SHIKEN

A Journal of Language Testing and Evaluation in Japan

Volume 29 • Number 1 • April 2025

https://doi.org/10.37546/JALTSIG.TEVAL29.1

Contents

Testing and Evaluation SIG

ISSN 1881-5537

Shiken: A Journal of Language Testing and Evaluation in Japan

Volume 29 No. 1
April 2025

https://doi.org/10.37546/JALTSIG.TEVAL29.1

Article
Shiken: A Journal of Language Testing and Evaluation in Japan
2025, Vol. 29(1) 1–30
DOI: 10.37546/JALTSIG.TEVAL29.1-1
ISSN: 1881-5537

Validating the assessment of an out-of-class listening task using Rasch analysis

Thomas P. Stones
thomas.p.stones@gmail.com
Kwansei Gakuin University

Abstract

There has been increasing interest in recent years in the use of out-of-class and extensive listening and various studies have explored its efficacy. However, there has been relatively little research into the appropriate methods of assessment for such tasks. Therefore, this study aims to explore the effectiveness of the assessment methods for out-of-class listening in use at a university in Japan. The test instrument used is a post-listening multiple-choice test, completed without the use of notes. A Rasch analysis was used to investigate test validation, and the results suggest that the test is valid, but there are areas for improvement in terms of reliability and distractor functioning. Implications for the assessment of out-of-class listening are discussed based on these findings.

Keywords: out-of-class listening, Rasch measurement, assessment validation, multiple-choice tests, listening assessment

Recently, there has been a growing interest in the use of extensive listening, and numerous studies have investigated its efficacy (e.g. Brown et al., 2008; Chang & Millett, 2014; Chang & Millett, 2016; Chang et al., 2018). Analogous with its more widely researched sibling, extensive reading, extensive listening is the practice of listening while focusing purposefully on content with the intention of enjoying it or learning something (Rost, 2011). Texts could include graded texts, authentic texts, academic texts, or general pleasure listening, such as podcasts.

Extensive listening is said to facilitate L2 listening fluency, whereby listeners can understand a text with relatively little effort. This contrasts with controlled processes where considerable cognitive effort is required by the user (Schmidt, 1992). Rost (2006), specifically discussing listening fluency, notes that a sufficient degree of understanding of the input is necessary for fluency but that this requires a significant amount of consistent practice and exposure to achieve. Nation and Newton (2009) suggest that activities aimed at developing second language fluency be focused on meaning and that they largely contain language within the listeners range of control. In addition to fluency, Renandya and Jacobs (2016) note numerous potential benefits including improved oral word recognition skills, ability to deal with a high speech rate, bottom-up listening skills, familiarity with common features of spoken English, and depth of comprehension through repeated listening (see also Onoda, 2014). Stephens (2011) similarly argues that extensive listening can enhance perception of intonation patterns common to spoken English.

Literature Review

Research on the effects of extensive listening has found that it leads to significantly improved comprehension (Chang & Millett, 2014) and that it can lead to comprehension of texts faster than those used for practice (Chang et al., 2018). Swan and Walter (2017) advocate for the development of listening fluency through exposure, combined with the type of remedial decoding and parsing activities that are advocated by Field (2009) and Wilson (2003). They suggest that class time may be spent introducing and setting up out-of-class listening activities or introducing the sort of resources that are suggested by Renandya and Farrell (2011) and Waring (2008), which include a range of easily accessible websites with

listening materials. Indeed, in Zhang's study (2007, cited in Renandya & Farrell, 2011), an extensive listening group who received 42 hours of listening instruction outperformed a group whose instruction focused on listening strategy training.

The above has focused on extensive listening and the associated research. However, many of the above researchers have adopted a fairly open view of the term 'extensive listening', with Renandya and Farrell (2011) even suggesting that it could be any form of listening that is enjoyable. Unfortunately, the reality of many EFL contexts is that learners may not be intrinsically motivated to listen for its own sake. In such cases, it can be useful to assign grades to submitted reports or otherwise base assessments around follow-up tasks (Nation and Waring, 2020). Qualitative assessment types, such as reports and summaries, are suitable to general monitoring of extensive reading or listening, but discrete item question-types, such as multiple-choice tests, are a popular choice due to their efficiency, accuracy and reliability across raters (Fulcher & Davidson, 2007). Such assessments, however, violate some of the principles of extensive listening, which are that the learners choose their own texts and are listening for pleasure, perhaps without even the need to prove comprehension (Field, 2009). That said, it should also be noted that multiple studies using the term extensive listening feature post-listening comprehension tests and often specify the texts to be listened to (e.g., Chang & Millett, 2014; Chang et al., 2018; Chang & Millett, 2016). In cases when learners are compelled to listen to specific texts, the term- out-of-class listening may be better, because even if the teacher makes substantial effort to select texts that seem to be relevant and interesting for learners, they will not be so for all.

As noted above, many learners in compulsory English classes, especially in Japan where this study takes place, are not motivated intrinsically to study but are often motivated to get good course grades (Nishida, 2013; Yamamoto & Ohba, 2018). As such, attaching a course grade to an out-of-class task can be a motivating factor. The intent in these situations should be for a positive 'study impact' (Fulcher & Davidson, 2007, p. 73) and to influence the participants behaviour in a positive way, aligning with Messick's original definition of Washback (1996), which is to encourage learners to do what they would otherwise not do as a consequence of the test but that is good for them. Alderson and Wall (1993) similarly note that tests can encourage out-of-class behaviour and increased effort, albeit mediated through extrinsic motivation. In the case of listening, the more learners listen out of class, the better their listening will become and should lead to outcomes similar to those accrued through 'pure' extensive listening.

Assessing out-of-class listening

The effective assessment of listening skills is a complex and challenging endeavor due to the complexity of the processes involved (Buck, 2001). This is because listening is an internal mental process and is not, therefore, directly measurable, thus requiring an additional task, such as a spoken summary, gap-fill task, or multiple choice-question to obtain a performance indicator. This need for an intermediary skill can introduce some construct irrelevant variance because proficiency in the additional skill can blur the measurement of the construct being targeted (Bachman, 1990; Buck, 2001). However, Rukthong and Brunfaut (2020) note that in listening to writing tasks, for example, the quality of the productive element is mediated by ability in listening comprehension. Furthermore, in post-listening assessment formats, memory can be a further serious source of construct irrelevant variance (Buck, 2001). Chang (2006) investigated such variance with reading by comparing a translation, written while reading, and a post-task summary. There was considerably more evidence of comprehension in the translation task than the recall task, highlighting the role of memory in demonstrating post-task comprehension and showing that comprehension does not equal remembering. To mitigate this, Alderson (2000) suggests assessment tasks be completed as soon as possible after listening. Despite this, Rukthong and Brunfaut (2020) found that better comprehension did lead to better text recall, a finding supported by Sawaki et al. (2013).

This challenge of text-recall is exacerbated for learners at lower levels of proficiency because when processing capacity is not fluent, working memory can be overloaded by the cognitive challenge of second language listening comprehension, leading to some content being forgotten (Just & Carpenter, 1992). Indeed, Kim et al. (2022) found that better working memory correlated with better listening comprehension when using post-listening multiple-choice tests, but that the biggest influencing factor was second language linguistic knowledge. Working memory relates mainly to only on-task processing, but other studies have found that available long-term memory also mediates comprehension as learners utilise background knowledge in comprehension and need to parse newly learned content to long term memory (Was & Woltz, 2007).

Furthermore, it has been shown that knowledge of the form of a post-task assessment will influence on-task learner performance and the cognitive strategies used. For example, Joh and Schallert (2014) found that when learners knew they would be asked to recall a text, they made conscious efforts to try to memorize the content and made predictions about the type of questions that may appear. Similarly, Rukthong (2021) found that the cognitive processes of word identification, parsing, pragmatic processing and local and global semantic processing were used while taking notes for MC tests but were utilized more often for listening to summarize tasks. The same study found that choice elimination, lexical matching, and guessing were used as test-wise strategies with multiple choice tests, which can also serve as construct irrelevant variance – those that are more test savvy may score better than others with equivalent listening skills. Other test-wise strategies include combining world knowledge and prominent language to make guesses (Wu, 1998), but this can also lead to incorrect guesses in spite of good comprehension (Wu, 1998). Therefore, in a post-task assessment model, both short- and long-term memory and test-wise strategy use will play a role in on-task and in-test performance, but the research suggests that overall listening comprehension ability should be the predominant trait that predicts good test performance.

Rasch Measurement

The 'Rasch Model' is a probabilistic statistical model that compares empirical test data to an idealized model and has become an established and reliable tool within language testing (McNamara & Knoch, 2012). Originally developed by Georg Rasch in 1960 (Rasch, 1960), it is a tool that can be used to investigate the effective functioning of an assessment by examining the relationship between the difficulty of items (i.e. test questions) and the ability of the test takers (McNamara, 1996). It can reveal, amongst numerous other things, the relative difficulty of items for a given sample of test takers. The primary benefit of the logit scale is that it represents a consistent scale with each increment being the same at all points on the scale, thus creating a unit measurement that can be used to objectively represent ability and difficulty across contexts (Wright & Stone, 1999). Central to this is the concept of specific objectivity (Rasch, 1977) which means that, once established, item difficulties and person abilities should remain stable across different assessment contexts when examining the same latent trait or skill (Wu & Adams, 2007).

Further, test designers can use the Rasch outputs to compare the response patterns and look for any items, persons or distractors that do not appear to be behaving as predicted and can be used to diagnose various issues in the assessment, such as poorly worded items (Bond & Fox, 2015). The Rasch model, therefore, provides a powerful set of highly sensitive tools for diagnosing the quality of test items and how well they function for individuals and is an excellent tool for test development (McNamara, 1996).

Methods

Research Context

This research took place at a university in Japan within a large, coordinated English program taken by around 1,000 first- and second-year students. Learners typically have TOEFL ITP scores ranging from around 320 to 550 (Approximately A2 ~ B1 on CEFR) with most scoring between 400 – 450. The students take four English classes per week, Reading, Writing, Listening and Speaking. Each class was 90 minutes and met 14 times throughout the semester. This study focuses on the listening component of the program, which featured four topic-based units, with each unit containing 3 sessions following a fixed structure. The lessons included pre-reading and language tasks, intensive listening to lectures and informal interviews, and bottom-up tasks focusing on connected speech. Homework for the third lesson was to complete the extensive listening tasks, the focus of this study. There were also introductory and wrap-up lessons in weeks 1 and 14.

Assessment Instruments

The out-of-class listening portion of the course involved students listening to one assigned and one self-selected text from the website Spotlight English. Spotlight's website states it uses a speech rate of 90 words per minute, 'half the normal speaking speed', shorter sentences and simplified vocabulary (Spotlight, 2022). Thus, it should provide material that is comfortably comprehensible to most participants. For both texts, the students completed a worksheet which involved writing the main points, the three most interesting points and commenting on how interesting they found it. This was referred to as the Extensive Listening Homework (ELH). On coming to class, students sat a 5-question, multiple-choice test about the assigned listening, named the Extensive Listening Quiz (ELQ) and could not refer to their notes during the quiz. The 4 topics of the assigned texts were national anthems, the Asian University for Women (AUW), workplace culture and graffiti. Each ELQ included one general overview question, typically about the overall gist of the listening, followed by 4 detail questions, targeting important information and occasionally some statistics. The full set of questions can be seen in Appendix AA.

Following the quiz, learners discussed their homework notes on the self-selected listening in groups, then submitted their worksheets. The ELQ was scored objectively, then the notes were awarded a score out of 10. Each quiz was worth 5% of the total grade and each set of notes was worth 4%. These combined made a total of 36% of the total course grade. Other assessments included four 'intensive' listening quizzes, worth 10% each, where students answered open questions on an audio passage and answered some bottom-up listening tasks including dictations and phonological recognition tasks. There were also quizzes on the preparation vocabulary and reading homework, at 4% each, and the remaining 4% was for in-class participation. As such, the ELQs that are the focus here are only a part of a much larger assessment battery that made up the course score. The variety in the assessments is to assess the wide range of elements that make up listening competence but also to overcome the construct irrelevant variance that may exist in any one question type (Buck, 2001).

Motivation for the Research

The motivation for the research stemmed from discussions among the teachers surrounding how effectively the extensive listening and assessments were functioning. On the one hand, given the emphasis in extensive listening on listening for pleasure and enjoyment, some teachers raised concerns as to whether it is appropriate to give a quantitively-orientated assessment that relies on recall, a potentially confounding variable. There was a concern that asking questions about specific details was contrary to

the principles of extensive listening and that it was unnecessarily demanding to recall specific details from a 6-minute text. They argued that the notes and a follow-up discussion would be sufficient. However, others noted that it would be easy for students to cheat by copying notes, which would significantly undermine attempts to expose learners to material in English, so argued that the ELQ would act as a form of accountability. This is on the grounds that the test itself is functioning correctly and is appropriately rewarding efforts to listen outside of class. If that is the case, there would be the possibility to scale up the out-of-class listening tasks to be completed more regularly, which would improve learner outcomes with minimal extra grading because the MC tests could be graded automatically.

Therefore, this research aims to explore the efficacy of the assessment instruments in place on the course and adjust accordingly. In order to realize tangible gains in listening fluency, the volume of listening needs to be substantial, so if the use of multiple-choice questions to assess the out-of-class listening appears valid, similar tasks could be added to this or other courses, potentially one a week, as a means to monitor out of class listening in a time-efficient yet valid manner, and significantly increase the amount of time the learners spend listening.

This project also locates within the practitioner research paradigm as exploratory practice and an attempt to better understand issues within the teaching context (Allwright, 2005). It is also an effort by myself to develop as a 'thinking professional' (Burns, 2010, p. 6) and drive positive change and professional engagement and development (Burns, 2010). I am also similar to the characterisation offered by McNamara and Knoch (2012) of a researcher who is entering the world of language testing with training in language teaching rather than in statistics and psychometrics and exploring the use of Rasch in my context, as per Smiley (2015). This research, therefore, is intended to develop this practitioner researcher's skills to supplement the normal tasks of the practising teacher (Burns, 1999).

In the current assessment design, the intended focus of the assessment was listening comprehension, but due to the fact that the assessment was delivered post-task, potentially sometime after listening, there was a substantial memory component involved, so learners also needed to make efforts to memorize the content; however, this likely drove time on task and increased focus, which has been shown to drive acquisition (Skehan, 1998). Furthermore, as noted above, learners with better comprehension skills should fare better on assessments that involve recall despite memory being an intervening variable. As such, the latent trait of the assessment was recall of content comprehended through auditory processing, with the underlying premise that listening comprehension was the primary trait being targeted. Therefore, the overall purpose was to explore the validity of the ELQs as an assessment of the recall of an out-of-class listening task.

Research Questions

1. Are the test items able to distinguish sufficiently between different levels of performance among test takers?
2. Is there a sufficient range of difficulty in the items?
3. Is the reliability and fit of the items sufficient that test scores are generalizable to another similar cohort?
4. Are the distractors functioning in an acceptable way?
5. Is the reliability and fit of persons sufficient to support the use of this style of multiple-choice test?

6. Are the test items unidimensional and locally independent?

The Sample

For this research, the ELQ results of 83 students of a cohort of 552 were analysed. This represented 4 of the 27 classes that made up the first-year cohort. Students were placed based on their TOEFL scores, and the classes selected were the lowest, highest, and two mid-level classes, representing the spread of abilities across the entire cohort. All 20 items from the 4 assessments (five items per test) were pooled together and analysed using the Winsteps software (Linacre, 2024, version 5.2.1.0).

Results

Research Question 1

Are the test items able to distinguish sufficiently between different levels of performance among test takers?

Table 1 suggests that the data generally fits the Rasch model due to the fact that the mean infit and outfit for both persons and items are very close to the model standard of 1.0. The separation ratio is 3.84 for the items, suggesting that the items can distinguish nearly 4 levels of difficulty, and the reliability is good at 0.94. This means it is likely that the same results would hold if the items were used with a different sample (Bond and Fox, 2015). However, the separation, 1.28, and reliability, 0.62, are low for persons. Person separation and reliability statistics less than 2 and 0.9 speak to the accuracy and reliability of the person statistics (Linacre, 2020a). This means that if these test takers were to take another similar test, they may not get a similar score as this one. These statistics are largely a consequence of the low number of items (20) relative to test takers (83). This also leads to the higher mean standard errors on the persons (0.7) compared to items (0.39).

Table 1

Overall Statistics for EFQ

Person	83		Measured		INFIT		OUTFIT	
	Total	Count	Measure	RealSE	INMSQ	ZSTD	OMNSQ	ZSTD
Mean	13.1	19.9	1.40	0.71	1.01	0.10	1.00	0.20
P. SD	3.5	2.3	1.19	0.20	0.33	0.90	1.13	0.70
Real RMSE	0.73	TRUE SD	0.94	Separation	1.28	Person reliability		0.62
Item			Measured		INFIT		OUTFIT	
	Total	Count	Measure	RealSE	INMSQ	ZSTD	OMNSQ	ZSTD
Mean	54.5	78.5	0.00	0.40	0.98	0.10	0.97	0.10
P. SD	18.3	1.8	1.79	0.20	0.16	1.20	0.54	1.60
Real RMSE	0.44	TRUE SD	1.73	Separation	3.90	Item reliability		0.94

Research Question 2

Is there a sufficient range of difficulty in the items?

This is perhaps best answered in reference to the Wright Map displayed in Figure 1. The Wright map shows the spread of persons and items relative to one another. The persons are labelled by class, with

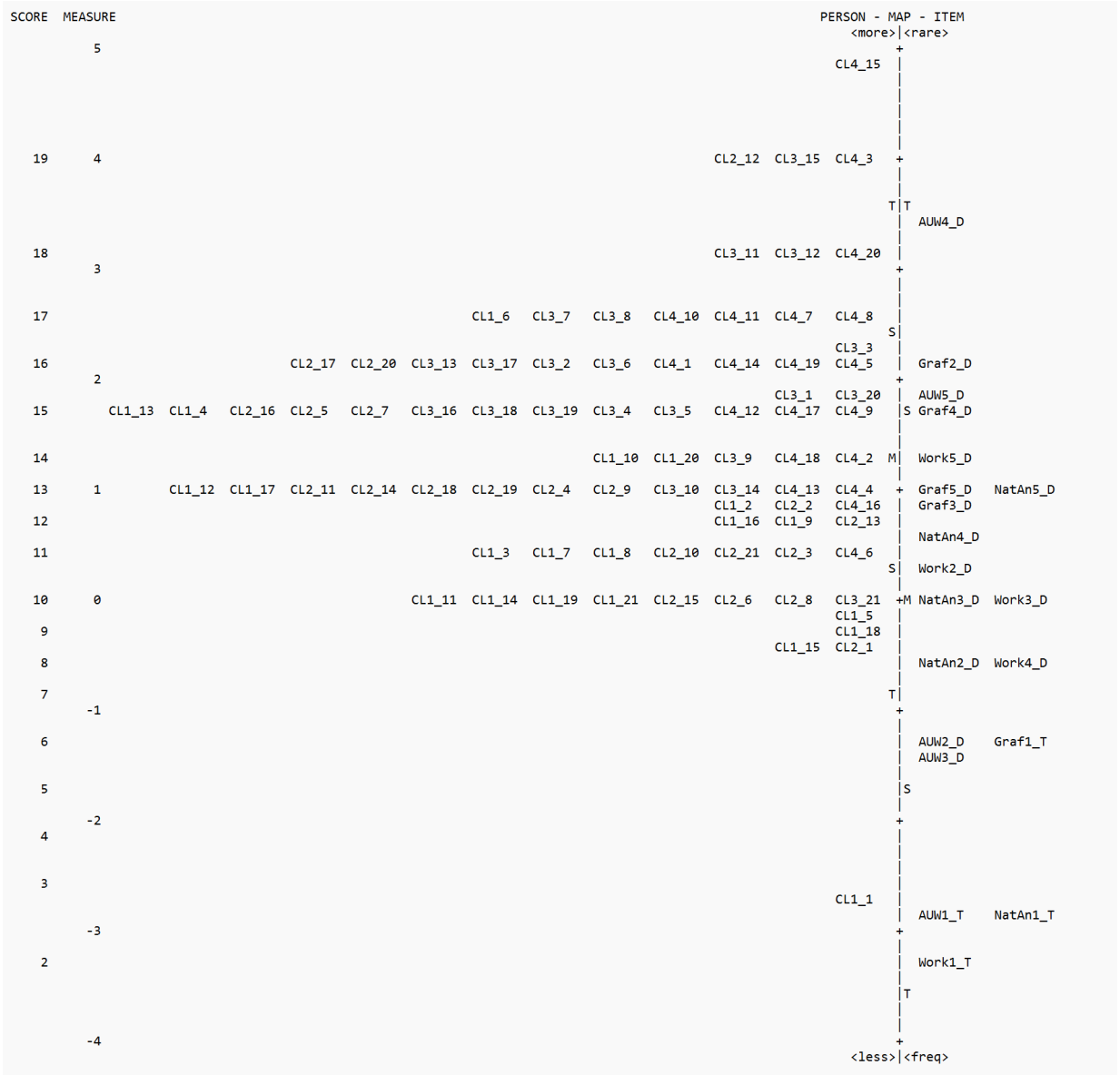
class 1 being the lowest and class 4 the highest. For example, CL1_4 is learner 4 from class 1. The ELQ topics were also given abbreviated labels, for example NatAn was used for questions relating to the national anthem topic. The suffix D was for detail questions and T for questions on the theme of the listening. The theme question was on the gist of the listening and was always question 1 of each set. This labelling allows for easier recognition of trends among classes, topics and item types. Indeed, it shows that classes 3 and 4 performed better than classes 1 and 2, broadly reflecting their class placement. In terms of listening topics, some appear harder than others with the graffiti topic appearing harder and the national anthem topic seemingly easier. The AUW topic appears to have the widest spread with 3 of the easiest questions and 2 of the hardest. However, there do not appear to be any extreme trends, with no topics significantly easier than others. In terms of question types, the 'what is the topic' questions are generally among the easiest, as expected, with other detail questions spread throughout the range of difficulty.

The measure column is in logits with the mean item difficulty at zero logits. The score column shows the raw scores for the quiz. The map reveals some interesting features of how the test is functioning. On first glance, it appears that some questions are too easy for this cohort because a number are clustered below most learners' ability level and the means, labelled M, do not align. However, there is an institutional requirement that most learners achieve a score of 70% and a pass is set to 60. The two means would only align if the mean score was 50%, meaning most test-takers fail. However, the ease of the lowest group of questions, Work1_T, NatAn1_T, AUW1_T, AUW3_D, Graf1_T is somewhat problematic for the person ability measures and is part of the cause of the lower person reliability score (0.62), due to an insufficient number of items targeted to the learner ability level, as well as the low number of items overall. In particular, the low number of targeted items contributes to this; six items are below the ability level of all but learner CL1_1, who only got two questions correct in total. This represents around 30% of the items, meaning we only actually have 14 items that are well targeted to most of the cohort. Also, there is only one item targeted to the top few students. Overall, although the institutional constraint to set a passing score at 60% requires some easier items, there needs to be a greater number of items targeted to the level of most students.

Figure 1 depicts all three facets modeled in the analysis. Wright maps use a logit—short for log odds—scale, which produces standardized interval measurements (as seen in the left-hand column) based on statistical probability. They provide a graphic illustration of the amount of variance within each facet, and the common scale allows for comparison with the other facets. Upon visual inspection it is clear that the greatest amount of variance is found among the students, followed by the raters, and finally the category items. Each facet is examined in detail below.

Figure 1

All-facet Wright Map for the MFRM Analysis



**CL = Class; Graf = Graffiti; Work = Work; AUW = University for Women; NatAn = National Anthems*

Research Question 3

Is the reliability and fit of the items sufficient that test scores are generalizable to another similar cohort?

Table 2

Item Fit Statistics

Item	Total Score	Total Count	JMLE Measure	Model S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
Graf2_D	28	79	2.12	0.26	1.21	1.76	2.91	6.02
Graf5_D	45	79	1.02	0.25	1.3	2.88	1.32	1.8
AUW1_T	79	81	-2.91	0.76	0.77	-0.17	1.28	0.6
Graf4_D	34	79	1.72	0.25	1.17	1.69	1.19	1.08
AUW5_D	33	81	1.87	0.25	1.15	1.48	1.18	0.96
Graf1_T	71	79	-1.25	0.4	1	0.11	1.18	0.5
Graf3_D	48	79	0.83	0.26	1.11	1.11	1.13	0.74
AUW4_D	13	81	3.47	0.34	1.1	0.53	1.03	0.22
AUW3_D	74	81	-1.41	0.42	1.06	0.3	0.69	-0.4
AUW2_D	73	81	-1.24	0.4	1.01	0.13	1.00	0.18
NatAn4_D	51	78	0.62	0.26	0.95	-0.47	0.84	-0.81
NatAn5_D	46	78	0.95	0.26	0.94	-0.56	0.87	-0.71
Work1_T	75	76	-3.29	1.02	0.94	0.26	0.23	-0.48
Work3_D	59	76	-0.02	0.3	0.93	-0.46	0.91	-0.23
Work4_D	65	76	-0.63	0.34	0.88	-0.5	0.61	-0.98
NatAn3_D	60	78	-0.07	0.29	0.84	-1.05	0.72	-0.99
Work5_D	40	76	1.34	0.25	0.82	-1.95	0.79	-1.63
Work2_D	55	76	0.3	0.28	0.79	-1.74	0.63	-1.74
NatAn2_D	65	78	-0.55	0.33	0.78	-1.14	0.53	-1.39
NatAn1_T	76	78	-2.87	0.76	0.73	-0.23	0.37	-0.44

In terms of the reliability, it would seem that the person reliability of 0.62 is too low to make stable inferences about the person abilities. This is partly due to the low number of total items as well as the low number of targeted items. This is a low stakes assessment, and each test is worth 5%, making a total of 20% of the course grade. However, it's still desirable to have better reliability, and if an assessment such as this is to be used as the major component of a course grade, more items would be needed. The Spearman-Brown prophecy can be used to calculate this (see Bobbitt, 2021 for an accessible explanation). This gives us a predicted reliability if an assessment is increased by a given number of items. In our case, increasing to 50 items, or a total of ten more listening tasks, would give a reliability of 0.80. This is achievable in a 14-week semester and would leave some space for additional assignments. Another option would be to increase the number of items per test. The texts are long, so this is possible, but we would need to take care not to violate the principles of local independence because questions could overlap.

In terms of fit, we could take the parameters of 0.5 – 1.5 to be an acceptable fit range because this is a low-stakes test (Linacre, 2020a). There are no items whose infit scores are beyond these boundaries, and most are within an even narrower range. Regarding Outfit, items 1 and 17 have scores of concern. Item 1

is the easiest, but underfit is less of an issues concern than overfit; Bond and Fox (2015) note this often does not have practical implications. However, when examining the misfitting response strings, Table 6.5 of the Winsteps output showed that this arose from one learner, CL1_8, getting this question wrong. A similar pattern exists for question 2 where two learners, 10 (CL1_10) and 76 (CL3_14), both of whom are clustered around the mean ability, got this question wrong. The most misfitting item is Question 17. However, the infit statistics are within acceptable bounds, so the issue is probably caused by unexpectedly right or wrong answers from learners not targeted by the specific question. Winsteps output Table 6.5 shows it was answered unexpectedly correctly by 6 learners: CL1_1, CL1_15, CL1_18, CL1_14, CL2_8 CL1_8, who are all below or around the mean ability of the cohort. Question 17 (Graf2_D) is in fact the second most difficult question at 2.2 logits and was most notably answered correctly by learner 1, one of their two correct answers of the 20 questions, which is the probable cause of the substantial misfit. Although the fit statistics are broadly within acceptable bounds, there do appear to be patterns within the test sets which broadly misfit in the following order: graffiti, AUW, national anthems, and work. That is, the graffiti and AUW test sets tended to have fit statistics over 1.0 and national anthems and work topics below 1. In overall answer to research question 3, it would seem that the fit statistics are broadly acceptable, but that reliability would need to improve, and this would be best done by increasing the number of items.

Research Question 4

Are the distractors functioning in an acceptable way?

The full distractor analysis can be seen in Appendix BB. Generally speaking, the distractors seem to be functioning effectively. Usually, the correct answer has the highest point-measure score, and average ability of persons choosing that item should be higher than for other items. This is largely the case with our data. Also, the ability does seem to ascend with the 'correctness' of the distractor options, and most distractors have acceptable fit statistics. However, a selection of distractors is reproduced below in Table 3: the three most misfitting items as well as items where the average ability does not ascend with the distractor response.

Table 3
Descriptor Analysis for Misfitting Items and Responses

Item	Data code	Value	Count	%	Ability Mean	INFIT MNSQ	OUTFIT MNSQ	PTMEA Corr
Graf2_D	1	0	6	8	0.86	0.60	0.60	-0.13
	3	0	36	46	1.15	1.10	1.00	-0.19
	2	0	5	6	1.49	1.00	1.10	0.02
	4	0	4	5	1.68	1.70	2.00	0.05
	5	1	28	35	1.77	1.40	3.90	0.23
Graf5_D	4	0	13	16	0.66	1.10	1.00	-0.27
	1	0	6	8	1.08	1.00	0.90	-0.08
	2	0	8	10	1.40	1.40	1.70	0.00
	5	0	8	10	1.61	1.60	1.60	0.06
	3	1	44	56	1.62	1.40	1.40	0.20
AUW1_T	3	0	2	2	-0.52	0.50	1.30	-0.26
	2	1	79	98	1.48	0.40	0.50	0.26
AUW2_D	2	0	3	4	-0.77	0.60	0.30	-0.36

	4	0	2	2	0.46	1.00	0.70	-0.13
	3	0	2	2	1.17	2.10	1.30	-0.03
	5	0	1	1	2.10	5.50	3.20	0.06
	1	1	73	90	1.54*	0.90	0.90	0.29
Graf1_T	3	0	1	1	-0.49	0.40	0.20	-0.18
	5	0	5	6	-0.01	0.90	0.80	-0.30
	1	0	2	3	1.90	4.50	2.70	0.07
	2	1	71	90	1.51*	0.90	0.90	0.27

The most misfitting item is the second question in the graffiti test (Graf2_D). The actual answer was salient in the listening itself, but as we can see, Banksy (option 3), an incorrect answer, was selected by the most students. Banksy was also mentioned in the text, but the general salience of Banksy overall in society would perhaps have led to more students choosing this option than otherwise would have and so has made it appear more difficult than it in fact is. Wu (1998) found a similar trend when exploring multiple-choice tests for listening assessment. However, this option could be maintained as a difficult question to help test whether learners are actually listening carefully. Also, it appears that mostly lower-ability learners selected this option because the mean ability for the item was 1.15 and the point-correlation scores align appropriately. Interestingly, a different option, option 4, had worse fit statistics, which was caused by higher ability students selecting that option, and is probably at the root of this question's misfit. Likewise, the correct option, has an outfit score of 3.9, which, as noted above, is caused by lower-ability students getting that correct. Turning to the Graf1_T item, there are similar issues with fit, as both infit and outfit score are well above acceptable levels. Once again this is due to two higher level students getting this wrong. A partial cause could be the relatively low number of participants, which makes the questions more sensitive to misfit.

In general, it seems that the distractors are generally acceptable, but there are a higher number of distractors than are generally recommended (Lord, 1977). Indeed, Haladyna and Downing (1989) note that the quality of distractors contributes to the difficulty of a question and non-functioning distractors do not add anything to an assessment. Kilgour and Tayyaba (2016) suggest that three-option questions are optimal (one correct answer and two distractors), and so one option moving forwards would be to delete the least-functioning distractors. Wu (1998) has also found that poorly worded or confusing distractors can lead to a test-taker getting the answer wrong, despite good comprehension, or right for the wrong reasons.

Research Question 5

Is the reliability and fit of persons sufficient to support the use of this style of multiple-choice test?

For this project, the item statistics are more important than the person statistics because large numbers of learners will take this test into the future, so having stable items is more important. However, it can be worth inspecting the person statistics because they can reveal anomalous patterns among the cohort. Table 4 shows the most underfitting persons.

Table 4

Underfitting Persons

Person	JMLE measure	Model S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
CL1_1	-2.70	0.89	1.8	1.46	8.71	2.72
CL3_4	1.69	0.61	0.99	0.08	5.38	2.55
CL1_15	-0.48	0.68	2.57	3.23	3.77	2.59
CL2_21	0.38	0.56	1.41	1.50	2.17	1.74
CL1_3	0.38	0.56	1.32	1.23	2.01	1.57

There are a number of misfitting persons in the data set. These mostly occur in the outfit statistics, so like the items, these are caused by unexpectedly right or wrong answers. The biggest misfit lies with low-ability student CL1_1 and their correct answer of item Q17, the second most-difficult item (mentioned above, Graf2_D). Learner CL1_1 is by far the weakest performing person in the class, and getting any moderately difficult questions right, let alone the second hardest item, would likely lead to misfit. In fact, learner CL1_1 got only 2 questions correct of 15 answered (they missed one test). The only other correct response was AUW3_D, a relatively easy item, and none of the easiest 3 were answered correctly. This likely accounts for the misfit in the infit statistic and the widely misfitting outfit statistics. The better performing student CL3_4 missed question 6, an easy item, which appears to account for their misfit. Similar patterns exist for learners CL2_21 and CL1_3, with individual instances of answers mismatching their ability level, possibly due to lucky guessing, slip-ups, or topic knowledge. However, learners CL1_1 and CL1_15 exhibit fairly unusual behaviour across the board because both infit and outfit statistics underfit. For learner CL1_15, this is caused by four items being answered unexpectedly correctly or incorrectly. Of the items themselves, 11, 10, 17, and 20, two items 17 and 20, do appear as the most underfitting items, discussed above. 11 and 10 appear to be functioning acceptably, making it hard to diagnose what exactly has gone wrong for them, but as there are few other learners exhibiting similar trends, we can probably put it down to chance rather than it being an issue inherent to test function. In addition, there are also a few overfitting persons, but because this is an achievement test with the content already known in advance, this is fairly unavoidable, and so will not likely cause any serious issues with the test.

Research Question 6

Are the test items unidimensional and locally independent?

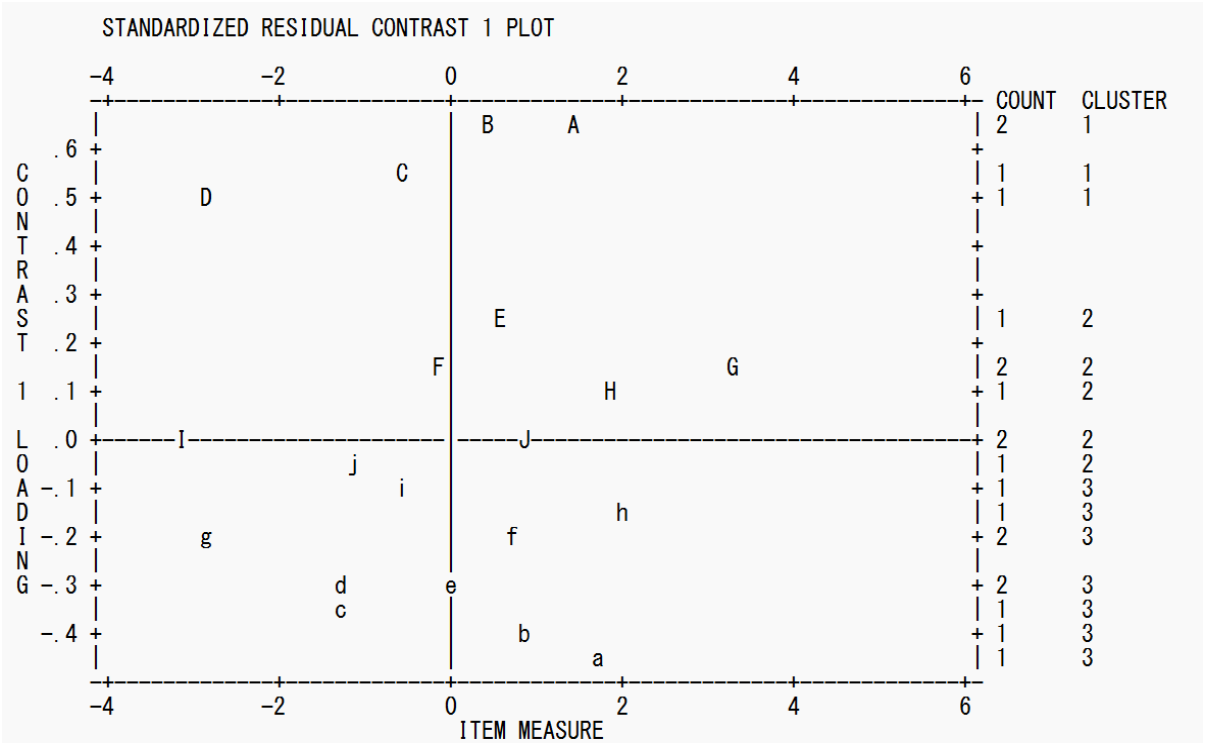
A fundamental Rasch concept is the concept of unidimensionality; that is, do the test items target a single construct (Wu & Adams, 2009), such as grammatical knowledge, or do certain items tap into additional constructs, such as grammatical and lexical knowledge. Aryadoust et al. (2020) have argued for the importance of reporting on the unidimensionality in assessment validation reports due to its importance for construct validity. Table 5 shows the residual data important for such an analysis. Overall, the results are good, because the observed variance explained by the Rasch measures, 36.7, is close to the expected value of 37.1. The person and item variances are also similarly close to the expected values. This percentage may appear low, but it is important to note that variance explained by Rasch measures is dependent on the spread of abilities and items, so if there is a narrow range of abilities among the test-takers and a narrow range of item difficulties, the variance explained will be small (Linacre, 2012). Therefore, low variance explained is not in itself an indicator of multidimensionality, or an indication of a bad test; what's important is closeness to model expectations.

Table 5
Standardized Residual Variance

	Eigenvalue	Observed	Expected
Total raw variance in observations	31.61	100.00	100.00
Raw variance explained by measures	11.61	36.70	37.10
Raw variance explained by persons	4.39	13.90	14.00
Raw variance explained by items	7.23	22.90	23.10
Raw unexplained variance (total)	20.00	63.30	62.90
Unexplained variance in 1st contrast	2.26	7.20	11.30
Unexplained variance in 2nd contrast	1.82	5.80	9.10
Unexplained variance in 3rd contrast	1.73	5.50	8.60
Unexplained variance in 4th contrast	1.51	4.80	7.50
Unexplained variance in 5th contrast	1.38	4.40	6.90

To examine whether or not there are any additional dimensions, or constructs other than the test construct, it is necessary to examine the unexplained variance in the 1st contrast. Any eigenvalues above 2 should be investigated further because this can potentially indicate a separate dimension (Linacre, 2020a). There is one such dimension with an eigenvalue of 2.26. This can be investigated using Figure 2 below.

Figure 2
Standardized Residual Variance Screen Plot



The four items that potentially make up the additional dimension are shown in the black box. These are clustered together and separated vertically from the rest. On a test of 20 items, this in fact represents 20% of the test items, a significant percentage, and have factor loadings of above 0.4, above the threshold for

further investigation. These 4 items are shown in Table 6 below. Additionally, they should be contrasted with any items with negative factor loadings of above -0.4. There is only one such item, item 20 with a loading of 0.42. However, three other items are close, items 7 and 19, so they could be investigated as a precaution.

Table 6

Item Contrast Analysis of Residuals

Contrast	Cluster	Loading	Measure	INFIT MNSQ	OUTFIT MNSQ	Item
1	1	0.68	1.34	0.82	0.79	Work5_D
1	1	0.65	0.30	0.79	0.63	Work2_D
1	1	0.56	-0.55	0.78	0.53	NatAn2_D
1	1	0.45	-2.87	0.73	0.37	NatAn2_T
1	2	0.18	3.47	1.10	1.03	AUQ4_D
1	3	-0.42	1.02	1.30	1.32	Graf5_D
1	3	-0.37	-1.24	1.01	1.00	AUW2_S
1	3	-0.36	1.72	1.17	1.19	Graf4_D
1	3	-0.31	0.95	0.94	0.87	NatAn5_D

All test items can be seen in Appendix AA, and on inspecting these items, it appears that they are not significantly related to one another in terms of their content and do not appear to target a common construct or be related to one another in any obvious or meaningful way. One possible explanation could be topic knowledge of the graffiti topic. This could account for the spate of unexpectedly correct responses. However, these learners are not art majors, and as far as I know, do not cover this theme in any other classes. A further explanation could be that the learners studied harder for this test than the others. The graffiti test was towards the end of the course, and these learners were at the bottom end of the ability chart with fewer correct answers overall, so potentially made an extra effort to bump their scores up to passable levels. Either way, testing with a larger sample would be advisable and would help to eliminate doubt surrounding issues such as this. Likewise, increasing the number and range of topics would be beneficial because the potential for topic familiarity influencing test functioning, and therefore reliability, would be reduced.

Item independence

One of the key assumptions in Rasch analysis is item independence. This is the idea that each item functions separately from one another and does not target the same piece of information or knowledge. This can be a challenge in language tests where one set of questions targets a single text (Baghei, 2008). Local item independence can be explored through correlations of items, where items that have a strong correlation of over 0.7 (Linacre, 2020c) are likely dependent on one another. Table 7 below shows that the strongest positive correlation is only 0.40, and the strongest negative correlation is -0.36. Therefore, it would appear that the items are functioning independently of each other.

Table 7

Correlations Among Test Items

Correlation	Item	Item
0.40	Work2_D	Work5_D
0.33	AUW3_D	Work1_T
0.29	Work1_T	Graf1_T
0.23	NatAn1_T	AUW1_T
-0.36	AUW4_D	Graf5_D
-0.32	Work3_D	Work5_D
-0.29	AUW3_D	Work5_D
-0.28	AUW3_D	Work2_D

Discussion

Broadly speaking, there are a number of findings that support the use of a post-listening recall test operationalized through multiple-choice questions. Items and people generally fit the Rasch model and local independence and unidimensionality appear to hold. This would support larger scale use of this assessment type with a greater number of out-of-class listening tasks to afford an increase in listening skill outcomes and improved reliability. However, there are a number of areas for improvement which are discussed below.

Item targeting, distractors and reliability

Overall, there is some way to go in terms of improving the reliability of the test and the item targeting. Typically, to increase the person reliability, it is necessary to either have a wider range of person abilities, or to have a higher number of items to make finer-grained judgements about the cohort's abilities; this becomes especially important with learners around the pass-fail point to avoid unnecessarily failing students who actually have the ability to pass. Furthermore, the items in this pilot study are stretched a little too 'thin' in the most important levels of ability, i.e. from around raw scores of 10 – 17 (around 0 – 2.5 logits), the area into which most learners fall and covering the pass-fail mark. The test is only one question deep in most of these locations. Bond and Fox (2015) note that reliability is driven by sample size, so, in our case, our 82 persons give more information about the items than the 20 (or perhaps only 14) give us about the people. Therefore, the low person reliability is largely a factor of the low number of suitably targeted items and the relatively narrow range of ability. Better item reliability comes from a high test-taker to item ratio. Additionally, there are proportionally too few items below most learners' level of ability and only one, AUW4_D, at the higher end, which covers all learners from 2.5 logits and above. Additionally, the four easiest items give little information about the learners – virtually everyone gets them right. However, they are necessary to a degree to ensure that the learners can earn enough points to pass or at least receive the required grade to pass the course/reach the 'standard' of 70%.

Thus, overall, for more reliable assessment, a greater number of items needs to be added, especially at the higher and mid-ranges of difficulty. This can be achieved by increasing the number of items to around 50, which would afford a reliability of 0.80 according to the Spearman-Brown prophecy. This could be done through a combination of additional items per listening and increasing the number of out-of-class listening tasks and therefore tests. This should lead to greater fluency gains due to the increased amount of listening done (Nation & Newton, 2009), without significantly increasing the grading burden. However, further

rounds of analysis would be necessary to explore item targeting and difficulty level of any new items and tests.

Another improvement to the test design would be decreasing the number of distractors. The five options given seem unnecessary, and above the generally recommended number of three or four (Haladyna & Downing, 1989; Lord, 1977). Examining the distractor performance (see Appendix BB) would help identify weaker performing distractors, and they could then be eliminated to make the test more efficient. Haladyna and Downing (1989) note that poorly functioning distractors add nothing to an assessment. A good approach could be that suggested by Linacre (2020c) where, in the case of four distractors, the options could be written according to the following scheme: completely wrong, mostly wrong, almost correct and correct. By doing so, it is possible to see whether the distractors intended to be wrong do in fact end up as such by comparing the point-measure correlations and determining whether there is a steady ascendence of person abilities for the incorrect distractors.

In terms of the overall test validity, the core construct being tested is recall of a listening task, rather than listening ability per se, but because the test appears to hold in terms of fit, unidimensionality and local independence, it seems the construct is stable. Although memory is clearly a mediating variable, it is likely that better listening skills would lead to better performance on the test, as has been found in other studies where other variables are at play (Kim et al., 2022; Rukthong & Brunfaut, 2020; Sawaki, et al., 2013). It is also likely that the intention of creating a positive study impact (Fulcher & Davidson, 2007) and positive washback (Messick, 1996) would be achieved through completion of the out-of-class listening tasks. The more deeply learners engage with the text, the more likely they are to improve their listening skills due to the increased attention and focus on the task (Field, 2009). Interestingly, brain imaging studies suggest that the brain is more active when the assessment is presented after listening (Aryadoust & Luo, 2022). Furthermore, Nation and Newton (2009), point out that repetition is a key component in acquiring fluency, so if learners do repeatedly listen to the audio to remember content and pass the later test, it will likely positively impact fluency skills. That said, in the delivery of the course, learners should be notified of the role of memory in good test performance. Alderson (2000) has noted where post-task assessments are used, it's better to administer them as soon as possible after listening. In our context, the learners decide when to listen, so pointing this out may help them to maximise their performance by listening or relistening close to the test. This would also help mitigate the memory factor.

Future research, limitations

In terms of limitations, using a greater sample of learners would have added greater reliability to the findings on the test items, although 83 participants is still a good number for a pilot study such as this one. Furthermore, getting input from the learners and teachers would also have added some depth to the findings. The learner input could have brought insight as to how learners are studying the assigned and independent listening as well as their thoughts on efficacy. Likewise, interviewing the teachers as to how the assessment may be better implemented and their views on its validity would also bring an interesting perspective. In addition, once new items or tests are added, a Rasch analysis should be done to explore their performance and their overall effect on the scores received by learners. Following this, exploring the impact on listening skills through a pre- post-design has potential. Investigating other aspects of the test battery also has potential, especially the notes taken on the extensive listening tasks. Many-facet Rasch measurement (MFRM) could be used to examine, in particular, the rating of the ELH and would shed light on how consistently and severely the teacher was rating and would reveal avenues for refinement of the scoring system.

A final point is whether this degree of scrutiny is necessary for a test with stakes as low as this. Doing a Rasch analysis is an in-depth process with a high barrier to entry (Smiley, 2015). Indeed, Schaefer and Martin (2023) discuss this matter in a recent interview with Dr Daniel Isbell, noting that the case for a particular assessment type may be quite informal, without complex validation procedures. However, the case is also made for validity arguments for low stakes assessments because they can clarify the rationale for a particular test type and the judgements made based on the test results (Schaefer & Martin, 2023). There are also substantial benefits in terms of professional development and teacher-led research, and this can lead to a deeper understanding of the local context (Allwright, 2005). It can also help identify a research agenda and other areas for course development. However, teachers are busy and do not often have additional time for validity studies, so they need to be judicious in selecting areas for attention. It is unlikely that all elements of a test battery can be analysed in any one term, so assessments could be analysed on a rotating basis, or test battery elements that have some doubt or controversy surrounding them could be targeted. Alternatively, educators could target assessments that have the potential for expansion with an initial pilot study before scaling them up for wider use.

Conclusion

Overall, the findings of the Rasch analysis suggest that using a discrete-item, multiple-choice test to assess assigned, out-of-class listening is a valid approach, even though there is the potential for memory to be a source of construct irrelevant variance. The ELQ used here is functioning relatively well, despite the low number of items. However, increasing the number of better-targeted items would improve test function, reliability, and, if it comes with additional listening tasks, better listening outcomes for the student cohort, which, ultimately is the intention of these tasks in the first place.

Declaration of competing interests:
T. Stones has declared no competing interests.

References

Alderson, J. C. (2000) *Assessing reading*. Cambridge University Press.

Alderson, J. C., & Wall, D. (1993). Does washback exist? *Applied Linguistics, 14*(2), 115–129. https://doi.org/10.1093/applin/14.2.115

Allwright, D. (2005). Developing principles for practitioner research: The case of exploratory practice. *Modern Language Journal, 89*(3), 353–366. https://doi.org/10.1111/j.1540-4781.2005.00310.x

Aryadoust, V., Ng, L. Y., & Sayama, H. (2020). A comprehensive review of Rasch measurement in language assessment: Recommendations and guidelines for research. *Language Testing*, *38*(1), 6–40. https://doi.org/10.1177/0265532220927487

Aryadoust, V., & Luo, L. (2022). The typology of the second language listening construct: A systematic review. *Language Testing, 40*,(2), 375–409. https://doi.org/10.1177/02655322221126604

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford University Press.

Baghei, P. (2008). Local dependency and Rasch measures. *Rasch Measurement Transactions, 21*(3), 1105–1106. https://www.rasch.org/rmt/rmt213b.htm#:~:text=Local%20independence%20of%20items%20is,local%20item%20dependence%20(LID)

Beglar, D., & Hunt, A. (2014). Pleasure reading and reading rate gains. *Reading in a Foreign Language, 26*(1), 29 – 48.

Bobbitt, Z. (2021, December 13). The Spearman-Brown Formula: Definition & example. https://www.statology.org/spearman-brown-formula/

Bond, T., & Fox, M. (2015). *Applying the Rach model: Fundamental measurement in the human sciences.* 3rd Ed. Routledge.

Brown, R., Waring, R., & Donkaewbua, S. (2008). Incidental vocabulary acquisition from reading, reading-while-listening, and listening to stories. *Reading in a Foreign Language*, *20*(2), 136–63.

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

Burns, A. (1999). *Collaborative action research for English language teachers*. Cambridge University Press.

Burns, A, (2010). *Doing action research in English language teaching*. Routledge.

Chang, Y-F. (2006). On the use of the immediate recall task as a measure of second language reading comprehension. *Language Testing, 23*(4), 520–543. https://doi.org/10.1191/0265532206lt340o

Chang C–S. (2009). Gains to L2 listeners from reading while listening versus listening only in comprehending short stories. *System, 37*(4), 652–63. https://doi.org/10.1016/j.system.2009.09.009

Chang, C-S. (2011). The effects of reading while listening to audiobooks: Listening fluency and vocabulary gain. *Asian Journal of English Language Teaching 21*, 43–64.

Chang, C-S. (2012). Gains to L2 learners from extensive listening: Listening development, vocabulary acquisition and perceptions of the intervention. *Hong Kong Journal of Applied Linguistics* 14(1), 25–47.

Chang, A., & Millett, S. (2014). The effect of extensive listening on developing L2 listening fluency: Some hard evidence. *ELT Journal, 68*(1), 31–40. https://doi.org/10.1093/elt/cct052

Chang, A., & Millett, S. (2016). Developing L2 listening fluency through extended listening-focused activities in an extensive listening programme. *RELC Journal*, *47*(3), 349–362. https://doi.org/10.1177/0033688216631175

Chang, A., Millett, S., & Renandya, M. (2018). Developing listening fluency through supported extensive listening practice. *RELC Journal, 50*(3), 422–438. https://doi.org/10.1177/0033688217751468

Field, J. (2009). *Listening in the language classroom*. Cambridge University Press.

Fulcher, G., & Davision, F. (2007*). Language testing and assessment: An advanced resource book*. Routledge.

Haladyna, T., & Downing, S. (1989). Validity of a taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education, 2*(1), 51–78. https://doi.org/10.1207/s15324818ame0201_4 .

Joh, J., & Schallert, D. J. (2014). How conception of task influences approaches to reading: A study of Korean college students recalling an English text. *TESOL Quarterly, 48*(4), 715- 737. https://doi.org/10.1002/tesq.147

Just, M. A., & Carpenter, P. A. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review, 99*(1), 122–149. https://doi.org/10.1037/0033-295X.99.1.122

Kilgour, J. M., & Tayyaba, S. (2016). An investigation into the optimal number of distractors in single-best answer exams. *Advances in Health Science Education, 21*, 571–585. https://doi.org/10.1007/s10459-015-9652-7

Kim M., Nam Y., Crossley S. A. (2022). Roles of working memory, syllogistic inferencing ability, and linguistic knowledge on second language listening comprehension for passages of different lengths. *Language Testing, 39*(4), 593–617. https://doi.org/10.1177/02655322211060076

Linacre, 2012. *Winsteps tutorial 4.* https://www.winsteps.com/a/winsteps-tutorial-4.pdf

Linacre, M.(2020a). *A users guide to Winsteps, Ministep: Rach model computer programs.* Winsteps.com.

Linacre, M.(2020b). *A users guide to Winsteps, Ministep: Table 23.99 Largest residual correlations for items.* https://www.winsteps.com/winman/table23_99.htm

Linacre, M. (2020c). *A users guide to Winsteps, Ministep: DISTRACTOR= output option counts in Tables 10, 13-15 = Yes*. https://www.winsteps.com/winman/distrt.htm

Linacre, M. (2024). *Winsteps version 5.2.1.0.* https://www.winsteps.com/winsteps.htm

Lord, F. M. (1977). Optimal number of choices per item: A comparison of four approaches. *Journal of Educational Measurement, 14*(1), 33–38. https://doi.org/10.1111/j.1745-3984.1977.tb00026.x

McNamara, R. (1996). *Measuring second language performance*. Longman.

McNamara, T., & Knoch, U. (2012). The Rasch wars: The emergence of Rasch measurement in language testing. *Language Testing, 29*(4), 555–576. https://doi.org/10.1177/0265532211430367

Messick, S. (1996). Validity and washback in language testing. *Language Testing 13*(3), 241–256. https://doi.org/10.1177/026553229601300302

Nation, P., & Newton, P. (2009). *Teaching EFL/ESL listening and speaking.* Routledge.

Nation, P., & Waring, R. (2020). *Teaching extensive reading in another language.* Routledge.

Nishida, R. (2013). The L2 ideal self, intrinsic/extrinsic motivation, international posture, willingness to communicate and Can-Do among Japanese university learners of English. *Language Education and Technology, 50*, 47–63. https://doi.org/10.24539/let.50.0_43

Onoda, S. (2014). Investigating effects of extensive listening on listening skill development in EFL classes. *The Journal of Extensive Reading in Foreign Languages, 1*(1), 43–55.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Danish Institute for Educational Research.

Rasch, G. (1977). On Specific Objectivity: An attempt at formalizing the request for generality and validity of scientific statements. *The Danish Yearbook of Philosophy*, *14*, 58–93.

Renandya, W. A., & Farrell, T. (2011). 'Teacher the tape is too fast!' Extensive listening in ELT. *ELT Journal, 65*(1), 52–59. https://doi.org/10.1093/elt/ccq015

Renandya, W. A., & Jacobs, G. M. (2016). Extensive reading and listening in the L2 classroom. In W. A. Renandya, & Handoyo, P. (Eds.), *English language teaching today* (pp. 97–110). Routledge.

Rost, M. (2006). Areas of research that influence L2 listening instruction. In E. Usó-Juan & A. Martínez-Flor (eds.). *Current trends in the development and teaching of the four language skills*. Mouton de Gruyter.

Rost, M. (2011). *Teaching and researching listening.* 2nd Ed. Pearson.

Rukthong, A. (2021). MC listening questions vs. integrated listening-to-summarize tasks: What listening abilities do they assess? *System, 97,* 1–12. https://doi.org/10.1016/j.system.2020.102439

Rukthong, A., & Brunfaut, T. (2020). Is anybody listening? The nature of second language listening in integrated listening-to-summarize tasks. *Language Testing, 37*(1), 31–53. https://doi.org/10.1177/0265532219871470

Sawaki, Y., Quinlan, T., & Lee Y. W. (2013). Understanding learner strengths and weaknesses: Assessing performance on an integrated writing task. *Language Assessment Quarterly, 10*(1), 73–95. https://doi.org/10.1080/15434303.2011.633305

Schaefer, E., & Martin, J. (2023). Language testing in changing times: An interview with Professor Daniel Isbell. *Shiken, 27*(2), 1–5. https://doi.org/10.37546/JALTSIG.TEVAL27.2-1

Schmidt, R. (1992). Psychological mechanisms underlying second language fluency. *Studies in Second Language Acquisition*, *14*(4), 357–385. https://doi.org/10.1017/S0272263100011189

Skehan, P. (1998). *A cognitive approach to language learning.* Oxford University Press.

Smiley, J. (2015). Classical test theory or Rasch: A personal account from a novice user. *Shiken, 19*(1), 16–31.

Song, M-Y. (2011). Note-taking quality and performance on an L2 academic listening test. *System, 29*(1), 67–89. https://doi.org/10.1177/026553221141537

Spotlight English. (2022). *About spotlight.* https://spotlightenglish.com/about-us/

Stephens, M. (2011). The primacy of extensive listening. *ELT Journal, 65*(3), 311–313. https://doi.org/10.1093/elt/ccq042

Swan, M., & Walter, C. (2017). Misunderstanding comprehension. *ELT Journal, 71*(2), 228–236. https://doi.org/10.1093/elt/ccw094

Waring, R. (2008). Starting extensive listening. *Extensive Reading in Japan*, *1*(1), 7–9.

Was, C. A., & Woltz, D. J. (2007). Reexamining the relationship between working memory and comprehension: The role of available long-term memory. *Journal of Memory and Language, 56*, 86–102.

Wilson, M. (2003). Discovery listening—improving perceptual processing. *ELT Journal, 57*(4), 335–343. https://doi.org/10.1093/elt/57.4.335

Wu, M., & Adams, R. (2007). *Applying the Rasch model to psycho-social measurement: A practical approach*. Educational Measurement Solutions.

Yamamoto, J., & Ohba, H. (2018). Motivational characteristics of lower-level Japanese university EFL learners. *International Journal of Curriculum Development and Practice, 19*(1), 37–50. https://doi.org/10.18993/jcrdaen.19.1_37

Appendix A

Test Items

Note: Items with positive loading: 1, 2, 12, and 15; items with negative loading: 7, 19, and 20.

Question 1 (NatAn1_T)
What was this report about?

1. National Pride
2. National Guard
3. National Olympics Teams
4. National Anthems (correct answer)
5. National Review

Question 2 (NatAn2_D)

Where did Philip Sheppard find the national songs?

1. The Internet
2. The British Parliament
3. The British Library (correct answer)
4. The country of Bhutan
5. The British Museum

Question 3 (NatAn3_D)

Where did the words of the Japanese anthem come from?

1. A very ancient novel
2. A very old poem (correct answer)
3. Japanese students
4. The country's leaders
5. Mozart

Question 4 (NatAn4_D)
Which country has the oldest anthem?

1. Holland (correct answer)
2. Bhutan

3. England

4. Spain

5. Denmark

Question 5 (NatAn5_D)

The longest anthem is from what country?

1. Bhutan

2. England

3. Spain

4. Denmark

5. Greece (correct answer)

Question 6 (AUW1_T)

What was this report about?

1. The American University for Women

2. The Asian University for Women (correct answer)

3. The Asian Development University

4. The Asian Women's College

5. Students working for social change

Question 7 (AUW2_D)

A UNESCO report said that in 2009 only ___% of women in Asia attended university

1. 28% (correct answer)

2. 25%

3. 38%

4. 22%

5. 18%

Question 8 (AUW3_D)

Where is AUW located?

1. Bhutan

2. the Palestinian territories
3. Bangladesh (correct answer)
4. China
5. Myanmar

Question 9 (AUW4_D)

According to the report, what is a board of trustees?

1. A group of people that helps make decisions (correct answer)
2. A group of people that go to university
3. A group of people that hopes to get jobs and begin working
4. A group of people that trains women to become leaders
5. A group of people that works for social change

Question 10 (AUW5_D)

What does Sumpa Sarkar want to do after she graduates?

1. Teach English
2. Teach at AUW
3. Return to the Palestinian territories
4. Teach political science (correct answer)
5. Teach in the United States

Question 11 (Work1_T)

What was this report about?

1. Job interviewing
2. The best companies to work for
3. Finding a job
4. Work culture (correct answer)
5. Government work

Question 12 (Work2_D)

According to Satheesh Kumar, what is one thing that can help you gain trust from your co-workers?

1. Ask your boss a lot of questions
2. Read the company website
3. Dress like them
4. Stay late every night
5. Eat lunch with co-workers (correct answer)

Question 13 (Work3_D)

According to the report, what factors influence workplace culture?

1. The president of the company
2. The size of a company and number of employees
3. The values of a company and the culture of the country (correct answer)
4. The average age of workers in a company
5. Your salary

Question 14 (Work4_D)

According to the report, one good way to learn work culture is to observe your….

1. Co-workers (correct answer)
2. Classmates
3. Boss
4. English teacher
5. Parents

Question 15 (Work5_D)

According to the report, which of the following is true about body language?

1. Standing too close is considered rude.
2. Standing too far away appears that you are not interested.
3. Canadians are most comfortable with 18 inches between each other.

4. Standing too far away will make you seem shy.

5. All of the above. (correct answer)

Question 16 (Graf1_T)

What was this report about?

1. Graffiti is done by artists.

2. Graffiti is a crime and an art. (correct answer)

3. Art in England.

4. Crime in the United States.

5. Graffiti is done by criminals.

Question 17 (Graf2_D)

Who did the biggest legal graffiti piece ever?

1. Doze Green

2. Abdal Ullah

3. Banksy

4. Thrash Lab

5. Saber (correct answer)

Question 18 (Graf3_D)

Graffiti is difficult to clean up because…

1. cities are large in size

2. the paint is very expensive

3. cities have no money for cleaning

4. criminal gangs don't like cleaning

5. the paint is hard to remove or cover up (correct answer)

Question 19 (Graf4_D)

What is a piece?

1. Paint used to spray on walls

2. When a person signs their name in large letters

3. Large, colorful graffiti that covers a whole wall (correct answer)
4. A portfolio of work from artists
5. An art show for graffiti

Question 20 (Graf5_D)

Why is it difficult to tell the difference between graffiti and other kinds of art?

1. The city pays a lot of money to paint over it
2. The letters are large, colorful and all joined together
3. Many graffiti artists have art shows and earn money (correct answer)
4. People think it ruins the environment
5. Rich, educated people and people on the street see it

Appendix B

Full Distractor Analysis

Item	Response Option	Correct Answer	Data Count	%	Ability Mean	INFT MNSQ	OUTF MNSQ	Pt. Mean Correlation
NatAn1_T	1	0	2	3	-1.18	0.5	0.4	-0.35
	4	1	76	97	1.5	0.4	0.8	0.35
	Missing	***	5	6#	0.99			-0.09
NatAn2_D	5	0	6	8	-0.53	0.5	0.3	-0.47
	2	0	3	4	0.06	0.6	0.4	-0.22
	4	0	4	5	0.69	1	0.8	-0.14
	3	1	65	83	1.72	0.8	0.8	0.53
	Missing	***	5	6#	0.99			-0.09
NatAn3_D	5	0	2	3	-1.33	0.4	0.2	-0.37
	3	0	2	3	0.06	0.5	0.3	-0.18
	1	0	13	17	0.63	0.9	0.8	-0.29
	4	0	2	3	0.69	0.9	0.6	-0.1
	2	1	59	76	1.77	0.9	0.8	0.49
	Missing	***	5	6#	0.99			-0.09
NatAn4_D	2	0	6	8	0.38	0.6	0.5	-0.25
	4	0	8	10	0.41	0.9	0.7	-0.28
	5	0	7	9	0.8	1	0.9	-0.16
	3	0	4	5	1.18	1.4	1.3	-0.05
	1	1	53	68	1.8	1	1	0.45
	Missing	***	5	6#	0.99			-0.09
NatAn5_D	2	0	5	6	0.11	0.8	0.6	-0.28
	1	0	8	10	0.48	0.6	0.5	-0.26
	4	0	9	12	0.88	0.9	0.9	-0.16
	3	0	11	14	1.06	1.2	1.2	-0.12
	5	1	45	58	1.94	0.9	0.9	0.49
	Missing	***	5	6#	0.99			-0.09
AUW1_T	3	0	2	2	-0.52	0.5	1.3	-0.26
	2	1	79	98	1.48	0.4	0.8	0.26
	Missing	***	2	2#	0.37			-0.13
AUW2_D	2	0	3	4	-0.77	0.6	0.3	-0.36
	4	0	2	2	0.46	1	0.7	-0.13
	3	0	2	2	1.17	2.1	1.3	-0.03
	5	0	1	1	2.1	5.5	3.2	0.06
	1	1	73	90	1.54*	0.9	0.9	0.29
	Missing	***	2	2#	0.37			-0.13
AUW3_D	2	0	1	1	-0.49	0.5	0.2	-0.18

	1	0	5	6	0.4	1	0.8	-0.22
	3	1	75	93	1.52	1.6	1.3	0.28
	Missing	***	2	2#	0.37			-0.13
AUW4_D	3	0	14	17	1.11	1.7	1.3	-0.12
	5	0	33	41	1.23	1.6	1.3	-0.13
	4	0	22	27	1.38	0.7	0.9	-0.03
	2	0	1	1	1.7	0.5	0.9	0.02
	1	1	11	14	2.51	1.1	0.8	0.35
	Missing	***	2	2#	0.37			-0.13
AUW5_D	5	0	5	6	0.18	0.3	0.3	-0.26
	2	0	13	16	0.58	0.9	0.8	-0.31
	1	0	2	2	0.86	0.6	0.6	-0.08
	3	0	27	33	1.48	1.3	1.3	0.03
	4	1	34	42	1.93	1.2	1.2	0.35
	Missing	***	2	2#	0.37			-0.13
Work1_T	3	0	1	1	-0.49	0.4	0.2	-0.21
	4	1	75	99	1.54	0.9	1	0.21
	Missing	***	7	8#	0.2			-0.3
Work2_D	4	0	2	3	-0.22	0.3	0.2	-0.25
	1	0	8	11	0.5	0.7	0.5	-0.31
	2	0	7	9	0.51	0.7	0.5	-0.28
	3	0	4	5	0.85	1	0.7	-0.14
	5	1	55	72	1.9	0.9	0.8	0.55
	Missing	***	7	8#	0.2			-0.3
Work3_D	5	0	1	1	0.06	0.5	0.3	-0.15
	2	0	7	9	0.6	0.8	0.9	-0.26
	4	0	6	8	0.66	0.9	0.8	-0.22
	1	0	3	4	1.26	1.5	1.3	-0.05
	3	1	59	78	1.75	0.9	0.9	0.38
	Missing	***	7	8#	0.2			-0.3
Work4_D	2	0	2	3	0.1	0.6	0.4	-0.21
	3	0	7	9	0.38	0.7	0.5	-0.32
	5	0	1	1	0.69	1	0.6	-0.08
	4	0	1	1	1	1.4	0.9	-0.05
	1	1	65	86	1.7	0.9	0.9	0.4
	Missing	***	7	8#	0.2			-0.3
Work5_D	2	0	5	7	0.58	0.7	0.6	-0.22
	4	0	6	8	0.65	0.7	0.6	-0.22
	3	0	24	32	0.91	0.8	0.8	-0.37
	1	0	1	1	1.7	1.4	1.3	0.02
	5	1	40	53	2.12	0.8	0.8	0.57
	Missing	***	7	8#	0.2			-0.3

Graf1_T	3	0	1	1	-0.49	0.4	0.2	-0.18
	5	0	5	6	-0.01	0.9	0.8	-0.3
	1	0	2	3	1.9	4.5	2.7	0.07
	2	1	71	90	1.51*	0.9	0.9	0.27
	Missing	***	4	5#	1.52			0.02
Graf2_D	1	0	6	8	0.86	0.6	0.6	-0.13
	3	0	36	46	1.15	1.1	1	-0.19
	2	0	5	6	1.49	1	1.1	0.02
	4	0	4	5	1.68	1.7	2	0.05
	5	1	28	35	1.77	1.4	3.9	0.23
	Missing	***	4	5#	1.52			0.02
Graf3_D	2	0	16	20	0.81	1	1.1	-0.24
	3	0	7	9	0.83	1.4	1.2	-0.14
	4	0	2	3	1	0.9	0.8	-0.05
	1	0	6	8	1.11	1.2	1.7	-0.07
	5	1	48	61	1.73	1.1	1.1	0.34
	Missing	***	4	5#	1.52			0.02
Graf4_D	5	0	6	8	0.69	0.6	0.5	-0.17
	1	0	10	13	0.95	0.9	0.8	-0.14
	2	0	22	28	1.11	1.3	1.3	-0.14
	4	0	7	9	1.5	1.3	1.4	0.03
	3	1	34	43	1.82	1.3	1.3	0.3
	Missing	***	4	5#	1.52			0.02
Graf5_D	4	0	13	16	0.66	1.1	1	-0.27
	1	0	6	8	1.08	1	0.9	-0.08
	2	0	8	10	1.4	1.4	1.7	0
	5	0	8	10	1.61	1.6	1.6	0.06
	3	1	44	56	1.62	1.4	1.4	0.2
	Missing	***	4	5#	1.52			0.02

** Average ability does not ascend with category score*

 # Missing & includes all categories. Scored % only of scored categories.

Interview
Shiken: A Journal of Language Testing and Evaluation in Japan
2025, Vol. 29(1) 31–35
DOI: 10.37546/JALTSIG.TEVAL29.1-2
ISSN: 1881-5537

Assessing for student success: An interview with Dr. Liying Cheng

Sachi Oshima[1] and Jeffrey Martin[2]
oshima.s@mc.cgu.ac.jp
martinjpsla@gmail.com
[1] *Chuo Gakuin University*
[2] *Daito Bunka University*

Bio

Dr. Liying Cheng is Professor and Dean of the School of Education at the City University of Macau. Before taking up this role, she served as Professor and Director of the Assessment and Evaluation Group at the Faculty of Education, Queen's University, Ontario, Canada. Dr. Cheng is internationally recognized for her research on washback illustrating the global impact of large-scale testing on instruction, the relationship between assessment and instruction, and research on how to align assessment practices with student success and language development.

Keywords: evolving practices in assessment, alignment, feedback, grading, educational and cultural contexts, teacher development

The Testing and Evaluation SIG of JALT was honored to help sponsor Dr. Cheng as a plenary speaker at the 50th JALT International Conference, held at the Shizuoka Granship in November 2024. In a joint opening plenary with Dr. Andy Curtis, she reflected on JALT's 50 years of contributions to language education and shared insights on Opportunity, Diversity, and Excellence. Dr. Cheng also conducted a workshop titled *Assessing for Student Success*, where she emphasized the importance of assessment as a process for supporting learning through alignment, fairness, and engaging students in discussions about their learning progress.

The following interview was conducted online and via email correspondence following the conference. Building on the insights that she shared, we explored a range of topics, including her initial interest in language assessment, the CARE framework for academic acculturation in supporting student success (Cheng, 2020), types and purposes of assessments, and the application of these ideas in diverse L2 learning contexts.

Dr. Cheng, thank you for taking the time to speak with us. It was a pleasure to hear your plenary and to be in your workshop at JALT2024. We'd like to know more about the ideas you shared at the conference.

It's my pleasure to do this interview with the Testing and Evaluation SIG of JALT following the 50th JALT International Conference! I'd like to thank you for making the time in your busy schedule to conduct the interview.

At the JALT2024 conference in Shizuoka, Japan, you mentioned differences between your formative education in Beijing, China, and your graduate education in the UK, and that they brought your attention to language assessment. For readers of Shiken, what were some of the formative experiences that helped shape your subsequent work in this area?

It's quite a big question for me to answer. And between the experience I had in China and the experience I had in the UK doing my master's degree, I have to say that most of the experience was really in what differential teaching and learning were like. I came from a very traditional testing background in China, where testing was used for selection, but it was not the case in the UK then. For example, instead of testing as a selection purpose, part of my master's research involved a test called TEEP (Teaching English for Educational Purposes; https://www.reading.ac.uk/isli/english-language-tests/teep). It is designed to show whether a student has a level of English sufficient for degree-level study in UK higher education. The test was designed for international students who wanted to study there and would start what was called the Pre-sessional Course. This was instead of a pass/fail situation but based on a criterion: depending on their mark, they would do a shorter or longer number of weeks in the pre-sessional courses, which were like ESL courses before students actually joined study courses in subject areas.

Studying in the UK was a shock for me. I think the biggest shock was that I had been learning through Chinese before then, but in the UK, I had to learn in English (the medium of instruction). The way of learning there and then was also different. At that time, the program was very tough. I was at the University of Reading, which had one of the top three applied linguistics programs in the UK then. Back at that time, an example of the strictness was that, while we were required to do group presentations, projects, and tests outside the classes, for the final course paper, we had to submit it by 4 o'clock to the departmental secretary. If we didn't submit it on time, we basically failed that course.

That academic testing situation was very tense for me. It reminded me of what I experienced in China around that time. The assessment field has really changed over the past 30 years. There was a major reform in education in 1998 when Black and Wiliam (1998, 2010) published their papers entitled *Inside the Black Box*. Their work was the breakthrough that brought the core value of assessment to support teaching and learning. Before that, testing done almost everywhere was primarily for selection.

In my initial education in China, testing was conducted at the national scale. At the University of Reading, assessment was really focused on measuring how much you had learned. It was based on criterion-referenced standards, meaning if you didn't meet the criteria, you would fail. It was a type of selection within the university's academic program. This was 1994, when fairness was defined differently. For example, nobody had personal computers. We had to go to a certain place to use shared resources. For example, I remember using a dot matrix printer at another department and running back to the other half of campus to meet the 4 o'clock deadline.

You mentioned your framework "CARE: Key to Academic Acculturation" in your plenary speech. Can you first elaborate on the components of this framework? C: Compassion; A: Acquisition; R: Respect; E: Evaluation. Then, could you describe differences in how this framework might be applicable in ESL contexts and EFL contexts? We thought of asking you about potential differences in applying CARE since you have taught both in Canada and in Macau. Many Shiken readers work in the EFL contexts of Japan.

Compassion is the ability to put ourselves into others' shoes. And so that means having the ability to feel what other people feel. This is crucial when it comes to the differences in teaching and learning in a diverse higher education context. During my time in Canada, for example, I was teaching a course which had both master's students and PhD students. We had domestic students who were born and grew up in Canada, international students, and short-term exchange students. They represent a range of diverse views regarding teaching and learning. In addition, Canada's situation is complex as it was not obvious from the outlooking surface as to who is a domestic or international student in a classroom due to immigration, so the differences in teaching and learning were embedded.

I use the term *Acquisition* to refer to both intentional and unintentional learning. I adopt this as a second language acquisition term, as used by Stephen Krashen (1981). *Respect* is really about how we deal with differences among others, particularly given the different experiences we all have. The *Evaluation* piece involves making a judgment, which I think is closely related to critical thinking skills. Critical thinking, to me, is the ability to be able to solve new problems, especially in the higher education context, where students come from many different cultural and educational backgrounds. The presence of diverse groups of people coming from different backgrounds brought a great deal of tension to the Canadian higher education classroom. So, I developed *CARE: Key to Academic Acculturation* (Cheng, 2020; Cheng & Fox, 2008).

You asked about the application of the four components of my CARE model by teachers in EFL and ESL contexts. I don't think there's a difference when we use the model between an EFL class or an ESL class. The idea of Compassion should work for all human beings. Compassion, Acquisition, Respect, and Evaluation should happen in and applied to any learning and teaching context. We need to support our learners in developing those four aspects of competences and skills, because it is our foundation to be a better human being.

About *Acculturation*, I think it is the context where I developed CARE. Acculturation is a psychological term. In Canada, acculturation is used to refer to the acculturation on both sides, i.e., I understand you as the host, and you understand me as the newcomer. We all have our own culture—ways of thinking and doing. So, acculturation is about understanding each other. When we talk about academic study, acculturation is also important. If I'm an international or immigrant student and want to succeed in Canada, I need to learn what good teaching or good learning looks like, because that's how my teachers and instructors assess me.

But at the same time, as instructors, we need to understand the backgrounds of our students. For instance, I ran a workshop in Canada on how to pronounce Chinese names, which is a big deal because Chinese names are often mispronounced. It's very difficult for English speakers. I think we had a chance to really help each other. And the more we understood each other, the better we could appreciate our shared humanity. Respecting every culture is important. We need to celebrate the differences, yet we need to recognize how challenging and hard it is to teach and learn in a diverse context.

Acculturation is the larger context where CARE can be used. I actually talk about CARE in Macau, even when all my students are Chinese students. There are still differences. Acculturation in this context means that students and teachers may have different perspectives but can come to a shared understanding.

In your workshop titled "Assessment for Student Success," you introduced three types of assessment. Could you describe them and their differences? For those interested in distinguishing and implementing these ideas, what research or teaching resources would you suggest?

Assessment *for* Learning refers to the process of seeking and interpreting evidence for use by students and their teachers to decide where the students are in their learning process, where they need to go, and how best to get there. Assessment *of* Learning refers to the assessment that happens after learning has occurred and aims to determine whether or not learning has taken place. These assessments are used to make statements about students' learning status at a particular point in time. Assessment *as* Learning occurs when students reflect on and monitor their progress to inform their future learning goals. It is regularly occurring, both formal and informal, and helps students take responsibility for their past, current, and future learning (Cheng & Fox, 2017).

You also asked whether self-assessment, peer-assessment, and peer feedback can be included in Assessment *as* Learning. Yes, but peer-assessment can also fall under Assessment *for* Learning, because that's what we do in the classroom on a daily basis. We do peer-assessments; we do self-assessments as well. This is also true in your teaching context. For example, in your own article (Martin & Oshima, 2024), you discussed how students can engage well in course objectives and enhance their sense of agency when they can help lead the learning of their peers. The practical tips you gave for productive and receptive L2 coursework also demonstrated adaptability in different teaching contexts.

These three types of assessment happen in our classrooms all the time, yet at different stages of the instruction. Assessment *of* Learning tends to happen at a specific period of time toward the end of an instruction period. So, it's not something that always occurs. Assessment *of* Learning tends to involve summative tasks, combination of assignments, a quiz, a test, a project, a presentation, or maybe a book report—something that evaluates cumulative learning. As we wrote in our book (Cheng & Fox, 2017), the idea is that students themselves understand their learning situation, their starting point, and how to improve themselves. These three assessment types overlap. In practice, they are all part of teaching and learning.

The definition of these three assessment types of assessment are not my own terms. The field of assessment has been working on these ideas for some time. The initial two terms were *summative* assessment and *formative* assessment (Scriven, 1967). These have been used in our field for a while, and now we've started to use these three terms: Assessment *for* Learning, Assessment *of* Learning, and Assessment *as* Learning.

In your workshop, you also mentioned aspects of the purposes and uses of instruction, diagnosis, and grading in assessment. How do you advise classroom teachers to use their limited time most effectively to best achieve these important aspects of language assessment?

Most teachers engage with these three purposes—instruction, diagnosis, and grading. Among these, instruction is the broader category, as we do that in our classrooms all the time. Diagnosis and grading, however, often require more attention, especially in terms of teacher professional development.

Diagnosis comes with feedback, which I think is very important. We are all teachers, and we know that providing effective feedback can be a challenge. Sometimes, we default to say, "Good job, you've done a good job." But most of the research, especially in the past 10 years on feedback, shows that the key for effective feedback is to focus on the task rather than the person. For example, if the task is essay writing, we should focus on specific objectives of the writing. Let's say the goal is to teach students to use linking words or improve transitions between paragraphs and sentences. You can center your feedback around that. Feedback is a technique, and if you work with a class of students, you don't have to give feedback on everything at once. For instance, in one session, you might focus on coherence—how ideas are linked and flow together. Next time, you might focus on the use of tense. This is especially challenging for Chinese students because, in Chinese, verbs do not change to reflect tense. In terms of diagnosis, for me, it's about knowing the students and identifying how to support them in moving forward in their learning.

Grading is also very important. I've been working on grading issues. I received a major SSHRC grant (the Social Sciences and Humanities Research Council of Canada). We spent five years investigating grading issues across countries (see Cheng et al., 2020). Grading is different from feedback; while feedback is part of the process, grading represents the end of an assessment period. That's when you give a mark. Often, this mark is a combination of various factors. For example, teachers often incorporate multiple components into grades. It's crucial that teachers ask themselves: "Does this mark—let's say, 76—accurately reflect the student's learning? Or does it also reflect something else, like their classroom behavior or participation?" Sometimes, the school or departmental policies require teachers to factor in attendance or participation when assigning grades. But grading is a much broader instructional issue, as schools and universities often have guidelines dictating how to grade. Grading is also a high-stakes process, as it directly affects students' wellbeing. Your decision can

cause students to feel rewarded for their hard work over the semester or, conversely, feel frustrated, upset, or demoralized. That said, teachers cannot give everyone an "A" mark, because that would defeat the purpose of grading. Grading is ultimately about decision-making, which Fox and I detailed in our 2017 book as a process that teachers need to approach carefully.

I believe all teachers need to learn more about feedback and grading. If we have the opportunity to offer more workshops, I would dedicate an hour to feedback and another to grading. More research has been published on grading than on feedback. Together with colleagues, we wrote a paper (Cheng et al., 2020) specifically about the grading differences in Canada and China. This was part of our comparative study that provided a detailed examination at grading practices across countries. This study was not specific to language classrooms but general education classrooms. It included data from teachers, students and parents. The data were very rich, though messy, as they reflected how teachers (and other stakeholders) have approached grading in various contexts.

For our last question, we were curious about how your concepts apply, or have been applied, with learners of other languages. Chinese, for example, is a foreign language studied by many people in other cultures in the world. What are your thoughts about how researchers, teachers, and language programs of other languages are implementing these ideas?

I don't see any differences in how to apply the concepts across different language teaching and learning contexts. I think what we've talked about mostly in terms of assessment applies to all classrooms. I work with teachers on the basic principles of how to use assessment to support students' learning and teachers' instructional practices. If we look at the field of education, we ask teachers to think about their unique classroom contexts and how they're going to apply these theories to their classrooms (Cheng, 2023; Cheng & Fox, 2017). I always start by explaining my own teaching and learning context when I talk about assessment. Then, I move on to provide guiding theories and practices for teachers to consider. I emphasize that teachers need to think about how they can adapt these ideas to their own classrooms, because every classroom is different. I always say, "Teaching is about context." Context is really important, and we need to trust teachers to make their own decisions based on their individual classrooms. They might say, "Okay, this concept might not work in my classroom, so I need to adapt it," or "I need a different pedagogy because my students or my subject matter are different."

To give an example of the importance of the context, respect is universal, but the way we express it can differ across cultural contexts. I'll give you one example. I spend a lot of time with my Chinese students talking about the difference between humility and humbleness. This humbleness is a huge concept, but humility is not talked about much in that context. I couldn't really find the corresponding word in Chinese to represent humility. So, I had conversations with my students, and we debated academically about which Chinese words might represent humility in a Chinese context. These debates were fascinating.

Another point that needs to be highlighted for all teachers is that our prior experiences with assessment are deeply influential and often shape our teaching. Research shows that our past learning experiences are one of the strongest predictors of how we teach. That's why one of the first things we emphasize in Canadian teacher education is reflective practice. We encourage teacher candidates to reflect critically—not just on what worked or didn't work in the past, but also on why something worked and whether it can be improved. Colleagues and I have conducted many qualitative studies about this. An example is Wu et al. (2014), where two of my students and I wrote about the relationship between assessment and motivation. The relationship is foundational in the teaching profession, not only in Canada, where I worked for 20 years, but also in other countries like China.

About the teaching of languages around the world other than English, we also have ethical guidelines for assessment and testing practices such as those from the International Language Testing Association (ILTA; https://www.iltaonline.com/page/CodeofEthics). At the end of the day, we have to remember that education systems are products of their societies. The world is far more internationalized than before, but each education system reflects the society it operates within. That's why compassion is so important. Compassion is about embracing differences. Whether we are teaching in Canada, China, or Japan, compassion allows us to see beyond cultural differences and focus on the shared human values embedded in education.

Thank you for sharing your valuable insights in this interview. We appreciate your contributions to the JALT2024 conference and we hope that this conversation adds meaningfully to the ongoing dialogue on assessment and language teaching.

Declaration of competing interests:
No conflicts of interest were reported.

References

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, *5*(1), 7–74. https://doi.org/10.1080/0969595980050102

Black, P., & Wiliam, D. (2010). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan*, *92*(1), 81–90. https://doi.org/10.1177/003172171009200119

Cheng, L. (2020). *Compassion, acquisition, respect, evaluation (CARE): Key to academic acculturation*. Human Rights and Equality Office, Queen's University. https://www.queensu.ca/hreo/together-we-are/compassion-acquisition-respect-evaluation-care-key-academic-acculturation/

Cheng, L. (2023). *Language classroom assessment* (2nd ed.). TESOL Press.

Cheng, L., DeLuca, C., Braund, H., Yan, W., & Rasooli, A. (2020). Teachers' grading decisions and practices across cultures: Exploring the value, consistency, and construction of grades across Canadian and Chinese secondary schools. *Studies in Educational Evaluation*, *67*, 100928. https://doi.org/10.1016/j.stueduc.2020.100928

Cheng, L., & Fox, J. (2008). Towards a better understanding of academic acculturation: Second language students in Canadian universities. *The Canadian Modern Language Review*, *65*(2), 307–333. https://doi.org/10.3138/cmlr.65.2.307

Cheng, L., & Fox, J. (2017). *Assessment in the language classroom: Teachers supporting student learning*. Palgrave McMillan.

Krashen, S. D. (1981). *Second language acquisition and second language learning*. Pergamon.

Martin, J., & Oshima, S. (2024). Using peer-led evaluation in productive and receptive L2 coursework. In B. Lacy, R. P. Lege, & P. Ferguson (Eds.), *Growth mindset in language education* (pp. 54–64). JALT. https://doi.org/10.37546/JALTPCP2023-07

Scriven, M. (1967). The methodology of evaluation. In R. W. Tyler, R. M. Gagne, & M. Scriven (Eds.), *Perspectives of curriculum evaluation* (pp. 39–83). Rand McNally.

Wu, Y., Cheng, L., & Bettney, E. (2014). Assessment and motivation: Perspectives from teacher candidates. In S. V. Nuland (Ed.), *Conference Proceedings of the 58th International Council on Education for Teaching (ICET) World Assembly: Moving forward in curriculum, pedagogy and leadership* (pp. 358–371). Oshawa, Canada: University of Ontario Institute of Technology. https://www.icet4u.org/docs/Canada_2014.pdf

Call for Papers

Shiken: A Journal of Language Testing and Evaluation in Japan is seeking submissions for publication in the October 2025 issue. Submissions received by 1 May, 2025 will be considered, although earlier submission is encouraged to allow time for review and revision. Shiken: A Journal of Language Testing and Evaluation in Japan aims to publish articles concerning language assessment issues relevant to classroom practitioners and language program administrators. This includes, but is not limited to, research papers, replication studies, review articles, informed opinion pieces, technical advice articles, and qualitative descriptions of classroom testing issues. Article length should reflect the purpose of the article. Short, focused articles that are accessible to non-specialists are preferred and we reserve the right to edit submissions for relevance and length. Research papers should range from 4000 to 8000 words, but longer articles are acceptable provided they are clearly focused and relevant. Novice researchers are encouraged to submit, but should aim for short papers that address a single research question. Longer articles will generally only be accepted from established researchers with publication experience. Opinion pieces should be of 3000 words or less and focus on a single main issue. Many aspects of language testing draw justified criticism and we welcome articles critical of existing practices, but authors must provide evidence to support any empirical claims made. Isolated anecdotes or claims based on "commonsense" are not a sufficient evidential basis for publication.

Submissions should be formatted as a Microsoft Word (.doc or .docx format) using 12 point Times New Roman font, although plain text files (.txt format) without formatting are also acceptable. The page size should be set to A4, with a 2.5 cm margin. Separate sections for tables and figures should be appended to the end of the document following any appendices, using the section headings “Tables” and “Figures”. Tables and figures should be numbered and titled following the guidelines of the *Publication Manual of the American Psychological Association, Seventh Edition.* Within the body of the text, indicate approximately where each table or figure should appear by typing “Insert Table x” or “Insert Figure x” centered on a new line, with “x” replaced by the number of the table or figure.

The body text should be left justified, with single spacing for the text within a paragraph. Each paragraph should be separated by a double line space, either by specifying a double line space from the Microsoft Office paragraph formatting menu, or by manually typing two carriage returns in a plain text file. Do not manually type a carriage return at the end of each line of text within a paragraph.

Each section of the paper should have a section heading, following the guidelines of the *Publication Manual of the American Psychological Association, Seventh Edition.* Each section heading should be preceded by a double line space as for a regular paragraph, but followed by a single line space.

The reference section should begin on a new page immediately after the end of the body text (i.e. before any appendices, tables, and figures), with the heading “References”. Referencing should strictly follow the guidelines of the *Publication Manual of the American Psychological Association, Seventh Edition.*