaining the full attention of students, which should be further investigated in the future. In sum, the path analysis supports the structural validity of the modified Path 2 model.

Future research

Although the ALSA as presented here appears to be reliable and content-wise, substantively, and structurally valid, it has not been externally validated. External evidence of validity is an important stage of validation in self-rating instruments (Messick, 1989), carried out by correlating students' self-assessed scores with their scores on another instrument that directly assesses the relevant latent traits. Significant correlation indices at the trait or item level are evidence of external validity.

Although we collected data on the performance of the students in our dataset on the IELTS listening test, we did not attempt an external validation of the ALSA using IELTS listening test scores, because the structure of the IELTS listening test does not seem to reflect the academic listening model which we proposed based on our review of relevant literature. For example, while our literature review shows that understanding inferred or implied messages is an important academic listening subskill, Aryadoust (2011a, c) and Geranpayeh and Taylor (2008) reported that IELTS listening tests taps only the ability to understand explicitly stated information. Several IELTS studies have reported weak

correlations between IELTS listening scores and the subsequent academic achievement of test takers (see Aryadoust, 2011b, c) and threats to its cognitive validity (see Field, 2009). A listening test with a similar underlying structure as the questionnaire would be most useful for investigating the criterion-referenced evidence of validity.

Two further concerns that should be addressed in the future research include parameter estimation in the CFA stage and participants' (potential) overestimation of their listening ability. First, estimating parameters in CFA modelling would to a great extent rely on the sample size. Although all CFA models converged in the present study, it would be important to further investigate and reconfirm the findings in the future with larger samples. In addition, it would be probable that students had somewhat overestimated their listening skills. To address this concern, future research should correlate students' self-ratings with their performance on objective academic listening tests (see Aryadoust, 2012). High correlations would indicate the relative accuracy of self-ratings. It is important to note that overestimation is different from deception or "faking" which is a deliberate attempt to falsify or mask information in self-assessment psychological studies in order to gain a benefit (see Kubinger, 2009; Seiwald, 2002). It seems that psychopathic patients are more prone to and successful at faking (Billings, 2004). Such behaviors in the present study are certainly very unlikely.

Implications for pedagogy and research

Learner-centered learning and assessment are attracting increasing attention among researchers and educators. The ALSA can be adopted into learner-centered curricula, especially in university contexts. Because its underlying structure fits the constituent structure of academic listening, we believe the ALSA has the potential to raise university students' awareness of their general listening skills, and of the elements of academic presentations that affect their academic achievement, such as their note-taking skills. This can allow students "take full account" of their own assessment (Little, 2005, p. 324); encourage teaching practices that develop lecture comprehension skills (see Nunan, 1988); and improve university awareness of weak and strong lecture comprehension skills (Ekbatani, 2000).

The ALSA is also well-suited to virtual educational environments, and can help improve independent learning and assessment in these environments (Dragemark, 2006). It can help transfer some assessment responsibilities of teachers to students, which is a new pedagogical goal in language for special purposes (Butler & Lee, 2006; Dragemark, 2006).

Educators who wish to use the ALSA (or other self-rating tools) can take several steps to obtain reliable and accurate results in self-rating. First, they must train students in the methods, objectives, and advantages of self-rating. They should also make efforts to develop students' autonomy, and should be sensitive to particular cultural variables that can affect self-rating outcomes.

Author Notes

Vahid Aryadoust is a lecturer at the Centre for English Language Communication of the National University of Singapore. Christine Goh is professor of linguistics and language education in the National Institute of Education, Nanyang Technological University, Singapore, with a special interest in L2 listening. Lee Ong Kim is associate professor of Measurement at the Policy and Leadership department.

References

Adams, C., & King, K. (1995). Towards a framework for student self assessment. *Innovation in Education and Training International, 32*, 336-43.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*, 561-73.

Armbruster, B. B. (2000). Taking notes from lectures. In R.F. Flippo & D.C. Caverly (Eds.), *Handbook of college reading and study strategy research* (pp. 175-199). Mahwah, NJ: Erlbaum.

Aryadoust, V. (2011a). Application of the fusion model to while-listening performance tests. *SHIKEN: JALT Testing & Evaluation SIG Newsletter, 15*(2), 2-9.

Aryadoust, V. (2011b). Constructing validity arguments for the speaking and listening modules of international English language testing system. *Asian ESP Journal, 7*(2), 28-54.

Aryadoust, V. (2011c). Differential item functioning in while-listening performance tests: The case of the International English Language Testing System (IELTS) listening module. *International Journal of Listening, 26*(1), 40-60.

Aryadoust, V. (2012). Reliability of second language listening self-assessments: Implications for pedagogy. *English Language Teaching World Online: Voices from the Classroom* (*ELTWO*), *5*. Retrieved from http://blog.nus.edu.sg/eltwo/about/.

Aryadoust, V., Goh, C., & Lee, O. K. (2011). An investigation of differential item functioning in the MELAB Listening Test. *Language Assessment Quarterly*, *8*(4), 1-25.

Bachman, L. F. (2004). *Statistical analyses for language assessment*. Cambridge: Cambridge University Press.

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

Bachman, L., & Palmer, A. (1989). The construct validation of self-ratings of communicative language ability. *Language Testing, 6*, 14-25.

Benson, M. (1989). The academic listening task: A case study. *TESOL Quarterly, 23*, 421-445.

Billings, F. J. (2004). Psychopathy and the ability to deceive. Dissertation Abstracts International. *Section B: The Sciences and Engineering, 65*(3-B), 1589.

Bodie, G. D. (2009). Evaluating listening theory: Development and illustration of five criteria. *International Journal of Listening, 23*, 81-103.

Bodie, G. D., & Fitch-Hauser, M. (2010). Quantitative research in listening: Explication and overview. In A. D. Wolvin (Ed.), *Listening and human communication in the 21st century* (pp. 46-93). Oxford, England: Blackwell.

Bodie, G. D., Worthington, D., & Fitch-Hauser, M. (2011). A Comparison of four measurement models for the Watson-Barker Listening Test (WBLT)-Form C. *Communication Research Reports, 28*, 32-42.

Bond, T. G. (1994). Too many factors? *Rasch Measurement Transactions, 8*, 347.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah, NJ: Lawrence Erlbaum Associates.

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrica, 52*, 345-370.

Brace, I. (2008). *Questionnaire design: How to plan, structure and write survey material for effective market research (2nd Ed.).* London and Philadelphia: Kogan Page.

Brantmeier, C. (2005). Nonlinguistic variables in advanced second language reading: Learners' self-rating and enjoyment. *Foreign Language Annals, 38*, 494-504.

Brantmeier, C. (2006). Advanced L2 learners and reading placement: Self-rating, CBT, and subsequent performance. *System, 34*, 15-35.

Brown, G., & Yule, G. (1983). *Teaching the spoken language: An approach based on the analysis of conversational English.* New York: Cambridge University Press.

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

Butler, Y. G., & Lee, J. (2006). On-task versus off-task self-rating among Korean elementary school students studying English. *The Modern Language Journal, 90*, 506-518.

Cameron, L. (1990). Adjusting the balance of power: Initial self assessment in study skills for higher education – a case study. In C. Bell (Ed.) *Assessment and evaluation* (pp. 63-72). London: Kogan Page.

Chaudron, C., Loschky, L., & Cook, J. (1995). Second language listening comprehension and note-taking. In J. Flowerdew (Ed.), *Academic Listening: Research perspectives* (pp. 75-92). Cambridge: Cambridge University Press.

Dragemark, A. (2006). Learning English for technical purposes: The LENTEC project. In T. Roberts (Ed.), *Self, peer, and group assessment in e-learning* (pp. 169-190). Hershey: Idea Group Inc.

Dudley-Evans, T. (1994). Variations in the discourse patterns favoured by different disciplines and their pedagogical implications. In J. Flowerdew (Ed.), *Academic listening: Research perspectives* (pp. 146-158). New York: Cambridge University Press.

Dunkel, P. A., & Davis, J. N. (1994). The effects of rhetorical signaling cues on the recall of English lecture information by speakers of English as a native and second language. In J. Flowerdew (Ed.), *Academic listening: Research perspectives* (pp. 55-74). New York: Cambridge University Press.

Ekbatani, G. (2000). Moving toward learner-directed assessment. In G. Ekbatani & H. Pierson (Eds.), *Learner-directed assessment in ESL* (pp. 1-11). Mahwah, NJ: Lawrence Erlbaum.

Eslami, Z. R., & Eslami-Raseck, A. (2007). Discourse markers in academic lectures. *Asian EFL Journal, 9*, 22-38.

Field, J. (2009). A cognitive validation of the lecture-listening component of the IELTS Listening paper. In P. Thompson, (Ed.), *IELTS research reports* (*Vol. 9*). Canberra: IELTS Australia, Pty Ltd.

Flowerdew, J. (1992). Definitions in science lectures. *Applied Linguistics, 13*, 202.221.

Flowerdew, J. (1994). Research of relevance to second language lecture comprehension: An overview. In J. Flowerdew (Ed.), *Academic listening: Research perspectives* (pp. 7-29). Cambridge: Cambridge University Press.

Flowerdew, J., & Miller, L. (1992). Student perceptions, problems and strategies in second language lecture comprehension. *RELC Journal, 23*(2), 60-80.

Ford, W. S. Z., Wolvin, A. D., & Sungeun, C. (2000). Students' self-perceived listening competencies in the basic speech communication course. *International Journal of Listening*, *14*, 1-13.

Ford, W. S. Z., & Wolvin, A. D. (1992). Evaluation of a basic course in speech communication. In L. W. Hugenberg (Ed.), *Basic communication course annual* (pp. 35-47). Boston: American Press.

Ford, W. S. Z., & Wolvin, A. D. (1993). The differential impact of a basic communication course on perceived communication competencies in class, work, and social contexts. *Communication Education, 42*, 215-223.

Frederick, P. (1986). The lively lecture – 8 variations. *College Teaching*, *34*, 43-50.

Gillham, B. (2006). *Developing a questionnaire (2nd Ed.)*. London and New York: Continuum.

Givón, T. (1979). *On understanding grammar*. New York: Academic Press.

Goh, C. (2005). Second language listening expertise. In K. Johnson (Ed.), *Expertise in Second Language Learning and Teaching* (pp. 64-84). UK: Palgrave Macmillan.

Goh, C., & Aryadoust, S. V. (2010). Investigating the construct validity of MELAB listening test through the Rasch analysis and correlated uniqueness modeling. *Spaan Fellowship Working Papers in Second of Foreign Language Assessment, 8*, 31-68. Ann Arbor, MI: University of Michigan English Language Institute.

Granville, S., & Dison, L. (2005). Thinking about thinking: Integrating self-reflection into an academic literacy course. *Journal of English for Academic Purposes, 4*, 99-118.

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). *Multivariate data analysis* (7th ed.). New Jersey: Pearson Educational Product.

Hansen, C. (1994). Topic identification in lecture discourse. In J. Flowerdew (Ed.), *Academic listening; research perspectives* (pp. 131-145). New York: Cambridge University Press.

Heilenmann, K. L. (1990). Self assessment of second language ability: The role of response effects. *Language Testing, 7*, 174-201.

Hodgson, V. (1984). Learning from lectures. In F. Marton, D. Hounsell, & N. Entwistle (Eds.), *The experience of learning* (pp. 103-123). Edinburgh: Scottish Academic Press.

Jafapur, A. (1991). Can naive EFL learners estimate their own proficiency? *Evaluation and Research in Education, 5*, 145-57.

James, K. (1977). Note-taking in lectures: Problems and strategies. In A. P. Cowie & J. B. Heaton (Eds.), *English for academic purposes* (pp. 89-98). London: BAAL/SELMOUS.

Janusik, L. (2007). Building listening theory: The validation of the Conversational Listening Span. *Communication Studies, 58*, 139-156.

Imhof, M., & Janusik, L. A. (2006). Development and validation of the Imhof-Janusik listening concepts inventory to measure listening conceptualization differences between cultures. *Journal of Intercultural Communication Research, 35*(2), 79-98.

Johns, A. M. (1981). Necessary English: A faculty survey. *TESOL Quarterly, 15*, 51-57.

Jordan, R. R. (1997). *English for academic purposes: A guide and resource book for teachers*. Cambridge: Cambridge University Press.

Jöreskog, K. G., & Sörbom, D. (2001). *LISREL 8.8: User's reference guide.* Lincolnwood, IL: Scientific Software International, Inc.

Kelly, P. (1991). Lexical ignorance: The main obstacle to listening comprehension with advanced foreign language learners. *IRAL, 29*, 135-149.

Kiewra, K. A. (1984). Implications for note taking based on relationships between note taking variables and achievement measures. *Reading Improvement, 21*, 145-149.

Kiewra, K. A. (2002). How classroom teachers can help students learn and teach them how to learn. *Theory into Practice, 41*, 71-80.

King, P. (1994). Visual and verbal messages in the engineering lecture: Note-taking by postgraduate L2 students. In J. Flowerdew (Ed.), *Academic Listening: Research Perspectives* (pp. 219-238). Cambridge: Cambridge University Press.

Kubinger, K.D. (2009). Three more attempts to prevent faking good in personality questionnaires. *Review of Psychology, 16*, 115-121.

Kunnan, A. J. (1994). Modelling relationships among some test-taker characteristics and performance on EFL tests: An approach to construct validation. *Language Testing, 11*, 225-52.

Liao, Y-F. (2009). *A construct validation study of the GEPT reading and listening sections: Re-examining the models of L2 reading and listening abilities and their relations to lexico-grammatical knowledge*. Unpublished doctoral dissertation, Teachers College, Columbia University.

Linacre, J. M. (1998). Rasch First or Factor First? *Rasch Measurement Transactions, 11*, 603.

Linacre, J. M. (2008). The expected value of a point-biserial (or similar) correlation. *Rasch Measurement Transactions, 22*, 1154.

Linacre, J. M. (2000). Comparing "partial credit" and "rating scale" models. *Rasch Measurement Transactions, 14*, 768.

Linacre, J. M. (2004). Optimizing rating scale category effectiveness. In E. Smith & R. Smith (Eds.), *Introduction to Rasch measurement* (pp. 258-278). Maple Grove, MN: JAM Press.

Linacre, J. M. (2010b). *A users' guide to WINSTEPS® MINISTEPS Rasch-model computer programs.* Wisteps.com.

Linacre, J. M. (2010b). *WINSTEPS: Rasch model computer programs* [computer program]. Wisteps.com.

Little, D. (2005). The Common European Framework and the European Language Portfolio: Involving learners in their judgments in the assessment process. *Language Testing*, *22*, 321-336.

Matsuno, S. (2009). Self-, peer-, and teacher-assessments in Japanese university EFL writing classrooms. *Language Testing, 26,* 75-100.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (pp. 13-103). New York: Macmillan.

Mowl, G., & Pain, R. (1995). Using self and peer assessment to improve students' essay writing: a case study from Geography. *Innovation in Education and Training International* 32, 324-35.

Nunan, D. (1988). *The learner-centered curriculum*. Cambridge: Cambridge University Press.

Olmos, O. L., & Lusung-Oyzon, M. V. P. (2008). Effects of prior knowledge and lesson outline on note taking and test scores. *Education Quarterly*, *66*, 71-86.

Olsen, L. A., & Huckin, T. N. (1990). Point-driven understanding of engineering lecture comprehension. *English for Specific Purposes, 9*, 33-47.

Orsmond, P., Merry, S., & Reiling, K. (1997). A study in self assessment: Tutor and students' perceptions of performance criteria. *Assessment & Evaluation in Higher Education, 22*, 357-67.

Oscarson, M. (1984). *Self assessment of foreign language skills: A survey of research and development work*. Strasbourg: Council of Europe, Council for Cultural Co-operation.

Oscarson, M. (1999). Estimating language ability by self assessment: A review of some of the issues. In *Papers on Language Learning Teaching Assessment.* Festskrift till Torsten Lindblad, Göteborgs universitet, Institutionen för pedagogik och didaktik.

Patri, M. (2002). The influence of peer feedback on self- and peer-assessment of oral skills. *Language Testing, 19*, 109-131.

Powers, D. E. (1986). Academic demands related to listening skills. *Language Testing, 3,* 1-38.

Richards, J. C. (1983). Listening comprehension: Approach, design, procedure. *TESOL Quarterly, 17,* 219-240.

Rivers, W. P. (2001). Autonomy at all costs: An ethnography of metacognitive self-rating and self-management among experienced language learners. *The Modern Language Journal, 85*, 279-290.

Rolfe, T. (1990). Self and peer-assessment in the ESL curriculum. In G. Brindley (Ed.), *The second language curriculum in action (Vol. 6)* (pp. 163-86). Sydney: NCELTR, Macquarie University.

Ross, S. (1998). Self assessment in second language testing: A meta-analysis and analysis of experiential factors. *Language Testing, 15*, 1-20.

Rost, M. (1994). On-line summaries as representations of lecture understanding. In J. Flowerdew (Ed.), *Academic Listening: Research Perspectives* (pp. 93-127). Cambridge: Cambridge University Press.

Rounds, P. (1987). Characterizing successful classroom discourse for NNS teaching assistant training. *TESOL Quarterly, 21*, 643-671.

Sawaki, Y., & Nissan, S. (2009). *Criterion-related validity of the TOEFL® iBT listening section* (TOEFL iBT™ Report No. iBT-08). Princeton, NJ: ETS.

Schumacker, R. E., & Linacre, J. M. (1996). Factor analysis and Rasch. *Rasch Measurement Transactions*, *9*, 470.

Seiwald, B. B. (2002). Replicability and generalizability of Kubinger's results: Some more studies on faking personality inventories. *Psychologische Beiträge, 44,* 17-23.

Shore, T. H., Shore, L. M., & Thornton III, G. C. (1992). Construct validity of self and peer evaluations of performance dimensions in an assessment center. *Journal of Applied Psychology, 77*, 42-54.

Smith, R. M. (1996). A comparison of methods for determining dimensionality in Rasch measurement. *Structural Equation Modelling*, *3*, 25-40.

SPSS Inc. (2007). *SPSS for Windows release 16.0 standard version.* Chicago: SPSS Inc.

Stefani, L. A. J. (1994). Peer, self, and tutor assessment: Relative reliabilities. *Studies in Higher Education, 19*, 69-75.

Strodt-Lopez, B. (1991). Tying it all in: asides in university lectures. *Applied Linguistics, 12*, 117-140.

Sullivan, K., & Hall, C. (1997). Introducing students to self-rating. *Assessment and Evaluation in Higher Education, 22*, 289-305.

Tauroza, S., & Allison, D. (1994). Expectation-driven understanding in Information Systems Lecture Comprehension. In J. Flowerdew (Ed.), *Academic listening: Research perspectives* (pp.35-54). Cambridge University Press: Cambridge.

Van Dieten, A. (2000). Alternative assessment: Self-rating beyond the mainstream. *Melbourne Papers in Language Testing, 9*, 18-29.

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

Weir, C. (1990). *Communicative language testing*. New York: Prentice Hall.

Williams, R. L., & Eggert, A. (2002). Notetaking predictors of test performance. *Teaching of Psychology, 29,* 234-237.

Willis, G. B. (2005). *Cognitive interviewing: A tool for improving questionnaire design*. Thousand Oaks, Calif.: Sage Publications.

Wolfe, E. W., & Smith, Jr. E. V. (2007a). Instrument development tools and activities for measure validation using Rasch models: Part I – Instrument Development Tools. *Journal of Applied Measurement, 8,* 97-123.

Wolfe, E. W., & Smith, Jr. E. V. (2007b). Instrument development tools and activities for measure validation using Rasch models: Part II – Validation Activities. *Journal of Applied Measurement, 8,* 204-233.

Wright, B. D. (1994a). Comparing factor analysis and Rasch measurement. *Rasch Measurement Transactions, 8*, 350.

Wright, B. D. (1994b). Local dependency, correlations, and principal components. *Rasch Measurement Transactions, 10*, 509-511.

Young, L. (1994). University lectures – macro structure and micro features. In J. Flowerdew (Ed.), *Academic listening: Research perspectives* (pp. 159-176). Cambridge, Cambridge University Press.