8th International Conference on Higher Education Advances (HEAd'22)
Universitat Politècnica de València, València, 2022
DOI: http://dx.doi.org/10.4995/HEAd22.2022.14514

Validation of a High-Stakes Test: GA IESOL Listening Unit

Snezana Mitrovic[1], Emanuela Botta[2], Giusi Castellana[3]
[1]Department of Medicine and Psychology, Sapienza University of Rome, Italy, Department of Medicine and Psychology, Sapienza University of Rome, Italy, [3]Department of Education, Roma Tre University, Italy.

Abstract

The purpose of this study is the validation of the GA Classic IESOL examination at CEFR level B2 (Reading, Listening, Writing and Speaking), as a criterion-referenced achievement test, using the multi-trait, multi-method approach. The data analysed at the moment pertain to the Listening unit and have been studied to examine the content validity, reliability of the scores and unidimensionality of the construct. The Listening unit consists of three tasks (22 items) in four different versions (D-E-F-G). Four separate statistical analyses were performed and all revealed excellent values of internal consistency. As for unidimensionality of the construct, excellent values were revealed for versions E, F and G and acceptable for version D. Similar analyses on the reading test are planned, with the aim of reaching a multitrait-multimethod analysis of the examination as a whole.Further steps of the study will be the IRT equating procedure of the four versions.

Keywords: *listening ability; language assessment; international examinations, validity, reliability.*

1. Introduction and Aims

Gatehouse Awards IESOL Classic exam is an internationally recognised exam of English as a foreign language, designed to assess all four language skills or abilities: reading, listening, writing and speaking, and consists of four units, each of them testing one of the abilities. The awarding body, as well as the examination, is recognised by Ofqual (The Office of Qualifications and Examinations Regulations)[1], a non-ministerial department that regulates qualifications, examinations and assessments in England. The exams are delivered in a number of countries, through the national distributors and approved examination centres.

The aim of the study is the validation of the GA Classic IESOL examination at CEFR level B2 (consisting of Reading, Listening, Writing and Speaking), as a criterion-referenced achievement test, using the multi-trait, multi-method approach (Fiske & Campbell, 1959; Bachman & Palmer, 1982). The data available at the moment pertain to the Listening unit and have been analysed to examine the content validity, reliability of the scores, as well as the unidimensionality of the construct.

2. Listening Ability as the Listening Test Construct

Most if not all internationally recognised examination boards test listening as part of their IESOL examinations. The listening unit is most often administered as a separate test. This division is likely to have been based on the first models for describing language proficiency, which distinguished skills (reading, listening, writing and speaking) from components of language (Bachman, 1990). According to Buck (2001), a description of listening ability can be used as the basis for defining the listening construct, especially when we believe that the test performance is an indicator of an underlying competence.

When it comes to foreign language testing, interest in test objectivity and internal consistency first appeared during the psychometric-structuralist era (Weir, 2005), which started in the 1960s when the first multiple-choice tests were used, as a result of the need to deliver a large number of exams each year. Theories of communicative competence and performance only appeared in the 1970s and helped overcome the shortcomings of these tests, which focused mostly on psychometric characteristics. This is when terms such as ‘authenticity’, ‘context’ and ‘criterion-referenced’ were first mentioned in relation to foreign language assessment (Morrow, 1981; Fulcher, 2000). A large number of different types of tests proposed over the years have stayed, and the existing tests most often combine the different types of items and testing approaches proposed over the years (Morrow, 1981).

[1] https://www.gov.uk/government/organisations/ofqual

The fact remains that whichever test type is used, there needs to be a well-defined test construct, which in the case of listening tests would be a description of the listening ability, since it is the ability of the test-takers that we are most often interested in (Buck, 2001). After the Council of Europe first recommended the Common European Framework of Reference as a standard of language proficiency in 2001, English language awarding bodies defined the listening ability as a listening test construct in accordance with the CEFR definition of listening and its levels. Specifically, the GA IESOL Classic B2 listening unit is mapped to the CEFR and its definition of the listening ability at the B2 level. CEFR provides global scales for each of the six levels of foreign language proficiency (A1, A2, B1, B2, C1 and C2), as well as illustrative scales for each of the abilities. The competences that successful B2 candidates need to demonstrate, as detailed in the GA Classic IESOL Specification[2], in the four domains of the CEFR (public domain, personal domain, educational domain, and occupational domain) reflect the illustrative descriptor for listening comprehension (Council of Europe, 2000, p. 66) and are as follows:

- can understand standard spoken language on both familiar and unfamiliar topics,
- can follow the essentials of lectures, talks and reports,
- can understand animated conversation between native speakers,
- can understand the main ideas of propositionally and linguistically complex speech on both abstract and concrete topics,
- can follow complex lines of argument, provided the topic is reasonably familiar,
- can understand speech delivered in a standard dialect and at normal speed.

Listening comprehension ability encompasses different types of listening (Wilson, 2008). One can listen for gist or for general idea of what is being said, or for specific information, which is when we only need to understand a very specific part. Furthermore, we can listen in detail, when, for example, we need to find errors and cannot afford to miss anything. Finally, there is inferential listening, called that way as it may involve inferring, which is what we do when we want to know how the speaker feels (Wilson, 2008, p. 10). For each of the items in the listening unit, candidates are required to employ one of the types of listening.

Content domain specification, that is examination specification, being considered a necessary requisite for the test construct and content validity (Bachman, 2002; Brown 1996) is provided in detail and includes functions and notions, grammar, discourse markers, topics and key language items for the B2 level.

[2] https://www.gatehouseawards.org/wp-content/uploads/GA-Qualification-Specification-Classic-IESOL-A1-C2-V3-1.pdf

3. Method and Participants

The B2 level listening unit consists of three tasks, with a total of 22 items. In Task 1, candidates listen to a conversation between two speakers and answer six multiple-choice questions (A, B or C) about the content of the conversation. The candidates hear the recording twice. In Task 2, candidates listen to a monologue (e.g., news, talk, presentation or instructions) and answer 8 multiple-choice (A, B or C) questions about the content of the recording. In Task 3, candidates listen to three different speakers presenting their opinions on the same subject. Candidates then match eight statements to the correct speaker. The candidates hear the recording for each of the tasks twice and are given one minute to read the questions before the recording is played.

The data set used for this study originates from the online examinations held from October 2020 to November 2021. Four forms of the listening test were administered in that period (VD, VE, VF and VG) and for each of the forms analyses have been performed. A total of 597 candidates completed a randomly assigned version: 145 candidates completed form VD, 147 form VE, 162 form VF and 143 candidates completed form VG.

Examination tasks and items are designed by trained writers, after which they are reviewed by two other trained item writers. At this stage, the item can be accepted without further changes, revised in line with the feedback provided by reviewers or rejected. At this stage, the accepted items enter the live question bank, after which test forms are designed.

4. Data Analyses

The approach to validity employed in ESOL Classic examinations is consistent with Messick's "unitary" view of validity (1995) and the principle that the validity of a test resides in the test scores and score interpretations. In addition, it is also in line with Weir's view that different types of evidence are needed to demonstrate the validity of test scores. These different types of evidence are seen as complementary and not as alternatives (Weir, 2005). While the definition of the construct addresses content validity, the statistical analyses performed address the issues of reliability, criterion-related and construct validity.

The forms were calibrated with respect to the difficulty parameter *b*, to verify the fit to the Rasch model and to identify items that might need to be amended or replaced. xCalibre 4.2 was used for the calibration. The following were estimated / calculated in this phase: the discrimination index of the items, R (Point Biserial Correlation), which expresses the correlation of an item with others of the same test version, assuming $R > 0.20$ as the lower limit value, and the main fit indices, such as the standardised residual (zResid), Infit and Outfit. According to Hambleton et al. (1991) and Agresti and Finlay (2012), items that are acceptable are the ones that have a zResid in the range of -2 to 2 and not significant, and Infit

and Outfit values between 0.80 and 1.20. In the analysed forms, only one item with an inadequate zResid value was identified. As explained in Botta (2021) the Infit index (InMSQ, Infit Mean Square) is a fit statistic that tends to assume high values when there are misfits due to students who respond correctly to difficult items but not to easy items, or alternatively to easy and difficult items, it is consequently high even in items with low discrimination. Index values greater than 1 indicate misfit to the model and undermine the validity of the measure. Values lower than 1 indicate a local deficit in stochastic variability; low but not extreme values do not disturb the significance of the measure. Values greater than 1 indicate underfit, the real data are not very predictable from the model and generally lower than that estimated by it, while values lower than 1 indicate overfit, the real data are higher than that estimated by the model. In our case, as can be seen in table 1, items with a high Infit index are very rare. The more frequent ones, which, however, do not influence the validity of the measurement significantly, are items with low mean-squares. We can therefore say that, despite a low number of items, as a whole, the examination versions fit the model.

Table 1. Data analyses summary

Form	Number of Items	Number of Candidates	b_{min}	b_{max}	Cronbach's Alpha	Number of Items with discrimination R < 0,20	Number of Items with Infit > 1,2
VD	22	145	-1.64	2.78	0.807	3	1
VE	22	152	-1.12	2.82	0.818	3	3
VF	22	162	-1.39	2.10	0.843	2	1
VG	22	143	-1.82	1.46	0.727	7	0

Since this IRT model, as well as many others (Bachman, 1990) assumes that the items in the test measure a single or unidimensional ability or trait, EFA (Exploratory Factor Analyses) were performed to evaluate the dimensionality of the data, to confirm that the listening unit item responses form a unidimensional construct according to the Rasch model. As explained by Botta (2021) and Barbaranelli and Natali (2005) for unidimensional constructs, Cronbach's alpha value needs to be high, however, the opposite is not always true. This means that a high value of Cronbach's alpha is not the only indicator of unidimensionality and does not guarantee the unidimensionality of the construct on its own.

EFA were conducted on the data set, one for each version. Considering that all the items are dichotomous, with two possible item scores: correct (1) and incorrect (0), MPLUS 7.1 software was used for EFA (Muthén e Muthén, 1998-2010), which uses a specific procedure for the analysis of categorical data. For all four forms of the test, the analyses performed confirm the hypothesis of a single factor.

Table 2. EFA Results for each of the four versions

Version	RMSEA	RMSEA CI INF	RMSEA CI SUP	RMSEA Probability <= 0.05	CFI	TLI	Ratio between first & second eigenvalue (L1 / L2)	Number of items loading 0.30
VD	0.062	0.049	0.075	0.067	0.843	0.826	3.115	1
VE	0.029	0.000	0.047	0.978	0.963	0.959	3.612	2
VF	0.023	0.000	0.041	0.995	0.986	0.985	5.126	1
VG	0.016	0.000	0.039	0.997	0.977	0.975	2.630	3

As we can see, only version D has marginally acceptable goodness of fit values, while the other three versions have excellent values of RMSEA (Root Means Square Error of Approximation, Steiger & Lind, 1980; Steiger, 1990), CFI (Comparative Fit Index) and TLI (Tucker and Lewis Index). An RMSEA lower than 0.05 indicates a low error of approximation and shows that the model can relatively predict the data accurately, while values in the range of 0.05 to 0.08 indicate an acceptable level of error of approximation are considered an indication of fair fit. As regards CFI and TLI, values greater than 0.95 indicate relatively good model-data fit, while a CFI and TLI lower than 0.90 indicate a poor model-data fit. Furthermore, in all four forms, there are only few items with low factor loading while the scree plot confirms that there is only one factor in the data set.

Considering it is a criterion-referenced test, it is also evident that the distribution of students' skills and item difficulty tend to be misaligned so we can assume that a large part of students decide to take the examination once they have prepared themselves appropriately and thoroughly for CEFR B2 level, which is awarded to the candidates who reach 55% in each of the units, listening included.

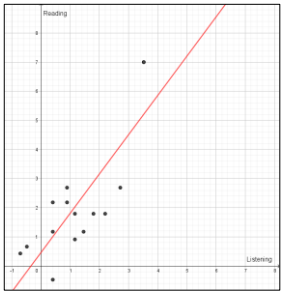

Figure 1. Correlation Listening – Reading Theta

With regard to criterion-related validity, a correlation analysis with reading tests was performed. The subsample of subjects who took the same reading test, (N = 16), was extracted for the F version. A comparison of the values of the ability parameter (θ) in listening and in reading, revealed a correlation coefficient equal to 0.811, as illustrated in Figure 2.

5. Conclusion

As already stated, the purpose of this study is the analysis and validation of a set of B2 level English language tests (listening, reading, writing and speaking). Validation of a test is essential whenever a new one is designed, especially when it comes to high-stakes examinations such as internationally recognised certifications. However, as both Messick (1992, as cited in Weir, 2005) and Weir (2005) maintain, not many test makers actually provide validity evidence or perform validation studies.

Validation is a continuous process meant to start together with the examination design. Schilling (2004) and Weir (2005) underline the importance of verifying the validity a priori, during the construct description phase, stressing that the more accurate the description of the construct one intends to measure, the more significant and reliable the validation statistical analyses will be and can consequently be used in the interpretation of the test results.

For this reason, it was decided to analyse the four versions of the examination in question in terms of the validity of content and construct, to verify their internal consistency and reliability. Four separate statistical analyses were performed and revealed excellent values of internal consistency for all four versions. As for unidimensionality of the construct, excellent values were revealed for versions E, F and G and acceptable for version D. The unidimensionality identified in the construct allows us to exploit all the properties of invariance that the Rasch model makes available and, since the sample was randomly selected and versions randomly assigned to the candidates, the identified differences between the four groups of candidates can be attributed to chance.

One of the next steps of the study will be the IRT equating procedure of the four versions. Similar analyses on the reading test and other units are planned, with the aim of reaching a multitrait-multimethod analysis of the examination as a whole.

References

Agresti, A. & Finlay, B. (2012). *Metodi statistici di base e avanzati per le science sociali. Milano:Pearson*

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

Bachman, L. F. (2002). Some reflections on task-based language performance assessment. *Language Testing*, *19*(4), 453-476.

Bachman, L. F., & Palmer, A. S. (1982). The construct validation of some components of communicative proficiency. *TESOL Quarterly*, *16*(4), 449-65.

Barbaranelli, C. & Natali, N. (2005). *I test psicologici: teorie e modelli psicometrici*. Roma, Carocci editore SpA

Botta, E. (2021). *Sperimentazione di un modello adattativo multilivello per la stima delle abilità in matematica nelle rilevazioni su larga scala*, Roma: Nuova Cultura

Brown, J. D. (1996). *Testing in Language Programs.* Upper Saddle River, NJ: Prentice Hall Regents.

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

Campbell, D.T. & D.W. Fiske., D.W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin* 56, 2:81-105

Council of Europe (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment.* Retrieved from: http://www.coe.int/t/dg4/linguistic/source/framework_en.pdf

Fulcher, G. (2000). The 'communicative' legacy in language testing. *System 28*(4), 483-497.

Hambleton, R. K., Swaminathan, H. & Rogers, H. J. (1991). *Fundamentals of Item Response Theory*, Sage Publ.

Messick, S. (1992). Validity of test interpretation and use. In M.C. Alkin (ed.), *Encyclopedia of Educational Research* (6th edition). New York: Macmillan.

Messick, S. (1995). Standards of validity and the validity of standards in performance assessment. *Educational Measurement: Issues and Practice, 14*(4), 5-8.

Morrow, K. (1981). Communicative language testing: revolution or evolution? In Brumgit, C. J. and Johnson, K. (Eds), *The communicative approach to language teaching* (143-57). Oxford: Oxford University Press.

Muthén, L. K. & Muthen, B. O. (1998-2010). *Mplus User's Guide. Sixth Edition,* Los Angeles, CA: Muthén & Muthén

Schilling, S. G. (2004). Conceptualizing the validity argument: An alternative approach. *Measurement, 2*, 178-182.

Steiger, J. H. & Lind, C. Statistically based tests for the number of common factors, *Annual meeting of the Psychometric Society*, Iowa City, IA, 1984.

Steiger, J. H. (1990). Structural Model Evaluation and Modification: *An Interval Estimation Approach, Multivariate Behavioral Research 25*(2), 173-180.

Weir, C. J. (2005). *Language testing and validation: An Evidence-Based Approach.* Basingstoke: Palgrave Macmillan.

Wilson, J. J. (2008). *How to teach listening*. Pearson Longman.