RESEARCH

Open Access

A communicative validation study on an English listening test in Korea

Ci Zhang[1], XiaoShu Xu[1,2*] and Yunfeng Zhang[3]

*Correspondence:
Lisaxu@wzu.edu.cn

[1] School of Foreign Studies, Wenzhou University, Wenzhou, China
[2] Stamford International University, Bangkok, Thailand
[3] Centre for Portuguese Studies, Macau Polytechnic University, Macau, Macao

Abstract

This study presents the validation process of a listening test based on a communicative language test proposed by Bachman (Fundamental considerations in language testing, 1990). It was administered to third-grade high school students by the sixteen Korean Provincial Offices of Education for Curriculum and Evaluation in September 2012 to assess their listening ability at the end of high school learning and compare it with the standard of the CSAT (College Scholastic Ability Test). The research questions were the following. First, to what extent does the test measure the listening comprehension construct? Second, what sub-skills does the test measure? Third, to what extent does the test measure communicative ability? To answer these three questions, a study was designed to examine the test's construct validity using classical test theory (CTT). Then, an exploratory factor analysis (EFA) and a confirmatory factor analysis (CFA) were applied to find a model that would fit the score of 400 examinees to explain the correlation between six divisible sub-skills of English listening comprehension and twenty listening items. R-program is used as a tool to analyze the above data. The results show that this test is not discriminatory, as the purpose of a summative assessment is not to level students into different groups. With an acceptable measurement of the construct, this research concludes that the test does not have a clear division of listening sub-skills but, on the other hand, sufficiently measures communicative ability.

Keywords: Factor analysis, Classical test theory, Listening comprehension sub-skills, Communicative validation

Introduction

English listening tests are widely used as components of English language proficiency tests in many tests, namely placement tests, achievement tests, and diagnostic tests. However, tests developed by teachers are often of poor quality (Anderson, 2005, p77). This is because the test developers do not have a clear idea of the reliability and validation of a test, or, even if they know the importance of reliability and validation, they copy and paste test items from published tests and textbooks (Coniam, 2009) due to time constraints. Moreover, being a receptive skill, listening constitutes a complex and multidimensional process that is difficult to assess. Therefore, developing a reliable and valid listening test is complex. In order to find a solution to better develop the reliability

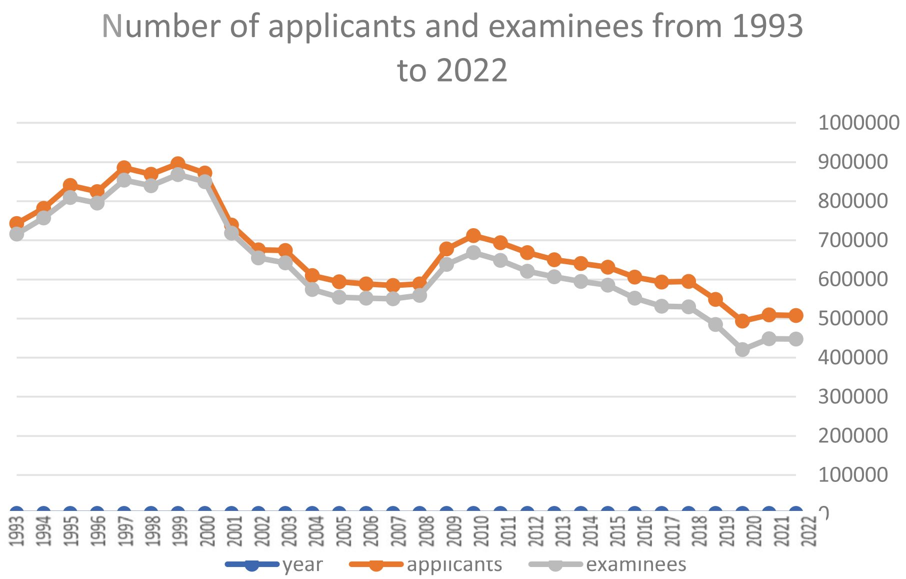


Fig. 1 Number of applicants and examines from 1993 to 2022

and validity of a listening test, the College Scholastic Ability Test (CSAT) in Korea was selected.

CSAT in Korea

CSAT, a large-scale English scholastic aptitude test with high stakes in Korea, is designed to assess English language skills and proficiency in accordance with National Curriculum Standards for Korean English learners. The Korea Institute of Curriculum and Evaluation (KICE) is a government-funded research institution that implements education evaluation programs for curriculum, teaching-learning, and providing assessments for the future of school education.

CSAT is the same as the National Entrance Exam in China and plays an important role in Korean education, being described as an opportunity for all high-school graduates to break their future (https://www.kice.re.kr/main.do?s=english). Since its inception in 1993, the CSAT has experienced rapid growth and expansion.

Figure 1 depicts the number of high school graduates who took the CSAT from 1993 to 2022.

As shown in Fig. 1, it is a large-scale test at the national level in Korea, with test takers ranging from 400,000 to 900,000 in a different year of the country. The number of applicants and examinees reached a peak around 1998–1999 but has decreased to 400,000–500,000 in recent years due to the decreasing birth rate in Korea in recent decades.

The impact of CSAT

CSAT contains 20 listening items and 25 reading items with a total score of 100. The listening part is composed of 20 multiple-choice questions from i1 to i20 which account for 44% of the total score. The CSAT is designed to test the candidate's English proficiency based on Korea's high-school curriculum for those who will study in colleges or universities in Korea. Due to its importance in standardizing high-school education and providing accurate, objective data for university admission, the CSAT is an efficient and important test for Koreans, who spend 12 years preparing for just 1 day. It does not only dictate whether the students go to university but can affect their job prospects, income, where they will live, and even their future relationships.

Focus of this study

Studies on test reliability and validation have used various approaches to remedy this situation. In the last century, researchers focused, on the one hand, on analyzing the covariance structure of two related factors, such as listening and reading (Bae & Bachman, 1998). On the other hand, there were studies concentrating on listening comprehension tests using a verbal report methodology (Buck, 1991). More recently, to improve test quality, experts analyzed the construct validation of specific types of listening tests common in their own education system (Chun, 2011; Cai, 2012). A trend from qualitative to quantitative research in construct validity uses different models and data supporting those models, among which classical test theory (CTT) and factor analysis (FA) are mentioned as the most persuasive. This paper uses both CTT and FA to investigate high school students listening comprehension skills as measured by a specific listening test. This study also attempts to determine the test's authenticity to shed light on the test's overall quality. The study provides test developers and teachers with information regarding listening comprehension sub-skills that can be measured as a result of the latest advances in research. More importantly, a set of sub-skills provides test designers with information about how well a test measures students' communicative ability and which items are of poor quality.

Methods/experimental

Research questions

Motivated by the aforementioned theoretical and practical considerations, the following research questions will be addressed:

1. How accurately does the test measure the construct of listening comprehension?
2. What sub-skills does the test measure?
3. How accurately does the test measure communication skills?

Literature review

According to Buck (1991), to divide listening skills, it is essential to know what listening comprehension is and why listeners use these skills. Although there is no generally accepted explanatory theory of listening comprehension on which to base L2 listening ability, research on listening comprehension has evolved from classroom objectives to the genuine nature of listening (Dunkel et al., 1993) and from language skills to communicative skills. Nevertheless, Glenn (1989) argues that the lack of a universal definition of listening comprehension limits research into the nature of listening comprehension.

A perspective on listening comprehension based on information processing

In language comprehension, "human working memory performs two functions: storage of information for later retrieval, and processing" (Just & Carpenter, 1987, 1992). In terms of listening comprehension, this means that listeners can remember the speaking content, retrieve what is essential, and process it by paraphrasing or inferring. Regarding information, Anderson and Lynch (1988) presented a scheme of information sources that included context knowledge, schematic knowledge, and systematic knowledge. Context knowledge and schematic knowledge refer to the context area of a text and the routines of linguistic interaction reflected in the rhetorical structure of the language (Margana, 2012). In addition, Anderson and Lynch classified systematic knowledge as phonological, morphological, syntactic, and semantic.

Language knowledge comprises both actual context and internal structures, each of which is indispensable. By grouping cognitive abilities into low (comprehension) and high (inference and evaluation) orders, Cai (2012) categorized knowledge of the language system as a lower-order source and background knowledge as a higher-order source. Situational knowledge and co-text would be in between. He also insisted that this categorization corresponded to top-down and bottom-up processing in cognitive psychology.

Listening comprehension sub-skills

Because of the limitations of human memory, cognitive skills are required to process information, especially when the task demands are high, as in a listening comprehension test. Otherwise, computation will slow down, and some results from working memory processing may be lost (Wu, 1998). Therefore, EFL listeners may hear everything but forget it easily or be incapable of constructing meaningful relationships from what they hear.

Goh (1999) also investigated students' difficulties with listening comprehension. Problems were most likely to occur during the cognitive processing phases of perception, parsing, and utilization, as well as in word recognition and attention failure during the perceptual process. Various levels of listening strategies are required to improve listening comprehension. Carroll (1972) proposed a two-stage taxonomy consisting of the capacity to comprehend linguistic information and relate it to the broader discourse. Richards (1983) developed Oakeshott-Taylor's (1977) model of listening macro- and micro-comprehension attributes by proposing a list of "micro-skills" for academic purposes, such as "the ability to identify purpose and scope of the lecture." Munby (1978) proposed a list of 250 sub-skills for the four language skills, which included recognition of intonation and discourse markers, as well as a selection of key points. However, his lengthy approach to sub-skills makes them difficult to define.

Van Dijk and Kintsch (1983) identified two primary listening strategies: local and global. One uses a local, bottom-up strategy when listening, relying on clues from phonology, vocabulary, and syntax levels. Global strategies, on the other hand, seek facts from texts and relate them to prior knowledge and general beliefs about the world, which are activated top-down. As Treiman (2001) suggested, both bottom-up and top-down processes work together to ensure accurate and rapid information processing.

Consequently, both lower- and higher-order knowledge will be required for successful comprehension. Top-down (global) and bottom-up (local) approaches are widely adopted in consideration of listeners' cognitive manners and strategies used for listening comprehension. Buck (2001) also puts forward a default listening construct, defining listening as the ability to (1) process extended samples of realistic spoken language, automatically and in real time; (2) understand the linguistic information that is unequivocally included in the text; and, (3) make whatever inferences are unambiguously implicated by the content of the passage (p. 114). Meanwhile, there was evidence for the validity of a two-factor model, related to the processing of (1) explicitly stated information and (2) implicit information (Vandergrift & Goh, 2009). White (2008) emphasized listening is an interactional process that is not only for factual information but also for social interaction. However, Wagner (2002) suggests that the implicit and explicit distinction may not be a clear-cut definition in that listeners need to understand the explicit to infer the implicit.

Compared above, Shin (2008) proposed a more inclusive listening repertoire that includes one or more of the following sets of sub-skills:

a. Understanding lexico-grammatical features (e.g., Shin, 2008)
b. Understanding explicitly stated information (e.g., Field, 2008)
c. Understanding paraphrase (e.g., Wagner, 2004)
d. Identifying intentions, attitudes, and rhetorical clues (e.g., Vandergrift, 2007)
e. Making inferences (e.g., Tsui & Fullilove, 1998)
f. Drawing conclusions (e.g., Liao, 2007; Sawaki et al., 2009)

Although numerous methods exist for decoding listening ability, researchers should combine specific tests when determining which sub-skills to assess.

Testing listening comprehension

Listening comprehension in L2 has been extensively assessed. Experts in test analysis offer various definitions of the listening comprehension construct. Weir (1993) includes direct meaning comprehension, inferred meaning comprehension, contributory meaning comprehension, and listening and writing in listening comprehension tests. Buck (2001) expands the definition to include the understanding of the sound system, local linguistic meanings, full linguistic meanings, inferred meanings, and communicative listening ability. In other words, it includes both lower-order processes, such as comprehending the local and complete linguistic meaning and sound systems, and higher-order processes, such as comprehending inferred meanings and having a communicative listening ability. As for assessment tasks, many have been mentioned, ranging from listening cloze, listening recall, and gap filling in summaries and dictation to short answers and multiple choice. In an MTMM study consisting of open-ended comprehension questions, short answer questions in listening were used (Buck, 1989). This form's test method includes time constraints, arbitrary marking, and different judging interpretations from the test developer (Buck, 1991). Cai (2012) asserts that partial dictation, similar to gap filling, measures the same construct but provides an easy-to-administer and -score valid test. While each of the above listening test formats has its advantages, some educators believe that the traditional multiple-choice format is more objective, reliable, and effective in terms of grading (Bailey, 1998; Brown & Hudson, 2002). However, it has limitations, such as being too dichotomous to measure listening ability accurately.

In conclusion, listening comprehension in a communicative setting is one of the most important language skills that listeners must acquire. To accomplish this, communicative language tests should replace traditional ones. Listening comprehension is such a process of handling information that sub-skills such as lower-level and higher-level processing may be used. According to Wang et al. (2014), higher-level processing is occupied by more communicative and challenging listening skills. In addition, listening is a process that interweaves crucial communication functions, such as social support and persuasion (Arnett & Nakagawa, 1983; Bodie, 2011). In this paper, the authenticity and validity of an English listening test will be investigated. Therefore, models must evaluate the test's validity per se, and listening skills must be clarified to determine if the test assesses the students' communicative interaction.

Participants

The 400 test participants came from sixteen provinces in Korea. They were all junior high school students going to participate in the CSAT. Their average age was 18. Their first language is Korean, and they had the same academic background.

Procedures

This national listening test included twenty multiple-choice questions. Students could listen to the listening material related to these twenty multiple-choice questions only once and select one of four provided responses. The examination was administered by trained personnel using a CSAT-like procedure. Regarding listening items, the questions and answers were written in both Korean and English. In addition, some items (i16–i20) required a sentence response based on dialogue. Some (i13) required selecting a corresponding dialogue based on a picture. The rest required selecting a correct answer according to the question. In order to analyze the responses to these questions, 400 groups of data were collected on whether they provided the correct answer by marking 0 (wrong answer) or 1 (correct answer), and then the R-program[1] CTT and factor analysis models were applied.

Assessment standard of test authenticity

Shin (2008) examined the construct validity of a web-based listening test to claim that knowledge of information structure was multidimensional, implying that the listening test may have contained the same hierarchy of ideas as the reading test (Kobayashi, 2002; Vongpumivitch, 2004). His research suggested that different constructed response formats can be indicators of information hierarchies. The summary task, for example, was more beneficial to overall understanding, and open-ended questions about major ideas can better indicate comprehension of major ideas than incomplete outline items. That is, because the sub-skills in the test were empirically divisible, it was thought necessary to

[1] Website of R-program: https://rstudio.github.io/r-manuals/

identify components in the listening construct for assessment purposes. Although sub-skill research yielded mixed results (see also Field, 2008; Flowerdew & Miller, 2006; Goh, 2010), supporting studies (e.g., Shin, 2008) indicate that listening comprehension sub-skills would generally include one or more sets of sub-skills.

To explain its communicative aspect, six sub-skills are used: understanding lexicon-grammatical features; understanding explicitly stated information; understanding paraphrasing; identifying intentions, attitudes, and rhetorical clues; making inferences; and drawing conclusions (Goh & Aryadoust, 2015).

These sub-skill sets were assessed in the study using EFA and CFA models to function in an interactive and interdependent manner in communicative listening events. Six assessors were required to number the sub-skill on each item in the listening comprehension test according to Shin's categorizations in order to determine whether the test in this study is authentic.

If more than one sub-skill was agreed upon among different accessors, the item can be cross-loaded on different sub-skills.

Establishing models to measure

Validation is now the accepted standard in language testing research (Bachman, 2000). Messick's (1989) validity framework provides comprehensive guidance covering content, substantive, structural, generalization, external, and consequential aspects. Kane (1992) defined validation as developing two types of arguments: the interpretive argument and the validity argument. Bachman and Palmer state (2010), "the Assessment Use Argument framework […] demands evidence for qualities as beneficial consequences, value-sensitive and equitable decisions, meaningful, impartial, generalizable, relevant and sufficient interpretations, and consistent assessment records." In terms of CTT, these equate to construct validity, content validity, and reliability (Cai, 2012). To conduct research on validation, researchers created theoretical models based on a criterion-based validity model, a content-based validity model, and a model that combines criterion and content aspects. However, the inadequacy of employing a single model soon became evident.

Similarly, a unified concept of validity (Michael, 2001) also has its flaws. Today, theoretical ideas are still prevalent, but researchers increasingly rely on alternative validity arguments. In this paper, different models, such as CTT models, EFA, and CFA, are used.

Alderson (1991) introduced methods of using internal correlations to evaluate construct validity, one of which is CTT, which describes the relationship between sub-tests (test items in this paper) and the entire test. It suggests that good construct validity is the correlation between the sub-tests (test items) and the entire test that goes beyond 0.70, and that between 0.30 and 0.70 is acceptable. The internal correlation only provides a general picture of how sub-tests (test items) are related to one another. A more sophisticated method is required to explain what ability (listening comprehension ability in this paper) they are designed to assess and how much they affect students' test performance. A latent variable approach, such as EFA, is required to investigate this further. EFA is useful for examining the underlying structure of test items, but it does not provide a method for specifying a factor structure beforehand. CFA accommodates this aspect and provides a method for testing the explanatory

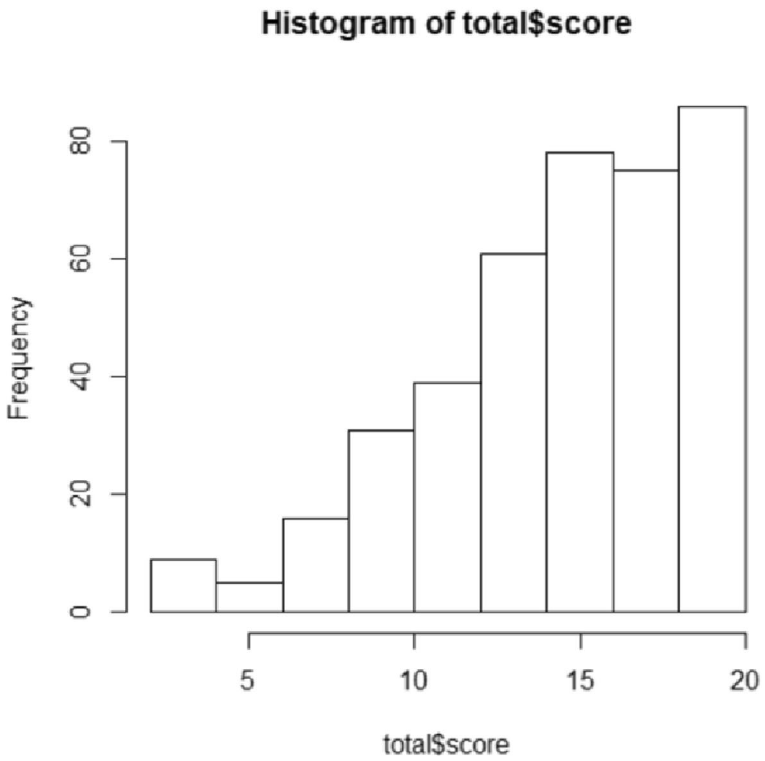


Fig. 2 Histogram of the total score and number of students

power of prior models based on theory-derived hypotheses (Bae & Bachman, 1998). In this paper, EFA is used to determine the optimal factor numbers, and CFA is used to determine which sub-skills of listening comprehension the listening test assesses and which factor model best predicts students' performance on different listening skills.

After unsuccessfully attempting a 4-, 5-, or 6-factor analysis, all test items fell into two separate categories. Originally, six sub-skills were applied (Shin, 2008) to analyze whether there is a correlation between the items, but this failed. The same result was given with 5 and 4 factors by combining the paraphrasing factor (c) with the intention factor (d) into five factors, as well as by conflating inference (e) into the former combination (c, d) to make four factors. Finally, the 2-factor model yielded an acceptable fit. Consequently, all items were divided into the two hypothesized listening levels: lower-level processing, using items that demand clearly stated information (b), and higher-level processing, using items that require inferences based on text information (a, c, d, e, f). In addition, the exploratory factor analysis result indicates that the 2-factor model is the optimal model.

Data analysis

In this section, data analyses are addressed in four aspects: classic test theory (CTT), item characteristic curve (ICC), exploratory factor analysis (EFA), and confirmative factor analysis (CFA).

CTT analysis

The preceding histogram illustrates the general trend of students' performance on the twenty listening questions. As illustrated in Fig. 2, the total score was twenty. Here, frequency represents the number of students. The listening test was generally too easy for students, as most achieved a score of 12 or higher out of a possible 20. The median score on this examination was 16, while the mean score was 14.88. In addition, the minimum score was 2 points. The highest possible score was 20. This distribution was not expected, indicating that the higher the score, the greater the number of students, except for those scoring 5 or 6 and 16 or 17 points.

The CTT data were operationalized so as to examine the difficulty and discrimination of the twenty items in greater detail. CTT is a traditional quantitative approach to testing the reliability and validity of a scale from its items (Cappelleri et al., 2014). Discrimination is the degree to which an item distinguishes between students with good and poor listening skills. In addition, difficulty points out the likelihood of having correct answers.

As shown in Table 2 in Appendix, the discrimination indexes regarding i1 (0.35), i3 (0.45), i4 (0.47), i8 (0.40), i12 (0.33), i13 (0.38), i16 (0.39), i19 (0.38), and i20 (0.33) are the lowest, showing that these items have low discriminatory power and cannot effectively distinguish between good and poor students. Similarly, i7 (0.51), i14 (0.49), and i18 (0.55), with relatively discriminatory results, can partially reflect students' listening ability. In contrast to the item's value, the item difficulty is the opposite of its value. If an item has a high mean score, it must be easy for students. Items i1, i12, i13, and i16, though too easy, can also distinguish to some extent, between students' abilities. Meanwhile, i3 (0.93), i4 (0.91), i8 (0.92), i11 (0.86), i15 (0.93), and i18 (0.88) are adequate items with strong discriminatory power. In particular, attentions should be paid to items i19 (0.42) and i20 (0.38) for which students scored low, i.e., difficult items whose discriminatory powers remain lower, instead. This indicated two possibilities: either some poor students accidentally got the right answer or the items were not very well designed and needed further moderating (Thorndike, 1976; Gui, 1986). Such a parameter also shows the value of Cronbach's alpha, which demonstrates the reliability of a test. A general rule of thumb for this value should range between 0.30 and 0.70. In this case, the parameter is 0.84, so the correlation between items is relatively high. In other words, high internal consistency ensures that almost all test items measure the same construct.

ICC The following figure explains the characteristics of each item based on the test score and item mean. As we know, the ICC is a mathematical representation of the IRT pattern that represents the likelihood of a test taker's correct answer (i.e., success rate) based on the test taker's ability parameters and item characteristic parameters obtained from the test. The corresponding item parameters for the same ICC are unique. The following are the ICC of typical logistic models (Fig. 3):

The ICCs shown above indicate the following two points: As we can see, first and foremost, the slope at the inflection point of the characteristic curve is the maximum value of the slope, which indicates the degree of discrimination (Hambleton et al., 1991).

The higher the value, the more the discrimination against the test taker. Take, for example, i3, where the slope is flat, particularly in the score range of 12 to 16, indicating that the item has a low discrimination degree.

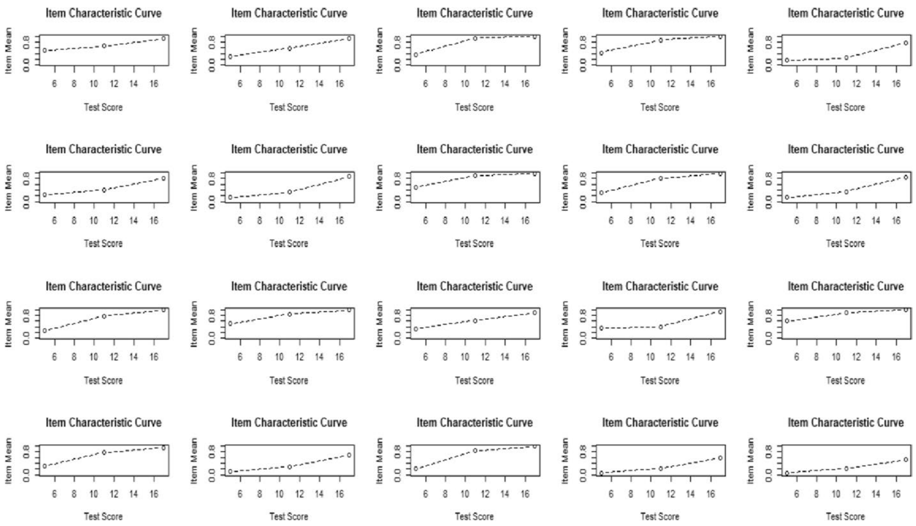


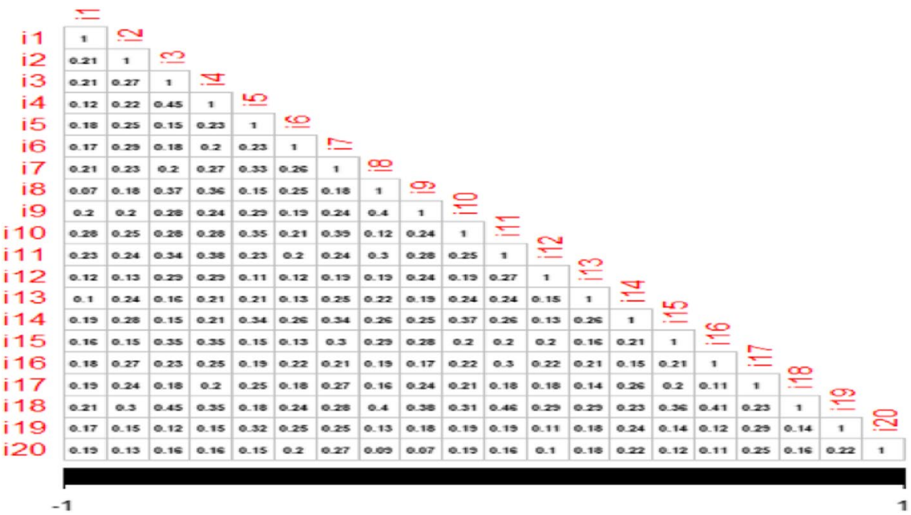
Fig. 3 Graphics depicting the mean of twenty items

Next, corresponding to the steepest point on the characteristic, the curve in the graphics indicates the difficulty of the test item (Baker & Kim, 2004). As the model is a three-parameter model, the intercept of the ICC represents the guessing parameter of the test item. The larger the value, the easier it is to guess, regardless of the ability of the test taker. For instance, the difficulty of i3 is lower than that of i2. The ICC of IRT clearly indicates the ability of the subject's relationship to items. It also indicates a certain ability and the probability of a certain item being answered correctly. Therefore, as long as the ability values of the test takers are known, the probability that they may answer a certain item correctly can be predicted.

The same conclusion can be drawn from these graphics. I3, i4, i8, i9, i12, and i15 are uncomplicated items. Conversely, i2, i5, i6, i7, i10, i13, i14, and i17 are relatively adequate items because students with a low item mean receive lower marks and students with good skills receive relatively higher marks. The preceding graphs do not indicate an unacceptable result because the probability of correct answers increases, and so does the test score, making the test relatively reliable.

EFA In this part, a more nuanced method is used to explain which items of listening comprehension skills have loadings and to what extent they are influenced by the factor on which they have loadings. A latent variable approach, such as EFA, is required to investigate this further. First, parallel analysis and eigenvalue were used to determine the optimal number of factors. Then, CFA will determine whether all items strongly correlate with the factors. In EFA and CFA, these items are referred to as latent variables, and their relationship to the factor (sub-skill) is referred to as loadings.

Figure 4 above is a paralot chart depicting the relationship between items. Correlations that are too strong or too weak imply that the items assess the same sub-skill of listening or that they assess irrelevant skills. So the rule suggests that a correlation between 0.25 and 0.8 is acceptable. In Fig. 4, i19 and i20 have obtained relatively low correlation indices, as have i6 and i17. On the other hand, i3, i4, i7, i8, i9, i10, i11, and



	i1	i2	i3	i4	i5	i6	i7	i8	i9	i10	i11	i12	i13	i14	i15	i16	i17	i18	i19	i20
i1	1																			
i2	0.21	1																		
i3	0.21	0.27	1																	
i4	0.12	0.22	0.45	1																
i5	0.18	0.25	0.15	0.23	1															
i6	0.17	0.29	0.18	0.2	0.23	1														
i7	0.21	0.23	0.2	0.27	0.33	0.26	1													
i8	0.07	0.18	0.37	0.36	0.15	0.25	0.18	1												
i9	0.2	0.2	0.28	0.24	0.29	0.19	0.24	0.4	1											
i10	0.28	0.25	0.28	0.28	0.35	0.21	0.39	0.12	0.24	1										
i11	0.23	0.24	0.34	0.38	0.23	0.2	0.24	0.3	0.28	0.25	1									
i12	0.12	0.13	0.29	0.29	0.11	0.12	0.19	0.19	0.24	0.19	0.27	1								
i13	0.1	0.24	0.16	0.21	0.21	0.13	0.25	0.22	0.19	0.24	0.24	0.15	1							
i14	0.19	0.28	0.15	0.21	0.34	0.26	0.34	0.26	0.25	0.37	0.26	0.13	0.26	1						
i15	0.16	0.15	0.35	0.35	0.15	0.13	0.3	0.29	0.28	0.2	0.2	0.2	0.16	0.21	1					
i16	0.18	0.27	0.23	0.25	0.19	0.22	0.21	0.19	0.17	0.22	0.3	0.22	0.21	0.15	0.21	1				
i17	0.19	0.24	0.18	0.2	0.25	0.18	0.27	0.16	0.24	0.21	0.18	0.18	0.14	0.26	0.2	0.11	1			
i18	0.21	0.3	0.45	0.35	0.18	0.24	0.28	0.4	0.38	0.31	0.46	0.29	0.29	0.23	0.36	0.41	0.23	1		
i19	0.17	0.15	0.12	0.15	0.32	0.25	0.25	0.13	0.18	0.19	0.19	0.11	0.18	0.24	0.14	0.12	0.29	0.14	1	
i20	0.19	0.13	0.16	0.16	0.15	0.2	0.27	0.09	0.07	0.19	0.16	0.1	0.18	0.22	0.12	0.11	0.25	0.16	0.22	1

Fig. 4 Correlation of different items

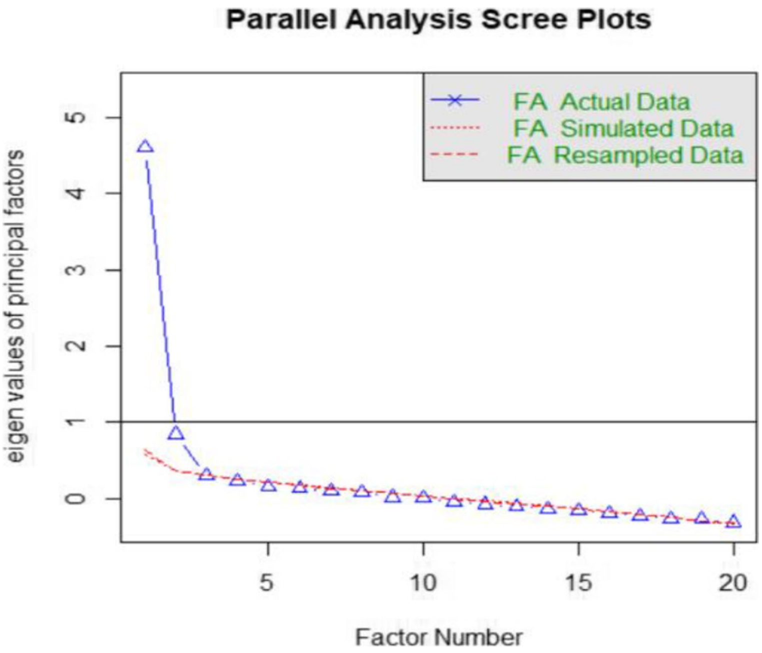


Fig. 5 Parallel analysis

i18 have correlation indices around 0.5 to 0.6, suggesting that they are quite relevant items but focused on testing dissimilar abilities.

Figure 5 shows a parallel analysis. The scree plot suggests that there should be two optimal factors for this listening test, as two eigenvalues are exceptionally high. Therefore, a two-factor model was tested in the confirmatory factor analysis section.

CFA In confirmatory factor analysis, it is crucial, first, to define the factors based on the sub-skill categorization and, second, to group 20 items into two distinct sub-skills

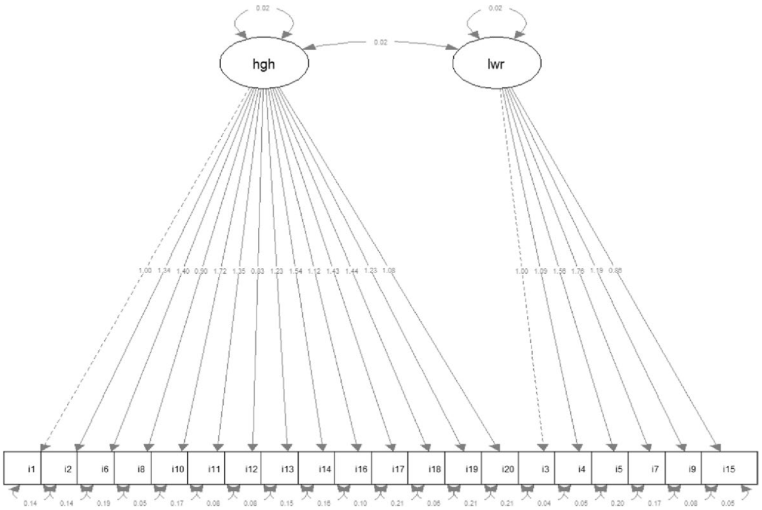


Fig. 6 semPlot of bifactor analysis based on EFA models (hgh, higher level; lwr, lower level)

to test the variables. As mentioned above, an initial attempt was made to analyze the results using Shin's (2008) six sub-skills, with no result. Combining closely related sub-skills yields the same outcome with five or four factors. In conclusion, the 2-factor model provides an acceptable fit. Based on the results, it was decided to classify all items into two hypothetical levels of listening: lower-level processing, using items that required clearly stated information (b), and higher-level processing, using items that required inferences based on textual information (a, c, d, e, f). The following data are provided for further consideration.

As seen in Fig. 6, items 1, 2, 6, 8, 10, 11, 12, 13, 14, 16, 17, 18, 19, and 20 are in the same group in terms of higher-level processing skills, composed of several subcategories, namely, understanding lexicon and grammar features, understanding paraphrases, identifying indications of intent, drawing inferences, and drawing conclusions. Meanwhile, items 3, 4, 5, 7, 9, and 15 were defined as lower level in terms of processing skills. They belong to the category of "explicitly stated information." In the last column of this figure, Std.all represents all the standard variable values. It indicates whether the factor explains the items. The better the item is explained, the greater its value. Except for i1, i19, and i20, the remaining variables are all greater than 0.4. This implies that items are highly relevant to the hypothesized sub-skills. Comparatively speaking, i8, i10, i11, and i18 are more closely associated with higher-level processing skills than i3, i4, i7, i9, and even i15. TFI, CFI, and RMSEA can be relied upon collectively for model analysis. This model's TFI is 0.839, CFI is 0.856, and RMSEA is 0.056, all of which are acceptable. However, the model has flaws because the covariance between the two skills is greater than.9. It suggests that, on this listening test, these two skills are too intertwined. In addition, it indicates that the test did not clearly target distinct listening sub-skills. Although the two skills developed from the EFA model are optimal as a result of the test from the R program, the listening test remains problematic in testing a set of clear-cut sub-skills.

Results and discussions

The section that follows will go over the validity and reliability of the listening test in terms of communicative skills. Several details will be examined to see if they explain the same results as the models shown above.

Validity and reliability

From the perspective of the models discussed above, the overall validation of this listening test was satisfactory. The reliability of the CTT and the validity of the CFA offer acceptable parameters. However, the difficulty and correlation covariance does not yield a satisfactory outcome. In other words, the test is too easy (total $M = 0.7$) to provide clear discrimination. Moreover, there is no clear division of the sub-skills assessed, as the covariance data between sub-skills A and B (higher- and lower-level processing) are relatively high. This result suggests that insufficient discrimination exists regarding the various sub-skills of listening comprehension. This is partially attributable to the fact that students do not use listening skills because they are not taught how to do so in class. The items are too simple to diagnose accurately the students' listening difficulties. CTT also determined that i3, i4, i8, i9, i12, and i15 are too simple items. As four of the six simplest items belonging to lower-level processing are i3, i4, i9, and i15, it can be concluded that students perform better on items requiring lower-level processing skills. Students are more likely to make mistakes when processing higher-level items such as i10, i14, i16, i19, and i20. If the test is intended to be authentic, it can be concluded that more communicative and challenging higher-level items should be included.

Test authenticity

Wang et al. (2014) pointed out that authenticity is the major characteristic that differentiates communicative language tests from traditional language tests. In Leung and Lewkowicz's (2006) research, authenticity can be explained "in terms of the extent to which a test or assessment task relates to the context in which it would be performed in real life" (p. 214). Hence, if a test or assessment reflects real-life English situations, it is communicative and authentic.

Shin (2008) researched the underlying structure of a web-based listening test and showed that a higher order could best capture the complexity of the underlying structure of the listening test. The traits were made up of the abilities to identify overarching main ideas, major ideas, and supports. Sawaki et al. (2009, pp. 200–201) claimed that the underlying test structure could be elaborated by three main subskills: understanding "general and specific information," "text structure and speaker intention," and "connecting ideas."

To further explore whether the test really is authentic, it seems advisable to use the following sets of subskills featured in recent discussions to explain its communicative aspect: more communicative skills should be included besides comprehending paraphrases, such as identifying intentions and attitudes, making references, and understanding grammar and lexical features such as stress or emphasis (Wang et al., 2014).

Table 1 Sub-skill distribution on each item

Item	Assessor1	Assessor2	Assessor3	Assessor4	Assessor5	Assessor6	Cross loaded	Only one
i1	1	6	1	1	1	2		1
i2	2	5	5	5	5	6		5
i3	2	6	4	6	2	4	2, 4, 6	
i4	2	5	2	1	3	4		2
i5	1	2	2	1	5	2		2
i6	2	4	6	6	2	6		6
i7	2	2	2	1	6	3		2
i8	3	5	2	2	4	4	2, 4	
i9	2	2	2	1	3	2		2
i10	2	5	5	1	5	5		5
i11		3	2	2	6	6		6
i12	4	4	4	4	4	1		4
i13	2	4	3	4	3	5	3, 4	
i14	1	6	6	5	1	5	1, 5, 6	
i15	2	3	2	1	2	2		2
i16	5	5	5	4	4	5		5
i17	4	5	5	4	4	5	4, 5	
i18	2	5	5	4	4	5		6
i19	4	5	5	4	4	5	4, 5	
i20	4	5	5	4	5	5		5

Wang, Zuo, and Liu did not include "drawing conclusions" on their list of communicative skills, as this was not included as a sub-skill in their research.

In addition, according to McNamara (2012), summarizing is also a communicative skill. In contrast, comprehension of explicit information is associated with noncommunicative behavior. Here, an attempt is made to explain the communicative perspective of this test by categorizing the following items.

As shown in Table 1, the sub-skills are analyzed according to Shin's definition (sub-skills 1 to 6 correspond to Shin's sub-skills a to f):

a. Understanding lexico-grammatical features (e.g., Shin, 2008) = 1
b. Understanding explicitly stated information (e.g., Field, 2008) = 2
c. Understanding paraphrase (e.g., Wagner, 2004) = 3
d. Identifying intentions, attitudes, and rhetorical clues (e.g., Vandergrift, 2007) = 4
e. Making inferences (e.g., Tsui & Fullilove, 1998) = 5
f. Drawing conclusions (e.g., Liao, 2007; Sawaki et al., 2009) = 6

Six assessors are six Ph.D. students majoring in English education. They were required to number the sub-skill on each item in the listening comprehension test according to Shin's categorizations. The sub-skill most students numbered was the only sub-skill on each item. If there were more than one sub-skill agreed upon among different accessors, the item can be cross-loaded on different subskills. The items numbered 3, 4, and 5 are less challenging in applying communicative ability since they only require students' understanding of explicit information. In some items, such as i3, i7, i11, i13, i14, i17, and i19, there may be more than one skill involved. In that case, the most influential or main skill should be considered. For example, i3 assesses three sub-skills, with understanding explicitly stated information being the most prominent.

As can be seen, most questions concern the comprehension of the explicitly stated information. It is a relatively straightforward aspect of listening comprehension. Students need only to comprehend the highlighted portions of the paragraph. Since this comprises nearly a third of the items, there is no doubt that the test is straightforward: I5 requests that students select the total rental fee and rental date. Although the texts are based on everyday topics, such as preparing for a conference, this item's primary focus should be numbers, which can be easily gleaned from the conversation. Therefore, item i5 is deemed irrelevant to communicative interaction, the same as i3, i4, i7, i9, and i15.

Next, the questions testing the ability to make inferences come second. Making inferences uses learners' prior knowledge to specify a text's true meaning, so it is related to one's communicative ability. Take i20 as another example. The text addresses a situation that any student may encounter. Susan was preoccupied with noises from upstairs at midnight while she had much work to do. Therefore, she contacted the apartment manager. The text then raised the question of what Susan would likely say to him. Students can relate this to their own experiences and determine the correct response. Everyday items like i2, i10, i16, and i19 refer to the following situations: school interviews, consulting a moving company, making suggestions to friends, and expressing one's willingness. The ability required to respond correctly to these questions is considered helpful in communicative interaction.

Conclusion drawing is a third frequent subskill. It is a higher-level processing skill that requires students to summarize information. To arrive at the correct answer, students must know both the essential details and the central theme of the listening material. Similar to i11, the question is which credit card a girl chooses. A figure is presented briefly, describing the functions of the playing cards, so that students can directly determine what bank card the girl needs. The dialogue involves nearly every banker and his or her client, and listeners can learn how to open a bank account. I6 and i18 are comparable, so this is indeed a communicative sub-skill.

The rest of the sub-skills, such as understanding lexical features, identifying intentions and attitudes, and understanding paraphrases, are equally frequent. Students need these essential communicative skills to express their ideas, demonstrate their attitudes, and explain something difficult to comprehend. Take i1 as an example. A girl is deciding what kind of flowerpot she wants by giving critical information like "a pot I can put on the windowsill," "that round one looks good," and "the one with a rim," so the question "what kind of flowerpot the girl wants to buy" can be easily answered. If crucial information is received or provided, the conversation will be more effective. Even if the entire idea is not understood, the conversation can still serve its purpose. Therefore, it is an essential communication skill. Item 12 is an illustration of the communication skill of identifying the intentions of others. Based on the monologue in which a man proposes giving away books and letting them travel the world, the text indicates that he proposes that listeners engage in the activity of book exchanging. Although it is not specified in the text, the question "what purpose does the man have" can be easily answered. This skill allows people to comprehend the meaning behind words, bridging the gap between them. Lastly, item 13 is presented in a unique manner. It provides a picture and four possibly related dialogues from which listeners must choose. The image depicts a mother requesting a picture of her daughter with a cartoon character. Students must recognize both the meaning of the image and the meaning of the dialogue to identify the correct answer. Gruba (2004) believed the integration of audio and visual information enriched the modality of input in listening comprehension, which may help students to understand more. On the other hand, it assesses students' ability to paraphrase pictures and words. Such a skill is used when explaining a scene to someone. The three sub-skills discussed in this section are all communicative skills applied in everyday life.

Based on the above analysis, it can be concluded that the test adequately measures communicative skills when referring to content, ranging from buying goods and opening bank accounts to overcoming problems in school life and giving suggestions to friends. However, authenticity should also be reflected in other aspects, such as the length of the listening material or the questions' format. Despite the absence of a clear division of listening sub-skills in the test paper, both the text and the questions are communication oriented.

The reason why the test is low discriminant is interpreted in the following three aspects: First, there are more easy items in the test that will lower the discrimination level. For example, if the item is too easy andalmost all the students got correct answers, such an item cannot distinguish high-ability students from low-ability ones. Next, though there are some hard items, they happened to be answered correctly by low-ability students by guessing. This would be most likely to weaken the item's discrimination power. Finally, items that are either correctly answered or incorrectly answered by all the test takers will also lower the discrimination level.

Conclusion

This document reports on a communicative validation study of a Korean listening comprehension test. Due to the importance of this test, it is important to ensure its reliability and validity. According to the CTT and CFA models, both reliability and validity are acceptable. However, the test is deemed to be an uncomplicated test for students who are going to attend CSAT. In addition, if viewed as a summative examination, this test should assess potential sub-skills. The sub-skills and divisibility of the listening construct into separate units have been discussed, only to discover that students' performance does not appear to be influenced by sub-skills. Maybe the items are too simple to assess accurately students' actual sub-skills or that students do not use listening skills because they are not taught in class. If the former is true, then difficult items should be added to the test to make it more discriminatory, as the proportion of difficult items in a test is approximately as follows, according to CTT: very difficult (VD) 5%, difficult (D) 15%, intermediate (I) 60%, easy (E) 15%, and very easy (VE) 5%. The number of simple items, such as i3, i4, i8, i9, and i12, should be reduced to achieve a balance. There should be more difficult items such as i19 and i20.If the latter is true, teachers and evaluation boards should take into account the washback of exams. In the end, instruction and assessment go hand in hand. Field (2008) suggests that successful listening comprehension requires attention to certain aspects of listening ability at lower-level and higher-level processing. The communicative validity of the test here examined was endorsed based on various listening sub-skills at both the item and text levels. On a content level, the test is, therefore, communicative and authentic, except for a few items that assess students' lower processing levels. To add difficulty to the test, test developers could focus on processes for various contexts, such as an academic context rather than a conversational context. In addition, it could be helpful to develop a variety of item formats, such as short answers and gap-filling tasks. These varieties could be added in different sections of the listening part. As part 1 is about conversation, multiple choices could be given. Part 2, for instance, is about an academic lecture for which multiple choices corresponding to a passage could be provided. Part 3 is about news and requires short answers. Thus, listening sub-skills may become more discriminatory. This study tried to look into validating a listening test closely related to CSAT. Modifications to these test items are warranted to equip students with additional listening sub-skills and enhance their communicative listening skills.

A validation study is essential for a widespread test because it can guarantee a more qualified assessment, both in the test's content and form. This could improve by focusing on test discrimination and the division of listening sub-skills.

However, two limitations remain. The first is that all skills each item assesses are based on model explanations without designers' confirmation. For further consideration, test designers' opinions on such explanations should be included to see the whole picture. The second one lies in the inadequate data support from the original test. Due to confidentiality terms, test items need to be kept confidential, as requested by the test paper supplier. In future studies, test items will need to be analyzed in detail to see if they fit the CFA model. The limitations of this study are partly caused by the return of the researcher from Korea, in that supplementary data cannot be guaranteed without official permission.

Appendix

Table 2 Table of item means and discrimination from EFA

	Item mean	Discrimination
i1	0.8000	0.3514847
i2	0.7675	0.4359167
i3	0.9300	0.4589655
i4	0.9175	0.4767874
i5	0.5550	0.4584574
i6	0.6300	0.4084156
i7	0.6375	0.5183886
i8	0.9250	0.4097258
i9	0.8725	0.4491314
i10	0.6350	0.4993749
i11	0.8600	0.4878493
i12	0.8975	0.3357101
i13	0.7575	0.3841545
i14	0.7075	0.4916279
i15	0.9300	0.4021931
i16	0.8450	0.3896152

	Item mean	**Discrimination**
i17	0.5125	0.4149311
i18	0.8850	0.5551910
i19	0.4275	0.3811711
i20	0.3850	0.3313076
>		
>ItemAnalysis$alpha 0.8419395		

Abbreviations

CFA	Confirmative factor analysis
CFI	Comparative fit index
CTT	Classical test theory
CSAT	College Scholastic Ability Test (Korean)
EFA	Exploratory factor analysis
RMSEA	Root-mean-square error of approximation
TFI	Tinnitus Functional Index

Acknowledgements

Not applicable.

Authors' contributions

CZ contributed to the conception of the paper, drafted the manuscript, and analyzed the data. XX contributed to the design of the work, and the interpretation of data, and was a major contributor to writing the manuscript. YZ substantively revised the paper. All authors read and approved the final manuscript. All authors have agreed both to be personally accountable for the author's own contributions and to ensure that questions related to the accuracy or integrity of any part of the work, even ones in which the author was not personally involved, are appropriately investigated, and resolved, and the resolution documented in the literature.

Authors' information

Ci Zhang, majoring in English education, gained her Ph.D. in Chonnam National University in Korea. Also, she is a lecturer at Foreign Language Studies Department, Wenzhou University. Her research foci are Instructed Second Language Acquisition, Teaching Methods, and Testing. She has published articles in domestic and oversea journals such as KCI, covering fields on grammar acquisition, readability, and POA teaching method.

Xiaoshu Xu (corresponding author, Ph.D.) is an Associate Professor at Wenzhou University and the Ph.D. Supervisor at Stamford International University in Thailand. She is the editor in chief of the Journal of Educational Technology and Innovation. She has published papers on higher education, personal learning environments, and teacher development in SSCI journals.

Yunfeng Zhang (Ph.D.), associate professor, is the Director of the Centre for Portuguese Studies of Macao Polytechnic University. He received Ph.D. in Linguistics at the University of Coimbra, Portugal. He has published books and journal papers in the fields of translation and machine translation, language testing, etc.

Funding

This research was financially supported by China's National Social Science Later-stage Project (21FJKB013).

Availability of data and materials

The datasets generated and/or analyzed during the current study are not publicly available due to ethical issues but are available from the corresponding author on reasonable request.

Declarations

Competing interests

The authors declare that they have no competing interests.

Received: 9 January 2023 Accepted: 13 May 2023

Published online: 25 May 2023

References

Alderson, C. (1991). Language testing in the 1990s: How far have we come? How much further have we to go? In S. Anivan (Ed.), *Current developments in language testing* (pp. 1–26). SEAMEO Regional Language.

Anderson, J. C. (2005). *Diagnosing foreign language proficiency: The interface between learning and assessment*. Continuum.

Anderson, A., & Lynch, T. (1988). *Listening*. Oxford University Press.

Arnett, R. C., & Nakagawa, G. (1983). The assumptive roots of empathic listening: A critique. *Communication Education*, *32*, 368–378.

Bachman, L. (2000). Modern language testing at the turn of the century: Assuring that what we count counts. *Language Testing*, *17*, 1–42.

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice*. Oxford University Press.

Bae, F., & Bachman, L. F. (1998). A latent variable approach to listening and reading: Testing factorial invariance across two groups of children in the Korean/English two-way immersion program. *Language Testing*, *15*(3), 380–414.

Bailey, K. M. (1998). *Learning about language assessment: Dilemmas, decisions, and directions*. Heinle & Heinle.

Baker, F. B., & Kim, S. H. (Eds.). (2004). *Item response theory: Parameter estimation techniques*. CRC press.

Bodie, G. D. (2011). The Active-Empathic Listening Scale (AELS): Conceptualization and evidence of validity within the interpersonal domain. *Communication Quarterly*, *59*, 277–295.

Brown, J. D., & Hudson, T. (2002). *Criterion-referenced language testing*. Cambridge University Press.

Buck, G. (1989). Written tests of pronunciation: Do they work? *ELT Journal*, *43*(2), 50–56.

Buck, G. (1991). The testing of listening comprehension: An introspective study. *Language Testing*, *8*(1), 67–91.

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

Cai, H. W. (2012). Partial dictation as a measure of EFL listening proficiency: Evidence from confirmatory factor analysis. *Language Testing*, *30*(2), 177–199.

Cappelleri, J. C., Lundy, J. J., & Hays, R. D. (2014). Overview of classical test theory and item response theory for the quantitative assessment of items in developing patient-reported outcomes measures. *Clinical Therapeutics*, *36*(5), 648–662.

Carroll, J. B. (1972). Defining language comprehensions: Some speculations. In R. O. Freedle & J. B. Carroll (Eds.), *Language comprehension and the acquisition of knowledge*. Wiley.

Chun, J. Y. (2011). The construct validation of ELI listening placement tests. *Second Language Studies*, *30*(1), 1–47.

Coniam, D. (2009). Investigating the quality of teacher-produced tests for EFL students and the impact of training in test development principles on improving test quality. *System*, *37*(2), 226–242.

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

Field, J. (2008). *Listening in the language classroom*. Cambridge University.

Flowerdew, J., & Miller, J. (2006). *Second language listening*. Cambridge University Press.

Glenn, E. (1989). A content analysis of fifty definitions of listening. *Journal of the International Listening Association*, *3*, 21–31.

Goh, C. (1999). How much do learners know about the factors that influence their listening comprehension? *Hong Kong Journal of Applied Linguistics*, *4*(1), 17–41.

Goh, C. C., & Aryadoust, V. (2015). Examining the notion of listening subskill divisibility and its implications for second language listening. *International journal of listening, 29*(3), 109-133.

Goh, C. C. M. (2010). Listening as process: Learning activities for self-appraisal and self-regulation. In N. Harwood (Ed.), *Materials in ELT: Theory and practice* (pp. 179–206). Cambridge University Press.

Gruba, P. (2004). Understanding digitized second language videotext. *Computer Assisted Language Learning*, *17*, 15–82.

Gui, Shichun. (1986). *Standardized test: Theory, principle and method*. 1st edition.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory* (Vol. 2). Sage.

Just, M. A., & Carpenter, P. A. (1987). *The psychology of reading and language comprehension*. Allyn & Bacon.

Just, M. A., & Carpenter, E. A. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review*, *99*, 122–149.

Kane, M. (1992). An argument-based approach to validity. *Psychological Bulletin*, *112*, 527–525.

Kobayashi, M. (2002). Method effects on reading comprehension test performance: Text organization and response format. *Language Testing*, *19*(2), 193–220.

Leung, C., & Lewkowicz, J. (2006). Expanding horizons and unresolved conundrums: Language testing and assessment. *TESOL Quarterly*, *40*(1), 211–234.

Liao, Y. (2007). Investigating the construct validity of the grammar and vocabulary section and the listening section of the ECCE: Lexico-grammatical ability as a predictor of L2 listening ability. *Spaan Fellow Working Papers in Second or Foreign Language Assessment*, *5*, 37–78. Ann Arbor, MI: University of Michigan English Language Institute.

Margana. (2012). Promoting schematic knowledge to English teachers of secondary school level. In *The 6^th^ international seminar in Satya Wacana Christian University*.

McNamara, C. (2012). *Organizational sustainability*. Free Management Library.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). Macmillan.

Michael, T. K. (2001). Current concerns in validity theory. *Journal of Educational Measurement*, *38*(4), 293–382.

Munby, J. (1978). *Communicative syllabus design*. Cambridge University.

Oakeshott-Taylor, A. (1977). Dictation as a test of listening comprehension. In R. Dirven (Ed.), *Hörverständnis im Fremdsprachenunterricht. Listening comprehension in foreign language teaching.* Scriptor.

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

Sawaki, Y., Kim, H.-J., & Gentile, C. (2009). Q-matrix construction: Defining the link between constructs and test items in large-scale reading and listening comprehension assessments. *Language Assessment Quarterly*, *6*, 190–209.

Shin, S. (2008). Examining the construct validity of a web-based academic listening test: An investigation of the effects of response formats. *Spaan Fellow Working Papers in Second or Foreign Language Assessment*, *6*, 95–129. Ann Arbor, MI: University of Michigan English Language Institute.

Thorndike, R. L. (1976). *Educational measurement*. 2nd edition.

Treiman, R. (2001). Linguistics and reading. In M. Aronoff & J. Rees-Miller (Eds.), *Handbook of linguistics* (pp. 664–672). Blackwell.

Tsui, A. B. M., & Fullilove, J. (1998). Bottom-up or top-down processing as a discriminator of L2 listening performance. *Applied Linguistics*, *19*, 432–451.

Van Dijk, T. A., & Kintsch, W. (1983). *Strategies of discourse comprehension*. Academic.

Vandergrift, L. (2007). Recent developments in second and foreign language listening comprehension research. *Language Teaching*, *40*, 191–210.

Vandergrift, L., & Goh, C. (2009). Teaching and testing listening comprehension. In *The handbook of language teaching* (pp. 395–411).

Vongpumivitch, V. (2004). *Measuring the knowledge of text structure in academic English as a second language (ESL)*. Unpublished doctoral dissertation. University of California at Los Angeles.

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*, 1–23. Ann Arbor, MI: University of Michigan English Language Institute.

Wagner, E. (2002). Video listening tests: A pilot study. Working papers in TESOL & applied linguistics, Teachers College, Columbia University, 2/1. Retrieved April 19, 2006 from http://www.tc.edu/tesolalwebjournal/wagner.pdf.

Wang, C., Zuo, Y., & Liu, B. (2014). Communicative validity of the new CET4 listening comprehension test in China. *Indonesia Journal of Applied Linguistics*, *4*(1), 109–121.

Weir, C. J. (1993). *Understanding and developing language tests*. Prentice Hall.

White, G. (2008). Teaching listening: Time for a change in methodology. In *Current trends in the development and teaching of the four language skills* (pp. 111–138). De Gruyter Mouton.

Wu, Y. (1998). What do tests of listening comprehension test? - A retrospection study of EFL test-takers performing a multiple-choice task. *Language Testing*, *15*(1), 21–44.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.