Reliability of Second Language Listening Self-Assessments: Implications for Pedagogy

by Vahid Aryadoust

National University of Singapore (Singapore)

Keywords: academic listening, rating scale model, self-assessment

Abstract

Language self-appraisal (or self-assessment) is a process by which students evaluate their own language competence. This article describes the relationship between students' self-appraisals and their performance on a measure of academic listening (AL). Following Aryadoust and Goh (2011), AL was defined as a multi-componential construct including cognitive processing skills, linguistic components and prosody, note-taking, rating input to other materials, knowledge of lecture structure, and memory and concentration. Participants (*n* = 63) were given a self-assessment questionnaire which is founded upon the components of AL presented by Aryadoust and Goh, and a test of academic listening developed by English Testing Service (ETS); subsequently, their performance on both measures were found to be correlated. Significant correlations were apparent, indicating that learners assessed their listening skills fairly accurately and precisely. Pedagogical implications and applications of self-assessment are discussed in this paper.

Introduction

Language self-assessment is a process by which learners evaluate their own language abilities. A number of language researchers have used self-assessment as an effective method in teaching and assessment (for example, Brantmeier, 2006; Little, 2005; Rivers, 2001) and found it to be a reliable method of improving students' skills and abilities (Ekbatani, 2000; Nunan, 1988). For example, Little (2005) argued that employing self-assessment procedures would "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).

Self-assessment can lead to learner autonomy. Several influential learning paradigms have recently advocated autonomous learning. For example, Little (2005, p. 321) reported that the European Language Portfolio (ELP) and the Common European Framework of Reference for

Languages (CEFR) have embraced learner-centered education and self-assessment systems, as these systems generate a learning context where students “take full account” of their own assessment. Similarly, Dragemark (2006) argues that self-assessment is useful in “virtual” and long-distance education: given that instructors are not physically present to provide feedback to long-distance learners, fostering student-led methods of learning and evaluation becomes highly crucial. To achieve these goals, researchers would need to develop reliable self-assessment systems and subject them to data analysis procedures to verify their underlying structure (Aryadoust, 2011).

Using self-assessment in (academic) listening comprehension classes, however, is not a well-researched arena, given that language researchers have not yet established a rigorous theory of listening comprehension (Aryadoust, in press). In recent years, useful attempts have been made to problematize and investigate the latent structure of listening comprehension (for example, Buck, 2001; Goh, 2008; Goh & Aryadoust, 2010; Vandergrift & Goh, 2009; Wagner, 2004), though language researchers have not reached a consensus on the definition and operationalization of listening. This is partly due to the nature of the skill as well as the medium through which the message is conveyed (aural) and the vast applications of the skill in academia and daily life (Bodie, 2009).

Nevertheless, teaching listening comprehension and engaging students in self-assessment have recently become two demanding responsibilities of language teachers in many language curricula. To help improve students’ listening skills through self-assessment, language teachers should initially develop a tentative listening theory (Buck, 2001) which reflects students’ needs and objectives of the course; the theory must be informed by contemporary scholarly literature which presents multiple perceptions of the skill (Bodie, 2009). Indeed, by engaging students in self-assessment, teachers would transfer a part of their responsibility and *knowledge* to learners and raise students’ awareness of the listening subskills, thereby fostering learner’s autonomy.

For example, as Bejar, Douglas, Jamieson, Nissan, and Turner’s (2000) model of the listening subtest of the Internet-Based Test of English as a Foreign Language (TOEFL iBT) suggests, listeners access their situational, linguistic, and background knowledge sources to process the auditory input and achieve comprehension. Situational knowledge pertains to the role of pictorial clues and gestures; linguistic knowledge refers to vocabulary, syntax, pragmatics, and discourse knowledge (Bachman, 1990); and background knowledge concerns listeners' schema and world knowledge. The result of the application of these knowledge sources to the oral input is a set of mental representations (or propositions), which aids in comprehension (see Kintsch, 2007; Kintsch & van Dijk, 1978). Second language learners are often unaware of such influential mechanisms, and instead they feel anxious and frustrated when they do not comprehend audio messages (see Graham, 2006; Rost, 2002). On this understanding, students who are studying towards achieving the language skills measured by, for example, the TOEFL iBT, would benefit from diagnosing their own weaknesses and strengths (Alderson, 2005)—the prime goal of self-assessment.

In practice, such awareness can be attained by many iterations of tests and self-assessments as well as direct instructions on listening subskills, which seem to be endorsed by language instructors who take a diagnostic approach (see Hayes & Read, 2004). In a module that takes a diagnostic approach, initially subskills have to be taught explicitly by teachers and there needs to be explicit engagement with students on matching self-appraisals with achievement (Aryadoust, 2011). The more the students learn about the value of self-awareness and the more emphasis teachers attach to it, the more precise the self-assessment becomes during the course.

While several studies have evaluated the efficacy of self-assessment procedures in educational and language measurement (for example, Ford, Wolvin, & Sungeun, 2000; Sawaki & Nissan, 2009), analysis of its utility in academic contexts has been critically limited. The present study investigates the relationship between second language academic listening ability measured by a listening test adapted from the *English Testing Service* (ETS) and self-assessment measured by an English academic listening self-assessment questionnaire (ALSAQ).

The ALSAQ is a 47-item self-assessment tool validated by Aryadoust and Goh (2011), to explore the reliability of listening self-assessment. Drawing on the results of an extensive literature survey, Aryadoust and Goh based the structure of the ALSAQ on a multi-componential construct comprising the following sub-skills or components:

1. cognitive processing skills (CPSs): ability to understand surface (explicitly stated) information and making inferences (16 items);
2. linguistic components and prosody (LCP): vocabulary and syntactic resources (13 items);
3. note-taking (NT): ability to take notes of main ideas and details of the aural message (4 items);
4. knowledge of lecture structure (LS): students' awareness and/or understanding of the framework upon which the structure of the lecture is founded (6 items);
5. relating input to other materials (RIOM): ability to form a mental connection between the information transferred through various modes (4 items); and
6. memory and concentration (MC): ability to keep important parts of the message in mind (3 items).

Using the Rasch model and structural equation modeling, Aryadoust and Goh (2011) investigated the psychometric features of the questionnaire and built a validity argument for it. They argued that the instrument would be most pertinent in academic contexts where English teachers / testers seek to use a reliable tool to raise students' awareness of their level of understanding, cognitive resources, and listening skills.

Finally, research into listening self-awareness has shown that students who take preparation courses for English exams (for example, the International English Language Testing System, or IELTS) would attain the self-awareness level to be able to answer the self-assessment inventories precisely (see Breeze & Miller, 2011). This assumption seems to hold regarding the participants in the present study.

Methodology

Participants

Sixty three (63) English as a second language (ESL) students participated in the study. Forty two (42) participants (66.5%) were pursuing master's degrees and the rest were undergraduate students (*n* = 21; 33.5%). They had taken English preparation courses and were familiar with the concepts tested by ALSAQ. Table 1 presents the distribution of their mother tongues.

Table 1

Distribution of the Participants' Mother Tongues

Language	Frequency	Percent
Chinese	22	35
Persian	17	27
Arabic*	13	20.5
Malay	11	17.5

Note. n = 63. * = Arabic countries include Jordan and Iraq.

Procedures

Participants filled out consent forms prior to participating in the study. They were given a test of academic English including a lecture on history followed by 14 questions. The lecture was 30 minutes long and had been selected by Sawaki and Nissan (2009) from a large pool of 60 lectures produced by *The Teaching Company* (http://www.thegreatcourses.com/). The test items are the property of the ETS and permission was obtained for the inclusion of these materials in this study. (Readers are referred to Sawaki and Nissan's 2009 research report for further information).

As part of the test administration, an outline of the direction of the lecture was provided to the participants. The lecture was played once and they were advised to take notes while listening to the lecture. After the test, participants filled in the questionnaire and submitted their answer sheets.

Data analysis

The psychometric features of the ALSAQ were initially investigated. Although this had been previously undertaken by Aryadoust and Goh (2011), it would be necessary to investigate the features of the items if the tool is administered to a*new sample* (Messick, 1989). The data was fit to Andrich's rating scale model (RSM) (Andrich, 1978) in an attempt to determine whether participants would perform on the questionnaire according to their estimated ability levels and to examine the features of scoring categories (i.e., four points on the Likert scale). For example, for the scoring category *2* to function appropriately, it must be chosen more often by the participants whose ability level (as estimated by the RSM) is greater than *2* and less often by the participants whose ability level is below *2*. The difficulty of scoring categories must increase "monotonically" from lower to higher categories (i.e., 1 < 2 < 3 < 4). The six ALSAQ components were subjected to the RSM separately, as they are regarded as separate yet interconnected *dimensions* of the ALSAQ (i.e., six integrated academic listening macro-skills). The reason for separate calibration of each dimension is that aggregating all dimensions into one general dimension would violate the assumption of unidimensionality of the data, which is a precondition of the RSM (explaining dimensionality would fall out of the scope of this article. Interested readers are referred to Aryadoust, Goh, & Lee, 2011).

The RSM is an extension of the Rasch model, which was developed to examine dichotomous data. Rasch model computer programs such as *WINSTEPS* (Linacre, 2012), which was used in this study, provide multiple *fit* indices (i.e., quality control statistics) to evaluate the quality of the data as well as the instrument. The most commonly used indices are infit / outfit mean square (MNSQ) and z-standardized (ZSTD[1]). Infit indices are sensitive to aberrations of the performance of average-ability respondents (as well as average-difficulty items) and outfit indices convey information regarding the aberration of high or low ability participants[2] (as well as high or low difficulty items). That is, they flag the persons and items whose psychometric features seem to be unusual, for example, a low-ability respondent who would answer a few difficult items accurately or a low-ability respondent who would endorse a difficult item highly (the term 'difficult item' is analogous to lowly endorsable items in the context of questionnaires). The expected MNSQ value is unity; with polytomous data, values below 0.5 are considered overfits whereas values greater that 1.5 are considered misfits or underfits (Bond & Fox, 2007).

Next, bivariate correlation coefficients were computed by using Rasch measures to assess the relationship between participants' scores on the ETS listening test and the subscales of ALSAQ. Bivariate correlation coefficients partial out (i.e., control for) the influence of other variables. Performance on the two instruments correlate significantly if test takers' awareness of their academic listening skills is relatively accurate.

Results

Table 2 gives the results of the RSM. For example, it can be said Item 1 was highly endorsed by participants (Difficulty measure = -0.74); that is, given that most participants perceived their ability to understand "*isolated words and short phrases in spoken English, such as numbers and commonplace names*" to be high, item difficulty (or endorsability) of this item is relatively lower than endorsability of Item 2 (Difficulty measure = 0.3). Item 19 was the least endorsable (most difficult) (Difficulty measure = 1.29) as most participants believed that they would have trouble modifying their "*understanding of the lecture if it is incorrect.*" Fit estimates of four items (i.e., 13, 22, 27, and 45) fell outside the range between 0.5 and 1.5, indicating unpredictability (noise) in the data. Due to the small sample size, it was decided to keep these items as the erratic fit statistics can be said to be indicative of potential problems (Bond & Fox, 2007).

Table 2

Item difficulty and Fit Indices Estimated by Using the Rating Scale Model

Item	Difficulty measure	Infit MNSQ	Outfit MNSQ
1	-0.74	0.94	0.96
2	0.30	1.07	1.01
3	0.30	1.56	1.42
4	-1.12	1.18	1.18
5	-0.19	0.98	0.95
6	-0.19	0.71	0.67
7	-0.07	0.86	0.83
8	-0.6	0.97	0.95
9	-0.67	1.45	1.46
10	0.81	0.97	0.94
11	0.00	0.81	0.78
12	-0.33	0.9	0.92
13	-0.89	0.9	1.92
14	-0.67	1.02	1.13
15	0.18	0.93	0.82
16	0.12	0.93	1.71
17	-0.53	1.01	0.91
18	-0.39	0.80	0.86

19	1.29	0.80	0.78
20	0.81	1.02	0.95
21	-0.07	0.75	0.69
22	1.70	1.85	1.81
23	0.53	1.05	1.25
24	0.36	0.73	0.97
25	0.12	0.84	0.73
26	-0.26	0.91	0.78
27	-0.26	1.43	1.79
28	-0.46	0.90	0.86
29	0.41	0.94	0.96
30	0.47	0.88	0.82
31	0.12	1.08	0.96
32	0.70	0.81	0.74
33	0.36	0.91	0.85
34	0.47	1.22	1.18
35	0.12	0.85	0.78
36	-0.26	0.83	0.84
37	-0.39	0.92	0.88
38	-1.21	1.17	1.58
39	0.53	1.09	1.03
40	0.64	0.78	0.78
41	-0.26	0.84	0.88
42	0.06	0.78	0.73
43	-0.33	0.94	0.94
44	0.53	1.12	1.08
45	-0.46	1.85	1.60
46	0.18	0.76	0.73
47	-0.74	0.66	0.52

Note. This table reports the results of the application of the Rasch Rating Scale model (RSM) to the data (*n* = 63). The RSM was applied to individual subscales. Difficulty 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.

Next, items tapping each dimension were aggregated and six aggregate-level variables (or super-items) were created and correlated. Table 3 presents the bivariate correlations of the variables. There were strong positive correlations between ALSAQ sub-skills and also between the ETS academic listening test and ALSAQ sub-skills ($p < 0.05$). That is, increase in self-appraisals was correlated with increase in ETS academic listening test scores.

Table 3

Correlation of ALSAQ Sub-skills and the ETS Listening Test

	CPS	LCP	NT	LS	RIOM	MC
CPS	1					
LCP	0.912^{**}	1				
NT	0.772^{**}	0.812^{**}	1			
LS	0.889^{**}	0.879^{**}	0.753^{**}	1		
RIOM	0.785^{**}	0.744^{**}	0.679^{**}	0.765^{**}	1	
MC	0.574^{**}	0.591^{**}	0.571^{**}	0.581^{**}	0.465^{**}	1
EST-Criterion	0.497^{**}	0.520^{**}	0.419^{**}	0.496^{**}	0.402^{**}	0.232*

Note. n = 63. LS = lecture structure; CPS = cognitive processing skills; LCP = linguistics component and prosody; MC = memory and concentration; RIOM = relating ideas to other materials; NT = note-taking.

* $p < 0.05$. ** $p < 0.01$

Discussion and Pedagogical Implications

This study investigated the relationship between academic listening self-appraisals as measured by the ALSAQ and scores achieved on the ETS academic listening test—a test of academic listening comprehension. Initially, the psychometric features of the ALSAQ were established and then the precision (predictive power) of participants' assessment of their academic listening performance was compared with their ETS criterion scores. Taken as a whole, it seems that students evaluated their academic listening competence relatively accurately, as their scores on cognitive processing skills (CPSs), linguistic components and prosody (LCP), note-taking (NT), knowledge of lecture structure (LS), relating input to other materials (RIOM), and memory and concentration (MC) correlated significantly with the ETS criterion scores. This finding is promising because as Little (2005) argued self-assessment procedures have the potential to tie learning with assessment and provide fine-grained diagnostic feedback to both students and teachers (Alderson, 2005). This feedback can help those learners who might be unaware of their weaknesses and strengths before using self-appraisal and feel frustrated when encountering comprehension difficulties (Graham, 2006; Rost, 2002).

ETS test scores' correlation with ALSAQ sub-skills further points to the predictive validity and precision of self-assessments. Predictive validity of measurement tools is supported when students' scores correlate significantly with the scores they achieved on a criterion instrument which assesses the same language ability; the more precise students' self-ratings, the higher the correlation between the two sets of scores. Indeed, familiarity with the concepts measured by the ALSAQ in the present study made learners' appraisal of their listening ability fairly precise. The accuracy and precision of such appraisals can be further improved if self-assessment is adopted as a constituent element of educational programs and student progress is regularly monitored by teachers (Dragemark, 2006). Because self-assessment breaks down the listening skill into several smaller sub-skills, it makes available a means by which students can closely examine their listening skills and thereby be more intentional and targeted about developing their skills (Alderson, 2005).

The results of the present study resonate with Bachman and Palmer's (1989), Dragemark's (2006), and Oscarson's (1999) studies which reported high correlations between self-assessment of language skills and objective tests. It is important to note that using Rasch measures in correlation studies confers an advantage over raw data. In raw data, test items (and similarly persons) are classified according to the extent to which they possess a set of characteristics, producing ordinal-level (or rank-order) data. Because the distances between rank orders are unequal (for example, the distance between 20 and 40 is not the same as that between 75 and 95), calculating mean scores or correlating them is regarded as an erroneous procedure by many commentators (see Mackintosh, 1998). In contrast, when the data fits the Rasch model, the instrument can be said to measure the targeted construct on an interval scale (see Bond & Fox, 2007, for a discussion); correlating interval data and calculating their mean score would be plausible when interval-level data are used.

As a diagnostic tool, ALSAQ's cognitive processing skills (CPSs) and linguistic components and prosody (LCP) subscales, which map onto subcomponents of linguistic knowledge, can facilitate students' cognitive development by highlighting their problem areas. Note-taking (NT), which is an important sub-skill in academic writing, can be stressed by applying the NT subscale in classrooms; and finally, knowledge of lecture structure (LS), relating input to other materials (RIOM), and memory and concentration (MC) contribute significantly to academic listening performance as well as test takers' performance (Bachman & Palmar, 2010). Therefore, enough time and space should be allocated to these subscales in language curricula.

Users of the ALSAQ should also note that the tool can be used as either a context-specific or general instrument; that is, students can respond to the questionnaire based on their estimate of their performance either after taking a specific test (specific use) or before doing a test (general use). The first way, which was applied in the present study, makes students gauge how well they were able to perform the specified task. It might be said that this method will furnish better estimates of ability which correlate significantly with performance scores. In contrast, the second way would help students gauge what they believe their listening skills ability to be, but possibly with lower precision.

The ALSAQ can further help foster skills of "one-way" listening for academic purposes (Lynch, 2011). Lynch argued that one-way listening is composed of an important set of language sub-skills which are applied in (mini-)lectures, seminars, and conferences by students. Such specific sub-skills are reflected in the ALSAQ (Aryadoust & Goh, 2011). The tool would likely raise students' awareness of discourse structure, metadiscourse, and discourse shift signals, if teachers who teach one-way listening and related courses lay enough stress on self-assessment. Development of precision and accuracy in self-assessment can be explored through an experimental study where participants receive continuous instruction on self-appraisal during a listening course. By administering the ALSAQ alongside an academic listening test at the beginning, in the middle, and at the end of the course, the teacher can build a growth model for the accuracy of self-appraisals and compare them with students' scores on the listening test. This method would render self-assessment suitable for measuring formative goals.

The current study has certain limitations that need to be taken into account. ETS academic listening test scores are aggregate-level scores, which is to say that contrary to ALSAQ, ETS academic listening test items are holistic and do not discriminate the underpinning sub-skills although the test engages these sub-skills. For example, the ETS academic listening test taps the note-taking sub-skill by allowing students to take notes, but notes are not marked separately. Students, however, can use their notes to answer the test items (see Goh & Aryadoust, 2010). Future research should address this limitation by separating and correlating listening sub-skills tapped by the ETS criterion (or similar tests) and corresponding components of the ALSAQ. Correlating the underlying sub-skills of both instruments would need a relatively larger sample.

Conclusion

The ALSAQ can be adopted into learner-centered assessment and pedagogy curricula, as it can raise students' awareness of their general listening abilities and of the constituents of academic listening that would affect academic achievement. This can encourage teaching techniques and methodologies that develop listening comprehension skills (Nunan, 1988).

The ALSAQ would also fit virtual educational environments and improve independent learning and assessment in these environments (Dragemark, 2006). Finally, educators who use the ALSAQ must train students on the goals and merits of self-assessment, attempt to develop students' independence, and be cautious about particular cultural factors that can influence self-appraisal outcomes.

References

Alderson, J. C. (2005). *Diagnosing foreign language proficiency: the interface between learning and assessment*. London: Continuum.

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

Aryadoust, V. (2011). Application of the fusion model to while-listening performance tests. *SHIKEN: JALT Testing & Evaluation SIG Newsletter, 15*(2), 2-9. Retrieved from http://jalt.org/test/ary_2.htm

Aryadoust, V. (in press). Using cognitive diagnostic assessment to model the underpinning structure of the lecture comprehension section of the IELTS listening test: A sub-skill-based approach. *Asian EFL Journal*.

Aryadoust, V., & Goh, C. (2011, June). *Developing an academic listening self-assessment questionnaire: An exploratory study of academic listening macro-skills.* Paper presented at the Applied Linguistics Association of Canada Conference, Fredericton, Canada.

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

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in the real world: Designing language assessments and justifying their use.* Oxford: Oxford University Press.

Bejar, I., Douglas, D., Jamieson, J., Nissan, S., & Turner, J. (2000). *TOEFL 2000 listening framework: A working paper* (TOEFL Monograph Series No. MS-19). Princeton, NJ: ETS.

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

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

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

Breeze, R., & Miller, P. (2011). Predictive validity of the IELTS Listening Test as an indicator of student coping ability in Spain. In L. Taylor (Ed.), *IELTS Research Reports* (Vol. 12) (pp. 201-234). www.IELTS.org. Retrieved from http://www.ielts.org/PDF/vol12_report_5.pdf

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

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.

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.

Goh, C. (2008). Metacognitive instruction for second language listening development: Theory, practice and research implications. *RELC Journal*, *39*(2), 188-213.

Goh, C., & Aryadoust, 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.

Graham, S. (2006). Listening comprehension: the learners' perspective. *System*, *34,* 165–182.

Hayes, B., & Read, J. (2004). IELTS test preparation in New Zealand: Preparing students for the IELTS academic module. In L. W. Cheng, Y. and Curtis, A. (Eds.), *Washback in language testing: Research contexts and methods* (pp. 97-111). Mahwah, NJ: Lawrence Erlbaum Associates.

Kintsch, W. (2007) Meaning in context. In T. K. Landauer, D. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of latent semantic analysis* (pp. 89-105). Mahwah, NJ: Erlbaum.

Kintsch, W., & van Dijk, T. A. (1978). Towards a model of text comprehension and production. *Psychological Review*, *85,* 363-394.

Linacre, J. M. (2012). WINSTEPS: Rasch model computer program [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.

Lynch, T. (2011). Academic listening in the 21st century: Reviewing a decade of research. *Journal of English for Academic Purposes*, *10*, 79–88.

Mackintosh, N. J. (1998). *IQ and human intelligence*. Oxford: Oxford University Press.

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

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

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.

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

Rost, M. (2002). *Teaching and researching listening*. London: Longman.

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

Vandergrift, L., & Goh, C. (2009). Teaching and testing listening comprehension. IN Long, M. & Doughty, C. (EDS.), *The handbook of language teaching (Handbook in Linguistics Series)* (PP. 395-411). Oxford: Blackwell Publishing.

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.

[1] The ZSTD statistics are the transformation of the fit indices to standard normal distributions with a mean index of zero and a standard deviation of one. The acceptable range of the ZSTD indices is between -2 and +2 (Bond & Fox, 2007). Given that ZSTD indices do not precisely reflect the quality of data in small samples, they are not reported in the present study.

[2] Ability level is defined as students' endowment of the language skill under assessment and estimated by the computer program merely on the basis of the responses that participants provide.

About the author

Vahid Aryadoust is a lecturer at the Centre for English Language Communication of the National University of Singapore (CELC-NUS). He currently teaches academic writing courses and is the principal investigator in several validation studies of language proficiency exams, including the International English Language Competency Assessment (IELCA).