LanguageCert

EXPLORING TEST GENDER BIAS IN THE LANGUAGECERT SELT IESOL SPEAKING AND LISTENING TESTS

Michael Milanovic

Tony Lee

Leda Lamproloulou

David Coniam

Abstract

This paper examines LanguageCert's two-skills Secure English Language Testing (SELT) International ESOL Speaking and Listening (IESOL) tests. These tests are offered at CEFR levels A1, A2 and B1 and aimed at candidates applying for a visa to migrate or work in the UK, providing evidence of ability to operate in English. The purpose of the current study explored the unbiased nature of the two-skills test, affirming that test results may be seen to be robust and reliable.

An overview is first provided of where two-skills tests are positioned in the broader picture of language skills assessment. An analysis of the A1, A2 and B1 tests is then presented over the period when the tests were administered, i.e., from 2020 to 2023. With the three tests graded in line with CEFR difficulty levels, a study of test bias from the perspective of gender which was explored via differential item functioning (DIF) reported negligible-to-no bias.

Within the constraints of high pass rates, the paper concludes that the three SELT IESOL Speaking and Listening tests, perform within operational expectations. The SELT IESOL Speaking and Listening tests are robust tests, are functioning as intended and returning reliable results.

Introduction

In an era of communicative language teaching and assessment, there is a general recognition that assessment should cover all four language skills (see e.g., Guerrero, 2000; O'Sullivan et al, 2022; Powers, 2010). In the majority of assessment situations, evidence of ability in all four skills is the norm – in school situations and in applying for entrance to university etc – in part to encourage washback and for integrated instruction to be provided in all four skills. The conventional four-skills testing approach, which has been widely used in language assessment for decades, aims to comprehensively evaluate learners' language abilities across all four modalities, providing a comprehensive picture of their overall language proficiency. As language teaching methodologies have evolved and our understanding of language acquisition has deepened, some educators have, however, begun to question the efficacy and practicality of assessing all four skills in a single test.

There is a case for two-skills tests, specifically speaking and listening, where such as authenticity, efficiency, and alignment with communicative language teaching approaches, and the ability to use language for real-life communication is seen as a key competence.

It has come to be accepted that different language learners will exhibit differing levels of ability in the different language skills. Bachman (1985) argued that a divisible model of language ability with a general factor plus distinct traits is a plausible model for how language ability may be compartmentalised. Bachman (1990) extended the earlier research, examining various aspects of language proficiency, including the ability to use language skills separately and in combination.

It has been argued that listening and speaking are theoretically and practically not easily separable (see Douglas, 1997) and that the two skills should be integrated in assessment. Children learn their first language almost exclusively through listening and responding to spoken input, with some estimations that at least half the time spent in communicative interaction involves listening (see Wagner, 2018).

The two-skills speaking and listening test format has the potential to address several key concerns associated with four-skills tests. In contexts where reading and writing skills are not seen as relevant, a focus on the testing of speaking and listening skills, can create more authentic and communicatively meaningful assessment experiences. In this context, testing these skills also aligns with the principles of communicative language teaching.

Frost et al. (2011) state that while language assessment has traditionally focused on measuring the four skills independently, such a focus may be problematic since many 'real world' communicative acts involve the integration of two or more skills, as well as other non-linguistic cognitive abilities.

Gender is considered a key variable in terms of gauging fairness and lack of bias in high stakes tests (see e.g., Ozdemir and Alshamrani, 2020; Song et al., 2015). Against this backdrop, in the current study, gender is explored via DIF in the context of the LanguageCert SELT IESOL Speaking and Listening tests.

Two-skills Tests

A number of two-skills tests have been developed; their main features and focuses are summarised below.

Tavil (2010) reports the successful implementation of an integrated two-skills listening and speaking test which assessed candidates' oral/aural skills through information-gap tasks at a Turkish university.

Frost et al. (2011) investigated how candidates integrate stimulus materials into their speaking performances on an integrated listening-then-speaking summary task. They conclude that the use of an integrated listening and speaking task together with its associated rating scale functions well as a measure of speaking proficiency.

Lion et al. (2013) describe the use of the ALTA Clinician Cultural and Linguistic Assessment, an oral/aural Spanish Speaking and Listening Test administered to physicians in the USA by the ALTA language testing service. The situation required solely an oral/aural test since the study wished explicitly to evaluate American physicians' ability to communicate directly with Spanish-speaking patients.

Cao (2019) outlines the Computerized English Listening and Speaking Test (CELST) which was developed in 2011 and assesses English pronunciation, listening proficiency, interactional competence. The author claims that the CELST meets the requirements of a good oral test, by focusing on information exchange, creating contextualised situations and authenticity to incorporate interaction into language communication.

Rukthong and Brunfaut (2020) investigated listening in the context of integrated tasks such as listening-to-speak. They conclude that the listening/speaking summarisation test task which they developed illustrates that test-takers use a range of cognitive processes strategies in processing listening input.

Four providers offer two-skills Speaking and Listening tests at CEFR levels A1-B1 for visa applicants to meet UK Home Office English language requirements. These are LanguageCert, the IELTS SELT Consortium, Trinity College London and Pearson (see https://www.gov.uk/guidance/prove-your-english-language-abilities-with-a-secure-english-language-test-selt).

An overview of the LanguageCert SELT IESOL Speaking and Listening tests follows.

The LanguageCert IESOL Speaking & Listening Tests

The LanguageCert IESOL Speaking and Listening Test (IESOL S&L) series of graded examinations provide 'steps up the ladder' of proficiency and are suitable for non-native speakers of English who in particular need to demonstrate that they have met the required level of English as specified by the UK Home Office.

The qualifications demonstrate a candidate's ability to communicate using English in real life situations, as may be seen to be appropriate at the respective CEFR levels (A1 to B1 in this case). For details see https://www.languagecert.org/en/language-exams/english/languagecert-esol-selt.

LanguageCert IESOL Speaking and Listening Test Test Makeup

The LanguageCert International IESOL Speaking and Listening (IESOL S&L) tests are structured such that candidates respond to speaking and listening tasks which elicit a range of skills. Table 1 elaborates.

Table 1: Speaking and Listening Test tasks

Test Parts	*Skill and Focus*	*Task*
Part 1: Respond to questions on familiar matters and communicate personal information	A1 and A2: Give personal information. B1: Express opinions and ideas in addition to the above.	Give and spell name Give country/place of origin Answer three to four questions
Part 2: Initiate and respond appropriately in social situations	A1, A2, B1: communicate in real-life situations using a range of functional language to elicit or respond as appropriate. The sophistication and length of the expected candidate output increases through A1 to B1.	Two situations are presented by the interlocutor at each level and candidates are required to respond to and initiate interactions.
Part 3: Exchange information and opinions	A1 and A2: Exchange information to complete a simple task . B1: Co-operate to reach agreement/decision. The sophistication and length of the expected candidate output increases through A1 to B1.	Exchange information to identify similarities and differences in pictures of familiar situations at A1 and A2 levels. Hold a short discussion to make a plan, arrange or decide on something using visual prompts at B1.
Part 4: (subparts a & b): Understand a short monologue delivered by the marking interlocutor; deliver a short, uninterrupted talk on a relevant topic	A1 and A2: Demonstrate the ability to understand and use sentences and produce a piece of connected spoken English B1: Narrate, describe or communicate ideas and express opinion(s). The sophistication and length of the expected candidate output increases through A1 to B1.	Listen to the monologue and answer the questions. After 30 seconds of preparation time, talk about a topic provided by the interlocutor. Preliminary – half a minute Access – 1 minute B1 – 1 and a half minutes Answer follow-up questions .

The format of the tests and the nature of the assessment criteria reflect the broad multi-faceted construct underlying the Speaking and Listening tests. Communicative ability is the primary focus, while accuracy and range become increasingly important as the CEFR level of the test increases. For example, a level '3' for A1 grammar is defined as:

- *control of a restricted range of A1 grammar*
- *several errors occur with some A1 grammar*

In contrast, a level '3' for B1 grammar is defined as:

- *good control of a restricted range of B1 grammar*
- *errors occur with some B1 grammar*

Against this backdrop, candidate responses are evaluated using an analytic mark scheme which matches the CEFR descriptors. Separate marks are awarded by marking examiners for five aspects of speaking / listening ability in the output produced by candidates. This set of criteria ensures that a wide range of oral/aural skills are considered, thus enhancing the reliability and representativeness of test scores. Table 2 lays out the rating scale criteria.

Table 2: Rating scale criteria

Rating scale	Short form	Constructs assessed
Listening and Responding	LR	Understand interlocutor prompts and respond appropriately
Interactive Communication and Task Fulfilment	TF	Understand and maintain the interaction, and manage the tasks adequately for the level
Accuracy and Range of Grammar	ARG	Demonstrate a range and control of grammar for the level
Accuracy and Range of Vocabulary	ARV	Demonstrate a range and control of vocabulary for the level
Pronunciation, Intonation and Fluency	PIF	Connect utterances, maintain the flow and engage in effective communicative exchanges

All scales are rated on a five-point scale; for Listening and Responding, the score awarded is doubled to 10 points to represent the fact that two skills are being assessed, as well as balancing the representation of the listening construct in the scoring. The maximum possible score is 30, with 18/30 being a pass. Results for the test awarded are Pass or Fail and are accompanied by a score out of 100. While the same five criteria are applied across the three levels at which the test is offered, the demands posed by the criteria at a specific level reflect speaking and listening language ability expectations at that level.

Following accepted practice for analysing multiple facets in a performance test such as speaking, the best analytical practice involves the use of Rasch measurement since this enables different facets (candidate ability, examiner severity, task difficulty, for example) to be modelled together (see e.g., Coniam and Falvey, 1999; Hidri, 2018). In the Rasch model, a unified interval metric for measurement is obtained where the units of measurement ('logits') are evenly spaced along the measurement scale (Wright, 1997). With a common scale established for the test facets (in this case, different features in assessing speaking), different features can be examined and their effects monitored or controlled. Against this backdrop, DIF via Rasch measurement was used primarily to investigate the question of gender bias in the three tests.

Test Data

The data in the current dataset was compiled from tests administered over the period mid 2020 to early 2023. Table 3 provides details of sample sizes over the period.

Table 3: Sample detail

CEFR level	Candidates
A1	12,868
A2	5,758
B1	22,968

The largest candidature is at B1 level, reflecting the popularity of the respective visa type.

Purpose of the study and its Research Question

As mentioned earlier, the purpose of the study was to investigate whether acceptable quality levels were maintained in terms of the two-skills tests in relation to gender bias or more accurately, lack of it.

Test data and the Global Scale

At LanguageCert, tests, items, and candidate test results are linked to the CEFR via the LanguageCert Global Scale (Milanovic et al., 2023). Global Scale ranges for the three CEFR levels explored in the current study are provided in Table 4.

Table 4: Global Scale (GS) ranges

CEFR level	*GS level cut point*
A1	10
A2	20
B1	40
B2	60
C1	75
C2	90

Examiner, task and candidate facets were explored using Rasch measurement. This involved investigating where the different facets are located on the Global Scale, and where they are located relative to each other. The results of these analyses are not reported here given that the main focus of this paper is gender bias.

Table 5 first presents details of sample sizes for the different test levels and pass rates.

Table 5: Sample sizes and pass rates

CEFR level	*Candidates*	*Pass rate (%)*	*Mean (max. 30)*	*SD*	*SEM*
A1	12,868	11,043 (85.82%)	23.75	6.33	0.06
A2	5,758	5,136 (89.20%)	24.99	5.95	0.08
B1	22,968	21,976 (95.68%)	27.20	4.45	0.03

KEY: SD=Standard Deviation; SEM=Standard Error of the Mean

As may be seen, pass rates are high for all test levels. The pass mark, as mentioned above, is 18/30. All tests have a mean score considerably above this. Measurement error is nonetheless small. Part of the reason for such high pass rates may be attributed to ‘candidate readiness’. With the IESOL S&L tests, the situation is somewhat different from how ‘candidate readiness’ may be perceived in a school situation. In the latter situation, a student generally takes a test when they are ready for it, often as recommended by their teacher. In contrast, on the IESOL S&L tests, the candidate profile is different by virtue of the fact that the majority of candidates need proof of ability in order to be eligible for the issuing or renewing of a visa. In this context, many candidates sit an IESOL S&L test that is considerably below their actual proficiency level. Many IESOL S&L candidates, for instance, have lived in the UK for many years and are virtual native speakers, i.e., at CEFR C2 level. Such candidates nonetheless need to pass a B1, or even an A1, level test as proof of ability. This is the main reason that such high pass rates emerge.

For many candidates, then, whether they take an A1 or a B1 test makes little difference: many are still going to be C1 or above. The issue is further complicated by the high-stakes nature of the test where a pass is required in order to obtain a visa. School students taking a test which is suggested to be at their level generally accept and live with the results – even a fail grade. In contrast, many IESOL S&L candidates who are marginal and who failed a test the first time around will often retake the test until they achieve a pass. Such a situation exacerbates the high pass rates. In the current study, regarding candidates who have taken a test multiple times, only the candidate’s best result has been included in the dataset.

On a methodological point, high pass rates, it should be noted, complicate analyses. Statistical analyses generally need ‘space’ – i.e., a range of test scores – to be able to conduct sufficient, yet accurate, computations. The lack of such space – as with the current tests with pass rates above 85% – somewhat constrains statistical analysis.

Differential Item Functioning Analysis

This section presents an investigation of differential item functioning (DIF) into the key variable, gender. DIF analysis involves an exploration of whether any subgroup of candidates in a test is being unfairly disadvantaged. In the exploration of potential bias among subgroup types, gender is a key variable that is seen to be worthy of investigation (Ferne & Rupp, 2007).

Rasch-based methods (Roznowski & Reith, 1999) have come to be the preferred statistical mode of analysis for DIF in terms of identifying latent traits. One extension of DIF which has been used in previous studies is Differential Group Functioning (DGF). DGF involves grouping items into sets that share the same latent trait (e.g., Gierl et al., 2001). DGF, which is used in the current analysis, reports biases between candidates' actual responses against the estimated Rasch-calibrated item locations. For ease of reference, however, given the general acceptance of the term "DIF", it is "DIF" that is used in the current study.

In analytic terms, the most demanding category – indicating moderate-to-large DIF strength – is stated as being greater than 0.64 logits (Zwick, 1999). In LanguageCert terms, 0.64 logits equate to approximately 10 Global Scale points. It is this threshold which is taken as the limit for indicating possible bias in the current study.

IESOL S&L candidates are not required to provide demographic detail when registering for the test. Consequently, certain detail is incomplete. Table 6 provides details of test sample sizes and the number of candidates who supplied details of their gender.

Table 6: Sample size and gender detail

CEFR level	*Candidates*	*Stating gender*	*Male*	*Female*
A1	12,868	7,167 (55.70%)	1,751 (24.43%)	5,416 (75.57%)
A2	5,758	1,207 (20.96%)	388 (32.15%)	819 (67.85%)
B1	22,968	6,457 (28.11%)	3,130 (48.47%)	3,327 (51.53%)

Among the three tests, more females than males provided their demographic details, with A1 candidates being the most responsive test group of the three. The available sample size is nonetheless sufficiently large to be able to conduct DIF analyses.

Following the analysis of rating scales above, a DIF analysis was conducted on gender against rating scale. DIF size differences between DIF and actual Global Scale values are provided in Table 7 below.

Table 7: DIF by gender

Gender	Rating scale	**DIF size**
F	TF	0.32
F	ARG	0.86
F	ARV	0.77
F	PIF	0.54
F	LR	-2.07
M	TF	-0.47
M	ARG	0.45
M	ARV	-0.17
M	PIF	0.00
M	LR	0.00

As can be seen from the table above, the largest DIF value was 2.07, considerably below the proposed threshold of 10 scale points. From this, it can be concluded that neither gender can be seen to be unfairly disadvantaged with ratings awarded on the IESOL S&L tests.

Conclusion

This paper has presented an examination of LanguageCert's two-skills Secure English Language Testing (SELT) International ESOL Speaking and Listening tests. The purpose of the study has been to explore the quality of the test and the robustness of results with particular reference to gender bias.

The two-skills tests are offered at CEFR levels A1 to B1, being aimed at candidates who are applying for a visa to migrate, work or study in the UK. The ability focus is on oral/aural skills as evidence of spoken English proficiency.

The data which was used in the study was obtained from tests administered in the period 2020 to 2023.

Pass rates were high, with all tests reporting pass rates of 85% or higher – a reflection of the generally high ability of the candidature and the requirement that candidates possess a pass on a particular test if they are to meet certain UK visa or study requirements. Within these constraints, the three LanguageCert IESOL Speaking and Listening tests have been shown to function reliably, with examiners, tasks and rating being seen to be within operational li mits.

In closing, we would therefore state that the SELT IESOL Speaking and Listening tests may be considered robust, that they function as intended, and provide unbiased results.

References

ALTA Language Testing Services. (2023). Clinician cultural and linguistic assessment (CCLA). http://www.altalang.com/language-testing/ccla.aspx.

Bachman, L. F. (1985). An examination of some language proficiency tests from a communicative viewpoint. ERIC.

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

Douglas, D. (1997). Testing speaking ability in academic contexts: Theoretical considerations. TOEFL monograph series, 8. Princeton, NJ: ETS.

Frost, K., Elder, C., & Wigglesworth, G. (2012). Investigating the validity of an integrated listening-speaking task: A discourse-based analysis of test takers' oral performances. Language Testing, 29(3), 345-369.

Guerrero, M. D. (2000). The unified validity of the four skills exam: Applying Messick's framework. Language Testing, 17(4), 397-421.

Hidri, S. (2018). Assessing spoken language ability: A many-Facet Rasch analysis. Revisiting the assessment of second language abilities: From theory to practice, 23-48.

Lee, T., Milanovic, M., & Pike, N. (2022). Equating Rasch values and expert judgement through externally-referenced anchoring. International Journal of TESOL Studies, 4(1), 187-202.

Lee, T., Papargyris, Y., Milanovic, M., Pike, N., & Coniam, D. (2023). Aligning LanguageCert SELT tests to the LanguageCert Item Difficulty (LID) scale. London, UK: LanguageCert.

Linlin, C. (2020). Comparison of Automatic and Expert Teachers' Rating of Computerized English Listening-Speaking Test. English Language Teaching, 13(1), 18-30.

Lion, K. C., Thompson, D. A., Cowden, J. D., Michel, E., Rafton, S. A., Hamdy, R. F., ... & Ebel, B. E. (2013). Clinical Spanish use and language proficiency testing among pediatric residents. Academic Medicine, 88(10), 1478-1484.

Milanovic, M., Pike, N., Papargyris, Y., Lee, T., & Coniam, D. (2023). The LanguageCert Global Scale. London, UK: LanguageCert.

O'Sullivan, B., Motteram, J., Skipsey, R., & Dunlea, J. (2022). The importance of the four skills in the Japanese context.

Ozdemir, B., & Alshamrani, A. H. (2020). Examining the Fairness of Language Test Across Gender with IRT-based Differential Item and Test Functioning Methods. International Journal of Learning, Teaching and Educational Research, 19(6), 27-45.

Powers, D. E. (2010). The case for a comprehensive, four-skills assessment of English-language proficiency. R & D Connections, 14, 1-12.

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.

Song, X., Cheng, L., & Klinger, D. (2015). DIF investigations across groups of gender and academic background in a large-scale high-stakes language test. *Papers in Language Testing and Assessment*, *4*(1), 97-124.

Tavil, Z. M. (2010). Integrating listening and speaking skills to facilitate English language learners' communicative competence. Procedia-Social and Behavioral Sciences, 9, 765-770.

Wagner, E. (2018). Assessing listening. In Ockey, G. J., & Wagner, E. (2018). Assessing L2 listening: Moving towards authenticity, pp. 29-44. John Benjamins Publishing Company.

LanguageCert is a business name of PeopleCert Qualifications Ltd, UK company number 09620926

DISCLAIMER
This publication is designed to provide helpful information to the reader. Although care has been taken by LanguageCert in the preparation of this publication, no representation or warranty (express or implied) is given by LanguageCert with respect as to the completeness, accuracy, reliability, suitability or availability of the information contained within it and neither shall LanguageCert be responsible or liable for any loss or damage whatsoever (including but not limited to, special, indirect, consequential) arising or resulting from information, instructions or advice contained within this publication.

Language
Cert