Conference Proceedings of the ALTE 8th International Conference

LANGUAGE ASSESSMENT FIT FOR THE FUTURE

ISSN: 2789-2344

ALTE Association of Language Testers in Europe

Conference Proceedings of the ALTE 8th International Conference, Madrid

Contents

Foreword

The ALTE 8th International Conference took place in Madrid in April 2023 under the title *Language Assessment Fit for the Future*. The overall theme was divided into three strands reflecting current innovative practices in the field: *Fit for the Digital Age; Diversity and Inclusion in Language Assessment;* and *Implementation of Frameworks*. Together they very much support the three main missions of the Association: to set standards for good quality in language assessment; to support the learning, teaching and assessment of a wide range of languages and promote the recognition of qualifications in these languages; and to maximise the positive impact of tests on society by making the connections between policy, research and practice.

The overall aim of the conference was to consider how language assessment shapes, and is shaped, by wider society, in order to be fit for the future. Out of the 98 refereed papers presented during the conference, 47 were offered and accepted for publication in the Proceedings, quite evenly spread over the three strands: *Fit for the Digital Age* (18 papers), *Diversity and Inclusion in Language Assessment* (17 papers), and *Implementation of Frameworks* (12 papers). The papers report on theoretical or empirical research, frequently linked to practical application and offering innovative perspectives on the given conference strand. Most of the authors represent academic research centres; quite a few however, also provide insights from the work of independent testing organisations, and some include valuable policy makers' observations. The reports offer a wide geographic perspective covering both European (13 countries) and global experiences (contributions from the USA, Japan, China, Mexico, and Brazil). Most papers focus on English but we can also find reports on other languages under investigation here – Czech, Finnish, French, German, Greek, Italian, Norwegian, Romanian, Spanish, Zhuang (China), and Sign Languages.

The Editor wishes to thank all authors for the considerable effort to elaborate on insights from their research and practice and for contributing to such a fascinating set of readings.

Waldek Martyniuk, December 2023

Fit for the Digital Age

Online testing: Investigating the candidates' attitudes and reactions

Letizia Cinganotto
University for Foreigners of Perugia, Italy

Abstract

The contribution aims at describing the ongoing research carried out at the Centre for Language Evaluation and Certification (CVCL) at the University for Foreigners of Perugia, Italy, related to the digital evolution of assessment and training after the COVID-19 pandemic. The main features of the different tests and certificates provided by the CVCL are addressed as a starting point, in particular the Italian language certificates named CELI, aligned with the CEFR and the methodological certificate named DILS-PG. The presentation highlights preliminary results of a study conducted with a sample of teachers and students aimed at investigating their reactions and attitudes towards different dimensions related to online testing, such as the candidates' timing, device use, anxiety level, digital literacy, etc. An online questionnaire as well as interviews with the respondents were arranged in order to collect useful information for further developments, in preparation for the possible future digitalization and innovation of the entire testing process at CVCL. The research begins with comparisons with 'pen and paper' version of the tests, in order to collect the participants' reactions and to get the best out of the traditional process. Main participants' comments are described and commented on according to the Framework Analysis.

Introduction

The Centre for Language Evaluation and Certification (CVCL) at the University for Foreigners of Perugia, Italy, has a long tradition in designing and delivering tests and issuing certificates of linguistic and methodological competence. CVCL is a member of CLIQ Association (Certificazione Lingua Italiana di Qualità), which includes the four institutions officially recognized by the Italian Ministries for issuing certificates of Italian as a Foreign or Second Language (University for Foreigners of Perugia, University for Foreigners of Siena, Roma Tre University, Dante Alighieri Society). CVCL is also a full member of ALTE and is actively involved in all the different activities and initiatives promoted at international level.

In particular, two main certificates are designed and delivered by CVCL in Perugia and in a wide network of institutions all over the world, which signed an agreement with CVCL:

- CELI: Italian Language Certification, which is aligned with the levels of the Common European Framework of Reference for Languages (CEFR, Council of Europe, 2001) and is delivered in two formats: 'CELI A', addressed to adolescents, and 'CELI I', addressed to migrants. Both exam formats includes a written part (reading comprehension, written production, listening comprehension, and use of language), which is assessed centrally by the CVCL, and an oral part, based on pictures, graphs and tables as input for communicative tasks which are assessed by local examiners.

  Test items are designed according to the Profile of the Italian language (Spinelli, & Parizzi, 2010), Perugia Corpus (Spina, 2014) and CELI corpus (Spina et al., 2002), among the main resources. They are also in line with the CEFR Companion Volume (CV) (Council of Europe, 2020) Action-oriented Approach (Cinganotto, 2023a; Piccardo, & North, 2019) and learning scenarios, suggesting authentic scenarios for real tasks involving meaningful use of the Italian language.

- DILS-PG: Certification in teaching Italian as a foreign language, divided into two levels (Level I and Level II), which is aimed at assessing the language and methodological skills required to teach Italian as a second or foreign language (Santeusanio, 2014).

  The exam format includes a written test, divided into different parts, aimed at assessing the following dimensions:

  1. Theoretical knowledge

  – Teaching strategies

  – Metalinguistic awareness

  – Socio-cultural knowledge

2. Methodological competence

– Analysis of teaching material

– Classroom observation

3. Operational skills

– Planning of teaching material

– Management skills

– Digital skills

The study: Research questions, sample and methods

Considering the lessons learned from the pandemic (Cinganotto, 2023b; Malagnini, & Cinganotto, 2023, in press), which led universities, schools and institutions to rethink and reshape teaching and assessment practices in the era of the 'new normal', embedding the use of learning technologies as an integral part of the curricula, researchers, scholars and Italian language experts (CEL: Collaboratori Esperti Lingistici) at CVCL are also engaged in studying and investigating possible solutions for the digitalization of the testing process.

In fact, during the pandemic, a wide range of online training initiatives were organized in order to prepare teachers and students for the different types of tests. Lessons learned from the emergency situation are being capitalized on in order to possibly improve language testing and assessment procedures, also taking into account the literature in the field. In fact, the study took into consideration recent studies in language testing (Barni, 2023; Masillo, 2019; Serragiotto, 2016) as a starting point.

The research was divided into two strands, with two different research questions (RQs) and different samples.

The research questions were the following:

Part 1

RQ1: How is online language testing perceived by non-Italian-speaking candidates? What are their perceptions and reactions towards online testing, compared to pen-and-paper testing?

The sample was made up of 32 international students attending language courses at University for Foreigners of Perugia who were delivered an online CELI test at A2 level, previously designed and implemented on CVCL Moodle platform.

Part 2

RQ2: How is online testing in teaching Italian as a second/foreign language perceived by teachers and educators? What are their perceptions and reactions towards online testing, compared to pen-and-paper testing?

The sample was made up of 25 teachers and educators from S. Egidio Community, in Italy, attending a DILS-PG course at CVCL. They were delivered in the online DILS- PG format (only the first part), previously implemented on CVCL Moodle platform, after having taken the regular pen-and-paper version of the same exam, in order to compare the two formats.

An online questionnaire was delivered to both samples of candidates and interviews were also organized after the delivering of the tests. Data were collected through qualitative methods and comments, and remarks were grouped according to the Framework Analysis (Ritchie, & Lewis, 2003).

Main findings from the online CELI A2 delivery

The test was made up of a listening part and a reading comprehension part, with mainly multiple-choice and matching questions. The students reported they found more problems with the listening part than the reading comprehension, probably due to the technical settings of the platform, the anxiety of the timing and the self-delivery mode of the test. In fact, half of the respondents did not complete the listening part.

They also found the images too small and not very clear to match with the text. At the same time, they found this mode useful and convenient for self-study and self-assessment.

Here are some of the comments:

Images are too small and it is difficult to identify the right answer.

I couldn't hear anything during the listening exercise.

It is convenient for students who want self-study and self-assessment.

It can be done anytime, anywhere, and the test will never be damaged.

Feasible.

It is convenient, but can cause some anxiety.

It is more immediate.

Main findings from the online DILS-PG delivery

The respondents were quite satisfied with the online format of the test compared with the pen-and-paper one. They found it useful and comfortable as easier to read and correct, more in line with current formats. Some respondents highlighted anxiety related to timing, but no particular difference with pen-and-paper format was generally identified.

The Strengths, Weaknesses, Opportunities and Threats (SWOT) analysis returned the following comments:

Strengths

Time counting and word numerator.

Possibility of taking the test remotely.

Perhaps more in line with current testing techniques.

I could copy the open-ended answers to a word file and count the words.

Helps to focus on essentials.

Ease of use, ability to correct.

The ability to focus on the single exercise.

I honestly don't find there is a difference between tests taken online or on pen and paper.

Having the results right away.

I found it to be smooth, easy to understand, practical.

Weaknesses

On-screen reading of the test for correction, which is more tiring.

In case of longer exercises that don't all fit on the screen, not being able to have an overview of the exercise.

If there is no control it is easy to copy.

The graphics part of the database has some problems on the 'grids' and 'texts' that can inadvertently be deleted if not moved.

I found no weaknesses.

If one must be found, a little bit of anxiety about time running out.

I do not find a substantial difference.

Lack of the overview.

Lack of control.

Opportunities

There is more space in writing open-ended questions and the ability to erase and rewrite.

Possibility to take it remotely.

Useful as training, you could use it at least to practise old tests.

Practise and check the correction of closed questions right away.

Effectiveness and focus on content.

Being able to take the test even in more inconvenient situations (distances, schedule problems for work).

More space to write, compared to paper.

The possibility to focus on the single exercise that is proposed by the system.

Develop synthesis skills to the test.

Faster, standard handwriting.

Practicality, autonomy.

Threats

External help.

Anxiety and ease of copying.

Difficult for those with few digital skills.

Thinking that it might be easier.

Possibility of cheating.

The risk might be to get caught up in the anxiety of time running out.

Speed in responses.

Not being able to correct once the test is submitted.

The lack of connection.

Conclusions

According to the Framework Analysis, the following conclusions can be drawn to answer the two RQs, in light of the comments and reactions of the participants, both students and teachers/educators.

- The online format of the test, both the language and the methodological one, should be revised and reshaped: it is not possible to use the same format and items of the pen-and-paper version. A new digital plan should involve not only the technological procedures and features, but also the test format and design in general.
- The online test is perceived as quicker and more immediate by both teachers/educators and students.
- Teachers and educators are used to working with platforms: they do not find much difference from pen-and-paper tests.
- Students have problems with listening comprehension, especially on their mobile phone.
- Time can cause anxiety but can also be an important indicator especially for students.
- Teachers find it useful to have the capability to count the number of words in open questions and to check and correct the text.
- Both teachers and students would prefer to take the actual test online.
- Online testing allows teachers to save time for planning and correction.

The main findings from this study are very interesting and encouraging towards the digitalization of testing procedures at CVCL. This process, in line with the needs of the post-pandemic and digital era, will be moved forward by researchers, scholars at CEL at CVCL, under the supervision of the CVCL Steering Committee and Director.

Acknowledgments

The author is grateful to the research group involved in the study, in particular to M. Valentina Marasco, Danilo Rini, Roberta Rondoni, Nicoletta Santeusanio (Italian language experts at CVCL) and to Giovanna Scocozza, CVCL Director.

References

Barni, M. (2023). *Valutare le competenze nelle L2*. Rome: Carocci.

Cinganotto, L. (2023a). L'approccio orientato all'azione del "Quadro Comune Europeo di Riferimento per le Lingue, Volume Complementare" nella didattica digitale dell'Italiano L2/LS: scenari e attività didattiche nella percezione degli studenti. *Italiano LinguaDue*, *15*(1), 915–928.

Cinganotto L. (2023b). Learning technologies for ELT during the pandemic in Italy: Teachers' attitude. In S. Kourieos & D. Evripidou (Eds.), *Language Teaching and Learning during the COVID-19 Pandemic: A Shift to a New Era* (pp.c37–56). Newcastle: Cambridge Scholars Publishing.

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge: Cambridge University Press.

Council of Europe. (2020). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Companion Volume*. Strasbourg: Council of Europe Publishing.

Malagnini F., & Cinganotto, L. (2023, in press). Nuove opportunità del digitale nell'era del "new normal". In *Strategie per lo sviluppo della qualità nella didattica universitaria"*. Bari: Pensa Editore.

Masillo, P. (2019). *La valutazione linguistica in contesto migratorio: il test A2*. Pisa: Pacini Editore.

Piccardo E., & North B. (2019). *The Action-oriented Approach: A Dynamic Vision of Language Education*. Clevedon: Multilingual Matters.

Ritchie J., & Lewis J. (2003). *Qualitative research practice: A guide for social science students and researchers*. London: Sage.

Santeusanio, N. (2014). *Prepararsi alla DILS-PG*. Turin: Loescher.

Serragiotto, G. (2016). *La valutazione degli apprendimenti linguistici*. Bologna: Bonacci Editore.

Spina S. (2014). *Il Perugia Corpus: una risorsa di riferimento per l'italiano. Composizione, annotazione e valutazione*. In R. Basili, A. Lenci, B. Magnini (a cura di), *Proceedings of the First Italian Conference on Computational Linguistics CLiC-it 2014. Volume 1* (pp. 354–359). Pisa: Pisa University Press.

Spina, S., Fioravanti, I., Forti, L., Santucci, V., Scerra, A., & Zanda, F. (2022). Il corpus CELI: una nuova risorsa per studiare l'acquisizione dell'italiano L2. *Italiano LinguaDue*, *14*(1), 116–138.

Spinelli B., & Parizzi F. (2010). *Profilo della lingua italiana. Livelli di riferimento del QCER A1, A2, B1, B2. Per le Scuole superiori*. Florence: La Nuova Italia.

Does the mode of delivery influence test-takers' performance? A comparative analysis

Robert Märcz
ECL Examinations, University of Pécs, Hungary

Réka Werner
ECL Examinations, University of Pécs, Hungary

Abstract

ECL Language Examinations Centre offers standardized, general-purpose, and monolingual foreign language proficiency exams consisting of four parts: reading, writing, listening, and speaking. The exam was originally offered in a paper-based format, but in 2020 a computer-based version of the exam was developed. To find out whether the medium (computer) and the circumstances (proctored online exam) have any influence on test takers' performance, a comparability study was conducted. German and English language tests (reading, listening, and writing) used previously during live exams were converted into a computer-based format for candidates who took a mock online exam. The statistical analyses of the previous live exam results and those of the online mock exam were compared. To test the null hypothesis that there was no significant difference between the performances of the two groups, the Independent Samples T-Test was applied.

Introduction

The ECL Language Examination Centre offers standardized, general-purpose, and monolingual foreign language proficiency exams consisting of four parts: reading comprehension, listening comprehension, written communication, and oral communication. The exam was originally offered in a paper-based (PB) format. In 2020 the ECL International Centre developed a computer-based (CB) version of its language exam with two modes of delivery: candidates can take CB language exams at designated examination sites and at home.

Integrated Language Testing System

The ECL Language Examination Centre uses an integrated language testing system called LEO to manage its language testing process. CB testing is a module in this system. The LEO Integrated Language Examination Management System is a cloud-based service that the system developer provides to the examination centre and its examination sites. The cloud-based system is operated from a high-capacity server park of Hungary's leading server hotel. The system consists of highly reliable, redundant systems, continuous (365/24) operation with short response times, customer support, system monitoring, and corporate backup services. To ensure redundancy, the service is physically provided from two separate server hotels. In addition to the server hotels, the backup data are also backed up to the service provider's independent secure storage, which is stored in European data centres using Microsoft OneDrive.

When candidates take language exams at their homes, exam security is of utmost importance. They must download and install the exam software (Safe Exam Browser, SEB) on their computers at home, which is launched by invigilators before the exam. The SEB application allows the exam to be taken in a secure environment, featuring the following security measures:

1. The browser program is only allowed to run in closed, kiosk mode during the exam.
2. The computer is only connected to the central server of the examination centre.
3. Candidates are not able to start other applications (e.g., other dictionary programs, word processors, web pages) on their computers other than the exam software.
4. Candidates are not disturbed by pop-up windows, built-in messages, or updates from the operating system.
5. There is no external access to the computers used for examinations.

6. It is not possible to run the SEB application in a virtual environment.
7. Candidates do not have access to other applications on their computers using the test interface other than the permitted dictionaries.

The proctored online examination can be taken under the supervision of two cameras. To take the exam, the candidates need two devices:

1. A computer (i.e., desktop or laptop) with a built-in web camera (primary device) that broadcasts the exam from the front of the screen.
2. A mobile phone, tablet, or laptop (secondary device) that is positioned on the side and behind the candidate for broadcasting video and audio, and monitoring the screen of the candidate's computer used during an examination.

According to Hungarian regulations, for every 15 candidates there should be one invigilator trained for the task who continuously monitors the candidates' activities on at least one of the cameras during the whole examination. During the exam, candidates are continuously monitored through their primary and secondary devices. The candidates' computers must be in continuous communication with the central computer. To meet this requirement, candidates' computers send a signal ('heartbeat') to the central computer every five seconds, which is monitored by the invigilators. A continuous log file of all examination events is created, containing all events that occurred during the examination (status changes, saves, biometric identification data, invigilator and candidate entries). The electronic log is stored on the examination system server.

Research question

We wanted to know whether the two different modes of delivery (PB vs. CB) have an impact on the performance of the candidates. 'When test results are from high-stakes testing, evidence from mean score differences between relevant subgroups should be examined and if such differences are found, an investigation should be undertaken to determine that such differences are not attributable to a source of construct underrepresentation or construct-irrelevant variance' (Kunnan, 2007, p. 110). The question arises: whether the mode of delivery can be a construct-irrelevant variance?

Findings

There is not a great variety of studies that compare the performance of test-takers completing CB and PB tests. A non-exhaustive list of studies are as follows:

- In a study (Yu & Iwashita, 2021) including 92 Chinese undergraduate students, researchers found that (1) test scores in CBT and PBT were comparable; (2) two items (i.e., comfort level of reading articles on the computer and forgetting time when using computers) positively correlated with CBT scores; and (3) participants' attitude towards CBT did not impact test performance.
- Saad's study (2007) explored the comparability of paper and CB testing in an L2 reading context and the impact of test takers' characteristics (i.e., computer familiarity, computer attitude, testing mode preference, and test-taking strategies on students' performance on CB tests compared with PB tests). There were 167 Saudi medical students who participated in this study. No significant difference was found between each testing mode and none of the factors examined had an influence on students' performance when doing the CB tests.
- Another study (Khoshsima, Toroujeni, Thompson, & Ebrahimi, 2019) was conducted to investigate whether test scores of Iranian English as Foreign Language (EFL) learners were equivalent across CBT and PBT modes, with 58 intermediate learners studying at a private language academy located in Behshahr, Iran. Participants produced similar scores across modes, although they insignificantly outperformed on the CBT version.

Brunfaut, Harding and Batty (2018) aimed to determine the effect of delivery mode on two writing tasks (reading-into-writing and extended writing) featured in the Trinity College London Integrated Skills in English (ISE) test across three proficiency levels (B1–C1 of the Common European Framework of Reference for Languages (CEFR, Council of Europe, 2001)). The researchers found that the *delivery mode had no discernible effect,* apart from the PB mode being slightly easier for the ISE reading-into-writing task. When analysing the results of a T-test to compare the means of two test modes, the results of an Iranian study (Hosseinia, Abidinb, & Baghdarniac, 1998) showed the priority of PB testing over CB testing.

These research studies illustrate that the results are mixed and inconclusive when it comes to the influence of different language assessment delivery modes on candidates' performance.

Methods

Research design

Our null hypothesis was that there is no significant difference between the performance of the two groups of test takers taking exams in PB and CB modes of delivery. To test our null hypothesis, we compared the actual performance of two groups: 1) test takers who completed reading comprehension, listening comprehension, and written communication tests in the context of live test administration; and 2) test takers who completed identical reading comprehension, listening comprehension, and written communication tests in an online mock exam.

Context and participants

An online mock exam was announced for those interested in completing an ECL language test online. Considering the circumstances under which the two groups of test takers took the exam, there is a marked difference (Table 1). In addition, there are several factors that may possibly have an impact on the performance of the test takers, such as the following:

- full page visible vs. scrolling through webpages during the exam
- writing (underlining parts of the text, taking notes, etc.) vs. navigating with mouse
- handwriting vs. typing (preference for handwriting vs. computer literacy)
- being alone vs. taking the exam with others
- being watched by two cameras vs. being watched by live invigilators
- being afraid of technical problems during CB exams vs. having a more secure environment during live PB exams
- silence vs. noise
- invasion of privacy which may stimulate anxiety.

Table 1: Differences between test takers completing tests in two modes of delivery

Live exam administration (PB)	*Mock online exam (CB)*
• High stakes	• Low stakes
• Constraint	• Voluntary
• Extrinsic motivation	• Intrinsic motivation
• Taken at an exam site	• Taken at home
• Live invigilator in the exam site	• Being watched by two cameras
• Possibly weaker IT knowledge	• Presumably stronger IT knowledge

Instruments

Each ECL test consists of two tasks. In the case of the tests measuring receptive skills, the first task is always an objective one and the second task is semi-objective. Accordingly, the first task of an ECL reading comprehension test is a gap-fill exercise, and the second task is a short-answer type. The first task of an ECL listening comprehension test is multiple choice, and the second is a short-answers task. The ECL written communication test also consists of two tasks. One is always related to the CEFR private domain, the other is connected to a public domain. Tables 2 and 3 illustrate the most important statistical data of the live tests and the online mock exams, the dates of their delivery, and the number of test takers.

Table 2: Number of candidates who took the English paper-based test and its online computer-based version

Skill	*Date of live exam*	*N*	*– Cronbach's alpha (α)* *– Mean % corr (MPC)*	*Date of online mock exam*	*N*	*– Cronbach's alpha (α)* *– Mean % corr (MPC)*
Reading	02 2020	1,852	α: 0,876 MPC: 55%	01 2021	491	α: 0,813 MPC: 63%
Listening	02 2020	1,440	α: 0,784 MPC: 64%	01 2021	557	α: 0,762 MPC: 63,5%
Writing	10 2020	587	Mean score: 62%	01 2021	491	Mean score: 68%

Table 3: Number of candidates who took the German paper-based test and its online computer-based version

Skill	*Date of live exam*	*N*	*– Cronbach's alpha (α)* *– Mean % corr (MPC)*	*Date of online mock exam*	*N*	*– Cronbach's alpha (α)* *– Mean % corr (MPC)*
Reading	10 2020	240	α: 0,863 MPC: 69%	01 2021	174	α: 0,824 MPC: 50%
Listening	02 2020	249	α: 0,790 62%	01 2021	176	α: 0,776 MPC: 61%
Writing	10 2020	240	Mean score: 50,2%	01 2021	174	Mean score: 54,8%

Data analysis

In order to find out whether there was any difference between the performance of the test takers in the two groups, we applied an independent samples T-test to assess the difference in the mean scores of the two groups and a 95% confidence interval estimation. The T-test revealed that there was a significant difference between test takers' performance in the English reading comprehension and the German reading comprehension and written communication tests. This implies that no significant difference was found in the English listening comprehension and written communication test and the German listening communication test. Figure 1 shows the result of the 95% confidence interval estimation which corroborated the results of the T-test.

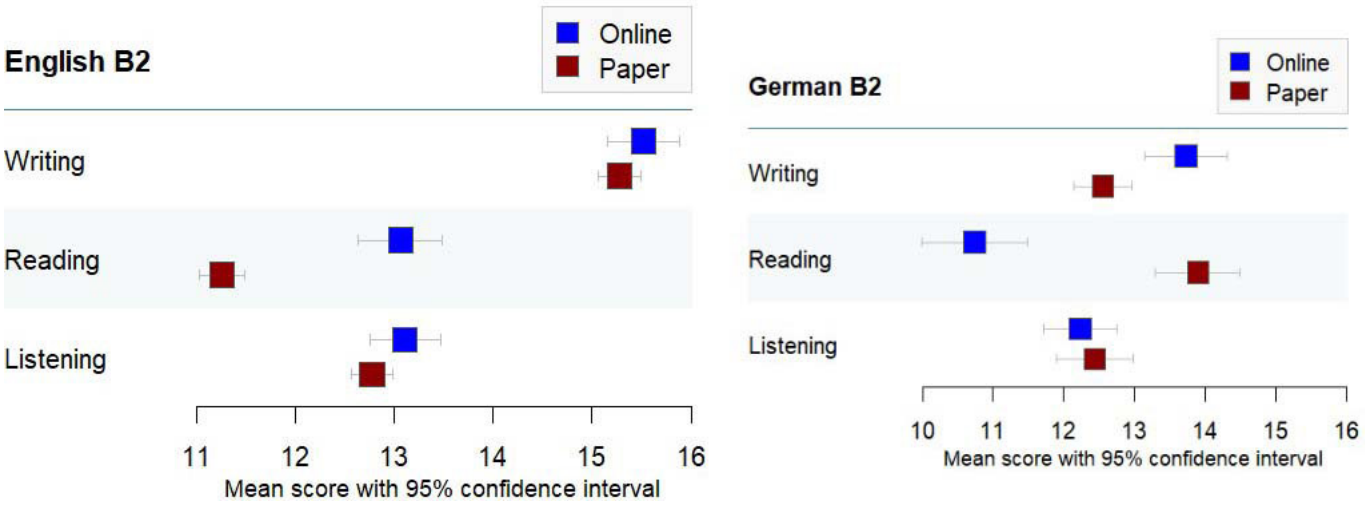


Figure 1 Results of the 95% confidence interval estimation

Summary and implications

On the basis of the results, our null hypothesis is to be rejected. The reasons for the significant differences between test takers' performance on the English reading comprehension and the German reading comprehension and written communication tests are related to factors that possibly impact test taker performance. To identify potential factors, additional comparative and quantitative studies should be implemented along with qualitative research to further explore the lived experiences of test takers and how they view the impact of PB and CB formats on their language exam performance.

References

Brunfaut, T., Harding, L., & Batty, A. O. (2018). Going online: The effect of mode of delivery on performances and perceptions on an English L2 writing test suite. *Assessing Writing*, *36*, 3–18.

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge: Cambridge University Press.

Hosseinia, M., Abidinb, M. J. Z., & Baghdarniac, M. (2014). Comparability of Test Results of Computer Based Tests (CBT) and Paper and Pencil Tests (PPT) among English Language Learners in Iran. *Procedia - Social and Behavioral Sciences*, *98*, 659–667.

Khoshsima, H., Toroujeni, S. M. H., Thompson, N., & Ebrahimi, M. R. (2019). Computer-based (CBT) versus paper-based (PBT) testing: mode effect, relationship between computer familiarity, attitudes, aversion and mode preference with CBT test scores in an Asian private EFL context. *Teaching English with Technology*, *19*(1), 86–101.

Kunnan, A. J. (2007). Test Fairness, Test Bias, and DIF. *Language Assessment Quarterly*, *4*(2), 109–112.

Saad, A. (2007). Computer-based vs. Paper-based testing: Does the test administration mode matter?. In *Proceedings of the BAAL Annual Conference 2007: 40th Annual Meeting of the British Association for Applied Linguistics 6–8 September 2007: Technology, Ideology and Practice in Applied Linguistics: with Extracts from the Joint Annual Meeting of BAAL and IRAAL, 2006* (pp. 101–110). London: Scitsiughil Press.

Yu, W., & Iwashita, N. (2021). Comparison of test performance on paper-based testing (PBT) and computer-based testing (CBT) by English-majored undergraduate students in China. *Language Testing in Asia*, *11*(32).

A comparative study of test-takers' perceptions of paper-based and computer-based language examinations

Krisztina Babos, Szilvia Dömők, Szilvia Gróf, Júlia Kissné Adorján, Magdolna Lehmann, Zoltán Lukácsi, Róbert Märcz and Zsuzsanna Soproni
Hungarian Accreditation Board for Foreign Language Examinations

Abstract

As a result of the COVID-19 pandemic, online proctored examination systems were granted accreditation in Hungary with the aim of ensuring that the same measurement principles are observed. Consequently, foreign language exams can now be taken in two modes of delivery: a) paper-based (PB) and b) computer-based (CB). The aim of the research conducted in the spring 2022 was to compare the lived experiences of test-takers taking language exams in the two different delivery modes. Applying the non-probability sampling method, an online questionnaire was sent to former candidates who had taken the exam in 2020 and 2021. Respondents were asked to give answers for Likert-scale items to get quantitative data and elaborate on their lived experiences, providing an emic perspective. The written answers were analysed using the constant comparative method to identify emerging themes. Of the 2015 respondents, 81% completed PB exams and 19% took CB exams.

Introduction

The aim of the research conducted by experts of the Hungarian Accreditation Board for Foreign Language Examinations (the Board) in the spring 2022 was to compare the lived experiences of test-takers who completed language exams in two different delivery modes. There were significant changes in the field of state-accredited language examinations in Hungary between 2018 and 2022. In parallel to the paper-based (PB) language exams, the computer-based (CB) delivery mode was also extensively regulated in 2021. Under the current regulatory conditions it is now possible to take language exams not only at accredited examination centres, but also at non-accredited examination sites (i.e., online and at the homes of test-takers). The introduction of the CB mode of delivery has raised the question of what impact they might have on the language examination systems and specifically on test-takers. Accordingly, the Board decided to carry out comprehensive, mixed-methods research to investigate the characteristics of each exam delivery mode, and to explore the lived experiences of candidates taking their exams in two delivery modes.

Research questions

To explore the issues mentioned above, the Board posited two research questions:

1. How do test-takers perceive their experiences of taking accredited language examinations?
2. How do test-takers' experiences differ according to the two delivery modes of their language exams?

Methodology

Design

Mixed-methods research was conducted with a qualitative focus. Applying the non-probability sampling method, the language examination centres accredited in Hungary were asked to forward an online questionnaire to their former candidates who had taken the exam in 2020 and 2021. The questionnaire consisted of three parts: 1) background information; 2) reasons for choosing a particular delivery mode and preparation method; 3) personal impressions of the different parts of the exam with

open-ended questions. Respondents were asked to rate their exam experience using Likert-scale items to get quantitative data and to elaborate on their lived experiences to provide an emic perspective.

Participants

The questionnaire was completed by 2,015 respondents in spring 2022. More than 80% of participants took PB exams. 85% of study participants who took the CB exam completed their exams online at home. In terms of age, adolescents and young adults completed the questionnaire: 42.94% were in high school, 31.01% were in university and 26.05% were either working or already finished university. 80% of respondents took their exams in English as a foreign language, 12% in German, and 8% in French, Italian and Spanish.

Instruments and procedures

The instrument used to carry out the mixed-methods study was an online questionnaire. Having drafted the first version of the questionnaire, interviews were conducted with candidates from each subgroup (test-takers who took PB and CB exams). On the basis of the interviews, the questionnaire was modified and validated. The questionnaire was then distributed through the exam providers, requesting them to send it to former test-takers. After receiving the completed questionnaires, the Likert-scale items were quantified and the answers to the open-ended questions were subjected to content analysis. To attain an in-depth understanding of the direct stories, ideas and meanings expressed by research participants, in vivo coding was applied. Having identified and categorised the various emerging themes, each answer was coded by two people. Finally, intercoder reliability was between 0.92–0.96 for the open-ended questions.

Results

In terms of reasons for their choices, 50% of candidates were more accustomed to taking PB exams, 40% thought they could concentrate better on the exam at an exam venue than at home, and 38% thought they would take PB exams in the future. Many participants also mentioned that the examination centres they selected only offered PB exams (31%) and that they did not want to be subjected to technical issues (30%) related to online exams at home. The main methods of preparation were using photocopies from language teachers (61%), studying from preparation books for the language exam (54%), and looking for practice exercises online (33%). Table 1 illustrates how respondents assessed each part of the exam.

Table 1: Likert-scale responses for the question: How would you rate this part of your exam?

1 (very bad) to 5 (excellent)	*Computer-based*	*Paper-based*
Reading comprehension	4.30	4.33
Written communication	4.43	4.14
Listening comprehension	4.13	3.75
Oral communication	3.92	4.27

There were nine open-ended questions:

1. In retrospect, what did you like about the delivery mode you chose (PB/CB)?
2. In retrospect, what did you NOT like about the delivery mode you chose (PB/CB)?
3. Overall, how would you describe your experience of taking the language exam (please do not write about the requirements, the structure of the exam, or your performance, but how you felt during the test)?
4. 5. 6. and 7. How did it feel to complete the reading comprehension/written communication/listening comprehension and oral communication tasks in the delivery mode you chose? Please describe in a few words.
8. Why would you recommend this delivery mode to other candidates?
9. Why would you NOT recommend this delivery mode to other candidates?

Our qualitative analysis (content analysis) yielded 15 categories each containing various themes. The frequency of categories and the order of importance of these categories found in the responses are shown in the Tables 2 and 3 below.

Table 2: Name and frequency of the 10 most important categories identified

1. JUDGEMENT (super/modern/adequate/negative/strange /easy/simple)
2. STRESS (normal/high/low/tech-related/privacy-related)
3. PRACTICALITY (urgency/constraint/travel+/travel– /comfort)
4. TECHNOLOGY (interface + or –/preparation/cameras/knowledge/operation/problems)
5. WAY OF WRITING (typing+/typing– /handwriting+/handwriting–/lack of paper)
6. SAFENESS (pandemic-related/safe environment/familiar environment/accustomed)
7. PERSONS (candidates + or –/being alone/staff members + or –/others)
8. QUALITY (sound quality + or –/visual clarity)
9. ENVIRONMENT (silence/online noise/PC noise/others' noise/outside noise)
10. TIME (not enough/waiting a lot/enough time)

Table 3: Order of importance according to the two groups of respondents

Computer-based (391/2,233 answers)			*Paper-based (1,624/8,141 answers)*		
CATEGORY	*ANSWERS (N)*	*% WITHIN ALL ANSWERS*	*CATEGORY*	*ANSWERS (N)*	*% WITHIN ALL ANSWERS*
JUDGEMENT	866	38.78	JUDGEMENT	3,309	40.65
STRESS	545	24.40	STRESS	1,221	15.00
TECHNOLOGY	439	19.65	SAFENESS	985	12.10
PRACTICALITY	374	16.74	PERSONS	875	10.40
WAY OF WRITING	164	7.34	QUALITY	419	5.14
CONTACT	157	7.03	WAY OF WRITING	315	3.86
SAFENESS	139	6.22	PAPER	231	2.83
PERSONS	132	5.91	PRACTICALITY	160	1.96
QUALITY	90	4.03	NOISE	123	1.51
TIME	83	3.71	TIME	104	1.27

The most frequent coded textual responses were related to rating the experience of taking the language exam. 26.98% of the answers provided by PB candidates included superlatives about the experience (e.g., 'I truly enjoyed the exam' or 'It was much more pleasant than I thought it would be') compared to 16.51% of the answers given by CB test-takers. The rating 'satisfactory' appeared in roughly half of the statements given by each of the two groups.

There was a significant difference in the answers regarding exam-related stress that test-takers experienced. For this construct, 35.29% of the answers given by the PB test-takers was related to candidates experiencing high stress (e.g., 'I was very nervous during the exam') compared to only 12.11% in the CB group (e.g., 'The fact that I could take the exam at home made me feel relaxed and less stressed, less nervous'). Lack of stress was reported in 66.05% of CB answers, compared to only 20.08% in the PB answers. It is interesting that only PB test-takers (20.06%) interpreted stress as a motivating factor.

When it comes to where test-takers feel safer and why, our analysis revealed that 85.61% of the CB test-takers preferred this delivery mode because they were able to stay at home (e.g., 'All in all it felt calm, I was in my own room, therefore I wasn't worried'). For the PB test-takers, being accustomed to the situation was the most decisive factor (e.g., 'A traditional exam that I am accustomed to').

Not surprisingly, taking the exam alone was the most important condition for CB test-takers as they mentioned this factor in 43.93% of their answers (e.g., 'I was alone, nobody bothered me, it was just great'). In contrast, 57.94% of the PB test-takers thought that seeing invigilators and the examiners had a positive effect on them (e.g., 'The invigilators were not really helpful when I had a problem'), and 26.40% specifically mentioned that watching the other test-takers (their 'fellow sufferers') made them feel more comfortable (e.g., 'It had a calming effect on me to see the other candidates').

To assess the different parts of the exam, the following differences can be seen. Regarding the reading comprehension test, the most frequent answer given by the PB test-takers was 'being accustomed to'. Almost twice as many PB test-takers rated the written communication exam as 'satisfactory' compared to CB test-takers (43.41% vs. 25.37% respectively). However, 24.24% of CB test-takers mentioned that they liked this part (written communication) of the exam because they prefer typing. Only 9.53% of PB test-takers said it was good because they prefer handwriting. There was a significant difference in the answers related to the audio files in the listening comprehension test. Every fourth (25.31%) PB test-taker complained about the sound quality compared to only 6.32% of the CB candidates. Finally, there was also a significant difference related to the oral exam, with 23.06% of PB test-takers stating that they liked that the exam was live, compared to 5.26% of CB test-takers who really liked the online form of the oral exam.

Conclusion

To sum up the most important differences, the following broad observations can be made: the CB mode of delivery is preferred when taking a listening comprehension test, whereas the PB mode is preferred when taking an oral communication exam. Additionally, participants who opted for the CB exam were less satisfied, preferred practical considerations (e.g., comfort, no travel), experienced less stress during the exam, preferred typing during the written communication sections of the exam, and considered the exam simple in this mode. In contrast, participants who chose the PB exam were overall more satisfied, selected this mode because it is what they are familiar with, considered stress a motivating factor, preferred handwriting during the written communication sections of the exam, and thought interpersonal relations were important in the context of taking the language exam.

It is also clear that the more common PB language examination delivery mode has a strong influence on the decisions and thinking of candidates in Hungary. Traditions change slowly and the existing education system has a strong influence on the test-takers. Language exams are high-stakes, serious events associated with many formalities. Simultaneously, there is also a growing tech-savvy group of candidates who prefer the online, CB delivery mode for language exams.

The comparability of computer-based and paper-based writing tests: A case study

Balázs Csizmadia
ELTE Origó Language Centre, Budapest

Abstract

The aim of this paper is to report on research into the comparability of computer-based (CB) and paper-based (PB) writing tests. It analyses the two types of writing tasks in English used by ELTE Origó Language Centre at B2 level of the CEFR. The PB and CB tests were identical except for the mode in which they were administered. The study examines possible differences in candidate performance and the kind of language produced in the two formats. In addition, the paper seeks to answer the question of whether the rating scales developed for the PB test work equally well for the CB test. In order to ensure the reliability of the comparison in spite of the small sample (12 test-takers), a within-subjects design or crossover design was used: half of the candidates were asked to do Writing Task 1 on paper and Task 2 on computer, and the other half vice versa.

Introduction

The aim of this paper is to report on research into the comparability of computer-based (CB) and paper-based (PB) writing tests. It analyses the two types of writing tasks in English used by ELTE Origó Language Centre at B2 level of the Common European Framework of Reference for Languages (CEFR, Council of Europe, 2001).

ELTE Origó Language Centre is a Budapest-based institution. Founded in 1967, it has been providing language courses as well as assessments and qualifications for over 50 years and is accredited by the Hungarian Accreditation Centre for Foreign Language Examinations. One of the key tasks of ELTE Origó is to standardize the assessment of the Hungarian language, but it also offers tests in many foreign languages, with English being the most popular among candidates.

Most of the tests offered by ELTE Origó are bilingual, and this in general applies to tests of the English language as well. This study will look at the two writing tasks used in these tests. In Task 1, candidates write a letter or email based on prompts that are given in Hungarian. In Task 2, there is a statement or opinion in English about an interesting or controversial topic, and candidates have to comment on this statement by arguing for or against it.

The present study was carried out in March 2020, when the CB format was being tested. Its main goal is to find out whether the PB and CB tests are interchangeable. Accordingly, two specific research questions can be formulated:

1. Does the format (PB or CB) have any influence on the performance of test-takers, and on the kind of language they produce in the writing tasks?
2. Do the rating scales developed for the PB test work equally well for the CB test?

Literature review

Fűköh has recently conducted an impact study (in Hungarian) looking into the interchangeability of the CB and PB exam formats at ELTE Origó (2020), taking into account all parts of the exam except for speaking. This study has served as a starting point for the present analysis, which looks more closely at writing across the two formats.

There is a vast amount of literature on CB examinations, but a detailed review of it is beyond the scope of this paper. Only a few studies comparing CB and PB examinations and exploring their differences and similarities are mentioned here, particularly ones that have examined writing tasks.

Jin and Yan (2017) found that CB writing elicited cognitive processes among test-takers that were similar to those of traditional PB writing, but a high level of computer familiarity had a positive effect on test-takers' performances.

However, a number of studies have also shown that the format itself can affect the writing process and consequently negatively affect the quality of the text. The findings of Chen, White, McCloskey, Soroui and Chun (2011), in a study of adult foreign language assessment, revealed that volunteers performed better overall on most aspects of the writing tasks when writing on paper than on computer. The results suggest that a computer mode of administration may disadvantage some subgroups (e.g., the unemployed) more than others (e.g., the employed).

In addition, a number of studies have shown that there are no significant differences in candidate performance between PB and CB tests. Endres (2012), in a comparative study of CB and PB writing tests, showed that although there are some differences between the kinds of texts produced in the two formats (number of words, paragraphs, mechanical errors), these do not affect scores in a significant way.

Green and Maycock (2004), analysing performance in IELTS tests, basically found no difference between scores in the different formats (CB and PB), and argued that the two formats could be used interchangeably. They found only a slight difference for writing tasks: candidates performed marginally better on PB writing tests than on CB, which could be due to differences in task content between versions, and other reasons.

A review of the literature suggests that there are some examinations and some populations for which the traditional PB or CB format was preferable, while many cases show that the two formats are interchangeable.

Method

The test-takers

The participants in the study were 12 students aged 14 to 21 from three different secondary schools in Budapest. The Origó English exam is very popular among this age group, and this is the reason why they were chosen for the study. The participants were asked to fill in a short questionnaire, providing their name and information about their age, the name of their school, when they started learning English, and whether they had a language exam in English already, and if so, at what level.

Their answers showed that they had been learning English for anything between 6 to 14 years, which means that the range was fairly wide. As for language exams, five of them stated that they had passed a B2 level exam in English already, three said they had a B1 exam, one said they had a C1 exam, and three participants had not taken any language exam at all by the time of the study. This variety reflects the number of years the test-takers had spent learning English.

The raters

There were two raters (both full-time employees of ELTE Origó), who regularly mark English language tests at ELTE Origó, including writing tasks, both on paper and computer. They are experienced raters who had received training in using the online marking platform to mark papers on the computer before the CB exam was launched at the institute.

They each marked 50% of the PB tests and 50% of the CB tests in the study, independently of each other. They were not told which performance belonged to which candidate in order to minimise the risk of bias in rating.

The methods used to gather data

A within-subjects design or crossover design was used to determine whether there are any differences in candidate performance between the PB and CB tests. Candidates were divided into two groups, with half of them asked to do Writing Task 1 on paper and Task 2 on computer, and the other half vice versa. Ensuring the reliability of the comparison in this way was especially important since only 12 volunteers participated in the study. The PB and CB writing tests were identical except for the mode in which they were administered.

In addition to the small sample, the present study has several other limitations as well; for instance, the test-takers were selected from within a relatively narrow age range; they were all Hungarians with a similar cultural background; and came from only three different schools, which are all located in Budapest.

Summary of findings

Report on interview with raters

I conducted an interview with the two raters involved in the study, asking them the two research questions mentioned earlier.

Both raters agreed that typos, a lack of (or, at least, a less consistent use of) paragraphs, and punctuation mistakes were more typical of the CB format. In the PB tests, they found that illegible or just messy handwriting could be a problem with some candidates. Compared to that, the clear legibility of the texts produced by candidates in the CB test could cause a positive bias in raters.

The rating scales

There are descriptors for four criteria in the rating scales for Writing Task 1: Content, Communicative Achievement, Vocabulary, and Grammatical Accuracy. In this paper, I will only focus on what is relevant in the scales in terms of possible differences between PB and CB tests.

In terms of Content, it is important that the candidate should not go significantly below or over the word limit and structure their text logically. As for Communicative Achievement, what matters most is that a coherent text needs to be produced. I have not found any specific issues regarding Vocabulary use in PB and CB tests, but in terms of Grammatical Accuracy, spelling and punctuation are problems that need to be considered in such a comparative study.

In Writing Task 2, there are only two criteria: Task Completion and Language. Task Completion is essentially the same as the Content and Communicative Achievement categories of Writing Task 1 combined, and Language is basically a combination of Vocabulary and Grammatical Accuracy.

Based on the above, the following potential issues emerge when applying the scales developed for PB tests to CB tests:

- being below or (especially) over the word limit
- problems with the logical structuring and coherence of the texts due to lack of paragraphs or inappropriate use of paragraphs
- problems with spelling and punctuation.

Analysis of the results

Based on a close analysis of the texts produced by the participants of this study, the results can be summarised as follows:

- There are typos in nearly all the CB tests, whereas this is not a problem in the PB tests.
- Also, candidates were somewhat more likely to make grammar mistakes, and more serious ones, in the CB tests, but there was no significant difference.
- There were somewhat more problems with punctuation (affecting especially the use of commas) in the CB tests, although the difference was not significant.
- There were two candidates who misspelt the personal pronoun 'I', using the lower-case letter, but only in the CB test. In fact, one of them did not consistently misspell the word in the CB test either, but they always spelt it correctly in the PB test. This is an interesting phenomenon which implies that the nature of writing is changing; typing text on the computer seems to be perceived as a different and less formal experience.
- Unclear handwriting was a problem with some candidates in the PB test, and many candidates also crossed out a word or words, which made these scripts look less neat.
- Candidates were twice as likely to use proper paragraphs in Writing Task 1 when they wrote it on paper than when they wrote it on the computer. On the computer, even if they used paragraphs, they did not always leave an empty line between the paragraphs and did not indent the first line of the new paragraph.
- The mode of delivery had no effect on the scores given by the two raters. There was a difference between Rater 1 and Rater 2, with one of them being a bit more lenient, but there was no difference between the CB and PB formats in this respect.
- As for the rating scales and descriptors, I have found no significant issues in applying them to the CB tests.
- There are basically no differences in the length of text produced in the two formats.

Conclusion

The results of the study show that there are no significant differences between the PB and CB formats, which means that they are interchangeable. However, candidates should be provided with the opportunity to be examined in the mode of their choice as this is clearly the best way to maximise fairness. In this context, Jones and Maycock (2007, p. 12) have talked about a 'bias for best' approach – the idea that candidates can choose the format that works best for them.

In addition, some adjustments to the descriptors in the rating scales for Writing Task 1 and Writing Task 2 are worth considering when applying them to CB tests. Finally, further research into the comparability of PB and CB tests would be needed, with a higher number of test-takers.

References

Chen, J., White, S., McCloskey, M., Soroui, J., & Chun, Y. (2011). Effects of computer versus paper administration of an adult functional writing assessment. *Assessing Writing, 16*(1), 49–71.

Council of Europe. (2001). *Common European Framework of Reference for Languages*. Cambridge: Cambridge University Press.

Endres, H. (2012). A comparability study of computer-based and paper-based Writing tests. *Research Notes, 49*, 26–33.

Fűköh, B. (2020). *Hatástanulmány a papír alapú és a számítógépes vizsgaformátum felcserélhetőségéről* [unpublished study]. ELTE Origó Nyelvi Centrum.

Green, A., & Maycock, L. (2004). Computer-based IELTS and paper-based versions of IELTS. *Research Notes 18*, 3–6.

Jin, Y., & Yan, M. (2017). Computer literacy and the construct validity of a high-stakes computer-based writing assessment. *Language Assessment Quarterly, 14*(2), 101–119.

Jones, N., & Maycock, L. (2007). The comparability of computer-based and paper-based tests: goals, approaches, and a review of research. *Research Notes 27*, 11–14.

Are humans redundant? Automated scoring in learning and assessment

Arum Perwitasari
ETS Global, Amsterdam, the Netherlands

Abstract

Automated scoring offers performance-specific feedback, which is not feasible under operational human scoring and helps score the responses efficiently and reliably, especially for test programs with large test-taker volumes. What are the roles of humans in automated scoring? Are humans redundant, and thus should be replaced by machines? The aim of this article is to present the use of automated scoring in assessing speaking and writing tasks of the TOEFL iBT® test as well as in selected test preparation tools. It provides an overview of the automated scoring capabilities, including the e-rater® automated scoring engine used for assessing writing and the SpeechRater® service for assessing speaking, and explains the design of the two to ensure scoring quality. This paper also highlights a hybrid approach in scoring constructed responses by combining human raters and automated scoring tools to maintain the validity, reliability, and fairness of the test scores.

Introduction

Artificial Intelligence (AI) has undeniably brought substantial value to automated scoring in learning and assessment. Automated scoring not only provides targeted feedback that might be overlooked in human evaluations but also streamlines the process of evaluation responses, especially when dealing with high volumes of test-takers. However, the role of humans in automatic scoring remains pivotal and far from redundant. Rather than advocating for the complete replacement of human raters with machines, the aim is to leverage automation to enhance the efficiency and reliability of the assessment process.

This article explores the use of automated scoring in assessing speaking and writing tasks of the TOEFL iBT®[1] test as well as in test preparation tools. It explores the automated scoring capabilities, specifically the e-rater® automated scoring engine used for assessing writing and the SpeechRater® service for assessing speaking. This article highlights a hybrid approach as an effective strategy ensuring the preservation of test score validity, reliability and fairness by discussing the limitation and strengths of both humans and machines in scoring.

Development of automated scoring systems in assessment

The development of automated scoring in assessment represents a significant evolution at the intersection of technology and education. Initially, automated scoring systems focused on objective multiple-choice tests where machines could readily compare answers to predefined correct responses. As technology advanced, Natural Language Processing (NLP) techniques enabled automated scoring to extend its reach to more complex tasks, such as written and spoken language proficiency. These systems could now analyse text and speech patterns, fluency, grammar and even nuances of communication.

The origins of automated scoring can be traced back to the mid 20th century wherein early attempts were made to streamline the grading process for standardized tests. One of the pioneering efforts was the development of computerized scanners that could read and grade multiple-choice answer sheets. This initial application addressed the need for efficiency in handling large volumes in assessments.

The real breakthrough came in the 1960s with the creation of the 'Project Essay Grade' (PEG) system by Page (1968). PEG aimed to evaluate essays through a combination of statistical analysis and linguistic features. While rudimentary by today's standards, PEG marked the inception of automated scoring for written responses, even though it focused on simpler language constructs.

[1] The TOEFL iBT test is a widely recognized standardized test designed to assess the English language proficiency of non-native speakers of English. The test measures all four academic English skills: reading, listening, speaking and writing. Official website: www.ets.org/toefl

Subsequent decades saw incremental developments in automated scoring, predominantly withing educational and language testing contexts. These systems grew more sophisticated by incorporating rule-based approaches and linguistic analysis to assess grammar, vocabulary, and sentence structure. However, these early systems were often limited to assessing specific aspects of writing, and their scope remained constrained.

The late 20th and early 21st centuries witnessed a surge in research and development of NLP and machine learning. This propelled automated scoring to new heights, enabling systems to evaluate more complex language phenomena, such as coherence, argumentation and fluency. Pioneering institutions like ETS (Educational Testing Service) and its TOEFL iBT test started employing automated scoring models to evaluate speaking and writing skills on standardized language proficiency tests.

AI components in automated scoring

Automated scoring relies on a combination of artificial intelligence (AI) techniques, primarily centred around NLP and Machine Learning (ML) (Yan, Rupp, & Foltz, 2020). Some key AI components used in automated scoring include:

a. Natural Language Processing (NLP)

NLP plays a central role in automated scoring by enabling machines to understand, analyse and generate human language. NLP techniques are employed to extract linguistic features from text to speech, such as grammar, syntax, semantics, coherence, and vocabulary usage. These features provide insights into the quality of written or spoken responses.

b. Machine Learning (ML)

ML algorithms enable automated scoring systems to learn patterns from large datasets of human-scored responses. Supervised learning models are commonly used, where the system is trained on examples with known human scores. The system learns to map features extracted from the responses to corresponding scores. Ensemble techniques, where multiple models are combined, can enhance the accuracy and robustness of scoring.

c. Feature extraction

AI-driven automated scoring systems extract a wider range of features from text or speech data. These features can include word frequencies, sentence lengths, syntactic structures, sentiment analysis and more. These features provide a nuanced understanding of the response's quality.

d. Scoring models

AI-powered scoring models use the extracted features and the patterns learned from training data to predict scores for new responses. These models are designed to evaluate specific linguistic and contextual dimensions, such as fluency, coherence and vocabulary richness and more.

e. Semantic analysis

Advanced NLP techniques allow automated scoring systems to perform semantic analysis. This involves understanding the meaning and intent behind the language used in responses. This capability helps in assessing higher-order language phenomena, such as argumentation, critical thinking and creativity.

f. Deep learning

Deep learning, a subset of ML, involves neural networks with multiple layers. Recurrent Neural Networks (RNNs) and Transformer models like the GPT series, have been adapted for automated scoring tasks. These models can capture context and dependencies within text more effectively, enhancing the system's ability to analyse and evaluate complex language constructs.

g. Feedback generation

AI-driven automated scoring systems can generate feedback for test-takers based on their responses. These systems can identify specific areas of improvement, highlight errors and suggest ways to enhance language proficiency.

h. Adaptive learning

Some automated scoring systems incorporate adaptive learning mechanisms, using AI to customize the assessment experience based on individual performance. This approach tailors the level of difficulty and feedback according to the user's skills and needs.

Automated scoring relies on a synergy of AI techniques, particularly NLP and ML to process, analyze and evaluate written and spoken language responses. This multifaceted approach enables the systems to assess diverse dimensions of language proficiency in an efficient and consistent manner.

Automated scoring engines for speaking and writing

The SpeechRater® service: Automated scoring engine for assessing speaking

ETS' SpeechRater service stands as a pioneering solution for assessing spoken proficiency. Powered by an automated speech scoring engine, it comprehensively analyses over 170 speech features pertinent to both impromptu and read-aloud speech. Through an ingenious process of automatic feature selection, a refined subset of the most impactful features is employed, optimizing compactness and comprehensibility. Notably, each fundamental facet of speaking proficiency finds representation through at least one pertinent speech feature. This sophisticated scoring model is meticulously honed through training on responses that have been human-scored.

With its integration starting in August 2019, the SpeechRater service seamlessly complements human assessment marking a significant stride. It is noteworthy that the application of the SpeechRater service dates back to 2006, wherein it has been an integral component of ETS test preparation materials, consistently advancing the landscape of speech assessment.

In the context of speaking tasks, particularly read-aloud exercises, automated scoring implementation encompass various dimensions to holistically evaluate a test-taker's performance. The specific details of these dimensions can vary depending on the particular speaking task. Read-aloud tasks will have different scoring dimensions compared to independent or integrated speaking tasks.

Some key scoring dimensions and their corresponding features for read-aloud tasks include:

Scoring dimension: Fluency

Fluency features include rate of speech (e.g., words/phonemes per second), frequency of silent pauses, average pause length, and number of long pauses). Fluency features are typically robust to Automatic Speech Recognition (ASR) errors and are included in most automated speech scoring systems (Yan et al., 2020)

Scoring dimension: Pronunciation

This dimension focuses on the correctness of articulation speech sounds in alignment with native speaker norms. It identifies pronunciation errors and deviations between native and non-native pronunciation patterns (Yoon, Hasegawa-Johnson, & Sproat, 2010).

Scoring dimension: Prosody

These features analyse the distribution of stressed syllables and tonal variation in the response, mirroring the natural cadence of speech. In the case of scripted speech like read-aloud tasks, the intonation contour in the test-taker's speech can be compared to the target model based on patterns observed in responses provided by native speakers (Wang, Evanini, & Yoon, 2015).

Each dimension contributes to the overall assessment, capturing critical aspects of speaking proficiency such as fluency, pronunciation accuracy, prosodic elements, and the fidelity of the spoken content to the provided prompts. By combining these dimensions and features, automated scoring systems aim to provide reliable and consistent evaluations of spoken responses.

The e-rater® engine: Automated scoring engine for assessing writing

ETS' e-rater engine offers a sophisticated solution for assessing writing proficiency. Driven by ETS research and harnessed with NLP technology, this scoring engine excels in the automatic evaluation of essay quality. It generates holistic scores of score predictions, aligning with the assessments of trained human raters. The use of e-rater engine extends to furnishing insightful feedback on essays, pinpointing specific errors. This tool has found integration in more than 20 products, including ETS Products and Services: GRE®, TOEFL®, Criterion® Online Writing Evaluation, TOEFL Practice Online, GRE® ScoreItNow!™. Remarkably, the e-rater engine has been an integral part of the TOEFL iBT test for over a decade.

The e-rater engine expedites the evaluation of student essays within moments, accessible from any-internet-connected computer. It provides a holistic score that often mirrors human-derived scores while also offering annotated feedback on essay quality. Despite its proficiency, the e-rater engine does not engage in comprehending the essays nor does it evaluate content in the manner of human raters. It is essential to recognize that the e-rater engine does not capture all conceivable errors and is not intended to replace human raters. Rather, its focus lies in measuring global aspects such as total word counts, sentence structure, paragraph count, vocabulary metrics and grammatical structures. In this way, the e-rater engine contributes significantly to the assessment process while respecting its specific scope and limitations.

Hybrid approach: Combining strengths of AI and human raters

Initiating the discussion with the recognition that assessing spoken and written responses present challenges to both human and automated machines is imperative. Humans exhibit certain strengths and weaknesses just as machines possess their own sets of advantages and limitations.

Humas excel in comprehending intricate linguistic nuances, including complex discourse structures that imbue meaning. Furthermore, human evaluators can anchor their judgment in real life contexts, mirroring the way humans naturally assess language. Nonetheless it's important to acknowledge that human raters are susceptible to biases, inconsistencies, and fatigue, which can affect the quality of their assessments.

In contrast, machines excel in maintaining consistent evaluations of specific linguistic aspects, such as fluency and pronunciation. Their efficiency, scalability and cost-effectiveness become evident when dealing with larger volumes. However, a notable limitation lies in the struggle for machines to grasp nuanced meanings and content with higher-order language phenomena. This nuanced interplay between the strengths and weaknesses of humans and machines underpins the complexities of automated scoring and highlights the need for balance.

Davis and Papageorgiu (2021) conducted a study exploring the possibility of human raters and machines score combination using responses from the speaking responses of the TOEFL iBT test. In their study, human raters adjudicated scores across three distinct subconstructs: delivery, language use and topic development. Meanwhile, the automated scoring system of speaking – the SpeechRater service – generated scores for delivery and language use.

The study found that composite scores, derived from three unique combinations of human and automated analytical scores, demonstrated equal or superior reliability in comparison to human holistic scores. This boost in reliability is likely attributed to the incorporation of multiple observations inherent in composite scores. However, it is crucial to note that composite scores exclusively derived from human analytical scores exhibited the highest reliability.

Intriguingly, the reliability exhibited a consistent decrease as machine-generated scores progressively replaced human-derived scores. This investigation underscores the intricate dynamics between human and machine-based evaluation and their collective impact on the scoring process.

Conclusion

As technology continues its rapid evolution, the trajectory of automated scoring promises increasing sophistication with the capacity to encompass a wider range of written and spoken responses. The hybrid approach stands as a strategic solution, poised to yield assessments that are both thorough and unbiased, all while upholding the overarching objectives of accuracy, reliability and efficiency.

References

Davis, L., & Papageorgiou, S. (2021). Complementary strengths? Evaluation of a hybrid human-machine scoring approach for a test of oral academic English. *Assessment in Education: Principles, Policy & Practice*, *28*(4), 437–455.

Page, E. B. (1968). The use of the computer in analyzing student essays. *International Review of Education*, *14*, 210–225.

Wang, X., Evanini, K., Yoon, S-Y. (2015). Word-level F0 modeling in the automated assessment of non-native read speech. *Proc. Speech and Language Technology in Education (SLaTE 2015*), 23–27.

Yan, D., Rupp, A. A., & Foltz, P. W. (Eds.). (2020). *Handbook of Automated Scoring: Theory into Practice* (First edition). London: Chapman and Hall/CRC.

Yoon, S.-Y., Hasegawa-Johnson, M., & Sproat, R. (2010). Landmark-based automated pronunciation error detection. In *Proceedings of the 11th annual conference of the international speech communication association* (pp. 614–617). Retrieved from: https://www.isca-speech.org/archive/pdfs/interspeech_2010/yoon10_interspeech.pdf

Automated scoring of spelling mistakes in short answers

Leska Schwarz
g.a.s.t./TestDaF-Institut, Bochum, Germany

Ronja Laarmann-Quante
Ruhr University Bochum, Faculty of Philology, Department of Linguistics, Germany

Abstract

When short-answer items are used in language proficiency tests, human raters mark them according to scoring guidelines for each item. Despite these guidelines, raters often find it difficult to judge whether a response containing orthographical errors should be marked as correct or incorrect, the question being whether the response remains comprehensible and shows sufficient understanding of the input text.

In this study, we analyse human decisions regarding the acceptability of very short, misspelled responses from a listening task of the digital TestDaF (*digitaler Test Deutsch als Fremdsprache,* Digital Test of German as a Foreign Language) in order to explain why an answer is classified as correct or incorrect. We explore possibilities to operationalise these decisions using features that can be applied by an automatic scoring system, through two approaches: a decision tree and assigning points to errors. Results show that the automated scorings are promising, but cannot yet explain all decisions.

Introduction

Short-answer items are a popular item type and commonly used in listening comprehension tasks. Raters of short-answer items are usually provided with scoring guidelines and samples of correct and incorrect responses (Buck, 2001; Harding, Pill, & Ryan, 2011; Leitner & Kremmel 2021). These scoring guidelines should also include borderline cases, i.e., responses containing mistakes that may be accepted as correct as long as they are comprehensible and show sufficient understanding of the input text. Mistakes in test-takers' responses may concern different domains, e.g., grammar, semantics or spelling. In this study, we focus on spelling mistakes.

In order to deal with spelling mistakes, a 'rule of thumb' is applied in a study by Harding et al. (2011) according to which a misspelled response is accepted if 'it is a reasonably close phonemic match of the target word and the meaning of the response remains clear' (p. 117). A very similar approach is used by Leitner & Kremmel (2021), where spelling variations are accepted if they 'appear to resemble the correct response phonetically, [and] the meaning is unambiguous in the context of the item' (p. 129). While human raters might find it difficult to apply these principles and decide whether a misspelled answer is to be rated as correct or incorrect, using automated scoring could assure that all answers are rated according to the same rules, providing explainable and reliable decisions. This leads to three research questions for this study:

1. Which features do human raters take into account when deciding whether a misspelled answer is correct or incorrect?
2. Can we build a model that applies these principles automatically? How well do the decisions of the model agree with human decisions?
3. Do the rules we find apply to answers to all items?

The data used in this study are test-taker responses to a listening comprehension task of the digital TestDaF. The digital TestDaF is a high-stakes language test for admission to higher education in Germany and assesses the language ability of participants at Common European Framework of Reference for Languages (CEFR, Council of Europe, 2001) Levels B2 to C1 (Kecker, & Eckes 2022).[1] TestDaF consists of four subtests: Reading, Listening, Writing and Speaking. The Listening section consists of seven tasks, four of which are selected-response tasks and three of which are short answer tasks. In the task type used in this study, participants listen to a conversation and fill in five gaps in a table or schedule by typing their answers while listening to

[1] See Drackert et al. in this proceedings for further research directions regarding TestDaF.

the input text. Each item can be answered with only one or two words. All responses in this study were rated by experts of the TestDaF Institute who have been entrusted with the assessment of TestDaF performances for at least four years.

In a previous study (Laarmann-Quante, Schwarz, Horbach, & Zesch, 2022), we found that using superficial features, such as the distance between the given answer and the target answer on the character and phoneme level, is not fully sufficient for automatically predicting the acceptability of an answer. Therefore, we take a more linguistically motivated approach in the two studies presented here.

Study 1: Decision tree

In our first study, we manually designed a decision tree in which the acceptability of an answer is determined by the number and types of spelling errors that are present in a given answer. Based on a set of 389 unique spelling variants from 27 different items in total, TestDaF rating experts analyzed how many and what kinds of errors would lead to acceptable or non-acceptable answers. Unlike in Laarmann-Quante et al. (2022), we only considered non-word errors.

To classify the spelling errors, we used the Litkey Spelling Error Annotation Scheme described in Laarmann-Quante et al. (2019). This scheme is based on the graphematic theory by Eisenberg (2006) and was originally developed for spelling errors produced by German primary school children. We adjusted the scheme in order to account for particularities in the spelling variants of non-native German speakers. For example, we added a new category for umlauts *ä*, *ö, ü* spelled as *ae*, *oe*, *ue*, which is the standard spelling when e.g., keys for umlauts are not available on the keyboard. Furthermore, there are now categories for missing or superfluous interfixes in compounds (e.g., **Strandshaus* for *Strandhaus* ['beach house']). The enriched scheme consists of 83 fine-grained error categories in total. Based on this scheme, we distinguish between *minor errors*, *systematic errors* and *unsystematic errors* in this study.

Minor errors comprise capitalization errors and word separation errors because these errors in general do not change the pronunciation, meaning or comprehensibility of a word. *Systematic errors* are errors where some principle of the German orthographic system is violated or overgeneralized, for example spelling *Wald* ('forest') as **Walt* (final devoicing) or *bunt* ('colorful') as **bunnt* (consonant doubling). Note that systematic errors typically do not change the pronunciation of a word. All remaining errors are called *unsystematic errors* (with regard to the orthographic system), e.g., a missing letter as in **Wad* for *Wald* or a substitution as in **bumt* for *bunt*.[2] These kinds of errors typically involve a change of pronunciation and distort a word more than systematic errors. Our motivation for this distinction was the hypothesis that minor errors and systematic errors are more likely to be accepted by human raters than unsystematic errors.

The resulting decision tree is shown in Figure 1.

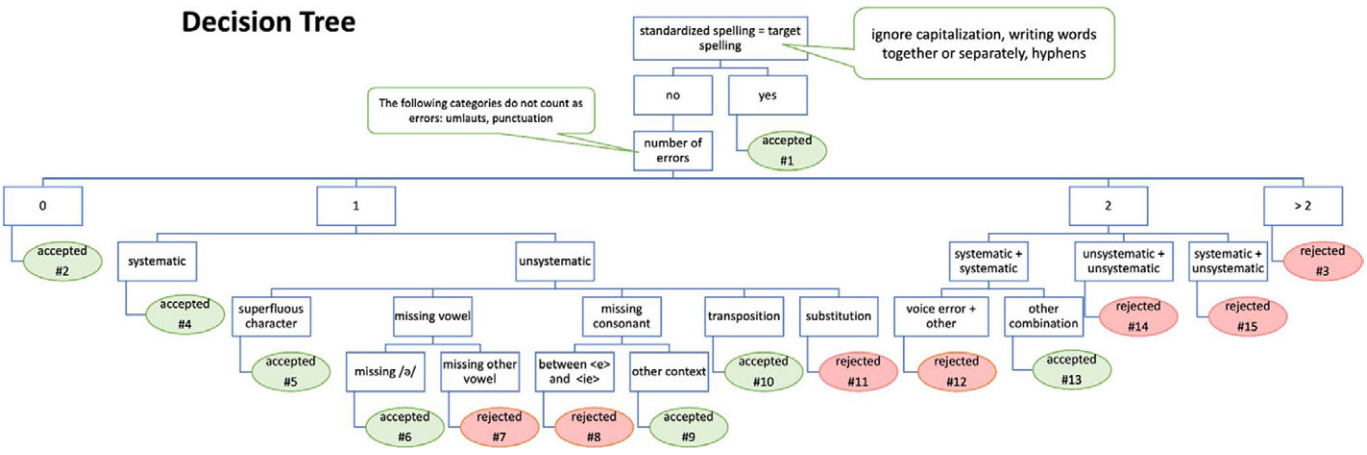


Figure 1 Manually constructed decision tree to decide about the acceptability of an answer

[2] In particular, following the scheme in Laarmann-Quante et al. (2019), minor errors are found under the levels SN ('Beyond single word spelling') and PC ('Punctuation') and unsystematic errors are found under level PGIII ('Edit operations'), while the other levels represent systematic errors.

The tree consists of five layers. The first layer deals with *minor errors* as described above, i.e., answers that only contain these kinds of errors are always accepted. The second layer looks at the number of spelling errors in the word[3]. Again, some categories do not count as errors, especially when umlauts are written as *ae*, *oe*, *ue* rather than *ä*, *ö*, *ü*. Answers with more than two non-minor errors are always rejected. For words with one or two errors, the third layer of the decision tree comes into play. Here, it is distinguished whether the errors are *systematic* or *unsystematic* as described above. The tree always accepts answers with one systematic error, whereas it always rejects answers with two unsystematic errors or one systematic plus one unsystematic error. For the remaining cases, the fourth and fifth layer of the tree define more fine-grained distinctions. For example, answers with a missing vowel letter are rejected, except when this vowel letter corresponds to the (hardly audible) schwa-sound (/ə/). These distinctions were the result of discussions with the human rating experts.

To apply this decision tree, the spelling errors are automatically analyzed as described in Laarmann-Quante (2016). We tested the decision tree on a subset of 224 unique spelling variants (which were used for training in Laarmann-Quante et al., 2022), and achieved an accuracy of 75%.[4] This result shows that using error categories is promising but there is still need for improvement.

Study 2: Assigning points to errors

When comparing the predictions of the model with the human ratings, we found that often the decision tree was too inflexible. Our discussion with the human raters showed that a) some more fine-grained distinctions were necessary and b) combinations of errors were not always treated adequately, especially with regard to *minor errors*. While minor errors by themselves are negligible, they contribute to the acceptability of an answer when further errors are present. As discussions with human raters revealed, errors 'add up' until an answer is regarded as too deviant from the target answer. To capture this notion, we decided to move away from the decision tree and to a different approach, i.e., assigning points to errors. When a certain threshold is surpassed, the answer is rejected, otherwise it is accepted by the model.

Methodology

To decide which errors would get how many points and where to set the threshold, we started off from the basic distinctions from the decision tree in Study 1: *Systematic errors* yield fewer points than *unsystematic errors* and up to two systematic errors in a word are accepted unless additional *minor errors* are involved. Based on the same set of 389 unique spelling variants that we also used in Study 1, we refined the categories and point assignments in an iterative process by comparing the model predictions to the human ratings, analyzing the prediction errors, discussing them with the human experts, adjusting the categories and point assignments, and so on. Table 1 gives an example for different errors in the word 'workshops' and their points. Table 2 shows the full categorization table. Answers with two or more points are rejected. We refer to the model that automatically makes a decision based on these points as the 'points model' in the following.

Table 1: Example for different error categories and their points

Error	*Description*	*Points*
workshops	Capitalization (minor)	0.50
work**sch**ops	*sch* for *sh* (special case)	0.50
worksho**pp**s	Systematic error regarding German orthographic system (consonant doubling)	0.75
worksh_ps	Missing unstressed vowel (unsystematic)	1.00

Results

First, we tested the model predictions on the 389 spelling variants that were involved in the development of the model ('development set'), and achieved a classification accuracy of 80% (79% when looking at the subset of 224 variants used for evaluating the decision tree, i.e., a notable improvement). These results have to be taken with caution, since the categories and points may overfit the development set. To test the generalizability of the model, we report its performance on new answers in the following.

We applied the points model to the test set from Laarmann-Quante et al. (2022), which contains 109 spelling variants from a new set of answers for five items. The same items had been used in the development set but only 22 of the answers in the test

[3] If the answer consists of more than one word, each word is assessed separately and the answer is only accepted if all words are accepted (erroneously separated words still count as one word).

[4] Note that the results in Laarmann-Quante et al. (2022) are not directly comparable to the results here because this study only focuses on non-word errors.

Table 2: Error categorization and points assigned to each category

Assigned points	*Error category*
0	• capitalization of first word (except for nouns) • capitalization if the whole word is written in uppercase • capitalization within the word • *ae, oe, ue* for umlauts *ä, ö, ü* • *ss* for *ß*
0.50	• other capitalization error • word-separation error • error concerning punctuation marks like hyphens • *a, o, u* for *ä, ö, ü* • *sh* or *ch* for *sch*, *sch* for *sh* • wrong but theoretically possible adjective ending • triple consonants • *r* for *er* (in case of [ɐ])
0.75	• systematic error regarding German orthographic system • missing vowel letter representing [ə] • *e* for *er* (in case of [ɐ]) • *a* for *e* representing [ə] • *k* for *ch*
1.00	• superfluous *e* at the end of the word if the target word has at least two syllables • superfluous vowel that does not change the number of syllables • missing vowel in unstressed syllable
1.50	• other superfluous consonant or vowel • transposition of letters (except if this introduces a new syllable) • missing consonant except between two vowel letters
2.00	• missing vowel in stressed syllable • *sc* for *sch* • other replacement error • superfluous consonant that turns an open into a closed syllable
2.50	• transposition of letters if this introduces a new syllable • missing consonant between two vowel letters • incorrect first letter with different pronunciation • error category 'diffuse'

set were present in the development set. On this set, the points model reaches an accuracy of 82%, showing that it does indeed generalize to new answers. While we evaluated the model based on the adjudicated human rating, we also wanted to see how well human raters agree in their decision. Therefore, two TestDaF experts rated all the answers in this test set anew, without knowledge of the previous ratings or the model predictions. With 92% (Cohen's κ = .82), their agreement was very high. When comparing the model predictions with the new human ratings, we found that the model agreed slightly more with Rater 1 (84%) than with Rater 2 (80%).

In order to test how well the points model generalizes to completely new items, we applied it to 187 spelling variants from 15 items that were not part of the development set. Based on the adjudicated human rating, the accuracy of the model was 73%, i.e., notably lower than on the previous test set. Again, the two experts rated these items anew and we found that their agreement was also notably lower on this set compared to the other test set (77%, Cohen's κ = .54), indicating that there are many borderline

cases in this set. The points model agreed slightly more with Rater 2 (78%) this time than with Rater 1 (73%). We see that human-model agreement is numerically comparable to human-human agreement. If we break down the results further, we see that in 119 cases, both human raters and the model agreed, in 43 cases, the human raters disagreed but the model agreed with one of them, and in only 25 cases, the human raters agreed but the model did not.

To look into this further, a next step could be to discuss the model's decisions with the human experts in order to decide whether the answers are borderline cases and the model's decision is acceptable, or whether the decision must be rejected and the model's error categories must be defined even more precisely.

Conclusion and outlook

This study has given us a better understanding on how human raters rate spelling mistakes in very short answers. Not only has the distinction between systematic vs. unsystematic errors proven useful, but within the unsystematic errors, we have identified more detailed error types that human raters seem to take into account when deciding if an answer is to be accepted or rejected.

We have built a model that can apply the rules we have found automatically. Model-human agreement is close to, but mostly lower than, human-human agreement. This leads us to the conclusion that there might remain features that need further investigation, e.g., word length, which might make a difference for human raters.

Application of the model to responses to new items showed that some features might have to be adjusted for specific target words.

References

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

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge: Cambridge University Press.

Eisenberg, P. (2006). *Das Wort* (Third edition). J.B. Metzler.

Harding, L., Pill, J., & Ryan, K. (2011). Assessor Decision Making While Marking a Note-Taking Listening Test: The Case of the OET. *Language Assessment Quarterly*, *8*(2), 108–126.

Kecker, G., & Eckes, T. (2022). Der digitale TestDaF: Aufbruch in neue Dimensionen des Sprachtestens. *Informationen Deutsch als Fremdsprache*, *49*(4), 289–324.

Laarmann-Quante, R. (2016). Automating multi-level annotations of orthographic properties of German words and children's spelling errors. *Proceedings of the 2nd Language Teaching, Learning and Technology Workshop (LTLT)*, 14–22. Available online: https://doi.org/10.21437/LTLT.2016-3

Laarmann-Quante, R., Schwarz, L., Horbach, A., & Zesch, T. (2022). `Meet me at the ribary' – Acceptability of spelling variants in free-text answers to listening comprehension prompts. *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, 173–182. Available online: https://doi.org/10.18653/v1/2022.bea-1.22

Laarmann-Quante, R., Ehlert, A., Ortmann, K., Scholz, D., Betken, C., Knichel, L., Masloch, S., & Dipper, S. (2019). *The Litkey Spelling Error Annotation Scheme: Guidelines for the Annotation of Orthographic Errors in German Texts*. Bochumer Linguistische Arbeitsberichte (BLA) 23. Available online: https://www.linguistics.rub.de/forschung/arbeitsberichte/23.pdf

Leitner, K., & Kremmel, B. (2021). Avoiding Scoring Malpractice: Supporting Reliable Scoring of Constructed-Response Items in High-Stakes Exams. In B. Lanteigne, C. Coombe, & J. D. Brown (Eds.), *Challenges in Language Testing Around the World* (pp. 127–145). Singapore: Springer.

Machine learning applications to develop tests in multiple languages simultaneously and at scale

Sarah Goodwin
Duolingo

Lauren Bilsky
Duolingo

Phoebe Mulcaire
Duolingo

Burr Settles
Fawm Labs[1]

Abstract

Spanish and French are two of the most commonly learned additional languages worldwide (Blanco, 2021). However, there exist few on-demand, low-cost, rapid assessments for learners of these languages to determine their proficiency level. By combining machine learning (ML), engineering, and materials development expertise, language learning applications are well positioned to tackle this challenge. We describe the development, difficulty estimation, and piloting of prototype tasks of listening and reading. Expert-rated linguistic features for CEFR and corpus frequency and usage data were used to train multilingual transformer-based ML models (RoBERTa; Liu et al., 2019), allowing estimation of item difficulty before any learner response data collection. English L1 adult learners (1,097 Spanish; 599 French) took pilots to validate ML item difficulty estimates. We found moderate correlations between ML estimates and empirical item difficulty (r = –0.521 ES; r = –0.404 FR). Finally, we discuss implications for scaling content creation and difficulty estimation.

Background

Learners acquiring languages for various purposes (school, online, tutoring, etc.) may need to certify their language proficiency. In the US, where multilingualism is not so much the norm as it is in Europe, learners seeking to describe their proficiency level often must self-assess, e.g., with frameworks such as the Common European Framework of Reference for Languages (CEFR; Council of Europe, 2001; 2020) or American Council on the Teaching of Foreign Languages (ACTFL, 2017) or estimate based on the number of semesters or years they have studied a language. Given the diversity of learning contexts, there is lack of interpretation across curricula and purposes for language use. Moreover, existing assessments are generally only available at limited times of day, take hours to sit, and cost 60 to 195 EUR. This presented an opportunity for Duolingo to harness our machine learning, engineering, and linguistics expertise to scale language assessments. Especially since 2018, large multilingual corpora are more widely available than previously, and there have been advances in large language models which can extend to content development (e.g., ChatGPT text generation). Existing processes helped us scale item generation procedures, content and fairness review, the user interface, and adaptive scoring methods from the Duolingo English Test (DET). The DET is a high-stakes computerized adaptive test of English proficiency, delivered online on-demand without testing centers, with cyclical, interconnected procedures involving humans and artificial intelligence throughout all item development, scoring, and administration (Burstein, LaFlair, Kunnan, & von Davier, 2022). We applied several of these DET procedures to prototype tests of Spanish and French (hereafter 'target languages').

[1] Research conducted at Duolingo.

Methods

This project's goal was to craft proof-of-concept prototypes toward the development of tests aiming at all levels of the CEFR (Council of Europe, 2001; 2020). Examinees would be acquiring target languages in locations where that language was not a primary language for education or commerce, but people would not need to be enrolled in a formal course of study at the time of testing. Envisioned use cases include placement into secondary/tertiary school language courses, and university readiness for instructional programs and study abroad.

Drawing on DET infrastructure available for computerized adaptive language test administration, we chose four of our existing item types. These were purposely chosen to balance among ease of development, fast administration, construct coverage, and strength of psychometric properties. The items did not require much special test preparation for examinees from diverse backgrounds to understand how to complete them. The prototypes were taken in a desktop browser; mobile devices were not permitted. Examinees did not need to be familiar with the exercise types from either the Duolingo language learning application or the DET. Listening and reading items consisted of the following types.

Table 1: French and Spanish prototype item types

Item type	*Description*
Dictation	Examinees transcribe a target language utterance they heard
Yes-no vocabulary, audio format ('audiovocab' hereafter)	Examinees listen and select real words from groups of real words and nonwords
Yes-no vocabulary, printed text format ('textvocab' hereafter)	Examinees see words and nonwords on screen and pick the real words
c-test	Gap-filling reading where passages' first and last sentences are intact, the second half of every alternating word in the remaining sentences is damaged, and examinees fill the blanks

The nonwords in audiovocab and textvocab items mimic target language phonology, orthography or morphology, but have no meaning and are not plausible words in the target language or English. We also followed item rules for the assessment format (e.g., Counsell, 2018; Riggs & Maimone, 2018); for instance, for the c-test we did not damage: the same content word after its first appearance, more than 50% of function words, or proper names identified using named entity recognition (a type of information extraction tool used to identify places, people, and other names).

We generated items based on several steps. To determine yes-no vocabulary stimuli, we supplemented existing datasets with tokens from open-source corpora. For nonword generation, we trained a character recurrent neural network on large sets of words. Both real word and pseudoword generation involved filtering out homophones, proper names, profanity, and words from other languages. For all items, we estimated the CEFR level based on statistical learning from app responses and also expert pedagogical decisions regarding when linguistic features should be introduced to learners. Because nonwords do not have instances in corpus data, we could not use part-of-speech tagging or frequency to estimate level, so we based difficulty on string length, number of syllables, and sequences of characters. For dictation, classification was done at the sentence level, and for c-test at the passage level. Items were also reviewed for fairness and bias to ensure they did not contain inappropriate, potentially-triggering, or regionally-specific content, and that nonwords were not confusable for real words. We applied multilingual transformer modeling (RoBERTa; Liu et al., 2019) and methods from Settles, LaFlair and Hagiwara (2020) and McCarthy et al. (2021) to determine machine-learned difficulty estimates, allowing us to predict the item level even before any test administration. Listening input audio was text-to-speech voices programmed by the Duolingo Speech Lab.

Vocab stimuli were drawn from the same pool as audiovocab and textvocab. The same token could occur in aural and visual form but not in two items of the same type, increasing the size of the item bank and allowing flexibility for whether the words and nonwords were presented as audio or printed text. Inflected forms of words, such as conjugations or plurals, appeared in yes-no vocabulary, but the same base or root word form could not appear in the same item (e.g., marche, marches, or marchons [I/you/we walk], or marcher for the infinitive form [to walk], are all permissible as real words). Not every form of a word was tagged at the same estimated CEFR level. In other words, examinees may not have known all inflections and derivations of a word (i.e., all its word forms, and related words and their forms; e.g., développer, développemental [develop, developmentally]), despite knowing the base form of the word. Tokens contained three characters minimum and 18 maximum. For audiovocab, nine words and nonwords were presented per screen, and for textvocab, 18 words and nonwords per screen. We grouped together real words

and nonwords of similar estimated difficulty, with some variation, to reduce the chance that beginner and advanced vocabulary appeared together. We avoided duplicate stems per each group of stimuli.

For audiovocab, examinees could click the stimulus an unlimited amount of times to relisten during the timeframe; clicking at the left of the rectangle replayed the audio, and clicking the checkbox at the far right of the rectangle toggled the word as selected or deselected. Examinees selected a word, and clicked again to deselect it if they changed their minds, as many times as they wanted before time expired. On dictation, examinees could begin typing as soon as the audio began playing. The dictation item stimuli could be played three times maximum; i.e., after the first audio play, two replays were permitted. Examinees clicked an orange circular button with a speaker icon to play the utterance again. Each stimulus was a complete sentence, containing at minimum a verb. The sentences contained no proper nouns. Stimulus sentences ranged from 3 to 20 words. For audiovocab and dictation, no printed text transcriptions of audio input appeared on screen.

C-test passages ranged from 30 to 160 total words. The first and last sentences of the paragraphs were left intact. For every other word, blanks corresponded to the number of letters the examinee was expected to type to finish each word. A minimum of eight words and a maximum of 25 per passage were damaged. Words two letters or longer were eligible to be damaged, and if the word contained an odd number of letters, (N/2)+1 (rounded up to a whole number) letters were damaged. Words with hyphens were not damaged, nor units of measurement. When two words were contracted with an apostrophe in French, the apostrophe was not counted toward the number of characters (e.g., l'ami became l'a _ _).

We conducted pilots in two phases. For the initial pilot, we recruited Mechanical Turk participants and administered language acquisition questionnaires, filtering by those who did not use the target languages extensively in their formative years. The items proved too difficult for this population, so we restricted the pool to easier sets. For the second phase of piloting, we recruited from Duolingo Spanish and French language learning app courses for English speakers, selecting only those in the US who opted into research opportunities (see Jiang, Rollinson, Plonsky, Gustafson, & Pajak, 2021, for similar learner populations). Examinees had to have reached the app skills that are mapped to A1.2 CEFR, and logged in recently enough in the 60 days prior to recruitment. Examinees in both pilots received an email invitation to participate in practice tests. As an incentive, examinees received a 30-day trial premium subscription to the language learning app.

Examinees (1,097 Spanish; 599 French) took six item sets. The first was the same anchor textvocab set given to all examinees, and the other five were always one c-test, two dictation, one textvocab, and one audiovocab, in any order. Once each item in the pool had approximately 30 responses, the pool was deactivated and replaced with a new one. This meant there was less chance of overexposure of items, and inter-item score comparisons within a pool became more meaningful because the item responses were likely from examinees at similar ability levels. Each set had an individual timer of three minutes for c-test, 90 seconds for audiovocab, and 60 seconds per dictation or textvocab screen. Tests were unproctored, but given the short task time limits, it was unlikely examinees could consult outside sources.

Data collection informed validation of the machine-learned difficulty estimates, described next.

Results and discussion

Average item scores and machine-learned item difficulties were moderately correlated ($r = -0.521$ ES, $r = -0.404$ FR). Values are negative because the more difficult the item, the fewer examinees scored it correctly, and the higher its estimated CEFR level. Mean item scores ranged from 0.062 to 0.976, indicating a good range across the item difficulty spectrum.

There were 156 unique Spanish items, and 87 unique French items that had 30 or more responses each. This suggested positive effects for pooling groups of items, because we gave only six items per examinee. This pooling resulted in gaining enough information about those items to estimate empirical difficulty without overexposing the test content.

We also used examinees' highest unit reached in the language learning app (Units 3 to 10) as a proxy for proficiency level. Unit level and session scores correlated at Pearson's $r = 0.325$ ($p = 0.000$). However, unit level from the language learning app is a measure with a great deal of noise, as learners are heterogeneous and their unit level may not be a good indication of their total amount of, or recency of, language practice.

Spanish c-test correlations were not as strong as correlations for the other three item types in Spanish, and French textvocab and dictation correlations were stronger than for the other two item types. Further investigation is needed, with more data collection and refinement of ML models, to explore possible reasons for these trends.

This project gave us insight into scaling content creation and difficulty estimation. When more learners participate in the assessments, the data collected can be used to refine the ML models, improving the accuracy of item difficulty estimates and leading to better overall assessment quality. This continuous improvement can be further scaled to enable cross-language applicability, allowing for the development of on-demand, low-cost, rapid assessments in various languages. Additionally, the

tasks from the English, Spanish, and French testing platform can be combined with data from the Duolingo language learning application, benefiting learners by giving them a way to track their progress and proficiency level.

References

ACTFL. (2017). *NCSSFL-ACTFL Can-Do Statements*. Available online: https://www.actfl.org/resources/ncssfl-actfl-can-do-statements

Blanco, C. (2021, December 6). *2021 Duolingo Global Language Report*. Duolingo. Available online: https://blog.duolingo.com/2021-duolingo-language-report

Burstein, J., LaFlair, G. T., Kunnan, A. J., & von Davier, A. A. (2022). *A theoretical assessment ecosystem for a digital-first assessment: The Duolingo English Test*. Duolingo Research Report DR-22-03 DRR-22-01. Duolingo. Available online: https://go.duolingo.com/ecosystem

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge: Cambridge University Press.

Council of Europe. (2020). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Companion Volume*. Strasbourg: Council of Europe Publishing.

Counsell, C. (2018). The C-test in French: Development and validation of a language proficiency test for research purposes. In J. M. Norris (Ed.), *Developing C-tests for estimating proficiency in foreign language research* (pp. 203–230). Berlin: Peter Lang.

Jiang, X., Rollinson, J., Plonsky, L., Gustafson, E., & Pajak, B. (2021). Evaluating the reading and listening outcomes of beginning – level Duolingo courses. *Foreign Language Annals*, *54*(4), 974–1,002.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*. Available online: https://arxiv.org/abs/1907.11692

McCarthy, A. D., Yancey, K. P., LaFlair, G. T., Egbert, J., Liao, M., & Settles, B. (2021). Jump-starting item parameters for adaptive language tests. In The Association for Computational Linguistics (Ed.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (pp. 883–899). Available online: https://aclanthology.org/2021.emnlp-main.0.pdf

Riggs, D., & Maimone, L. L. (2018). A computer-administered C-test in Spanish. In J. M. Norris (Ed.), *Developing C-tests for estimating proficiency in foreign language research* (pp. 265–294). Berlin: Peter Lang.

Settles, B., T. LaFlair, G., & Hagiwara, M. (2020). Machine learning-driven language assessment. *Transactions of the Association for Computational Linguistics*, *8*, 247–263.

New test format – new research agenda: An overview of the technology-related research at g.a.s.t.

Anastasia Drackert
g.a.s.t. e.V./TestDaF-Institut

Sonja Zimmermann
g.a.s.t. e.V./TestDaF-Institut

Daniela Marks
g.a.s.t. e.V./TestDaF-Institut

Denis Korflür[1]
g.a.s.t. e.V./TestDaF-Institut

Abstract

The portfolio of the Society for Academic Study Preparation and Test Development (Gesellschaft für akademische Studienvorbereitung und Testentwicklung, g.a.s.t. e.V.; see www.gast.de) includes language tests and scholastic aptitude tests, as well as online German courses for international study applicants. In the last years, digital versions of our high-stakes tests have been introduced, e.g., in late 2020 the digital Test of German as a Foreign Language (Test Deutsch als Fremdsprache, TestDaF). The introduction of the new test has consequently called for new validation studies. This paper summarizes different stages in the development and research on the digital TestDaF, starting from the needs analysis, research before the initial administration of the test, as well as current research regarding the different aspects of the exam.

Introduction

The Test of German as a Foreign Language (Test Deutsch als Fremdsprache, TestDaF) is a language admission test for German universities. Since its development, its aim is to assess whether test-takers have an adequate level of language proficiency to effectively accomplish a range of tasks within a university setting (Norris & Drackert, 2018). The exam was introduced in 2001 and until 2020 only existed in a paper-based format.

The increasing prominence of digitalization in German universities, along with the recognition of the manifold advantages presented by computerized language proficiency tests encompassing diverse task types and more efficient evaluation procedures, particularly in the context of global test administration, underscored the imperative for the digitalization of the TestDaF in its future trajectory. Given the advantages of computerized testing, the decision was made to meticulously revise the initial test format, rather than merely transfer the paper-based version of the exam into a computerized format. All in all, it took about 10 years to come up with a solid digital test with empirically tested and proved qualities.

After the needs analysis, several years were spent on the development of the test format and the in-house technical infrastructure. Both the new task types and the user interface were tested and revised in several try-outs and empirical studies (e.g., using thinking aloud, stimulated recall, questionnaires and other methods) until the test format was finally defined in 2017. The extended timeframe is, in part, a result of the ongoing development of a software for presenting the new task types, onscreen grading, and other essential modules. After the test format had been finally defined, the following years were spent on developing items and conducting field tests to compile a sufficient item bank, and investigating different

[1] We would like to acknowledge further colleagues who were actively involved in the development of the new test format and the research described below: Günther Depner, Thomas Eckes, Gabriele Kecker, Anja Peters, Leska Schwarz, Frank Weiss-Motz as well as colleagues from the IT department.

aspects of the test, e.g., the qualities of new task types or scoring procedures. The first exam took place in fall 2020 in the middle of the pandemic. [2]

Overview of the research studies

In the following, we provide a summary of studies that focused on the novel aspects of the exams, the ones that differed significantly from the paper-based format and thus formed the central focus of the initial research agenda: novel task types, a new scoring rubric, and a different approach to the preparation materials.

Needs analysis and new task types

The development of the new test format began with a needs analysis involving over 1,300 international students and 120 university teaching staff. We conducted focus groups and interviews with experts and students, used surveys and analyzed curricula in different academic disciplines at various German universities. The overarching aim of the needs analysis was to define which communicative tasks are particularly frequent, relevant and/or challenging for (international) students during the first year of university study in Germany. As reported in Arras (2012) and Marks (2015), we found that the four test sections of the paper-based TestDaF adequately cover many of the task types identified by expert informants. However, several types of real-world tasks identified in the needs analysis as frequent and relevant were not included in the paper-based version of the exam, partially due to the format restrictions. All in all, to assess academic language competence in a more authentic way, several integrated tasks were included into the digital exam.

In particular, in one of the listening tasks of the digital TestDaF students have to watch a video of a lecture accompanied by the PowerPoint slides and take short notes on the computer. In a speaking section, students have to give a short presentation based on a PowerPoint slide or react to an oral statement of another student in a seminar using information from a visual source on the same topic. In the reading comprehension section, we included a task in which students have to point at mistakes in a short summary based on a longer academic text and a graph on the same topic. Given the novelty of these tasks in the field of German as a Foreign Language testing, leading to unfamiliarity among both learners, teachers and raters, the decision was made to prioritize the investigation of the integrated tasks.

Construct of the integrated writing task

Integrated tasks are broadly defined as 'test tasks that combine two or more language skills to simulate authentic language-use situations' (Plakans, 2013, p. 1). For instance, academic writing often requires students to synthesize information from various sources and incorporate it into their own texts. To test this ability in the digital TestDaF, we ask test-takers to summarize information from a written text and a graphical input in relation to a given question (see Figure 1).

While integrated writing tasks are commonly employed in university admission examinations in English speaking countries, the foundational construct of such tasks has remained an unresolved matter, particularly in the context of assessing integrated writing tasks using graphical data as in the integrated task included in the digital TestDaF. Specifically, the factors influencing performance, such as the extent to which writing ability or reading skills contribute to test outcomes, remained unclear (e.g., Cumming, 2013). These research gaps were addressed by Zimmermann (2022).

With the aim to shed light on the construct underlying the integrated writing task, Zimmermann (2020, 2022) used a mixed methods design (eye-tracking, stimulated recall, analysis of the written texts, etc.) and investigated the cognitive processes involved in the task completion, the quality of the written products, and the reliability of the scoring of the integrated writing task. Among other things, she found that even though there were individual differences among participants, most of them approached the task in a similar way, engaging in cognitive processes relevant for writing and reading that varied at different stages of the writing process.

As expected, findings showed that test-takers' responses relied to a great extent on the source material as aimed by test developers. However, contrary to expectations regarding summary writing, participants also included information that was based on their own background knowledge and not on the information from the sources. Presumably, due to the novel task type, test-takers encountered challenges in applying their writing strategies to the completion of an integrated task. As a response to this finding at the pre-operational stage, short tutorials for every single one of the 23 test tasks of the digital TestDaF were produced, informing prospective test-takers about the task demands. Furthermore, a decision was made to develop new types

[2] For further information on the development of the digital TestDaF see Kecker and Eckes (2022) and Kecker, Zimmermann and Eckes (2022).

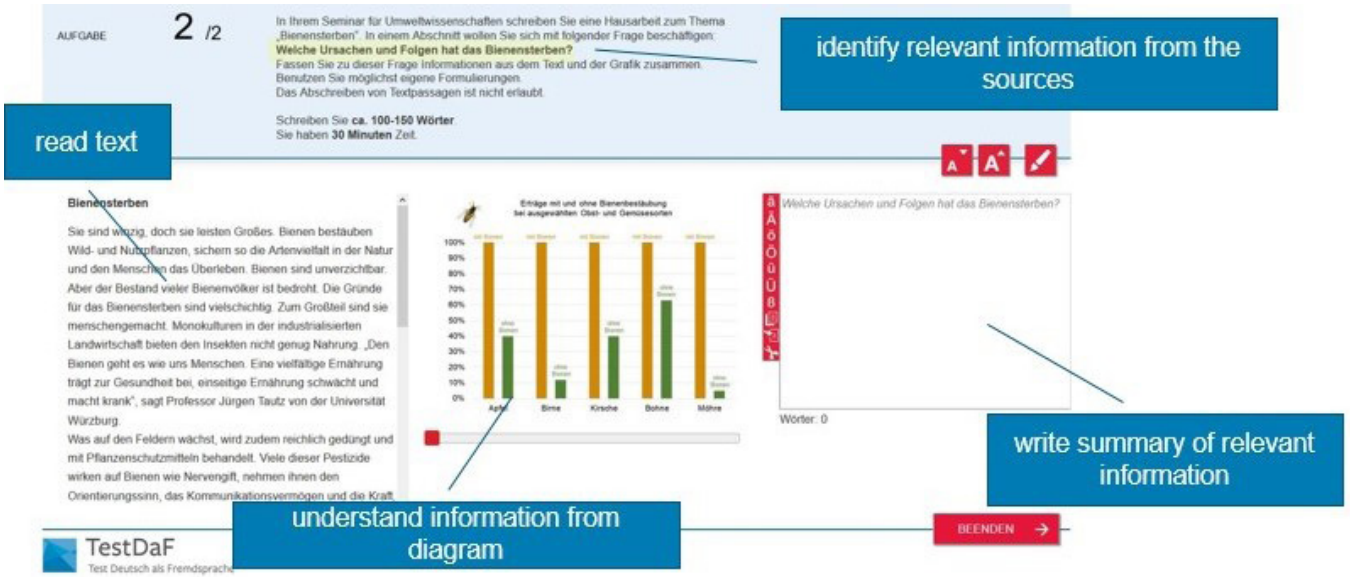


Figure 1 Integrated writing task

of preparation materials (see the section after the next one) not only to facilitate comprehensive exam readiness but to prepare students for successful handling of comparable tasks in their studies.

Scoring of the writing tasks

One of the decisions we had to make during the developmental stage is which scoring rubric to use for the evaluation of the written and oral responses. For the paper-based exam, we use analytic scales, which according to previous research offers several advantages: it allows for the distinct evaluation of specific aspects, it is easier to use and therefore requires less training for the reliable use (cf. Weigle, 2002). At the same time, clearly defined independent rubrics for each criteria are difficult to develop, the criteria correlating highly with each other and when combined into an overall grade, there is a risk that experienced raters align their assessment of the individual criteria with the expected overall grade. Likewise, a holistic scale has its advantages and disadvantages. On the positive side, a holistic scale focuses more on the strengths of a text, is more authentic in terms of text processing and is more efficient in terms of its use (cf. White, 1989). On the negative side, an extensive training of raters is crucial for the reliable use of a holistic scale.

To investigate which scale works best for the digital TestDaF we conducted a comparative study with the use of both rubrics for writing (Korflür, Marks, & Weiss-Motz, 2022). In particular, we first analyzed the analytical ratings of 20 raters of the total of 382 written texts from the first field test (2017) and then compared those with the holistic ratings of the same raters (nine raters, 127 texts) completed in 2019. We found high intercorrelations between the six criteria on the analytical scale (from 0,76 to r = 0,83 (p <.001). Furthermore, the regression analysis showed that each criterion has almost the same high influence on the total score, with two criteria chosen at random as predictors already explaining 82% to 86% of the variance of the total score. We also found a high intra-rater consistency between the rating (weighted Kappa 0,71) with a tendency to rate the written performances more severely using a holistic scale. Based on the findings, the decision was made to use the holistic scale that was slightly revised after the study. Furthermore, we made some changes into the rater training and extended it as necessary. Subsequently the use of the revised scale was investigated in Zimmermann (2022) which found that the scale was able to reliably distinguish among levels of proficiency, i.e., test-takers were placed at a certain level based on characteristics of their performance.

Preparation materials

In the piloting phase, it became clear that test preparation should comprise more than the format-focused familiarization with test tasks and test-taking strategies. Within the framework of learning-oriented approaches to assessment (Carless, 2007) that we chose to adapt for developing test preparation materials, understanding the real-life contexts and communicative significance of tasks in German Higher Education is crucial not only for improving test success rates but primarily for advancing long-term goals of studying. Furthermore, the approach entails active involvement with performance criteria and quality, either one's own or that of peers. Consequently, the preparation materials that we developed for our language centres were not structured according

to the sections of the digital TestDaF or the individual tasks. Instead, they focused on the competencies that underlie the test tasks across the different test sections as well as on raising test-takers' awareness of the requirements of the test tasks and how these are related to the target language use domain (Depner & Peters, 2022).

After the initial implementation, we conducted a study to evaluate the perceptions and effectiveness of this approach from a teacher's perspective (Zimmermann, Schwarz, Peters, & Depner, forthcoming 2024). To this aim, we interviewed teachers who used the materials in their test preparation courses. The interviews showed that the teachers appreciated the materials that helped them in gaining a better understanding of the concept, focusing on the underlying competencies rather than on single task requirements. They especially stressed the importance of the awareness-raising activities that helped learners to build on their previous experience and reflecting on useful strategies to cope with the requirements of the new tasks. At the same time, similar to the findings in other educational contexts (e.g., O'Sullivan, Dunn, & Berry, 2021), the teachers stressed that many test-takers were mainly interested in passing the test, and hence preferred the traditional approaches for test preparation like teaching to the test using downloadable practice material. An investigation of the learners' perspectives on the developed preparation materials would be the next logical step in this direction.

Conclusion and next steps

The studies summarized above targeted the most novel aspects of the new digital test: new task types, the new scoring rubric and a different approach to test preparation materials. Based on the empirical findings, substantial changes have been introduced to the rater-training process and to how we advise learners and teachers to approach the preparation for the test.

At the same time, we have started research that genuinely makes use of the advances in the field of language technology, in particular for the development of automated scoring systems as reported in Schwarz and Laarmann-Quante (this volume) for the evaluation of short responses. The next step would be to extend this research and develop automated scoring systems for the evaluation of test-takers' written and oral responses.

With the introduction of a new digital exam that more precisely mirrors the study requirements, we hope for a positive washback for language learning. Whether this washback will take place and make students better prepared for the successful studies is open to investigation.

References

Arras, U. (2012). Im Rahmen eines Hochschulstudiums in Deutschland erforderliche sprachliche Kompetenzen - Ergebnisse einer empirischen Bedarfsanalyse. In T. Tinnefeld (Hrsg.), *Hochschulischer Fremdsprachenunterricht: Anforderungen – Ausrichtung – Spezifik* (S. 137–148). Saarbrücken: HTW Saar.

Carless, D. (2007). Learning-oriented assessment: conceptual bases and practical implications. *Innovations in Education and Teaching International*, *44*(1), 57–66.

Cumming, A. (2013). Assessing integrated writing tasks for academic purposes: Promises and perils. *Language Assessment Quarterly, 10*(1), 1–8.

Depner, G., & Peters, A. (2022). Sprachkompetenzen entwickeln und trainieren: Ein Konzept für eine kompetenzorientierte Prüfungsvorbereitung. *Informationen Deutsch als Fremdsprache*, *49*(4), 325–345.

Kecker, G., & Eckes, T. (2022). Der digitale TestDaF: Aufbruch in neue Dimensionen des Sprachtestens. *Informationen Deutsch als Fremdsprache, 49*(4), 289–324.

Kecker, G., Zimmermann, S., & Eckes, T. (2022). Der Weg zum digitalen TestDaF: Konzeption, Entwicklung und Validierung. In P. Gretsch & N. Wulff (Hrsg.), *Deutsch als Zweit- und Fremdsprache in Schule und Beruf* (S. 393–410). Paderborn: Brill Schöningh.

Korflür, D., Marks, D., & Weiss-Motz, F. (2022). Analytisch oder holistisch? Welchen Einfluss hat die Beurteilungsmethode auf das Verhalten von Beurteilenden und die Ergebnisse von Prüfungsteilnehmenden?. *Informationen Deutsch als Fremdsprache 49*(4), 346–368.

Marks, D. (2015). Prüfen sprachlicher Kompetenzen internationaler Studienanfänger an deutschen Hochschulen – Was leistet der TestDaF?. *Zeitschrift für Interkulturellen Fremdsprachenunterricht, 20*(1), 21–39.

Norris, J., & Drackert, A. (2018). Test review: TestDaF. *Language Testing, 35*, 149–157.

O'Sullivan, B., Dunn, K., & Berry, V. (2021). Test preparation: an international comparison of test takers' preferences. *Assessment in Education: Principles, Policy & Practice*, *28*(1), 13–36.

Plakans, L. (2013). Assessment of integrated skills. In C. Chapelle (Ed.),*The Encyclopedia of Applied Linguistics* (pp. 1–8). Malden: Wiley-Blackwell.

Weigle, S. C. (2002). *Assessing Writing*. Cambridge: Cambridge University Press.

White, E. (1989). *Developing Successful College Writing Programs*. San Francisco: Jossey-Bass.

Zimmermann, S. (2020). „Das ist doch Leseverstehen!" – Eine empirische Untersuchung zum Konstrukt von integrierten Schreibaufgaben. In A. Drackert, M. Mainzer-Murrenhof, A. Soltyska & A. Timukova (Hrsg.), *Testen bildungssprachlicher Kompetenzen und akademischer Sprachkompetenzen. Zugänge für Schule und Hochschule* (S. 187–213). Frankfurt am Main: Peter Lang.

Zimmermann, S. (2022). *Validating integrated writing tasks – A mixed-method approach to investigate the construct of summarization.* [Dissertation, Universität Bremen].

Zimmermann, S., Schwarz, L., Peters, A., & Depner, G. (forthcoming 2024). Enhancing teachers and test takers' assessment literacy? Insights from test preparation for the digital TestDaF. In B. Baker & L. Taylor (Eds.), *Language Assessment Literacy and Competence Volume 1: Research and Reflections from the Field*. Studies in Language Testing Volume 55. Cambridge: Cambridge University Press & Assessment.

Using multi-level tests in benchmarking projects in Iberia

David Bradshaw
Cambridge University Press & Assessment, United Kingdom

Victoria Peña Jaenes
Cambridge University Press & Assessment, United Kingdom

Abstract

The importance of speaking at least one foreign language has led to the implementation of language programmes to improve students' foreign language ability. Once the programmes are established, assessment tools are used to offer diagnostic information and to measure advances in proficiency levels.

Cambridge University Press & Assessment has developed a Multi-Level Test (MLT) to offer an assessment solution to support such language programmes. The test assesses all four skills and reports results in alignment with the levels of the CEFR.

This presentation focused on the use of the MLT in benchmarking projects in Spain. The MLT solution measured the English language proficiency levels of primary and secondary students, and the results informed the recommendations to capitalise on existing strengths and mitigate weaknesses of these programmes.

Introduction

Following the European Union and the Council of Europe's lead on languages, recognising them as crucial elements of modern society, member states put into place different measures and policies to foster foreign language learning and attain the goal proposed by the European Council in Barcelona (2002), that students should study two languages in addition to their first language – the so-called 'mother tongue +2 goal' (European Commission, 2003, p. 7). In this context, regions and school groups in Spain established bilingual/plurilingual programmes (e.g., Consejería de Educación, Formación y Empleo, 2018), and interest has gradually grown in these sectors in using assessment tools to obtain diagnostic information and to measure advances in proficiency levels. The information obtained on students' performances may help inform the choice of learning materials and professional development requirements among other possible improvements to the programmes, as well as helping establish exit levels.

The benchmarking test

Although standard proficiency tests can and have been used to measure progress in such language programmes, the high stakes nature of these tests and the costs and security restrictions inherent in this have meant that they are less fit for purpose as benchmarking tools for large cohorts, being deemed too expensive by many possible clients. Cambridge University Press & Assessment (hereafter 'Cambridge') have developed a Multi-Level Test (MLT) to offer an accessible assessment solution to support language programmes by measuring performance at a cohort level, without necessarily providing individual results. This benchmarking test has three different versions to meet the needs of different age groups (11 to 13, 14 to 17 and adult – although the adult version is not routinely used currently), and has been designed to measure ability and report results aligned to the Common European Framework of Reference for Languages (CEFR, Council of Europe, 2001). The test for upper-primary (11–13-year-olds) measures abilities from Pre A1 up to B2 or above, while for upper-secondary (14–17-year-olds) and adults, the test measures abilities from Pre A1 up to C1 or above. The term 'or above' in the case of primary describes a performance at B2 (C1 in the case of secondary) or one that is stronger than what is necessarily expected at B2 but which the test, due to its construct, cannot place with confidence at C1 (or C2 in secondary).

The test is modular, and clients can decide which skills to focus on depending on their needs. The core test consists of multi-level computer-adaptive listening and reading components. 'Computer-adaptive' means that the system uses an AI algorithm to adjust the difficulty of each question for each individual candidate, based on their previous responses. So, if a candidate gets an item

correct, the following item will be more challenging, whereas if they get an item wrong the following item will be less difficult. The system continues to adjust levels until a stable estimate of the candidate's ability has been reached and the system is confident that the right level has been identified. The algorithm ensures that every candidate is given items covering a range of task types so that a full range of subskills is tested in every case. This means that each candidate has an individual pathway through the test, not necessarily seeing the same items as other candidates in the same session. It also means that some candidates may take slightly longer than others to complete the test and some may need to answer more items. Task types include multiple choice and multiple matching. All items in the reading and listening components require objective answers and are automarked by the system.

Clients can also choose to include speaking and/or writing components in the test. The speaking and writing components are not computer-adaptive, but are computer-based and designed to be accessible at different levels, while being age-appropriate. The automarker, in the case of writing, or the examiner for speaking, will determine the level of language used by the candidate and assign a CEFR level accordingly. The writing automarker has been trained using hundreds of thousands of annotated and tagged candidate scripts drawn from the Cambridge Learner Corpus and the standards used are based on an extensive standard-setting project. Over an extensive period, the automarker proved highly consistent (UCLES, 2018), with different studies checking the correlation between automarker rank scoring and the combined expertise of a panel of expert examiners typically reporting a rank order correlation of 0.90 or above. Speaking is currently marked by human examiners, with different examiners marking separate tasks independently.

The test is web-based and does not require any software to be installed on the clients' computers. On agreeing to carry out a benchmarking project, the client is provided with minimum system requirements and demo material to ensure that their equipment is suitable to run the test. Candidates take the test in exam conditions in their school, with invigilation provided by school staff, arranged in conjunction with the team in the local Cambridge office (see below for more detail on how the test is organised and administered).

The MLT solution offers diagnostic information at group level, helping institutions understand students' current level with an overall CEFR level and skills profile. This information can help identify strengths and areas requiring specific support within the cohort of students tested as well as highlighting possible differences in performance within the cohort (comparing performance in different schools or streams, for example). In addition, it can inform realistic exit levels in the shorter and longer term for different stages. The data can also contribute to making the right decisions in terms of certification, to ensure that students sit the exam at the most suitable level for them and have a positive exam experience. The benchmarking test is typically used as part of a wider diagnostic exercise, and as such it allows Cambridge to propose a tailor-made solution for the client. For example, it can help identify areas where a change of methodology may be appropriate or enable specific training programmes to be developed for the teachers in the schools involved. Some clients have opted to carry out benchmarking projects regularly to understand if the measures implemented are having the desired impact on students' learning and where necessary adjust their approach.

Administering the benchmarking test

The administration of the MLT solution is the main part of a larger diagnostic project where the client and Cambridge are in close contact. The project can be divided into three main stages: pre-exam, exam window, and post-exam.

The pre-exam stage starts with conversations between Cambridge and the client to understand their needs and the objectives as well as any potential constraints. With that information in mind, there is agreement of the cohort, the variables to analyse, the modules to administer (which will necessarily determine the results obtained), and the level of detail in the reporting. The conversations will also help establish exam versions, key dates and the project team, which ideally should include Cambridge staff as well as collaboration from the client. At this point, the client carries out an analysis of their facilities and equipment with Cambridge support, confirms the research questions, and shares the relevant data. After that, Cambridge produces the logins and organises the project team.

At the beginning of the exam window, Cambridge provides logins and attendance lists. Meanwhile, it is advisable that the client shares detailed information about the sessions and any incidents that may have occurred. Cambridge recommends that the staff participating in the exam administration be made up of invigilators provided by the client and a Cambridge supervisor, who offers on-site support to complement the ongoing contact with the wider project team. As the exam progresses and the first answers come through the system, quality checks are carried out to ensure that the data obtained are of sufficient quality to be assessed and to suggest changes in subsequent exam sessions if necessary.

The post-exam stage involves marking answers, carrying out additional quality control checks as well as the analysis of data. The results can be reported in several formats to be agreed with the client although, in general, they are included in a report produced by Cambridge. This option offers a thorough analysis of the context, the methodology and the results. The level of

detail allows Cambridge to answer the research questions and look into the data on the basis of the variables agreed with the client, which usually involves a number of comparisons within the cohort. Finally, a number of recommendations for learning, professional development and future research are made. The submission of the report can be followed by a meeting with the client where views are shared and future steps are discussed.

Conclusion

An assessment solution of some kind is essential in order to measure the effectiveness of any language learning programme introduced, be it by the state or by a privately run school group. While it is possible to use internationally recognised proficiency exams for this purpose, many of the attributes of such exams are not necessary in order to provide diagnostic evidence of this effectiveness, and may indeed prove to be a hindrance particularly when testing large cohorts. The MLT assessment solution contributes to a better understanding of the current situation and enriches the decision-making process based on students' performance at cohort level and the client's context, while proving easier to administer than standard high-stakes alternatives. At the same time, benchmarking projects are valuable opportunities for collaboration between school groups or governments and Cambridge. Cambridge continues to develop the MLT to meet clients' evolving needs based on our experience and research. This work also forms part of wider development of multi-level computer-adaptive tests at Cambridge and automarker technology.

References

Consejería de Educación, Formación y Empleo (2018, June). *Orden EDU/31/2018, de 1 de junio, por la que se regulan los centros de Educación Secundaria, Bachillerato y Formación Profesional bilingües en la Comunidad Autónoma de La Rioja*. Available online: https://www.larioja.org/educarioja-centros/es/lenguas-extranjeras/bilinguismo.ficheros/1026443-Biling%C3%BCismo%20secundaria.pdf

Council of Europe. (2001). *Common European Framework of Reference for Languages. Learning, Teaching and Assessment*. Cambridge: Cambridge University Press.

European Commission. (2003). FINAL *Promoting Language Learning and Linguistic Diversity: An Action Plan 2004–2006*. COM (2003) 449. Brussels: European Commission. Available online : https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52003DC0449

UCLES (2018). *Linguaskill Writing Trial Report. June 2017*. Available online: https://www.cambridgeenglish.org/images/466042-linguaskill-writing-trial-report.pdf

Using a learner corpus to refresh rating scales of CELI exams

Fabio Zanda
University for Foreigners of Perugia, Italy

Danilo Rini
University for Foreigners of Perugia, Italy/Centro per la Valutazione e le Certificazioni Linguistiche (CVCL) – Centre for Language Evaluation and Certification of the University for Foreigners of Perugia, Italy

Abstract

In the last few years, we have witnessed an increase in the use of corpora to inform language testing and assessment practices. Among other purposes, the analyses of well-designed collections of real learner performances may be used as an effective counterpart to more traditional methods for the development and revision of rating scales.

In this contribution, we briefly present a learner corpus which consists of over 3,000 written texts produced by candidates of Italian L2 CELI exams, the CELI corpus (Spina et al., 2022; Spina, Fioravanti, Forti, & Zanda, 2023). (CELI stands for *Certificati di Lingua Italiana*, 'Certificates of Italian Language', issued by the University for Foreigners of Perugia, Italy). We then present a project of the Centre for Language Evaluation and Certification of the University for Foreigners of Perugia, where we explore the potentiality of the CELI corpus in informing the revision of CELI rating scales, combined with the consultation of assessment reference resources and the opinion of expert raters.

Introduction: The use of corpora to inform language testing and assessment

Corpora can generally be defined as large digital collections of authentic language productions sampled according to specific criteria to represent a certain language variety (McEnery, Xiao, & Tono, 2006). Being stored in electronic format, corpora allow for a wide range of computer-assisted queries and analyses, as well as systematic linguistic features' comparison with other similar corpora. Preliminary discussions about the potential applications of corpora in language testing and assessment (LTA) commenced in the mid-1990s since Charles Alderson outlined prospective uses to inform the development and validation of language tests with the aid of corpus data (Alderson, 1996). Following his intuitions and the noteworthy impact of corpus linguistics in linguistic analysis and pedagogy (Taylor, & Barker, 2008), corpus methods were introduced in LTA practices, signaling a steady increase in the exploitation of corpora for the development of new tests and in the maintenance and revision of existing tests (Barker, 2010; Cushing, 2021; Gablasova, 2020; Park, 2014). Another extension in the contribution of corpora to LTA was implemented with the advent of large collections of near-authentic learner texts compiled according to explicit design criteria (Granger, 2008), i.e., learner corpora. In fact, it has been reported that reliable learner corpus data 'have the potential to increase transparency, consistency and comparability in the assessment of L2 proficiency, and in particular to inform, validate, and advance the way L2 proficiency is assessed in the CEFR' (Callies, & Gotz, 2015, p. 3). Among other uses (cf. Barker, Salamoura, & Saville, 2015), learner corpus data analysis can be employed – often in combination with native corpora – for specific purposes in the testing cycle, such as to inform the development of word, phrases or structure lists (Capel 2010; 2012; La Russa, D'Alesio, & Suadoni, in print), to identify specific lexical units to inform new task types or ameliorate existing test formats (Hargreaves, 2000), provide plausible performance-based distractors for multiple choice tasks (Gyllstad & Snoder, 2021), or to supply an empirical basis to test developers when constructing or reviewing rating scales and descriptors for learner production (e.g., Barker, 2013; Hawkey & Barker, 2004).

Approaches in developing (and revising) rating scales

Fulcher (2003) and Fulcher, Davidson and Kemp (2011) oppose two major methodological approaches in developing rating scales: a *measurement-driven approach* and a *performance data-driven approach*. Both approaches present pros and cons (Fulcher et al.,

2011) and can be summarised as follows. On the one hand, the measurement-driven approach is based on intuitive methods in elaborating rating criteria, thus involving judgements of experts in language teaching and assessment. It engages in favouring clearness and usability of scales and is the most widely used. However, among the points of criticism that have been highlighted in opposition to this approach, the lack of concreteness and objectivity in the language employed in the descriptors stands out, as it may cause potential subjective misinterpretations of the scores and their meaning for raters. This raises questions about the reliability and validity of score inferences which, in addition to the absence of examination of real performances, make post-hoc quantitative or qualitative analysis of the resultant scales indispensable (Banerjee, Yan, Chapman, & Elliott, 2015). On the other hand, the performance data-driven approach is based on empirical methods, being derived from the analysis of real performance data (Fulcher, 2003, p. 92). Therefore, as a first step, this approach adopts a bottom-up method, as it 'identifies traits or features that characterize and discriminate written texts or writers across proficiency levels' (Banerjee et al., 2015, p. 6). In other words, in the performance data-driven approach, the development of rating scales is preceded by linguistic analyses of real performance data, which may be found in learner corpora purposely annotated and collected from exam data (cf. Barker et al., 2015). The scales derived from this approach do have the advantage of mirroring real performance data, which yet need to undergo time-consuming thorough analysis that 'tend[s] to generate linguistic constructs that either bear complex mathematical formulae or become extremely difficult to operationalize by human raters' (Banerjee et al., 2015, p. 6).

In light of the above, it would appear reasonable to opt for a mixed approach for the rating scale review process, relying both on real performance analysis of corpus data and on expert intuitions to improve usability.

Reviewing rating scales of the CELI exams

In this paper, we present a new research project of the Center for Language Evaluation and Certification (CVCL – Centro per la Valutazione e le Certificazioni Linguistiche[1]) of the University for Foreigners of Perugia (Italy) which aims to analyse and potentially revise the current rating scales of the *Certificati di Lingua Italiana* ('Certificates of Italian Language') (CELI), since constant monitoring and evaluation of existing scales is vital in standardised testing and assessment (Banerjee et al., 2015). As per the CELI exams corresponding to B2, C1 and C2 proficiency levels of the Common European Framework of Reference for Languages (CEFR, Council of Europe, 2001), i.e., CELI 3, CELI 4, and CELI 5 exams, there exist analytic rating scales which are uniformly structured for all levels. In fact, all rating scales of CELI 3, 4, and 5 include the same four assessment criteria for the evaluation of written production tasks: *vocabulary control*, *grammar accuracy*, *sociolinguistic appropriateness*, and *text coherence and cohesion* (Grego Bolli, 2004).

The starting point of the project is to focus primarily on the vocabulary control criterion, check for updated vocabulary descriptors in assessment reference materials, i.e., in the CEFR Companion Volume (CEFR CV, Council of Europe, 2020), analyse the vocabulary actually produced by learners in written productions by candidates of the CELI at each proficiency level under scrutiny, and subject the existing scale descriptors of CELI 3, 4, and 5 (CEFR Levels B2, C1, C2) to a critical examination by CELI expert raters.

CEFR CV vocabulary descriptors

The publication of CEFR CV and the presence of entirely newly released or accurately refreshed descriptors reflects how recent studies and considerations over second language proficiency tend to stress the importance dedicated to word combinations and phraseological units in both language acquisition and L2 production (Ebeling, & Hasselgård, 2015; Siyanova-Chanturia, & Pellicer-Sánchez, 2019). In the CEFR CV descriptors concerning 'vocabulary range', in the levels of interest (B2, C1 and C2), reference is made to 'idiomatic expressions' for C1 and C2, but, very interestingly, at B2 level, the production of 'appropriate collocations of many words in most contexts fairly systematically' (Council of Europe, 2020, p. 132) is introduced as being characteristic to the level. Such examples clearly show how L2 assessment cannot set these features aside.

By comparing CEFR CV descriptors and CELI rating scales, a few expressions turned out to be overlapping, but a few others seemed to be missing in scales, whereas others were introduced there, probably with the aim of facilitating the work of raters. For instance, in CELI scales reference is made to the presence of errors in written production by candidates, a reference which, by the very nature of the approach chosen by the CEFR, is absent in the latter.

[1] CVCL webpage: https://www.unistrapg.it/en/certification-of-italian-as-a-foreign-language

Real exam data: The CELI corpus

In order to also base our reviewing process on real performance data, we chose to rely on the CELI corpus (Spina et al., 2022, 2023). The CELI corpus has been designed to systematically compile the written texts produced by different candidates of Italian L2 who have passed the CELI exams at B1, B2, C1 and C2 proficiency levels of the CEFR. Over 3,000 texts, elicited out of more than 60 comparable task assignments, were included in the corpus, with a balanced distribution of the tokens in terms of proficiency level, totaling c.600,000 tokens, thus featuring a pseudo-longitudinal design (Meunier, 2015), with 150,000 tokens per proficiency level (Spina et al., 2022, 2023).

Preliminary analysis of the CELI corpus

Research has shown that vocabulary is a key component in overall language competence development (Milton, 2013) and that phraseological competence plays a crucial role in language acquisition, processing, fluency and idiomaticity (Ellis, Simpson-Vlach, & Maynard, 2008; Wray, 2002). In view of this, we based our preliminary data analysis on recent vocabulary studies performed on CELI corpus data. First, we referred to a recent study (Forti, Fioravanti, & Zanda, 2022) which investigates one of the most popular constructs to analyse vocabulary knowledge, i.e., lexical complexity (Bulté, & Housen, 2012). Lexical complexity is defined as a multifaceted construct that includes the main dimensions of lexical diversity (the number of different words in a sample), lexical sophistication (the number of less frequent or unusual words in a sample) and lexical density (the ratio of content words on total words in a sample) (Kyle, 2019). Second, we also resorted to the investigation of phraseological competence, i.e., the ability to use phraseology and word combinations, operationalised as phraseological units, that is

> the co-occurrence of a form or a lemma of a lexical item and one or more additional linguistic elements of various kinds which functions as one semantic unit in a clause or sentence and whose frequency of co-occurrence is larger than expected on the basis of chance (Gries, 2008, p. 6).

We computed different measures of phraseological complexity, which, echoing Ortega (2003), is defined as 'the range of phraseological units that surface in language production and the degree of sophistication of such phraseological units' (Paquot, 2019, p. 124). The dimensions of phraseological complexity taken into account are phraseological diversity and phraseological sophistication for three typologies of co-occurrences that appear in specific syntactic relations (verb+direct object, adjective+noun, adverbial modifier+verb).

As for lexical complexity, the results of the analysis of B2, C1, and C2 sub-corpora indicate that there are differences in the development of complexity across proficiency levels, with a statistically significant linear development of lexical diversity. Concerning lexical sophistication and lexical density, although there are significant differences between the proficiency bands (B and C), these are not significant between the C1 and the C2 levels (Forti et al., 2022). With regard to phraseological complexity, we found that there are significant differences in the development of phraseological diversity across levels for all syntactic relations considered, while, again, for measures of phraseological sophistication, the results show significant differences for the relation of verb+direct object between the B and the C bands[2], yet not between C1 and C2. Conversely, for the adjective+noun and the adverbial modifier+verb relations the development appears not to be linear.

In a nutshell, according to the CELI corpus data, we could say that empirical research showed that learners present a development in the variety and originality of words and word combinations used in their texts as the proficiency level grows, but not necessarily in the 'rarity' of the lexical units employed in the context of a written production task in a standardised certification exam.

Impressions of raters on existing rating scales

On the basis of the qualitative analysis conducted on CEFR CV descriptors and CELI rating scales, and of the quantitative analysis of the CELI corpus, we proceeded to involve expert raters in the reviewing process of the current CELI 3, CELI 4, and CELI 5 rating scales[3]. Five texts per each level investigated were selected from CELI corpus and eight experienced raters were asked to assess them, using the vocabulary control criterion only. Afterward, questionnaires were submitted to raters. Questionnaires were built by dividing them into two main sections, the first one concerning the usage of CELI scale descriptors in assessing papers, and the second one concerning the structure and wording of CELI scales themselves.

---

[2] The B band comprises B1 and B2 levels together, while the C band includes C1 and C2.

[3] Current CELI 3 rating scale: https://www.unistrapg.it/sites/default/files/docs/certificazioni/competenze-punteggi-CELI-3-B2-scritto.pdf
Current CELI 4 rating scale: https://www.unistrapg.it/sites/default/files/docs/certificazioni/competenze-punteggi-CELI-4-C1-scritto.pdf
Current CELI 5 rating scale: https://www.unistrapg.it/sites/default/files/docs/certificazioni/competenze-punteggi-CELI-5-C2-scritto.pdf

From the first section it turned out that raters, in assessing the paper, put particular emphasis on the use of a lexical repertoire coherent with input and expected register (above 87% of cases), but also gave importance to the presence of errors in lexical usage (above 62%). The use of idiomatic expressions was not considered as important, but it is worth noting that, when asked, raters considered the appropriateness of phraseological units as being very important in assessment, just as vocabulary control and appropriateness, and its extent and variety. On the other hand, they underlined how those aspects were often not present in scales, and possibly should be included.

From the second section, mainly concerning the wording of vocabulary control descriptors in CELI scales, it turned out that scales were considered generally clear, but some of the terms used there are considered ambiguous, such as 'adequate', and the reference to the number of errors present in scales is considered as 'misleading'. Scales are considered generally exhaustive, but the absence of reference to appropriateness and metaphoric use of language was stressed, while, when it comes to easiness of use, raters underlined how too much is left to raters' interpretation, and too many aspects are to be taken into consideration.

Further comments stressed the difficulty of using scales while referring to a single component/criterion in assessment, without any reference to other aspects of written production, thus arising the ever-present questions about the use of analytic vs holistic scales in language assessment. Moreover, issues such as variety, originality, and appropriateness were mentioned as relevant features to be included in scales, while a few raters considered referring to errors particularly misleading in assessment.

Conclusion

In summary, corpora derived from learner productions can be indeed helpful to inform language testing and assessment practices. In the project that we presented, we chose to adopt a mixed approach to the review of the existing rating scales of the CELI exams, starting from the vocabulary control criterion. In this context, the CELI corpus was used effectively as an empirical basis: being a collection of real exam performances and thanks to its pseudo-longitudinal design, it served to identify and compute several vocabulary features across levels. In combination with corpus data, we resorted to the analysis of CEFR CV renovated descriptors concerning vocabulary and to expert raters' judgement on the existing descriptors in CELI rating scales. The analysis of rater questionnaires and of the opinion of CELI raters represent a precious instrument in determining how existing scales may be amended in order to eventually rephrase descriptors and thus achieve scales with easier applicability, possibly leading to a fairer assessment. Future steps in the project involve a within-level quantitative corpus analysis in order to possibly identify the features that actually discriminate between higher quality and lower quality productions of the same CEFR level; an in-depth analysis of raters' behaviour when assessing with the current scales; and the creation of a large database with in-text examples of lexical features indicated by raters as determining in their assessment, which could be included in future scales.

Acknowledgements

We would like to extend our gratitude to Daniela Alessandroni, Maria Cristina Bricchi, Claudia Fedeli, Marina Mancinotti, Elisabetta Marchetti, Franco Romano, and Roberta Rondoni (Centre for Language Evaluation and Certification, University for Foreigners of Perugia) for their contribution in the rating process and raters' questionnaires, and to Irene Fioravanti (University for Foreigners of Perugia) for the preliminary corpus analyses used in our paper.

Author contributions

The present paper is a joint effort by the co-authors. Zanda contributed to all sections except 'Impressions of raters on existing rating scales', which Rini wrote alone. Both authors contributed to the study design and to the final manuscript.

References

Alderson, C. (1996). Do corpora have a role in language assessment? In J. Thomas & M. Short (Eds), *Using Corpora for Language Research. Studies in Honour of Geoffrey Leech* (pp. 3–14). New York: Longman.

Banerjee, J., Yan, X., Chapman, M., & Elliott, H. (2015). Keeping up with the times: Revising and refreshing a rating scale. *Assessing Writing*, *26*, 5–19.

Barker, F. (2010). How can corpora be used in language testing?. In A. O'Keeffe & M. McCarthy (Eds.), *The Routledge Handbook of Corpus Linguistics* (pp. 633–645). New York: Routledge.

Barker, F. (2013). Using corpora to design assessment. In A. J. Kunnan (Ed.), *The Companion to Language Assessment* (pp. 1,013–1,028). Hoboken: Wiley-Blackwell.

Barker, F., Salamoura, A., & Saville, N. (2015). Learner corpora and language testing. In S. Granger, G. Gilquin & F. Meunier (Eds.), *The Cambridge Handbook of Learner Corpus Research* (pp. 511–534). Cambridge: Cambridge University Press.

Bulté, B., & Housen, A. (2012). Defining and operationalising L2 complexity. In A. Housen, F. Kuiken & I. Vedder (Eds.), *Dimensions of L2 Performance and Proficiency Volume 32* (pp. 21–46). Amsterdam: John Benjamins.

Callies, M., & Götz, S. (2015). Learner corpora in language testing and assessment: Prospects and challenges. In M. Callies & S. Götz (Eds.), *Learner Corpora in Language Testing and Assessment* (pp. 1–9). Amsterdam: John Benjamins.

Capel, A. (2010). A1–B2 vocabulary: Insights and issues arising from the English Profile Wordlists project. *English Profile Journal, 1*, E3.

Capel, A. (2012). Completing the English Vocabulary Profile: C1 and C2 vocabulary. *English Profile Journal, 3*, E1.

Council of Europe. (2001). *Common European Framework of Reference for Languages. Learning, Teaching, Assessment*. Cambridge: Cambridge University Press.

Council of Europe. (2020). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Companion Volume*. Strasbourg: Council of Europe Publishing.

Cushing, S. T. (2021). Corpus linguistics and language testing. In G. Fulcher & L. Harding (Eds.), *The Routledge Handbook of Language Testing* (pp. 545–560). New York/London: Routledge.

Ebeling, S.O., & Hasselgård, H. (2015). Phraseology in learner corpus research. In S. Granger, G. Gilquin & F. Meunier (Eds.), *The Cambridge Handbook of Learner Corpus Research* (pp. 207–230). Cambridge: Cambridge University Press.

Ellis, N.C., Simpson-Vlach, R., & Maynard, C. (2008). Formulaic language in native and second language speakers: Psycholinguistics, corpus linguistics, and TESOL. *TESOL Quarterly*, *42*, 375–96.

Forti, L., Fioravanti, I., & Zanda, F. (2022, September 22–24). *Lexical complexity across proficiency levels in L2 Italian: some preliminary findings* [Poster presentation]. 6th International Conference for Learner Corpus Research (LCR 2022), University of Padua, Padua, Italy.

Fulcher, G. (2003). *Testing Second Language Speaking*. London: Pearson.

Fulcher, G., Davidson, F., & Kemp, J. (2011). Effective rating scale development for speaking tests: Performance decision trees. *Language Testing*, *28*(1), 5–29.

Gablasova, D. (2020). Corpora for second language assessments. In P. Winke & T. Brunfaut (Eds)., *The Routledge Handbook of Second Language Acquisition and Language Testing* (pp. 45–53). New York/London: Routledge.

Granger, S. (2008). Learner corpora. In A. Lüdeling & M. Kytö (Eds.), *Corpus Linguistics. An International Handbook*. *Volume 1* (pp. 259–275). Berlin/New York: Walter de Gruyter.

Grego Bolli, G. (2004). Measuring and evaluating the competence in Italian as a foreign language. In M. Milanovic & C J.. Weir (Eds.), *European Language Testing in a Global Context: Proceedings of the ALTE Barcelona Conference, July 2001* (pp. 271–83). Studies in Language Testing Volume 18. Cambridge: UCLES/Cambridge University Press.

Gries, S. T. (2008). Phraseology and linguistic theory: A brief survey. In S. Granger & F. Meunier (Eds.), *Phraseology: An Interdisciplinary Perspective* (pp. 3–25). Amsterdam/Philadelphia: John Benjamins.

Gyllstad, H., & Snoder, P. (2021). Exploring learner corpus data for language testing and assessment purposes: The case of verb + noun collocations. In S. Granger (Ed.), *Perspectives on the L2 Phrasicon: The View from Learner Corpora* (pp. 49–71). Bristol: Multilingual Matters.

Hargreaves P. (2000). How important is collocation in testing the learner's language proficiency? In M. Lewis (Ed.), *Teaching collocation – Further Developments in the Lexical Approach* (pp. 205–223). Hove: Language Teaching Publications.

Hawkey, R., & Barker, F. (2004). Developing a common scale for the assessment of writing. *Assessing Writing*, *9*, 122–159.

Kyle, K. (2019). Measuring lexical richness. In S. Webb (Ed.), *The Routledge Handbook of Vocabulary Studies* (pp. 454–475). London/New York: Routledge.

La Russa, F., D'Alesio, V., & Suadoni, A. (in print). Designing a corpus based syllabus of Italian collocations. Criteria, methods and procedures. *Revue Roumaine de linguistique*.

McEnery, T., Xiao, R. & Tono, Y. (2006). *Corpus-based Language Studies: An Advanced Resource Book*. London: Routledge.

Meunier, F. (2015). Developmental patterns in learner corpora. In S. Granger, G. Gilquin & F. Meunier (Eds.), *The Cambridge Handbook of Learner Corpus Research* (pp. 379–400). Cambridge: Cambridge University Press.

Milton, J. (2013). Measuring the contribution of vocabulary knowledge to proficiency in the four skills. In C. Bardel, C. Lindqvist & B. Laufer (Eds.), *L2 Vocabulary Acquisition, Knowledge and Use. New Perspectives on Assessment and Corpus Analysis* (pp. 57–78). Amsterdam: Eurosla Monographs Series.

Ortega, L. (2003). Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Applied Linguistics*, *24*, 492–518.

Paquot, M. (2019). The phraseological dimension in interlanguage complexity research. *Second Language Research*, *35*(1), 121–145.

Park, K. (2014). Corpora and language assessment: The state of the art. *Language Assessment Quarterly, 11*(1), 27–44.

Siyanova-Chanturia, A., & Pellicer-Sánchez, A. (Eds.) (2019). *Understanding Formulaic Language: A Second Language Acquisition Perspective*. New York/London: Routledge.

Spina, S., Fioravanti, I., Forti, L., & Zanda, F. (2023). The CELI Corpus: Design and linguistic annotation of a new online learner corpus. *Second Language Research*.

Spina, S., Fioravanti, I., Forti, L., Santucci, V., Scerra, A., & Zanda, F. (2022). Il corpus CELI: Una nuova risorsa per studiare l'acquisizione dell'italiano L2. *Italiano LinguaDue*, *14*(1), 116–138.

Taylor, L., & Barker, F. (2008). Using corpora for language assessment. In E. Shohamy and N. H. Hornberger (Eds.), *Encyclopedia of Language and Education Volume 7. Language testing and assessment* (Second edition) (pp. 241–254). New York: Springer.

Wray, A. (2002). *Formulaic Language and the Lexicon*. Cambridge: Cambridge University Press.

Cross-country comparisons of English-speaking ability with PROGOS test

Masuyo Ando
PROGOS, Inc.

Yukio Tono
Tokyo University of Foreign Studies, Japan

Abstract

PROGOS is a CEFR-based English speaking test which provides two ways of evaluation: AI-driven automated evaluation or human rating. Since its launch in 2020, the test has been widely accepted by Asian countries. In this presentation, we share the cross-country comparisons of 110,000 test-taker results in 11 countries. Our Correspondence Analysis shows that there are three clusters of country groups, in which some countries do not match the Education First (EF) English Proficiency Index (EPI) rankings. We plan to collect more data to make further analysis with the five analytical aspects of speaking ability.

Introduction

PROGOS, introduced in 2020, is a Business English Speaking test designed to offer English learners worldwide a reliable, swift, convenient, and scalable evaluation method. The test assesses oral interaction and production skills through diverse scenarios that mirror general business environments. Administered online, the 20-minute test can be taken on a PC, tablet, or smartphone. Participants are prompted to provide spontaneous responses to open-ended questions. Two versions of the PROGOS test exist: one involving manual evaluation and the other, automatic evaluation. We pioneered an AI-driven automated assessment system, utilising a speech recognition application programming interface (API) coupled with machine learning. This innovation enables us to offer the test to individuals anytime and anywhere. The examination is divided into five segments: (i) short questions and responses, (ii) reading aloud, (iii) presentation, (iv) graphic presentation, and (v) role play. Test results are presented in accordance with the Common European Framework of Reference for Languages (CEFR, Council of Europe, 2001) displaying both an overall CEFR level and insights into six qualitative aspects of spoken English. This paper reports a preliminary analysis of data acquired from over 110,000 PROGOS test-takers across 11 nations.

Sample distribution across countries by CEFR levels

Table 1 presents a summary of the number of test-takers from 11 countries and their distribution across CEFR levels. Cells highlighted in yellow indicate the CEFR levels with the highest proportion for each country. For instance, in Japan, the A2+ level encompasses 25% of the test-takers, making it the most prevalent CEFR level in the country. Among the 11 nations, Singapore and the Philippines lead the pack. Singapore boasts the highest proportion of individuals at B1+ and above, with 89%, followed closely by the Philippines at 77%. China, Cambodia, and India trail with 70%, 63%, and 56% respectively. Vietnam and Indonesia both have B1+ as their most common CEFR level; however, their cumulative proportions of learners at B1+ and above are significantly lower than the previously mentioned countries.

Korea and Malaysia predominantly exhibit the B1 level as their most representative. When considering the cumulative proportion of B1+ and higher scorers, Malaysia stands at 44%, outpacing Korea's 29%.

Lastly, the group comprising Japan, Vietnam, Indonesia, and Thailand presents an intriguing finding: Japan has 35% of its test-takers at B1+ and above, a percentage higher than Korea's and on par with Vietnam's. For these countries, A2+ stands out as the most common CEFR level. As you can see, the number of test-takers was skewed, with Japan being the highest, followed by the Philippines. In fact, these two countries represent more than 90% of all the examinees, thus more balanced data should be obtained in the future.

Figure 1 visually summarises the rank and the most representative CEFR level for each of the eleven countries.

Table 1: Test-takers from eleven countries and their distribution across CEFR levels

Country	No. of test takers	Pre-A1	A1	A1+	A2	A2+	B1	B1+	B2	B2+ above
Japan	76,542	6%	2%	6%	7%	25%	21%	24%	10%	1%
Philippines	27,413	3%	0%	1%	1%	5%	12%	36%	36%	5%
Thailand	1,490	8%	3%	9%	8%	32%	21%	15%	4%	0%
Indonesia	688	10%	3%	7%	7%	22%	20%	24%	6%	1%
Cambodia	490	2%	1%	2%	3%	11%	17%	35%	26%	2%
Singapore	249	0%	0%	0%	0%	4%	7%	45%	39%	5%
China	146	3%	1%	0%	0%	9%	18%	55%	14%	1%
India	81	2%	7%	4%	4%	10%	16%	33%	21%	2%
Korea	60	5%	5%	3%	7%	20%	32%	22%	7%	0%
Malaysia	51	2%	0%	0%	0%	16%	39%	22%	22%	0%
Vietnam	40	10%	0%	10%	2%	25%	18%	25%	10%	0%

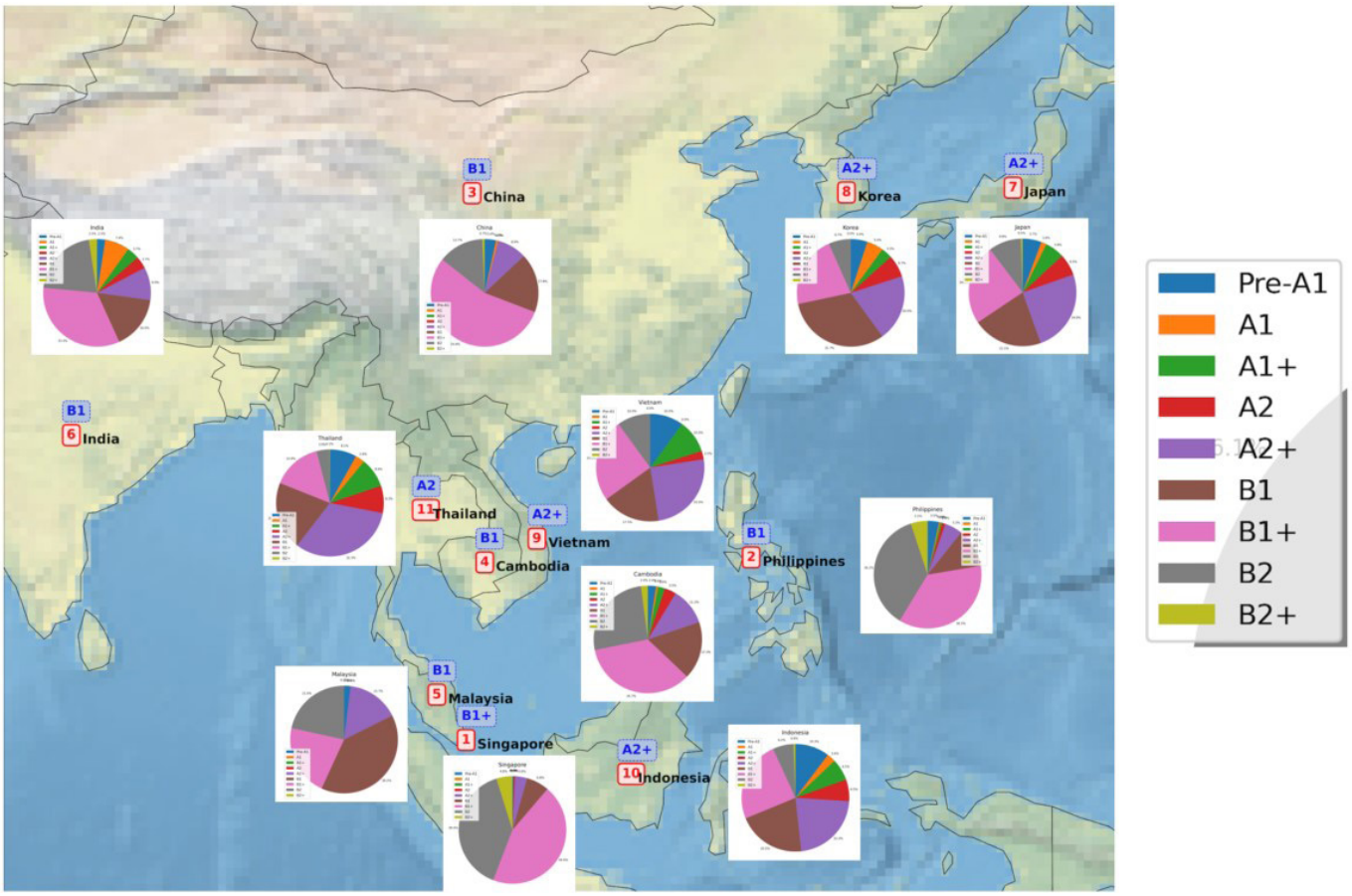


Figure 1 The rank and the most representative CEFR level in each country taking PROGOS

Visualising the relationship between the countries and the CEFR levels

We also analysed our results using a data reduction technique called Correspondence Analysis (CA). We used corresp() in R's MASS package. Figure 2 shows the results of CA over the countries and their representative CEFR levels.

Figure 2 shows the two-dimensional plots, where Dimension 1 (a vertical axis) explains more than 98% of the variances between the two variables, countries vs. their representative CEFR levels. There seem to be three clusters along Dimension 1. The

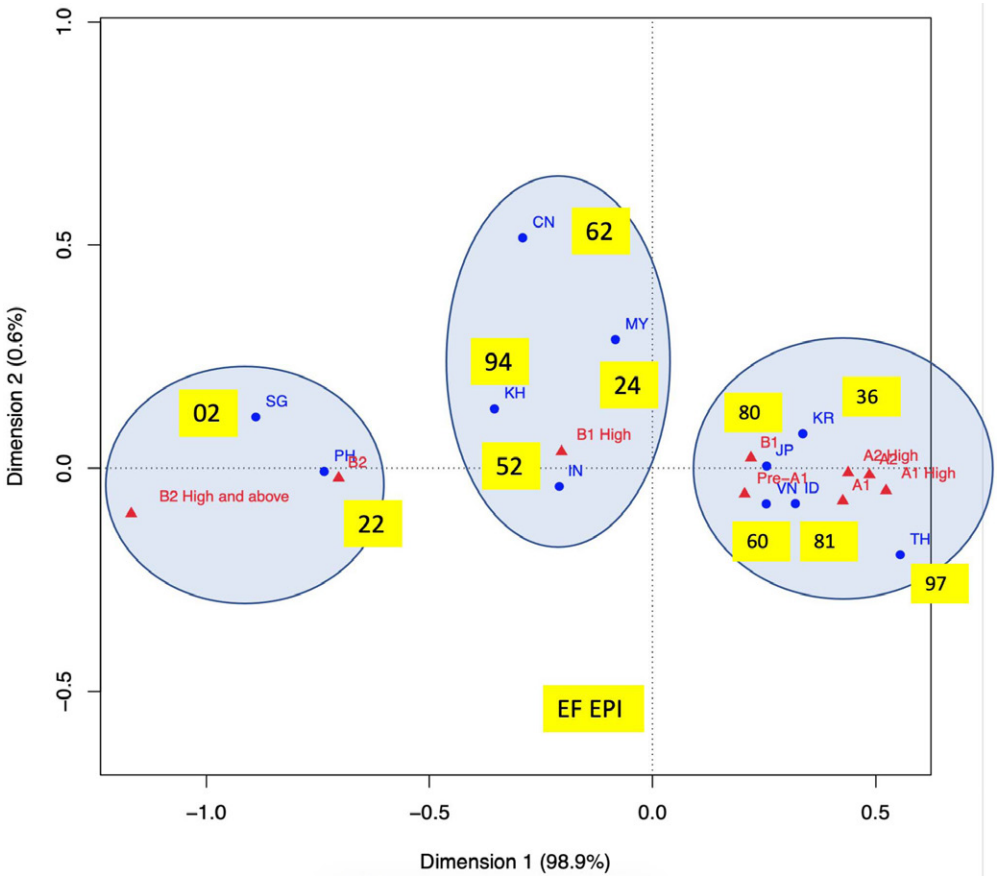


Figure 2 Correspondence Analysis between the countries and the CEFR levels

Notes: 1. CN=China; IN=India; ID=Indonesia; JP=Japan; KH=Cambodia; KR=Korea; MY=Malaysia; PH=Philippines; SG=Singapore; TH=Thailand; VN=Vietnam.

2. The numbers in yellow box show the English Proficiency Index (EPI) by Education First (EF).

leftmost cluster shows B2 and above groups, where Singapore and the Philippines belong. The second cluster in the middle of the plot features B1+ level, where China, Cambodia, Malaysia, and India are plotted. Then the rest of the countries belong to all the rightmost clusters. It is interesting that, when looking at Table 1 in isolation, it was difficult to determine Cambodia and Malaysia belong to the same group with China and India because we primarily looked at the most prominent CEFR levels. However, CA reveals that the frequency distribution patterns of the CEFR levels are similar in these four countries, where the number of 'below A2 level' was extremely small. This is, on the contrary, not the case with Japan, Korea, Vietnam, Indonesia and Thailand.

For comparison purposes, the English Proficiency Index (EPI), reported by Education First (EF, https://www.ef.com/wwen/epi/), is plotted to show each country's international ranking of English language proficiency. There are some cases where the EPI ranking does not match the level we found in PROGOS. For example, Cambodia ranks 94th in EF EPI, but it belongs to the second group here. On the other hand, Korea ranks 36th in EF EPI, but it belongs to the third group in PROGOS. These differences might be due to the sampling or could be the reflection of the test characteristics. For example, PROGOS has scoring criteria based on the five qualitative aspects of spoken language use, proposed by the CEFR (Council of Europe, 2001, pp.28–29): Range, Accuracy, Fluency, Interaction, and Coherence. We also added 'Pronunciation', following the CEFR Companion Volume's (Council of Europe, 2020) new descriptors. These aspects might have influenced the different results between PROGOS and EF EPI. More data will be needed to find out more about the relationship between different aspects of speaking ability, their communicative goals according to the CEFR, and their background in education and culture in each country.

Conclusion

We have presented the preliminary analysis of results from approximately 110,000 English learners who took the PROGOS, a CEFR-based English speaking test. Given that the data collected predominantly represents Japanese and Philippine English

learners, future studies will require more balanced sampling. Nonetheless, this study demonstrates the potential of a CEFR-based speaking test in its design principles.

References

Council of Europe. (2001). *The Common European Framework of Reference for Languages: Learning, Teaching, Assessment.* Cambridge: Cambridge University Press.

Council of Europe. (2020). *The Common European Framework of Reference for Languages: Learning, Teaching, Assessment: Companion Volume.* Strasbourg: Council of Europe Publishing.

Teaching the teachers: Designing digital assessment for language teachers which both evaluates and educates

Nicole Busby
Norwegian University of Science and Technology (NTNU)

Anja Angelsen
Norwegian University of Science and Technology (NTNU)

Abstract

This study describes student responses to a new digitally based formative assessment design in a linguistics course for in-service language teachers in Norway. The goals of the redesign were to increase flexibility by transitioning to a fully digital format and to increase the relevance and validity of the assessment, as well as introducing assessment approaches that the students could implement in their own teaching. We created an assessment package with four elements that integrated formative and summative assessment. Online surveys and students' reflective texts were used to collect data about students' perceptions of the assessment as part of investigating the validity of the new design. Feedback indicated that students were motivated, and that the combination of formative and summative assessment gave them greater opportunities to demonstrate what they had learned from the course, suggesting that this approach may also be suitable for other types of courses.

Introduction

New digital learning environments require a rethink of assessment to secure a good learning and course design. We wanted to create good constructive alignment (e.g., Biggs, 1999) by having a clear connection between the learning outcomes, course material, activities, and assessment.

Ensuring test validity, reliability and credibility can be a major challenge in adapting to unsupervised online assessment (see Carless, 2009). Since students are unsupervised and have access to all materials during the assessment process, we wanted to create a valid assessment design for this context which meant we needed to shift the focus from assessing memorised knowledge to assessing understanding of concepts and the ability to apply this understanding to new examples. Another challenge of digital unsupervised assessment is ensuring that students submit individual work that reflects their understanding (rather than their ability to search for immediate answers).

Context and goals for the assessment design

The course in question, ENG6025: Linguistics and Language Acquisition, is an online course for in-service teachers as part of a package that gives teaching qualifications in English in Norwegian secondary schools. It gives students an introduction to central areas related to the study of modern English language, including phonetics and phonology, morphology, syntax, and an introduction to first and second language acquisition. The course aims to provide students with the concepts and ideas used to describe and analyse linguistic phenomena of the English language, with the goal that they can use this knowledge in their own teaching.

One of the challenges with teaching this group is that they are working (almost) full-time as teachers and are therefore very busy, which means they tend not to be motivated to spend time on activities that they do not perceive as relevant for their work. Consequently, it is especially important to ensure and clearly demonstrate the relevance of this course to their teaching work.

The course content can be perceived as quite abstract and theoretical, and it is important to work through linguistic examples to understand it properly. We wanted to encourage students to work together on discussing the examples because this helps

with understanding the material and creating a sense of community where students can share their struggles and help each other to understand. In order to enable students to work together while still submitting individual work, we designed tasks that required them to choose their own examples to analyse. Focusing on analysis of self-chosen examples was also a way to increase motivation because it enabled students to work on things that they found interesting and relevant.

We created a new digital assessment package which required students to submit a portfolio of tasks at the end of the semester. The main goals for the new portfolio design were to:

- Create an assessment package that increased opportunities for assessing students' knowledge and understanding of the course material and decreased the possibilities for students to collaborate with each other or to look up answers to questions without understanding them.
- Increase motivation for learning the course content and to distribute the workload evenly throughout the semester.
- Increase relevance by introducing new teaching and assessment techniques that students could adopt in their own classrooms.

The portfolio design

The course content is divided into four units: phonetics and phonology, morphology, syntax, and language acquisition. At the end of each unit, students are required to complete an assignment involving applying linguistic concepts to self-chosen examples. The students receive constructive feedback that helps them to refine tasks for final submission in the portfolio.

The portfolio assessment for the course consists of four elements that count towards the final grade:

1. Refined tasks on three chosen topics.
2. Additional tasks related to the topics chosen.
3. A multiple-choice quiz with a limited time frame.
4. A reflective text.

Each of these elements has a different role to play in helping to ensure the validity and reliability of the portfolio assessment. The goal of the refined tasks is to enable students to revisit parts of the assignment and/or course material that they did not understand the first time and to learn from their mistakes. This is a formative element, where the assessment is used *for* learning. Research has shown that feedback is most useful when students perceive it as relevant, particularly when it contributes to longitudinal development (feed-forward) and includes information that can be immediately applicable to ongoing work (Price, Handley, Millar, & O'Donovan, 2010). Having the assignments due throughout the semester ensures that students are working with the material already from the first weeks and means they have the flexibility to revise their assignments immediately after feedback or at the end of the semester when the portfolio is due.

Students receive the additional tasks (the second element) shortly before the portfolio is due and they are not able to ask for feedback on these tasks. This is designed to ensure that they are able to understand the concepts from the course well enough to apply them to new examples, and without help from teaching staff.

The third element, the multiple-choice quiz, includes a more summative approach. The quiz needs to be completed within a limited time frame in order to reduce the possibility of looking up answers. The idea is that students really need to understand the content before the quiz to be able to answer the questions accurately because there is not enough time to look up answers once the quiz is underway.

Finally, the students also submit a reflective text which describes what they have learned from the process of doing the other tasks in the portfolio and what lessons they will take away from the course overall. The goal of this is to encourage students to reflect on their learning, to think about the process of learning from the student perspective (and whether this will influence their own teaching in the future), and to promote sustainable learning (e.g., Boud, 2007) by encouraging students to consider the relevance of the course material and also the experience of being a student and how this relates to their work as teachers.

Assessing the portfolio assessment design

In order to evaluate our assessment design and to reflect on the extent to which it is fit for the purpose, online surveys were conducted, and students' reflective texts were analysed for relevant themes. Our own perceptions as teachers are also reported.

Surveys

Online surveys were used to collect data about students' perceptions of the assessment as part of investigating the validity of the new design. Students reported finding the assessment design relevant for their own learning and as a model which could be used in improving their own teaching. They reported increased motivation and greater understanding resulting from being able to work on their assignments following feedback, and having smaller tasks throughout the semester enabled them to distribute their workload more evenly. Students also reported feeling that the combination of formative and summative assessment gave them greater opportunities to demonstrate what they had learned from the course.

Reflective texts

A qualitative analysis of the reflective texts that students submitted as part of the portfolio was also conducted. Approximately 30 students gave permission for their portfolios to be used for research purposes, so we worked on identifying themes from these reflective texts. We chose to focus on looking for reflections from the students that gave insight into their experiences and learning from the portfolio design. We identified several themes that were present across multiple texts. The first of these was the idea of receiving feedback and revisiting material as a learning opportunity. Many students mentioned that they benefited from the opportunity to have feedback on their assignments that enabled them to identify parts of the course content that they had not fully understood and that they were better able to understand the concepts after working further with the same assignment tasks. For example, one student mentioned that 'working with the assignments, getting feedback, and revising them has been a good way of learning this semester' and others mentioned the benefits of being able to learn from their mistakes. Several also mentioned that the process of revising the assignments was motivation to revisit the course material and therefore offered a good opportunity to revise for the exam.

Secondly, many reported that they found the course and its many deadlines to be challenging, but that they also found that being forced to engage with the material throughout the course helped them gain a deeper understanding of the content (e.g., 'I learned a lot while struggling with this'). In addition, they reported that they better understood the effects of challenging material and tasks (with sufficient help from teachers) and how the struggle can be part of the learning process and the benefits of working through examples ('reading the course material would not have been sufficient').

Finally, some reflective texts also demonstrated that they had used the opportunity to think about how their own students learn and gained an insight into the student perspective. Many described their own process of learning and how they would use this information to try new approaches in their own teaching. For example, some expressed that they felt they had learned a lot from getting feedback on their assignments and were more interested in incorporating more formative approaches to assessment in their own teaching. Some also identified their own shortcomings as students and felt that they understood their own students better as a result.

Teachers' experience

From our own observations, it seems that most students demonstrate a good understanding of relevant concepts and topics, based on the assignments and quiz results. We have also seen that students have worked together in groups to understand the course content but have still submitted individual assignments because they have chosen different examples. As teachers, it's encouraging to see progress in the students' understanding and it was also reassuring to have evidence that students were engaging with material throughout the semester. We also observed that having these assignments throughout the semester and giving detailed, personalised feedback encouraged interaction between students and teachers because they were able and motivated to ask specific questions about the feedback and it helped them to identify things that they didn't understand yet.

One downside to having multiple assignments throughout the semester that all require personalised constructive feedback is that it is very time-consuming and cognitively demanding for the teacher. It has been quite challenging to figure out what each student understood and did not understand, and how to give feedback that is constructive without giving away the answers, and is honest but encouraging.

Limitations and considerations

One potential limitation or consideration of the assessment design is that we have observed that grades are generally higher than they used to be compared with when the course was run with traditional school exams. This is not necessarily a problem, however, since it likely reflects the fact that students who are approved to take the exam have been demonstrably working continuously throughout the semester (rather than just showing up to the exam without having engaged with the course).

Designing tasks requires careful consideration to make sure that students get to demonstrate their knowledge and analytical skills, and not just look up answers. The tasks were designed to not be Googleable, but advances in AI technology require a continuous evaluation of the specific formulation of tasks, with the new opportunities and challenges that AI tools provide in mind.

Conclusions

From both the student and teacher perspectives, this assessment design seemed to work well for our objectives of creating a relevant and suitable portfolio design for this course. Having multiple deadlines and a variety of different tasks was perceived by the students as challenging but ultimately rewarding, and teachers appreciated being able to see progress in understanding. The combination of different assessment tasks enabled students to show their strength and the reflective text seems to show evidence of critical reflection and learning for the students that went beyond the course content and also contributed to new career-relevant perspectives on the learning process. Overall, the combination of summative and formative assessment is seen as positive and useful by both students and teaching staff, and we believe that this assessment design might also be applicable to a wide range of disciplines and courses.

References

Biggs, J. (1999). What the student does: Teaching for enhanced learning. *Higher Education Research & Development*, *18*(1), 57–75.

Boud, D. (2007). Reframing assessment as if learning were important. In D. Boud & N. Falchikov (Eds.), *Rethinking Assessment in Higher Education* (pp. 24–36). New York: Routledge.

Carless, D. (2009). Trust, distrust and their impact on assessment reform. *Assessment & Evaluation in Higher Education*, *34*(1), 79–89.

Price, M., Handley, K., Millar, J., & O'Donovan, B. (2010). Feedback: all that effort, but what is the effect? *Assessment & Evaluation in Higher Education*, *35*(3), 277–289.

Decision making in standard setting

Jane Lloyd
Cambridge University Press & Assessment, United Kingdom

Abstract

This paper reports on an ongoing study of operational standard setting for the reading component of Linguaskill, an adaptive test of EFL for adults. Three standard setting panels were convened online, using a combination of self-access materials and video-conferencing. This paper summarises findings to date on how these standard-setting panellists arrived at their judgements. It investigates factors that influence decisions and resources drawn on, and how these relate to decision making.

Introduction

An issue for any organisation producing tests, or for users of test results, is deciding what constitutes an adequate score or pass mark. Exams may have only one pass mark, or they may classify examinees into several possible score levels or grades. These grades or passing marks are called *cut scores* (Frey, 2018). The process used to establish one or more *cut scores* is called standard setting. There are many standard-setting methods in common use, but regardless of the method chosen, the goal of those implementing the standard setting is to facilitate a shared understanding of what a minimally competent candidate (MCC) can be expected to achieve on the test, and of the minimum requisite skills and abilities needed for a test-taker to pass. The panel is engaged in individual and collective decision making, and this decision-making aspect is the focus of the study.

The context of this paper is the ongoing programme of alignment of tests produced by Cambridge University Press & Assessment (Cambridge English) to the CEFR. In particular it concerns standard setting for the reading component of Linguaskill, an online multi-level test of English for adults. Three standard-setting panels were convened online by Cambridge English to recommend cut scores to distinguish between A1 through to C2 levels of performance for the reading component. In response to the COVID-19 pandemic, a computer-mediated approach to standard setting was developed in place of the previous face-to-face approach. Reflections on the suitability of this approach are discussed elsewhere in this proceedings (*An online flipped classroom approach to standard setting* by Jane Lloyd).

Study overview

A systematic literature review investigated decision making in standard setting, item writing and rating, and studies on collective decision making by expert teams. A basic model of decision making derived from the literature envisions the decision-maker attending to or influenced by internal and external factors. External factors consist of data and evidence, peer feedback and support, resources and time. In an assessment context, data would be item-level data from testing, and an example of evidence would be information on Common European Framework of Reference for Languages (CEFR, Council of Europe, 2001) levels of lexis provided by language databases. Peer feedback refers to the other participants in the standard setting. A resource used during standard setting could be the illustrative CEFR descriptors, and time refers to the time available for decision-making for the duration of the standard setting process. Internal factors comprise individual decision-making style or preference, assessment attitudes and beliefs, domain knowledge and expertise, and experience. In this assessment context, domain knowledge refers to knowledge of English as a Foreign Language, of the CEFR and of assessment. Experience refers to teaching experience or experience of working as an item writer for reading assessments.

This study employed a parallel convergent design where qualitative and quantitative data were collected and analysed independently and brought together for interpretation (Creswell & Plano Clark, 2018). Data collected specifically for the study consisted of selected response data from a) a questionnaire on factors affecting judgements and decision-making processes, b) open-ended data from individual retrospective reports on decision making on individual reading items and, c) 1–1 interviews carried out following the standard setting. This data collection was replicated for each of the three panels. In addition, data were collected which forms part of routine standard setting at Cambridge English. For example, an evaluation questionnaire was completed by panellists immediately following each panel. Use of the video conferencing software facilitated small breakout

rooms, and this feature was exploited extensively during the panels. Although these are not usually recorded, panellists gave their consent for study purposes. In total 43 small group discussions were recorded for inclusion in the study.

Participation in the additional data collection required for the study was optional and 21 panellists took part. 12 of these panellists attended all three panels, i.e., took part in the standard setting for all cut scores from A1 to C2, and two attended two panels. Participants contributed an interview and a set of recordings for each panel attended. This resulted in a total of 38 sets of individual retrospective reports on five reading items and 38 1–1 interviews.

Initial results

In the decision-making questionnaire, panellists were asked what resources they drew on to make their judgements, and to rank these in terms of importance using a Likert scale. The options were:

- My own professional experience
- My own experiences with real students
- My experience taking the test
- My own performance on the items
- The CEFR level descriptors
- The CEFR judgements I had previously made
- The definition of the MCC
- The group discussions
- Other participants' ratings
- Information on the relative difficulty of the items

All options were selected and provide evidence for the internal and external factors in the decision-making model derived from the literature. Of these, the four ranked as most important were:

1. My own professional experience
2. Information on the relative difficulty of the items
3. The definition of the MCC
4. The group discussions

In line with the convergent design of the study, coding of qualitative data from the interviews, the retrospective verbal reports and the group discussions was based on the questionnaire item options. As listed above, these options refer to standard-setting resources such as the definition of the MCC, standard-setting procedures such as the provision of collated judgement data of all panellists, and the holding of group discussions, plus individual panellist characteristics such as professional expertise. Analysis of the qualitative data is ongoing, but it is possible to identify instances in the qualitative data which support these findings from the questionnaire. Examples citing the four most commonly used resources are given below.

My own professional experience

In the coding of the qualitative data, several aspects of professional experience have emerged. These include beliefs about and knowledge of test design, including item design, identification of reading subskills required, and how items work. Examples of these beliefs are excerpts from the group discussions:

'It's been interesting doing this, these standard settings for reading, because I think where it's a very general opening question that does increase the load.'

'This is not a theory or theoretical point, but it always helps if the options have a similar pattern or similar grammatical shape. Sometimes the really difficult ones they just have four completely discrete ideas you know, going off in four directions.'

'One of the things that makes this question difficult is the fact that it actually tests the gist rather than, you know, detailed information.'

'So basically what I see here is a pattern. They need to understand basic synonyms in order to scan and find out phrases and particular words.'

Information on the relative difficulty of the items

In the interviews, panellists were asked whether they changed any of their decisions as a result of seeing the item difficulty data. In the way standard setting is implemented at Cambridge English, judges do not receive information on the difficulty of the items until after they have made their initial judgements and discussed these with other panellists. Once shown the information on the relative difficulty, the qualitative findings show that judges react to the item difficulty data differently, depending on how closely they reflect their judgements on the items. They are pleased when it confirms their decisions. When it differs, judges may or may not change their decision, and so far, findings suggest this depends on how confident they feel about their previous judgement on an item. This is usually based on their professional experience.

'It was nice to see because it kind of confirmed the vast majority of the judgement I'd made in round one. But I didn't let it influence me too much if it's already ok.'

'I wouldn't say it influenced my judgement massively because, you know, I'm a really experienced teacher.'

'I went back into it and I kind of thought, well, is there anything here that's making it more complex than I thought?'

'So it was reassuring to some extent, but where there were items in doubt. Or where we had discussed things and I had changed my mind. Then that list of live data was quite compelling.'

The definition of the MCC

Panellists seem to use the description of the MCC at different stages of their decision-making process for different reasons: to analyse the test, as a convenient summary of the relevant CEFR scales, as a supplement to their professional experience, and as a resource when they need to reconsider their decision.

'So you're looking at [the MCC description] and looking at the text and going right, which of these features here in this CEFR descriptor or this MCC descriptor are present in the text?'

'it's all distilled into one neat thing.'

'So I kind of drew on my experience as a B1 examiner and a B1 teacher to say that an A2 MCC probably wouldn't be able to answer that question correctly. I also looked at the description of a typical A2 MCC candidate and I just thought that through using those things together it was quite easy for me to make my judgement there.'

'After going back to the text and having another look at it and looking in a bit more detail at the description of a minimally competent A2 candidate, I changed my judgement to No.'

The group discussions

Comments indicate that the group discussions seem to be used as a resource in the same way as the item difficulty data. Judges who are confident in a decision find the discussions interesting and useful but are unlikely to be swayed. In cases where the judgement is not a confident one, they may change their mind, if other judges in the group put forward an alternative point of view or some aspect of the text or item they have not yet considered.

'The online discussion maybe didn't influence much, but it helped me clarify what should be considered when deciding on the cut-off point.'

'If you hear one other person, I think, say the same thing, you think, right, that's how we feel, and then you kind of feel a bit more certain that you'll stick to your judgement, but where the other two didn't agree, it made you think, so what's going on here? And you really have to go back, and I think on that occasion, that's probably why I changed that one item.'

There is also evidence that the judges are more sceptical of the item-level data than they are of arguments put forward by fellow panellists:

'There are so many variables that can affect that data related to the difficulty of the item. And because I don't know what the factors are. You know, I can't really trust it, if you see what I mean. But when I'm discussing with someone, and someone tells me "Look, I think this is this because" then OK, I've got reasons to believe what they are saying.'

Conclusion

As the study and analysis is ongoing, conclusions are tentative. However, it seems that the judges are making use of external resources as expected in standard-setting contexts, and are drawing on the skills, expertise and knowledge in the domain of language assessment which led to their selection as panellists in the first place. These are encouraging findings.

References

Council of Europe. (2001). *Common European Framework of Reference for Languages. Learning, Teaching, Assessment.* Cambridge: Cambridge University Press.

Creswell, J. W., & Plano Clark, V. L. (2018). *Designing and Conducting Mixed Methods Research* (Third edition). London: Sage Publications.

Frey, B. B. (Ed.). (2018). *The SAGE Encyclopedia of Educational Research, Measurement, and Evaluation*. Thousand Oaks: Sage Publications.

An online flipped classroom approach to standard setting

Jane Lloyd
Cambridge University Press & Assessment, United Kingdom

Abstract

The setting of this paper is the ongoing programme of alignment of tests produced by Cambridge University Press & Assessment (Cambridge English) to the Common European Framework of Reference for Languages (CEFR, Council of Europe, 2001). In response to the COVID-19 pandemic, a computer-mediated approach to standard setting was developed to replace the face-to-face approach previously used for standard setting by Cambridge English. This paper includes an overview of the key elements of standard setting, and how these can be accommodated in an entirely online setting. It describes the benefits and improvements to operational practice, and shares lessons learned in the setting up and running of panels in an online, global, flipped classroom environment.

Introduction

The setting of this paper concerns the standard setting for Linguaskill, an online multi-level test of English for adults. In response to the COVID-19 pandemic, a computer-mediated approach to standard setting was developed in place of the previous face-to-face approach at Cambridge English. The standard setting for Linguaskill (2021–2022) was the first time this completely online approach was deployed by Cambridge English.

What is standard setting?

An issue for any organisation producing tests, or for users of test results, is deciding what constitutes an adequate score or pass mark for a defined purpose, such as successful completion of a course of study. Exams may have only one pass mark, or they may classify examinees into several possible score levels or grades. These grades or passing marks are called *cut scores* (Frey, 2018). The process used to establish one or more *cut scores* is called *standard setting*.

Basics of standard setting

There are many standard setting methods in common use, all aimed at providing a structured and reasoned approach to identifying cut scores, and procedural and internal validity evidence. Regardless of the method chosen, standard setting requires qualified, well-trained panellists who are experts in a relevant area. An important concept in standard setting is that of *minimal competence*. The goal of those implementing the standard setting is to facilitate a shared understanding of what a minimally competent candidate (MCC) can be expected to achieve on the test, and of the minimum requisite skills and abilities needed for a test-taker to pass. The panel aims to identify the score point at which the MCC has a greater chance of passing than failing the test and may also identify other grade boundaries if these are an aspect of the test. The score point arrived at by the panel is the *recommended cut score*.

Common elements of standard setting

Standard-setting methods can be broadly classified into those that are test-centred and those that are candidate-centred (Jaeger, 1989). Both types of method may use performance data alongside panel judgements. Whatever method is used, it usually includes an initial round of individual decision-making and collation of the data. This is generally followed by a discussion between judges. Further rounds of decision-making follow until a reasonable consensus is reached. What is considered a reasonable level of agreement or an acceptable result is an institutional decision, implemented by those running the session. The workshop is rounded off by asking the judges for their feedback on the panel's final recommendation, and frequently includes

qualitative feedback from the panel members on their perceptions of the process, and recommendations for improvement (Cizek & Bunch, 2007). Standard setters need to consider how to facilitate the following, both for face-to-face and online or remote settings:

- Access to test materials, performance criteria and performance data
- Familiarisation activities
- Individual judgements
- Discussions of judgements, items, performances, and perceptions
- Collection of judgement data, of feedback and of views

How these elements were managed in an online flipped classroom approach is summarised in the next section.

A digitally-mediated approach

Standard setting for each skill was carried out using a dedicated panel. All panellists attended remotely, as did those running the standard setting. All sharing of materials, data collection and discussion activities were computer-mediated. Activities were spread across several days to allow for the demands of working from home and carrying out cognitively challenging tasks in a computer-mediated environment. Panellists took part in a series of three 90-minute online meetings for each panel they attended. Meetings were spread across a week and took place on Mondays, Wednesdays, and Fridays. Several weeks of panels were necessary to cover all cut scores, as Linguaskill is a multi-level test.

Preparation of and access to materials

Access to and dissemination of test materials and performance data was controlled through a secure sharing platform, Kiteworks, Version 8 (Kiteworks Inc. ©2023). For those managing the standard setting, the flipped classroom approach meant that the majority of preparation was done prior to the online meetings. In this approach, panellists worked individually and asynchronously before and between meetings, so there was a need for self-access materials. These included written instructions, MCC descriptions, introductory videos, online workbooks containing CEFR descriptors, online CEFR quizzes and evaluation surveys.

Familiarisation activities

In a flipped classroom approach, panellists completed the familiarisation activities in advance, and came to the first meeting already prepared for a discussion of their ideas on what minimal competence would look like for a specific cut score on a specific skill. Managing the familiarisation activities was achieved by using a dedicated area in Microsoft Teams and also via email. It was relatively straightforward to provide panellists with written instructions for the familiarisation activities, and to provide access to the Teams area and the secure platform. Teams was used to share non-confidential materials and resources. The social aspect of a face-to-face meeting was addressed by panellists introducing themselves on a discussion board in Teams prior to the first meeting.

Individual judgements

For the panellists, arriving at their judgements on items and performances happened outside and between the meetings. Some time was spent in the first meeting training panellists how to complete and record their judgements using online spreadsheets. Individual spreadsheets were provided to each panellist. Panellists completed their judgements in advance of the second meeting and submitted these for collation.

Plenary and group discussions

The flipped approach meant that meeting time could be devoted to small group and plenary discussions on judgements, items, performances, and perceptions. The facility provided by the software to hold small group discussions was one of the biggest differences to what was possible in a face-to-face setting, and this was made use of extensively in the scheduled online meetings. Collated judgement data were discussed in the second meeting. The process of review, submission and collation was repeated in advance of the third meeting, where the collated data and final cut score were discussed in small groups and in plenary.

Collection of data and feedback

Collection, sharing and review of individual judgement data was facilitated by the conference software, by the use of email, and by spacing the meetings out across a week. Incidental asynchronous communication took place via discussion boards, for example when there was a need to clarify an instruction or make note of a deadline. The software facilitated the design and dissemination of the evaluation survey.

Advantages and disadvantages

Overall, there were more advantages than disadvantages to the digitally-mediated approach, but possible downsides are also noted.

Benefits and improvements

Financial benefits of an online approach were as expected, as there was no need to provide food, transportation, or hotels, and there were no costs related to a venue, or printing. There was also more flexibility for who attended, when this was scheduled, and for how long. The online approach meant that panellist selection was not limited to those resident in the local area but could include language experts based all over the UK and those working and teaching in other countries and time zones. In the case of the Linguaskill exam, panellists were included who were based in Spain, India and Vietnam. Therefore the move to an online environment improved operational practice through increased inclusion and diversity of participants. The online software made it easier to collect evidence of discussions, judgements, and views.

Panellists worked on standard-setting tasks between online meetings, and this resulted in a reduced cognitive load for participants, with more time available to participants for considered judgement and quiet individual reflection. In other words, because much more time could be spent in discussion, the focus was on outcomes and concepts rather than deadlines. In evaluative surveys, panellists commented that they appreciated the time they had to concentrate on judgements between meetings, and how the sharing of data was supported by the technology used. The majority of panellists were in favour of carrying out future standard setting online, rather than face-to face, and have continued to participate in online panels with marked enthusiasm.

Possible drawbacks

The main drawback was the need to train people in the technology, both for the panellists and the staff running the standard setting. There was an initial need for staff to upskill rapidly, so they were confident setting up and hosting meetings and breakout rooms, designing online surveys, collating data, and designing the materials and online spreadsheets for use remotely. These are, however, operational demands that have had to be addressed in many aspects of language testing and are not unique to standard setting.

Participants who did have technical difficulties with the software or access issues were not always able to articulate their problem, and in some cases it was an issue that could not be resolved in real time. There was a need to limit the amount of screen time, which is why we adopted the solution of three meetings spread over a week. This meant that standard setting took longer to complete overall than might be the case in a face-to-face environment. The final drawback was the impact of other factors related to the COVID-19 pandemic. When we were all working from home or in lockdown, panellists were perhaps working in less-than-ideal settings at home and may have been combining other responsibilities such as childcare or self-isolation. Over time, these have become less prominent issues as we all adapt to new ways of working.

Issues to consider

Two of the main issues for anyone considering implementing this approach are time and technology. Issues related to time are finding the best common time to hold meetings, given that the panellists may not be working in the same time zone, and that the meetings will likely be held over several days. There is a need to measure time, in order to work out how to calculate the workload for those involved, in order to pay participants appropriately, and to estimate the overall duration and the likely timeframe for the outcomes to be available to the test provider.

In terms of technology, standard setters need to consider security of materials, and their file size (for example with audio or video) as this impacts how easy it will be to share these with panellists. While some platforms are good for sharing and dissemination, they may not be secure, and so a combination of platforms and resources may be needed. The design and amount of material

needs to be considered, as panellists may be working on one small screen at home, so resources need to be manageable in that scenario.

Conclusion

In a flipped approach, the majority of preparatory work happens before the meetings take place, and this has made the workload more manageable for the standard setters. One of the key benefits has been the smooth running of the meetings. This is because of opportunities for all involved to concentrate on their tasks between meetings; additional time for discussion; ease of data collection; and ample time to collate and analyse the data. The greater choice of who can be included has contributed to the depth of the discussions. Participants have come forward for multiple panels, and as we continue to run online standard setting, we build up both technical expertise and a community of practice.

References

Cizek, G. J., & Bunch, M. B. (2007). *Standard Setting: A Guide to Establishing and Evaluating Performance Standards on Tests.* Thousand Oaks: Sage Publications.

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment.* Cambridge: Cambridge University Press.

Frey, B. B. (Ed.). (2018). *The SAGE Encyclopedia of Educational Research, Measurement, and Evaluation*. Thousand Oaks: Sage Publications.

Jaeger, R. M. (1989). Certification of student competence. In L. R. Linn (ed.), *Educational Measurement* (Third edition) (pp. 485–514). New York: MacMillan Publishing.

Killing a flock of standard-setting judgements with one digital stone

Anniken Telnes Iversen
Norwegian Directorate for Higher Education and Skills

Torbjørn Olseth
Norwegian Directorate for Higher Education and Skills

Rønnaug Katharina Totland
Norwegian Directorate for Higher Education and Skills

Abstract

In April 2022, the Norwegian test for adult immigrants carried out our first digital standard setting of six cut scores for our listening and reading tests. Teachers from different parts of the country took part, contributing their knowledge of the candidates and the CEFR. Online seminars meant less time off teaching and reduced expenses. Using the principles of flexible learning, we made training, familiarization and group discussions more flexible and pedagogical. We offered short films and tasks to do at home, in addition to regular session before the group discussions. Group deliberations were conducted in breakout rooms with the help of screen sharing.

Using the same standard-setting method as before, we achieved very similar standard error and cut scores, despite having more and more diverse judges. There was, however, a lot less time pressure, and the teachers reported gaining a better understanding of language testing.

Introduction

The purpose of this paper is to share our experiences of an online standard-setting project. March 2022 was the first time the official Norwegian test for adult immigrants (*Norskprøven*) conducted the standard setting of listening and reading tests online rather than in person. The aim was to solve some problems from earlier sessions.

The most important thing we learned was that given careful planning, much successful digitalization can be achieved at little cost and without specialized tools and technical expertise. We found standard setting lends itself to digitalization. This paper will only relate how some specific changes functioned, not the entire process of standard setting.

Background

The Manual for relating language examinations to the Common European Framework of Reference for Languages [CEFR] defines standard setting as 'a process of establishing one or more cut scores on examinations' (Council of Europe, 2009, p. 7). *Norskprøven*, the official Norwegian test for adult immigrants, has four parts: speaking, listening, reading, and writing. The reading and listening tests use standard setting to decide scores for achieving Levels A1, A2, B1 and B2 on the CEFR.

The reading and listening tests are multi-level, testing Levels A1 to B2. They are taken online and scored automatically. There are thus eight cut scores for the reading and listening tests. Last year's standard setting set six of these, from A2 to B2.

April 2022 was towards the end of the COVID-19 period. Digital meetings and online collaboration, in addition to a webinar on flexible learning, inspired us to try out digital standard setting with workshops conducted on Zoom, using the principles of flexible learning.

Flexible learning is characterised by digitalization, small modules and most of all, a focus on learners, or, in our case, participants. Flexibility can be introduced in:

- Content
- Time
- Place
- Presentation
- Progression
- Interaction

The digital stone of the title refers to the fact that we were trying to achieve many things at once, using digital means. We also wanted to make the procedure more efficient, less demanding for participants, and to use standard setting to increase teachers' knowledge of *Norskprøven* and language testing.

The challenges we wanted to address can be summed up as: people, space, time, and mental capacity.

Challenges: People, space, time, and mental capacity

Our first challenge was finding good judges. Konrad, Spöttl, Holzknecht and Kremmel argue that the 'participation [of teachers] in standard setting can have additional advantageous effects by helping to increase public acceptance of a test and spreading awareness of principles and good practice of language testing among teachers' (2018, abstract).

For years we have wanted to include more teachers as standard setters, both for their benefit, and for the quality of the standard setting. While teachers of Norwegian as a second language serve as examiners and raters for the speaking and writing tests, they have much less knowledge of the reading and listening tests. We re-use piloted reading and listening tasks and maintain strict confidentiality. Taking part in standard setting would, we hoped, increase teachers' knowledge of the test, and enable them to prepare better.

Previously, few local teachers participated, due to the cost of travel and accommodation, and because teachers cannot take time off to take part. Digital meetings make it much easier for teachers from around the country to participate.

In addition to our 14 internal judges, 54 teachers were recruited as judges. Most had long experience of teaching at the levels they judged; many had also served as oral examiners and 21 as raters of the written test.

The second problem was space. Earlier sessions were held in our own offices, which are relatively small, where the rooms are not suitable, and it was difficult to organize group discussions.

Zoom, however, was perfect for this purpose. We could shift easily and quickly between plenary meetings and group discussions in breakout rooms. Instead of walking between rooms, we could take proper breaks, and participants could get coffee.

Online workshops thus solved the space problem, and we were able to include more participants than we otherwise would. It also removed stress from the procedure.

Before, we conducted all-day standard-setting workshops, covering two cuts a day. In 2017, for example, we set two reading cuts one day and two listening cuts the next. Some familiarization tasks were completed before the sessions, but the workshops included training and introductions to the procedures. Despite lasting a whole day, the workshops were stressful, and we felt pressed for time throughout.

This time, we cut down to six four-hour days, setting one cut and skill each time. Teachers set aside preparation time at home, and attended two workshops. They should only take part in levels that they knew well, and they had to do both reading and listening at that level. We thus had three groups of standard setters, each responsible for one cut.

Innovations and flexible learning

Most of the changes we made compared to earlier years had to do with time in some sense, but also mental capacity. Panellists have much to learn before they can evaluate items, and workshops are full of information and demanding cognitive work. To remove some pressure, explanations and introductions that had previously been part of the workshops were sent to participants in advance, much of it on video.

We created four videos of less than 7 minutes. The first explains what standard setting is and why it is done. Relevant language testing procedures and concepts are explained, such as piloting of listening and reading tasks, and the concept of 'the minimally

competent candidate'. We also discuss the link between the CEFR, pilot results and cut scores. This video was made in Microsoft Teams and shows only a person talking.

The next two videos outline the phases of the standard-setting process and explain the practical steps. These are quite technical and were made in PowerPoint with plenty of illustrations. The video plays the presentation with a voiceover.

The fourth video contains training tasks, both familiarization tasks, and practical exercises for participants to try out procedures explained in the videos. This material was also sent out in a Word file for printing.

The videos meant less work during workshops. For judges, other advantages included the possibility to:

- see the material one or more days before the workshops;
- choose when and where to watch them;
- decide how long to spend on videos and how often they watched;
- discuss content with colleagues;
- sleep on it (a night's sleep helps reinforce new learning):
- reflect and ask questions in the workshops.

The training became more flexible for participants, as they could decide when and how to do it, alone or with a colleague.

For the organizers it was important to save time in the workshops. Making videos felt particularly meaningful, since we knew they could be reused in later standard settings. They can also be altered and improved next time.

For the standard-setting workshops themselves, the biggest difference from earlier was the breakout rooms. We organized participants into groups of about five. Groups were set up beforehand, with one staff member in each, and experienced judges were distributed across different groups. These groups were the same over the two days that we used on each CEFR level. We saw that participants became comfortable with each other, and trust and communication grew over the sessions.

Breakout rooms enabled screen sharing, so everyone could see each other's judgements, and tasks and items were studied more closely. Our impression was that group discussions worked better than before.

What the process taught us

Planning the online event and making material taught us some important principles. First, that acceptable videos can be made without expensive equipment or expert help. We used PowerPoint and Zoom.

PowerPoint is particularly flexible because slides can be made individually and the audio for each one recorded separately. Only after all the slides are made is the whole presentation saved as a video. It can then be shared with an audience (e.g., on YouTube). Viewers see one seamless video presentation, while creators have the option of changing, adding or re-recording parts, and saving it again as a new video. Compared to filming a talk in a studio, this method is flexible and allows for easy revision.

Secondly, flexible learning made us see that making short videos and cutting materials into smaller thematic chunks makes it easier to use. Users can easily get an overview of the material, and they are more likely to watch each one to the end. If they want to repeat one topic, it is easy to find.

Short videos also create variation for viewers – especially if different people present them and they're made in different ways. Not all videos need illustrations or PowerPoint slides. A person talking gives your organization a face, and viewers have a person to relate to. People might also concentrate better if they don't read as they listen.

Results

The cut between A2 and B1 for listening can serve as an example of the results. The workshop had 35 judges, 11 internal and 24 external (mostly teachers). There were three rounds of individual judgement and discussion. When the same cut was set in 2018, we had 17 judges taking part in a face-to-face workshop, and there was only time for two rounds of judgements. The standard deviation was a little lower in 2018 (which is natural with fewer participants), whereas the standard error was a little lower last year.

The analyst's report stated that 'the standard was set with a low standard error and very close to the former standard' (our translation, 2022 p.10). Some other cuts changed a little, but his conclusion is that there is no indication that the quality of standard setting was poorer than previously.

After the event, participants were asked how they would rate 1) the content of the workshops (information material, videos and texts for preparations, group discussions, summaries, etc.); and 2) technical aspects of the workshops (information in advance, Zoom meetings, breakout rooms, score submissions, etc.). The options were 'good,' 'ok', and 'not very good'. More than 75% thought the content had been 'good', and 13% said it was 'ok'. Nobody answered 'Not very good'. Only two people said the technical aspects were 'ok', the rest said 'good'.

Table 1: Comparison of standard-setting processes 2018 and 2022

	Round 1		*Round 2*		*Round 3*	
Listening B1	*2018*	*2022*	*2018*	*2022*	*2018*	*2022*
N – no. participants		35	17	35	17	35
Average score		25.38	23.81	25.53	24.47	25.52
Standard Deviation		3.79	1.75	1.91	1.06	1.40
Standard Error		0.64	0.42	0.32	0.26	0.24

Written responses were also very positive. One teacher said 'I found the seminar very useful. It was challenging, but I learned a lot. I'd be happy to take part again later.' Another answered 'It's interesting and useful for me as a teacher to get an understanding of the complex process that standard setting is. I'm glad I had the chance to participate.'

We thus had confirmation that the workshops had been useful for the teachers, and had served as further education in language testing.

The smooth running of the event and positive comments on the technical aspects, we put down to very thorough planning. Digital events need more planning than physical ones, and two groups planned the technical and content parts respectively. The content group created videos and training tasks and wrote scripts for presentations and workshops.

The technical group created detailed plans for all activities before, during and after the workshops. The technical schedule for the workshop included roles for everybody on the organizing team, with instructions about what each should say and do and when. Everything was written down, including who would press the button to turn microphones on and off. Having so many short parts (presentations, individual tasks, forms to be submitted, coffee breaks and discussions) makes timing everything in advance and sticking to the schedule extremely important. Had we gone a minute over time on each part, the whole event would have been a disaster.

We also planned for things that might go wrong, and sent participants technical instructions and information on who to contact in case of technical problems.

Conclusion

In this project, where we tried online standard setting for the first time, we set the cut scores, and the quality was as good as before. We recruited 54 new judges, due to online meetings and more and shorter workshops. The combination of preparation at home and online workshops solved space problems. Group discussions in breakout rooms worked well and saved time. Participants were largely satisfied, and teachers who took part felt they had learned a lot. The organizing team have also gained valuable experience and are confident we can do even better next time.

One participant left the following comment: 'Given that this was the first time the standard setting was done online, I think everything was fantastic. . . . I think this was interesting and great fun to do digitally.'

References

Council of Europe. (2009). *Relating Language Examinations to the Common European Framework of Reference for Languages: learning, Teaching, Assessment (CEFR): A Manual*. Strasbourg: Council of Europe.

Konrad, E., Spöttl, C., Holzknecht, F., Kremmel, B. (2018). The Role of Classroom Teachers in Standard Setting and Benchmarking. In Xerri, D. & Vella Briffa, P. (eds.), *Teacher Involvement in High-Stakes Language Testing* (pp, 11–29). Springer, Cham.

Rossow, A. O., og Eriksen, J. (2022). *Rapport nr. 13/ CREATEDATE \@ "yyyy" * MERGEFORMAT 2022, Standardsetting Norskprøven, april 2022. Delprøve i lytte- og leseforståelse, nivå A2, B1 og B2.*

Exploring anti-plagiarism tool effects in the assessment of academic reading-into-writing

Valeriia Koval
University of Bremen, Germany

Abstract

The assessment of paraphrasing competence in integrated writing is somewhat challenging (Gebril & Plakans, 2014). A possible improvement in the assessment accuracy and efficiency can be attained by employing a tool that highlights copied passages from the source in the test-taker's writing. This study aims to investigate raters' perceptions of the application and efficiency of one such tool. The research includes the analysis of focus group interviews (FGIs) and retrospective interviews with five raters who shared their experience of rating with and without the tool. Apart from revealing the facilitative effects of the tool, the results provide insights into the effects of the tool on locating source text information and raters' processes under the two conditions. The research provides insights that the application of the tool contributes to an objective and efficient assessment of academic reading-into-writing.

Introduction

One of the core competencies in reading-into-writing is the ability to transform language from a source by employing effective paraphrasing. To assess this competence validly, raters need to know the input material well. Long and complex input may threaten the objectivity of the rating as it creates a cognitive load for the raters, who are expected to differentiate between language taken from a source text and language produced by the test-taker. This study investigates raters' perceptions regarding the rating of paraphrasing in integrated writing. In particular, I focus on how the raters perceive the assistance in the form of a tool that highlights textual borrowings from the source text.

Literature review

Research on source-text-use (STU) in integrated writing is abundant, especially from the writers' perspective (e.g., Hyland, 2009; van Weijen, Rijlaarsdam, & van den Bergh, 2019). Writers' paraphrasing practices have also been investigated in detail (Keck, 2006; Shi, 2004; Weigle & Parker, 2012). To the best of my knowledge, however, research on the perception of integrated writing assessment, in particular, paraphrasing assessment, is relatively scarce (e.g., Chan, Inoue, & Taylor, 2015; Gebril & Plakans, 2014; Weigle & Montee, 2012; Wang, Engelhard, Raczynski, Song, & Wolfe, 2017).

One thread in this research concerns different perceptions and attitudes toward textual borrowings. In their exploratory inquiry, Weigle and Montee (2012) employed Focus Group interviews (FGIs), stimulated recalls, and a rater judgement task to investigate how the raters perceive and evaluate textual borrowings in integrated performances. The results show that raters sometimes disagreed when assessing paraphrasing competence (Weigle & Montee, 2012). Gebril and Plakans (2014), conducting Think Aloud Protocols and follow-up interviews with two raters, came to a similar conclusion. These results are also depicted in the study by Wang et al. (2017), who investigated rater accuracy and perception of integrated writing assessment. The researchers pointed out that varying attitudes toward textual borrowings may be avoided by more detailed consideration of this aspect in the rater training (Wang et al., 2017).

In the previous research, the raters differed in their ability to notice textual borrowings (Weigle & Montee, 2012). They found it challenging to differentiate between the writer's language and source materials (Gebril & Plakans, 2014). The factors influencing the ability to recognize textual borrowings included knowledge of the source text, rating and teaching experience, and the writer's proficiency (Gebril & Plakans, 2014; Weigle & Montee, 2012). On the one hand, it was challenging for raters to identify liftings when the writers were proficient enough to integrate the borrowings into their language production (Weigle & Montee, 2012). On the other hand, it was easier for raters to identify unattributed citations in the low-proficiency scripts as the writer's language production differed significantly from the copied passage (Gebril & Plakans, 2014).

When considering rating processes reported in previous studies (Gebril & Plakans, 2014; Weigle & Montee, 2012), the raters often had to switch between written script and source materials to differentiate between the writer's language and citations from the source. The self-monitoring strategy was also broadly employed by the raters in the study by Gebril and Plakans (2014).

All in all, the challenging nature of the assessment of textual borrowings is evident. In this context, Gebril and Plakans (2014) pointed toward the facilitative effect that a computer-based tool may have to assist raters in source-text-use detection. This recommendation will be addressed in the current study by answering the following research questions:

1. How do raters perceive the effects of the tool and its features on the rating?
2. What effects does the tool have on locating source text information?
3. What are the reported rating processes with and without the tool?

Research context

This study is part of a broader research project (MASK[1]) that investigated integrated academic-linguistic competencies in the German higher education context and was funded by the German Research Foundation. In the project, four integrated reading-into-writing tasks were developed: two summary tasks and two opinion tasks. The tasks included authentic source texts of approximately 1,000 words.

For the assessment of the written integrated products (n = 674), an analytic rating scale was developed containing the following criteria:

1. Mining ST (source text) for ideas
2. Ideas correctness and precision
3. Linguistic processing
4. Attribution
5. Synthesis of ideas
6. Thematic development
7. Cohesion
8. Vocabulary range & accuracy
9. Grammar range & accuracy

Within the project, five novice raters were trained. The raters were senior students in programs with English as a Medium of Instruction at the University of Bremen.

Additionally, a source detection tool was launched by the MASK cooperation partners. The tool's purpose was to support raters in the third rating criterion – the linguistic processing of ST or writers' ability to paraphrase ideas from the source texts. The tool highlights passages, both in the source and student texts, that were borrowed from the ST. It is an online application accessible via a browser, where raters could upload ST and a student script for comparison. The length of the detected copied passage can be modified. The tool also highlights the keywords and allows the search of corresponding highlighted segments.

Methodology

The study employs retrospective interviews and FGIs with five raters from the MASK project. The combination of the two data sources was chosen because it allows access to a more in-depth understanding of raters' perceptions in different contextual settings: individual – immediately after rating, and interpersonal – in a group. During the individual retrospective interviews, participants answered questions about their rating experience and the idiosyncrasies of rating with or without the tool. In the FGIs, the raters discussed the perception of liftings in the text products, the tool's functions, and rating processes with and without the tool. The retrospective and FGIs were conducted with participants via Zoom, recorded, and transcribed.

The data from retrospective interviews and FGIs were coded with the same coding scheme, as the insights from both sources contribute to the data triangulation. The initial coding scheme was developed based on themes emerging from previous research

[1] Modeling of academic-linguistic competences

and issues discussed earlier (Chan et al., 2015; Gebril & Plakans, 2014; Wang et al., 2017; Weigle & Montee, 2012), and later adjusted to the context of assisted/unassisted rating. The coding scheme included the following broader categories: rating strategies, perception of the tool, and tool effects on the identification of borrowings. The percentage of inter-coder agreement was 90%. After coding, the coded passages were analyzed employing content analysis (Krippendorff, 2004).

Results and discussion

How do raters perceive the effects of the tool and its features on the rating?

All raters perceived the tool as helpful and useful and agreed that rating without the tool was more challenging than rating with the tool. According to three raters, it was easy to get used to the tool (Raters 4, 2 and 6). However, raters also pointed out that they needed to learn to use it correctly without overly relying on it (Raters 6 and 4). Despite the prevailing positive feedback, raters also mentioned issues regarding possible limitations of the tool. For instance, the inability of the tool to recognize spelling mistakes was often criticized by the raters (Raters 3, 2 and 5), which represents a potential danger to accurate lifting identification. Another effect mentioned by one rater concerned the perception of the highlighting in the tool. Rater 2 explained:

> *I would say that with the tool, sometimes it can be a little bit misleading because you have this color coding of the source text. And sometimes that can influence . . . like you have this visual in front of you. And it kind of influences your perception of how many instances there are.*

Regarding the tool features, all the raters found them helpful. For instance, the tool's highlighting of the keywords was perceived positively as it helped to confirm liftings more efficiently (Raters 2 and 5) and made the tool usage more targeted (Rater 6).

What effects does the tool have on locating source text information?

The findings show that the tool facilitated raters in the identification of liftings. Specifically, Rater 2 reported finding fewer liftings without the tool, and Rater 3 pointed out that identification of the liftings without the tool was time-consuming. These raters explained that it was difficult to spot liftings without the tool because of a writer's transformations to the source text. As Rater 2 pointed out:

> *. . . so, it's not really that easy to spot (lifting) when you read it without the tool because kind of these words inserted can be distracting.*

Weigle and Montee (2012) stated that it was easier for raters in their study to identify liftings in the lower-level scripts, while proficient writers could 'hide' the borrowed words in their constructions. In the present study, raters also reported that it was easier to identify 'hidden' liftings with the tool (R3, R4). In future rater training, nonetheless, it is important to include examples of various paraphrasing transformations to prepare raters.

What are the reported rating processes with and without the tool?

When examining the reported rating processes, the following outcomes can be observed:

Processes with the tool

- Investigation of the highlighted parts to confirm or deny the lifting (n = 5)
- Occasional comparing of specific highlighted segments to the source text for a detailed consideration (n = 4)

Processes without the tool

- Identifying potential liftings (e.g., by re-reading the script) (n = 5)
- Marking potential, most obvious liftings in a script (e.g., by underlying the passages) (n = 3)
- Searching source text to confirm or decline the liftings (n = 5)

According to the reported rating processes, rating with the tool allowed raters to focus on the identified liftings, while in the unassisted condition, raters had to identify liftings first. After identifying potential textual borrowings, the raters had to go back and forth between the script and source text/content expectations to locate and confirm the liftings when rating without the tool. Similar procedures were also reported in the studies by Gebril and Plakans (2014) and Weigle and Montee (2012). Hence, the tool application improved the efficiency of the rating by allowing raters to focus on already identified liftings. Therefore, the study results suggest that raters should be explicitly trained to follow the most efficient rating steps and not overly rely on rating assistance. Additionally, future research may benefit from a more detailed investigation of rating processes by, for example, observing actual rating behaviors.

Conclusion and implications

Investigating the perceived tool effects and their influence on the rating of integrated writing emphasizes the complex nature of source-based rating and the importance of assistance for rating. Nevertheless, the raters in the study reported both positive and negative effects of the tool. With positive outcomes of the tool application prevailing, the raters warn of overreliance on it.

Raters' perceptions about tool application and possible rating bias create a basis for better future rater training and monitoring practices, especially in the context of assisted rating.

All in all, the tool improved the efficiency of textual borrowing identification and can be recommended for assessment practices. It can also be recommended for classroom usage to raise students´ awareness of appropriate paraphrasing practices.

References

Chan, S., Inoue, C., & Taylor, L. (2015). Developing rubrics to assess the reading-into-writing skills: A case study. *Assessing Writing*, *26*, 20–37.

Gebril, A., & Plakans, L. (2014). Assembling validity evidence for assessing academic writing: Rater reactions to integrated tasks. *Assessing Writing, 21*, 56–73.

Hyland, T. A. (2009). Drawing a line in the sand: Identifying the borderzone between self and other in EL1 and EL2 citation practices. *Assessing Writing, 14*(1), 62–74.

Keck, C. (2006). The use of paraphrase in summary writing: A comparison of L1 and L2 writers. *Journal of Second Language Writing, 15*(4), 261–278.

Krippendorff, K. (2004). *Content Analysis: An Introduction to Its Methodology* (Second edition). Thousand Oaks: Sage Publications.

Shi, L. (2004). Textual borrowing in second language writing. *Written Communication*, *21*(2), 171–200.

van Weijen, D., Rijlaarsdam, G., & van den Bergh, H. (2019). Source use and argumentation behavior in L1 and L2 writing: a within-writer comparison. *Reading and Writing, 32*(6), 1,635–1,655.

Wang, J., Engelhard, G., Raczynski, K., Song, T., & Wolfe, E.W. (2017). Evaluating rater accuracy and perception for integrated writing assessments using a mixed-methods approach. *Assessing Writing*, *33*, 36–47.

Weigle, S.C., & Montee, M. (2012). Raters' perceptions of textual borrowing in integrated writing tasks. In M. Tillema, E. Van Steendam, G. Rijlaarsdam & H. van den Bergh (Eds.), *Measuring Writing: Recent Insights into Theory, Methodology and Practices* (pp. 117–151). Leiden: BRILL.

Weigle, S. C., & Parker, K. (2012). Source Text Borrowing in an Integrated Reading/Writing Assessment. *Journal of Second Language Writing*, *21*(2), 118–133.

Modelling information-based academic writing: A domain analysis focusing on the knowledge dimension

Chengyuan Yu
Kent State University, USA

Abstract

In this digital era of information explosion, academic writers face challenges in meeting information needs and effectively utilizing information for writing. While academic writing is believed to inherently involve information literacy, current assessment methods mainly rely on prompts or limited readings, failing to capture the information-driven nature. To fill this gap, this study conducted a domain analysis of information-based academic writing that could inform the development of a new academic writing assessment in Higher Education (HE). Focusing on knowledge, I interviewed graduate students, subject instructors, writing experts, and information specialists in the field of education. Findings revealed three key knowledge types: (1) knowledge of the discipline, (2) knowledge of the information environment, and (3) knowledge of the language. Implications for information-based academic writing were therefore discussed.

Introduction

In this digital age of information explosion, it is difficult for academic writers to identify information needs, and to locate, access, evaluate, organize, communicate, and use the needed information for writing. While it has been increasingly recognized that academic writing involves the processes of information literacy and is inevitably an information practice (Yu, 2023a; Yu & Zhao, 2021), existing writing assessments are still predominantly prompt-based (Yu, 2023b) or only integrate the reading of one or two passages (e.g., Test of English as a Foreign Language) and fail to reflect the information-involved nature of academic writing. To fill such a gap, the present study reports a domain analysis of information-based academic writing which can inform the development of a new generation of authentic writing assessment in the Higher Education (HE) context. This domain analysis specifically focuses on the knowledge, one of the most important dimensions of domain analysis (Chapelle, 2008; Mislevy, Steinberg, & Almond, 2003).

Methods

Participants

Four groups of participants were interviewed in a semi-structured way, including 24 graduate students (GS), four instructors in disciplines (IiD), four writing experts (WE), and six information specialists (IS). For anonymity concerns, they were coded in numbers, for example, WE1. All groups were interviewed about their understanding of academic writing practice, but with different foci. For example, WE were directed to share their teaching practice of writing and their most updated scholarly understanding of academic writing; IS reported how they used their expertise to facilitate students' academic writing development and how information literacy figures in the process of writing; IiD added to the disciplinarity of academic writing through reflecting on the characteristics of writing expectations for students; GS, as the major stakeholders, showed us the real scenario and practice in academic writing. The perspectives from different groups of participants can at best approximate the current practice of academic writing in the higher education context. As academic writing is a discipline-specific task and different disciplines may follow different practice in writing (Hyland, 2004; 2008; Wingate, 2006), this study selected students and instructors in the discipline of education to focus on one discipline. All the participants speak Mandarin Chinese as their first language or primary language in daily communication.

Data collection and analysis

The participants were interviewed individually in a private meeting room. The interviews were audio-recorded with the participants' consent. Mandarin was used as the language for interview to allow smooth and comfortable communication and the best representation of the interviewees' ideas. However, as the participants were all fluent English language users, they occasionally code-switched between Mandarin and English when they felt English may better represent their thoughts, especially academic terms. The interviews were transcribed verbatim and coded in a thematic way. Special attention was paid to the different types of knowledge. Words, phrases, and sentences relating to knowledge were identified and categorized iteratively. The coding was conducted again approximately three months later and shared with some participants to keep the data analysis from possible biases.

Findings

Three aspects of the knowledge important for information-based academic writing were identified in this study: (1) knowledge of the discipline, (2) knowledge of the information environment, and (3) knowledge of the language.

Knowledge of the discipline

Information-based academic writing first requires knowledge of the discipline. Regarding the information literacy practice, individuals with more knowledge in the discipline can more efficiently identify the venue to find useful information and evaluate the accessed information. For example, IS5 exemplified the importance of knowledge of the discipline as demonstrated in their interview transcripts:

> Knowledge of the discipline can help information users to have a better understanding of the advantages and disadvantages of different information sources . . . it is very very difficult to provide service to advanced-level information users. For a librarian, their academic backgrounds and disciplinary knowledge may vary . . . How can we provide services to a professor? It is very very difficult. (IS5)

IS5 demonstrated that knowledge of the discipline is the foundation for information literacy practice. Even an experienced information specialist like an academic librarian at a research university library can feel it difficult to provide information services that are closely associated with the discipline. IS4 further highlighted the importance of the knowledge of the discipline by categorizing it into the explicit and the tacit, with the explicit being familiarity with the discipline (e.g., evaluative judgement for the accessed information) that is developed through gradual socialization and the implicit being content knowledge:

> The more familiar is an individual with their own discipline, the more likely he knows where to find the information. He can also base on (content) knowledge in his own discipline to effectively evaluate the quality and usefulness of the information. (IS4)

The importance of the knowledge of the discipline was shared by writing experts teaching academic writing to students from a different research field. WE1 believed that 'the graduate students that she taught should be already ready for writing for their thesis, as they should have had a strong mastery of knowledge in their research fields'. WE4 believed that 'students should have a sound knowledge base which is the blood and flesh of their writing'. As the content experts in their disciplines, instructors in this study also expressed that knowledge of the discipline is critical. For example, when asked about the knowledge important for academic writing, IiD1 immediately told the researcher that 'the emerging scholars in this field first need to know the up-to-date theories or perspectives as well as the major findings in their field'.

Knowledge of the information environment

The information specialists emphasized that students should have a good knowledge of the available academic resources, electronic and hardcopy, that the university library houses, and students should have awareness of referring to these resources when they feel it necessary to read more on the topic. When asked about students' difficulties and problems in information literacy, IS6 told the researcher that 'students tend to rely on everyday search engines, for example, Google, rather than academic resources.' Instead, IS6 believed that 'students should know the valuable and recognized databases in their disciplines that they have access to at the university library' (IS6). Similar views were also held by other information specialists. For example, IS5 shared with the researcher that 'when new students come to the university, the library will organize orientation activities to introduce all the available databases' and she believed that 'if students know better about the available resources, they can find information resources for their writing more effectively and efficiently'. In addition, 'students should also know the useful information tools available in our university library, for example, EndNote to manage your references' (IS6). The researcher also found that students have difficulty accessing and identifying information due to their lack of knowledge of academic databases

that they can use at their current university. For example, GS7 only knew how to use the search engine offered at the university library portal and complained that 'the search can sometimes result in a huge number of entries, and it is tiring to pick up the useful articles from them by manually selecting sources from those prestigious journals'.

Knowledge of the language

Although writing in the HE context is no longer merely a language act, it is still language-intensive in more than one aspect. The understanding of writing tasks and the content of information as well as the act of producing texts requires a comprehensive and systematic linguistic knowledge base. The importance of linguistic knowledge can, for example, be demonstrated by WE1's descriptions of her academic writing courses. In her classes, she 'directed the students to use corpus tools like AntConc to analyze the expressions in academic articles'. She particularly highlighted that she 'guided students to pay attention to the stance markers and the overall discourse structures'. The special attention to language not only highlights the importance of linguistic knowledge but also suggests that knowledge of the language can be complex and multidimensional, from the micro-level subtle use of functional expressions to the organization of the entire text. Similar ideas were shared by WE3 who 'taught the students useful phrases in academic writing, for example, phrases to introduce the study or the differences between seemingly synonymous phrases'.

The importance of language knowledge can also be gleaned from the students' perspectives because many of them reported their lack of confidence in using English for academic writing and language-related difficulties when asked about challenges in academic writing. Related to information literacy practice, students can have difficulty finding the exact keywords to form the search query. For example, one doctoral student who received his Bachelor's and Master's degree in mainland China shared with the researcher that '[he] did not know the proper keywords in English, so [he] adopted a word-by-word translation strategy to translate Chinese terms directly into English and used it as keywords for searching.' Because of the lack of language knowledge, he 'needed to spend a lot of time in searching' (GS24). When asked about their difficulties in academic writing, many of the students told the researcher that language is one major problem:

> Now, language is the main problem, because I feel that I haven't read enough amount of literature in English and don't know some phrases and how to describe some statistical methods. (GS24)

Instructors also agreed that language is the foundation of academic writing. For example, IiD3 mentioned that '[he] sometimes cannot understand [his] students' writing because of improper word choices and grammatic errors'. IiD1 shared with the researcher that 'some students [he] supervised had problems in language. Students assumed that they know every word in the sentence, but they did not know the exact meaning in academic discourse.' This points to the importance of the metadiscourse aspect of academic writing, in addition to the propositional aspect.

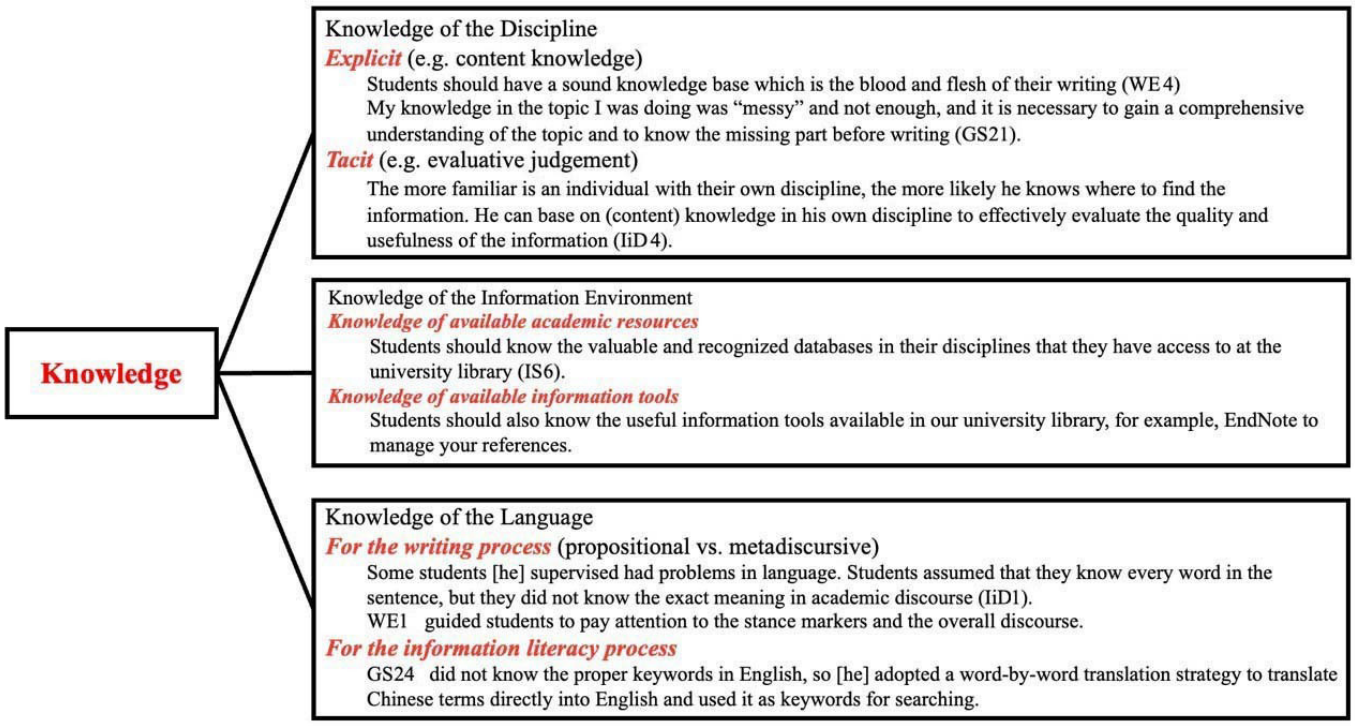


Figure 1 The three aspects of the knowledge important for information-based academic writing

Summary of findings

This study identified three types of knowledge important for information-based academic writing: (1) knowledge of the discipline, (2) knowledge of the information environment, and (3) knowledge of the language (see Figure 1). The knowledge of the discipline comprises (1a) explicit knowledge, for example, content knowledge, and (1b) tacit knowledge, for example, evaluative judgement that should be acquired through accumulating experience in the discipline. The knowledge of the information environment can be categorized into (2a) knowledge of available academic resources, and (2b) knowledge of available information tools. The knowledge of the language consists of (3a) knowledge of the writing process, which can be further categorized into propositional knowledge and metadiscursive knowledge, and (3b) knowledge of the information literacy process by which academic writers can know the keywords to search in the databases.

Conclusion

Based on the stakeholders' perspectives and experience, this domain analysis argues that the assessment of information-based academic writing should be situated in specific disciplines to incorporate disciplinary practice. More specifically, the task should first be related to content and conventions to assess students' familiarity of disciplinary contents and conventions. Second, the assessment should be situated in a specific information context, for example, to let students actually search in a university information context. In terms of the language aspect, writing assessments should assess both propositional and metadiscursive aspects of language. Language use should be assessed throughout the information-based academic writing process, including both the information literacy and the academic writing parts.

References

Chapelle, C. A. (2008). The TOEFL validity argument. In C. A. Chapelle, M. E. Enright & J. Jamieson (Eds.), *Building a validity argument for the Test of English as a Foreign Language* (pp. 319–350). New York: Routledge.

Hyland, K. (2004). *Disciplinary Discourse: Social Interactions in Academic Writing*. Michigan: University of Michigan Press.

Hyland, K. (2008). Academic clusters: Text patterning in published and postgraduate writing. *International Journal of Applied Linguistics, 18*(1), 41–62.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research & Perspective*, *1*(1), 3–62.

Wingate, U. (2006). Doing away with 'study skills'. *Teaching in Higher Education, 11*(4), 457–469.

Yu, C. (2023a). *Understanding the role of information literacy in academic writing in higher education* [Unpublished Ph.D. Thesis]. University of Macau.

Yu, C. (2023b). Testing issues in writing. In H. Mohebbi & Y. Wang (Eds.), *Insights into teaching and learning writing: A practical guide for early-career teachers* (pp. 56–70). Melbourne: Castledown.

Yu, C., & Zhao, C. G. (2021). Continuing the dialogue between writing experts and academic librarians: A conceptual model of information-based academic writing in higher education. *The Journal of Academic Librarianship*, *47*(6), 102454.

Using dynamic assessment of writing to promote technology-enhanced learning in higher education

Eleni Meletiadou
London Metropolitan University, United Kingdom

Abstract

Responding to students' requests for the use of digitally-enhanced formative assessment practices, this project used a dynamic assessment approach that has been developed within the Vygotskian sociocultural theory of learning. It aimed to: (a) develop students' professional skills, i.e., digital skills, in blended learning by promoting experiential learning, (b) improve students' writing performance and favourable attitudes towards learning, and (c) support their well-being in higher education institutions (HEIs) especially in the post-COVID-19 era. Fifty final year students participated in this project as part of their module. Adopting a process approach to writing, the lecturer/researcher used three rounds of mediation. This project aimed to foster inclusion of the increasingly diverse student cohorts due to globalisation, develop students' digital, academic, and professional skills, and innovate in tertiary education.

Introduction

Responding to students' requests for the use of formative assessment practices which would involve them more actively in their own learning process and allow them to develop their academic skills, the current project used a dynamic assessment (DA) approach that has been developed within the Vygotskian sociocultural theory of learning (Vygotsky, 1978).

This project explored the value of peer mediation in the context of academic writing skills' development among undergraduate Management students in blended learning. Previous implementations of this approach in other settings I have conducted indicated that its use can include students as partners, develop their professional skills by promoting experiential learning, i.e., collaboration, improve students' writing performance and favourable attitudes towards learning, and support their well-being in higher education institutions (HEIs) especially in the post-COVID-19 era (El Said, 2021).

Literature review

In HE, first year students, especially international multilingual and non-traditional students, often find it difficult to acquire disciplinary knowledge as students need to understand and use different concepts and theories from written texts in their assignments and meet differing criteria for academic excellence (Becher, 1994). Helping these students to increase their academic achievements is a major problem nowadays in HE institutions in the UK (Ivanic & Lea, 2006).

In terms of this project, a ground-breaking method, DA, was used. DA is an 'approach to understanding individual differences and their implications for instruction . . . [that] embeds intervention within the assessment procedure' (Lidz & Gindis, 2003, p. 99). In DA, student skills are transformed through dialogic cooperation between the learner and the lecturer (Poehner, 2007) but also between the learner and another learner/assessor. In this intervention we combined tutor with peer mediation which refers to digital and/or text-based interaction about the assignment text between the tutor and the learner but also between learners. DA was used as a kind of alternative formative assessment geared towards learning and writing enhancement based on assessment at different times and by different assessors (lecturer and students) during a module of study (Huot, 2002). Therefore, its main aim was to support students during the different phases of process writing and help lecturers make any necessary adjustments and/or provide remedial teaching to cater for all students' needs, tastes, and skills, aiming to 'provide feedback on performance to improve and accelerate learning' (Sadler, 1998, p. 77).

However, despite the recognition of the value of formative assessment of writing skills in HE (Walker, 2009), the use of alternative formative writing assessment which fosters enhanced student learning is under-researched. Previous research which has

explored the impact of lecturers' feedback on students' written assignments (Walker, 2009) revealed that lecturers' feedback referred either to micro-level aspects of writing i.e., grammar (Stern & Solomon, 2006) or were not taken into consideration when students were asked to revise their work (Walker, 2009). These findings indicate that lecturers should experiment with other forms of feedback, possibly integrating lecturers with students' feedback to offer more feedback and engage students more actively in their own learning process, increasing their engagement and overall academic performance.

This intervention study aspired to make a contribution to this area of research by exploring how DA may offer an innovative framework to support management students' academic writing performance by offering developmental feedback combining lecturers and students' comments in a unique way.

DA is based on Vygotsky's Sociocultural Theory (SCT) of Mind (1978), according to which human cognition and learning is regarded as a social and cultural – rather than a personal – process. In particular, the Vygotskian notion of the Zone of Proximal Development (ZPD) and mediation is closely related to DA. According to Vygotsky (1978, p. 86), the notion of ZPD refers to 'the distance between the actual development level as determined by independent problem solving and the level of potential development as determined through problem solving under adult guidance or in collaboration with more capable peers'. DA is based on process writing that does not focus exclusively on the final product and enables students to improve their performance by receiving feedback from their peers and lecturer. They are required to take small steps every time they are asked to revise their work, thus developing their reflective skills. DA allows students to detect their weaknesses so that they can overcome their challenges, receiving support from their peers and their lecturer until they realize their potential.

In this project, the researcher combines ideas from ZPD, which is about the individual student's potential development, and mediation, which offers one or more chances for further development. The term mediation refers to a process that human beings use to regulate the material world, others' or their own social and mental activity by employing 'culturally constructed artifacts, concepts and activities' (Lantolf & Thorne, 2006, p. 79). Therefore, taking into consideration the Vygotskian SCT point of view, any human task (i.e., higher mental operation) is mediated by objects (e.g., laptops), psychological tools (e.g., text) or another human being (e.g., lecturer, peers) (Wertsch, 2007). In this project, mediation refers to the deliberate and reciprocal interaction between a lecturer/student (and/or written texts/digital texts) and the learner in relation to the challenges students face and the developmental support provided by their peers and lecturers, taking into consideration their ZPDs. Therefore, mediation in terms of assessment enables the lecturer to work on a given assessment task more closely with the student and also allow other students to exchange ideas and solutions to problems they face in terms of academic writing, thereby enabling the lecturer to move them to the next level of their ZPDs with the help of their peers who may be able to more readily detect some of the challenges they face and provide valuable advice using language they understand better.

The literature indicates that very few studies – mainly in the USA – have explored the impact of mediation on student learning mainly in a face-to-face context, exploring its influence on students' speaking and listening skills in a modern foreign language (e.g., Ableeva & Lantolf, 2011; Antón, 2009; Poehner, 2005). The current study tried to address this gap in the literature.

Method

Fifty first-year Management students were asked to participate in this project as part of their module. They were taught various management theories and were then asked to prepare and submit their draft written assignment. The aim was to help students improve their work by raising their self-awareness, scaffolding, and responding to their individual needs (Poehner, 2018). Since students have been complaining about their inability to improve their writing performance and their low motivation to engage in writing, the researcher decided to adopt a process approach to writing which involved three rounds of mediation.

- Round 1 (implicit) consisted solely of a scored and highlighted rubric, not identifying the location or nature of the erroneous parts, and were provided by randomly chosen peers via Padlet to ensure anonymity and allow students to experiment with digital platforms. All students worked in groups of five and used an iPad at this first stage of the intervention. The researcher supervised the whole procedure and made amendments if necessary. She offered training and continuous support to students involved in the project.
- Round 2 (relatively explicit) consisted of narrative explanations of problems provided at the end of each participant's report provided by peers and the lecturer via Padlet. Students again worked in groups of five and used an iPad to offer their feedback. The lecturer supervised the whole procedure closely and intervened only when necessary.
- Round 3 (most explicit) consisted of comment bubbles/specific comments that showed each student the location of the most significant problems, explained the issues, and included recommendations for repair. The tutor provided this form of feedback.

Students provided anonymous feedback via Mentimeter twice during the implementation to explore their perceptions of the benefits and challenges related to this intervention. Students were also asked to write a short report before the implementation.

The researcher compared students' marks in the pre-test and post-test reports (final assignments). Descriptive statistics were used for the analysis of quantitative data and thematic analysis was used for the analysis of the qualitative data.

The aim of this project was to explore the impact of DA on undergraduate Management students to address students' complaints about their low performance and lack of motivation as literature indicates that it can improve students' academic skills and attitudes towards learning.

Summary of findings and discussion

The quantitative findings of this study indicated that students increased their writing performance by 35% in one academic year. The qualitative findings of this study confirm that students were very positive about the DA scheme. At first, students were reluctant to devote the extra time and effort to participate in the implementation and were disappointed by the fact that they could not see drastic improvements right from the beginning. With careful guidance, they built their confidence in academic writing in a stress-free environment with the support of their peers and their lecturer. They had many opportunities to reflect on and improve their work and receive more feedback than when involved in traditional forms of assessment. They developed valuable professional skills i.e., negotiation, metacognitive skills, collaboration as they indicated in their feedback via Mentimeter.

As Daniels (2007) points out, it therefore seems crucial to recognise this affective aspect in order to obtain a complete picture of any pedagogic practice, including assessment procedures. The learners' feedback indicates that DA adds a new dimension to assessment which may turn it into an enjoyable and rewarding experience and at the same time assist them to improve their academic (writing) performance considerably.

DA gradually developed students' self-regulation skills as they used self-assessment actively to plan what they had to do in order to improve their texts, emphasising enjoying the process rather than being intimidated by the end product (also in Fox & Riconscente, 2008; Nicol & Macfarlane-Dick, 2006). The lecturer also confirmed a dramatic change in most students' attitudes as they felt more responsible in their attempt to make sense of their shortcomings and plan their actions in order to resolve any problems they faced in terms of their assignment, offering and receiving help and guidance from their peers and lecturer. The dual feedback they received seemed to be complementary and valuable in order to improve their performance and motivation to engage in writing tasks.

Implications and conclusion

According to students' feedback, the current project enhances the creation of learning communities among students, promoting tolerance and enhancing student collaboration. It explored the beneficial impact of DA practices on: student outcomes, overall experience, and continuous professional development catering for all learners' needs; fostering inclusion of the increasingly diverse student cohorts due to globalisation; developing students' digital, academic, and professional skills; and innovation in tertiary education.

As traditional assessment methods were unable to sufficiently support these Management students, DA's focus on interactive and reflective learning and development helped students receive the kind of individual support they needed, urging them to develop their metacognitive skills to be able to improve their writing skills based on dynamic, tailored and on-going assessment feedback provided by their peers and their lecturer. With traditional assessment methods, this ongoing interaction and focus on process writing is not possible; this DA scheme offered students at least three opportunities to improve their writing performance and develop their self-assessment and reflective skills.

As DA is an intensive form of intervention, the researcher decided to use a combination of group DA and lecturer feedback to support students (Poehner, 2009). However, there is a need for purely experimental studies which will compare DA with non-DA students' academic writing development. It would also be interesting to explore the impact of DA on students in other fields and also conduct longitudinal studies to explore its long-term impact.

In the meantime, however, while recognizing our study is specific to a particular sociocultural context in HE and therefore the findings cannot be generalised, our study suggests that focused tutor mediation (in the form of wikis and exchanges) is an effective way of providing the kind of reflective, dynamic mediation that is able to effectively support students' academic writing development in a distance learning context.

References

Ableeva, R., & Lantolf, J. P. (2011). Mediated dialogue and the microgenesis of second language listening comprehension. *Assessment in Education: Principles, Policy and Practice, 18*(2), 133–149.

Antón, M. (2009). Dynamic assessment of advanced second language learners. *Foreign Language Annals, 42*(3), 576–598.

Becher, T. (1994). The significance of disciplinary differences. *Studies in Higher Education, 19* (2), 151–161.

Daniels, H. (2007). Pedagogy. In H. Daniels, M. Cole, & J. V. Wertsch (Eds.), *The Cambridge Companion to Vygotsky* (pp. 307–331). Cambridge: Cambridge University Press.

El Said, G. R. (2021). How did the COVID-19 pandemic affect higher education learning experience? An empirical investigation of learners' academic performance at a university in a developing country. *Advances in Human-Computer Interaction*, *2021*, 1–10.

Fox, E., & Riconscente, M. (2008). Metacognition and self-regulation in James, Piaget, and Vygotsky. *Educational Psychology Review, 20*(4), 373–389.

Huot, B. (2002). *(Re)Articulating writing assessment: Assessment for teaching and learning*. Utah: Utah State University Press.

Ivanic, R., & Lea, M. R. (2006). New contexts, new challenges: The teaching of writing in UK higher education. In L. Ganobcsik-Williams (Ed.), *Teaching academic writing in UK higher education* (pp. 6–15). Basingstoke: Palgrave Macmillan.

Lantolf, J. P., & Thorne, S. L. (2006). *Sociocultural theory and the genesis of second language development*. Oxford: Oxford University Press.

Lidz, C. S., & Gindis, B. (2003). Dynamic assessment of the evolving cognitive functions in children. In C. S. Lidz, B. Gindis, A. Kozulin, V. S. Ageyev & S. M. Miller (Eds.), *Vygotsky's educational theory in cultural context* (pp. 99–116). Cambridge: Cambridge University Press.

Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education, 31*(2), 199–218.

Poehner, M. E. (2005). *Dynamic assessment of oral proficiency among advanced L2 learners of French* [Unpublished PhD]. Pennsylvania State University.

Poehner, M. E. (2007). Beyond the test: L2 dynamic assessment and the transcendence of mediated learning. *Modern Language Journal, 91*(3), 323–340.

Poehner, M. E. (2009). Group dynamic assessment: Mediation for the L2 classroom. *TESOL Quarterly*, *43*(3), 471–491.

Poehner, M. E. (2018). Probing and provoking L2 development: The object of mediation in dynamic assessment and mediated development. In J. P. Lantolf, M. E. Poehner and M. Swain (Eds.), *The Routledge Handbook of Sociocultural Theory and Second Language Development* (pp. 249–265). New York: Routledge.

Sadler, D. R. (1998). Formative assessment: Revisiting the territory. *Assessment in Education Principles, Policy and Practice, 5*(1), 77–84.

Stern, L. A., & Solomon, A. (2006). Effective faculty feedback: The road less traveled. *Assessing Writing, 11*(1), 22–41.

Walker, M. (2009). An investigation into written comments on assignments: Do students find them usable?. *Assessment and Evaluation in Higher Education, 34*(1), 67–78.

Wertsch, J. V. (2007). Mediation. In H. Daniels, M. Cole & J. V. Wertsch (Eds.), *The Cambridge Companion to Vygotsky* (pp. 178–192). Cambridge: Cambridge University Press.

Vygotsky, L. S. (1978). *Mind in Society: The development of higher psychological processes*. Cambridge: Harvard University Press.

Diversity and Inclusion in Language Assessment

Assessing receptive skills development in deaf children who use Swiss German Sign Language as their primary language

Tobias Haug
University of Teacher Education in Special Needs (HfH), Zurich

Franz Holzknecht
University of Teacher Education in Special Needs (HfH), Zurich

Regula Perrollaz
University of Teacher Education in Special Needs (HfH), Zurich

Abstract

In the German-speaking part of Switzerland, there is a large demand for instruments to assess the development of Swiss German Sign Language (*Deutschschweizerische Gebärdensprache*, DSGS) in deaf children in the (pre)school context. So far, no operational test for DSGS is available. In response to this need, we developed and piloted two DSGS receptive skills tests targeting deaf children between 4 to 14 years old (N = 37, M_{age} = 9.00). One test is designed to assess the comprehension of morphological constructions (N_{items} = 46) and the other test targets narrative comprehension (N_{items} = 17). Preliminary statistical analysis revealed good results to inform the revision of the tests. The tests are currently being revised and used in a new project that will collect additional data over a period of five years to validate the tests with deaf children using DSGS as their primary language.

Brief background to sign languages

Sign languages are fully-fledged languages used primarily by deaf people. They are distinct from their surrounding spoken majority language, for example spoken Swiss or High German in the case of Swiss German Sign Language (*Deutschschweizerische Gebärdensprache*, DSGS). Each sign language has their own grammar and lexicon – they are not international – and the same linguistic categories are used to describe sign languages as in spoken languages, such as phonology, morphology, and syntax. In Switzerland, three different sign languages are used: Swiss French Sign Language, Swiss Italian Sign Language, and DSGS. There is also evidence of different dialects in DSGS (Boyes Braem, 2001).

An important feature of any sign language is the distinction between *manual* and *non-manual* components. Manual components are produced with the hands and arms; non-manual components are produced with the mouth, the face (e.g., with cheeks, eyes, eyebrows, etc.), the head, and the upper torso (Boyes Braem, 1995). For example, eye gaze can be used to re-establish reference in signing space, or raised eyebrows can be used to differentiate between a declarative and an interrogative sentence (Pfau & Quer, 2010). Another important aspect of sign languages is the use of *signing space*, i.e., the physical space in front of the signer's body, which serves various purposes (Johnston & Schembri, 2007), such as introducing and maintaining reference. For example, with the first mention of a referent (e.g., a person or an object) signers can use an index finger to locate the referent at a specific point in space. With gaze or index finger at this same locus the signer can then establish pronominal reference at a later point (Boyes Braem, 1995). The signing space is also important in representing how an object (e.g., a car) moves from A to B.

Context of the study

For most sign languages, research on their structure and acquisition is sparse. This lack of studies makes it challenging to develop or adapt instruments that assess sign language development in deaf children (e.g., see Haug, 2011a). In addition, the scarcity of research results in a lack of resources, such as sign frequency lists or learner corpora that could inform test item development. This becomes even more pronounced in small countries with more than one sign language, such as Switzerland where three sign languages are used.

One of the sign languages in Switzerland is DSGS – the primary language of approximately 5,500 deaf people in the German-speaking part of the country (Boyes Braem, Haug, & Shores, 2012). The need for DSGS tests that could be used in the (pre) school context is large. As a response to this need, several tests were developed between 2014 and 2015 as part of the EU funded Framework 7 project 'SignMET: Sign Language Methodologies and Evaluation Tools'. The project partners were universities in Italy, France, German Switzerland, and Catalonia. The goals of the project were to develop sign language tests in each partner country/region. Each partner institution developed the same version of the tests in their respective sign languages, i.e., Italian, French, Catalan, and Swiss German Sign Language. Four different tests were developed within the SignMET project, targeting deaf signing children aged 4 to 11 years. Two of these tests will be the focus of this paper: the DSGS Receptive Skills Test and the DSGS Narrative Comprehension Test.

Methodology

Test instruments

The DSGS Receptive Skills Test is an adaptation of the British Sign Language (BSL) Receptive Skills Test (Herman, Holmes, & Woll, 1999). The BSL test has already been adapted into several other sign languages, among others also into German Sign Language (Deutsche Gebärdensprache, DGS; Haug, 2011b). The DGS version served as the basis for the adaptation of the test into DSGS. The test targets deaf signing children aged 4 to 11 years and assesses receptive morphological skills, for example, the comprehension of negation, plural formation, spatial constructions, and classifiers. Classifiers refer to a grammatical category of DSGS which describe (1) how animate and inanimate objects are positioned in relation to each other or how they move around in space, (2) how sizes and shapes of objects are realized, and (3) how objects are used or manipulated by the hands. The test consists of 46 items which use three- and four-option multiple-choice format with one correct response. All responses are child-friendly colored drawings. The test is presented via a PowerPoint presentation on a laptop and takes about 30 minutes to complete.

In the DSGS Narrative Comprehension Test, the children first watch a four-minute signed narrative, before the narrative is repeated in three shorter sections. Each section is followed by a set of three-option multiple-choice items with one correct response (with a total of 17 items). All response options are signed DSGS videos. The test assesses aspects of morphology (e.g., plural, classifiers), syntax (e.g., locations of referents, actions of referents), vocabulary, discourse strategies, and narrative structure. Like the Receptive Skills Test, the Narrative Comprehension Test is presented via a PowerPoint presentation on a laptop and takes about 30 minutes to complete.

Participants

To pilot the two tests, we administered them to 37 deaf signing children aged 4 to 14 (M_{age} = 9.00). Children were recruited from three different schools of the Deaf in the German-speaking part of Switzerland.

Analysis

To investigate how well the two tests performed statistically, we calculated descriptive statistics and conducted an item analysis using Classical Test Theory (CTT). We decided not to use Item Response Theory (IRT) as the sample size was relatively small (37 students). Kremmel, Eberharter and Holzknecht (2022) recommend using CTT only if the participants are truly representative of the target population of the test, which was the case as all participants were deaf signing students at schools of the Deaf. For the CTT analysis, we combined all items in one analysis to reach more statistical power, as both tests were measuring DSGS reception and thus the same overarching construct.

After running the CTT analysis, we correlated students' raw scores with their age. The tests are designed in a way that older students should perform better on them than younger students, as older students have generally had more exposure to DSGS. Low or negative correlation coefficients between students' raw scores and their age would indicate potential flaws in the items, so it was important to study this.

Results

Table 1 displays descriptive statistics and Cronbach's alpha across all items of the two tests. Overall, students performed relatively well on the tests, with a mean score of 40.49 out of 63 points (SD = 12.96). Reliability of the two tests was very high (Cronbach's α = 0.94, see also Pallant, 2007).

In terms of individual items, we flagged an item as flawed when it matched both of the following criteria: a discrimination index (corrected item-total correlation, CITC) of 0.25 or less (see Henning, 1987) and an increase in the overall reliability of the test if the item were deleted (Cronbach's alpha if item deleted). As shown in Table 2, 6 out of 46 items (for the DSGS Receptive Skills Test) and 1 out of 17 items (for the DSGS Narrative Comprehension Test) matched these criteria and thus did not perform well statistically.

Table 1: Descriptive statistics and Cronbach's α

N test-takers	37
N items	63
Min	1
Max	56
M	40.49
SD	12.96
Cronbach's α	0.94

Table 2: Number of flawed items for each test

	N items	
	Receptive skills	*Narrative comprehension*
Total	46	17
Flawed*	6	1

*CITC < 0.25 and α >0.94 if item deleted

Finally, Table 3 shows the correlation coefficients between students' raw scores and their age, separately for the two tests. We ran the correlational analysis twice: once for all items and once without the statistically flawed items (see above). As can be seen in the table, correlation coefficients ranged from 0.56 to 0.74 and were thus generally quite high. Without the flawed items, correlation coefficients were higher, indicating that these items should either be dropped or revised. Scores on the Narrative Comprehension Test correlated more strongly with students' age than scores on the Receptive Skills Test (0.74 vs. 0.61 without the flawed items).

Table 3: Correlation between scores and age

	*Correlation coefficients**	
	All items	*Without flawed items*
Total scores – age	0.65	0.68
Receptive skills scores – age	0.56	0.61
Narrative comprehension scores – age	0.74	0.74

*Pearson correlation, p <0.001

Discussion and conclusion

Overall, both DSGS tests show good (preliminary) statistical properties. The CTT analysis revealed only seven flawed items (out of a total of 63) that are currently being revised in an ongoing project. In addition, both tests are 'age sensitive', i.e., older children outperformed younger children, which was important as the tests can then also be used to study language development in deaf children (e.g., Haug, 2011b).

In an ongoing project the tests will be used in schools where they will be administered to a larger group of deaf students. The data will be made available to us and will enable us to validate the tests more comprehensively, using more sophisticated statistical methods, ideally also including additional data such as test-takers' perceptions.

References

Boyes Braem, P. (1995). *Einführung in die Gebärdensprache und ihre Erforschung* [*Volume 11*]. Signum-Verlag.

Boyes Braem, P. (2001). A multimedia bilingual database for the lexicon of Swiss German Sign Language. *Sign Language & Linguistics*, *4*(1–2), 133–143.

Boyes Braem, P., Haug, T., & Shores, P. (2012). Gebärdenspracharbeit in der Schweiz: Rückblick und Ausblick. *Das Zeichen*, *90*, 58–74.

Haug, T. (2011a). Methodological and theoretical issues in the adaptation of sign language tests: An example from the adaptation of a test to German Sign Language. *Language Testing*, *29*(2), 181–201.

Haug, T. (2011b). *Adaptation and Evaluation of a German Sign Language Test: A Computer-based Receptive Skills Test for Deaf Children Ages 4–8 Years Old*. Hamburg: Hamburg University Press.

Henning, G. (1987). *A Guide to Language Testing: Development, Evaluation, Research*. London: Longman ELT.

Herman, R., Holmes, S., & Woll, B. (1999). *Assessing BSL Development: Receptive Skills Test*. Coleford: Forest Books.

Johnston, T., & Schembri, A. (2007). *Australian Sign Language: An Introduction to Sign Language Linguistics.* Cambridge: Cambridge University Press.

Kremmel, B., Eberharter, K., & Holzknecht, F. (2022). Pre-operational testing. In G. Fulcher & L. Harding (Eds.), *The Routledge Handbook of Language Testing* (Secon edition) (pp. 415–429). New York: Routledge.

Pallant, J. (2007). *SPSS survival manual* (Third edition). New York: Open University Press.

Pfau, R., & Quer, J. (2010). Nonmanuals: Their grammatical and prosodic roles. In D. Brentari (Ed.), *Sign Languages* (pp. 381–403). Cambridge: Cambridge University Press.

Acknowledgements

The project was co-funded by the EU Framework 7 project 'SignMET: Sign Language Methodologies and Evaluation Tools' (Education, Audiovisual & Culture Executive Agency: 543264-LLP1-2013-1-IT-KA2-KA2MP).

We would like to thank all the schools, the teachers, the parents, and especially the children for participating in this study.

Investigating potential bias in testing migrants' language proficiency in Switzerland

Hrisztalina Hrisztova-Gotthardt
Secretariat fide,[1] Berne, Switzerland

Gábor Szabó
University of Pécs, Hungary

Abstract

The *fide test* is aimed to assess the ability of migrants to master everyday communication in Switzerland. The communicative tasks, which candidates must accomplish during the test, reflect common real-life situations in which migrants and Swiss residents interact. Since most migrants living in Switzerland are, most probably, familiar with these communicative situations, it is assumed that the test reflects the same construct for all test-takers and does not advantage or disadvantage any individuals or test-taker groups.

To check this hypothesis, the results of the French and German *fide tests* completed between March and July 2022 are statistically analysed and checked for potential systematic differences across candidate groups. The analysis is performed with the goal of examining whether – with regard to the overall test results – there is any difference (i) between male and female test-takers and (ii) across test-taker groups with different first languages.

Introduction

Similarly to many other European countries (cf. ALTE, 2016, p. 11), Switzerland has introduced formal linguistic requirements for the purposes of migration such as first entry, residency, and citizenship. In 2018, language proficiency requirements became part of the Swiss national naturalization law. As defined in *Art. 12 Criteria for integration* of the *Federal Act on Swiss Citizenship*, '[s]uccessful integration is demonstrated in particular by: [. . .] being able to communicate in a national language in everyday situations, orally and in writing [. . .]'. Additionally, *Art.4 Integration* of the *Federal Act on Foreign Nationals and Integration*, the revised version of which has been in force since 1 January 2019, states that '[f]oreign nationals are required to familiarise themselves with the social conditions and way of life in Switzerland and in particular to learn a national language'.

In accordance with the above legal directives, the State Secretariat for Migration (SEM) has specified the minimum language proficiency requirements for different groups of migrants and types of residence permits, and has captured them in a so-called phase model (see Figure 1).

In this context, an obvious need emerged for a reliable tool to assess the migrants' communicative competence in three of Switzerland's official languages, namely French, Italian, and German. In response to this need, the *fide*[2] *test* was commissioned by the State Secretariat for Migration.[3] The test was designed particularly to access the language ability of migrants to master everyday communication in Switzerland. Its specific features include, among others:

- *Accessibility*: the test is intended to be low threshold, i.e., accessible for a very diverse group of test-takers, including those with a low literacy level.

[1] The acronym 'fide' in 'Secretariat fide' stands for 'Français, Italiano, Deutsch in der Schweiz'.

[2] *fide* stands for 'français', 'italiano', and 'deutsch', the languages migrants in Switzerland need to learn depending on the region where they establish themselves. In 2022, a test in Switzerland's fourth national language, Rhaeto-Romance, was introduced. This test is, however, not structurally identical to the *fide test*.

[3] The *fide test* was designed and developed within the framework of the *fide project*. The *fide project* was initialised by SEM to support the linguistic integration of migrants in Switzerland. It aimed to create both well-structured resources for institutions and schools providing language courses for migrants, and a sound framework for assessing migrants' communicative competence in the three languages in question.

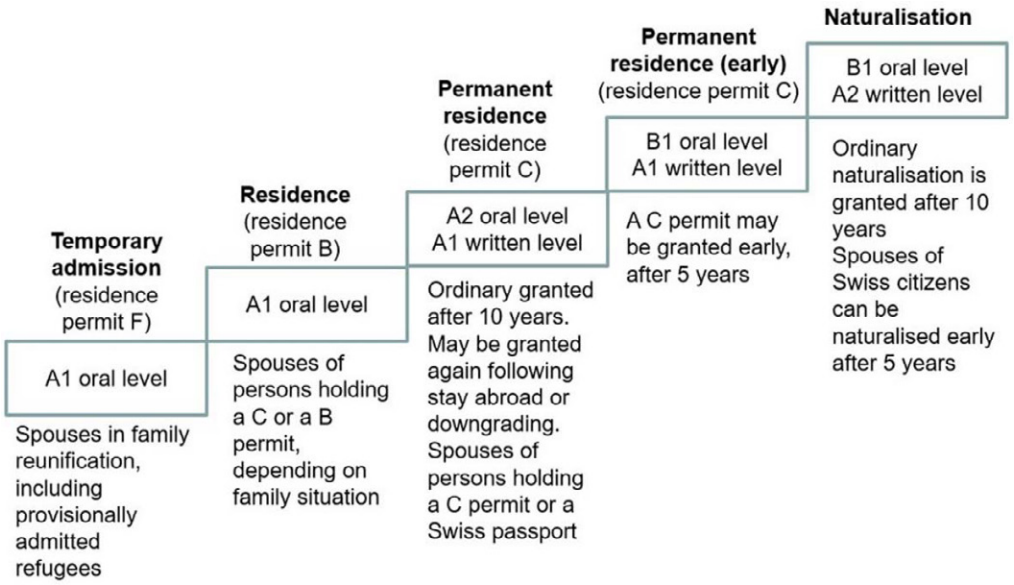


Figure 1 Minimum language requirements in Switzerland: A phase model

(Source: https://www.sem.admin.ch/sem/en/home/integration-einbuergerung/mein-beitrag/zugewandert/sprache.html)

- *Separate assessment of oral and written skills*: test-takers do not need to demonstrate any written skills to pass the oral part of the test.
- *Adaptivity*: based on their performance in the first speaking task, test-takers are advised to complete the oral part of the exam at A1 to A2 or at A2 to B1 levels. As to the written part, test-takers are directed towards the A1 to A2 or A2 to B1 ranges in a separate short pre-test.
- *Closeness to everyday communicative tasks*: The test tasks cover everyday life contexts and contact situations that were identified as most crucial in a needs analysis.

Prior to designing and developing the *fide test*, a needs analysis[4] was carried out by a working group under the supervision of the Institute for Multilingualism at the University of Fribourg. Between 2010 and 2012, interviews were conducted with representatives of different stakeholder groups, e.g., cantonal administration dealing with migration, employment services, social services, professors and teachers, employers, etc. with the aim of identifying frequent contact situations and related communicative challenges between them and migrants. Besides, interviews were held with migrants with the aim of identifying their language learning and communicative needs.

Based on the results of the needs analysis, a so-called scenario database was created. The *fide* scenario database contains over 120 real-life scenarios. Each scenario corresponds to series of verbal and non-verbal actions involving general knowledge and competences that are designed to lead to successfully carrying out the activity in question (see Council of Europe: *Linguistic Integration of Adult Migrants (LIAM)*; Müller & Wertenschlag, 2013, p. 28; Piccardo & North, 2019, pp. 140–142, 258–259; Schleiss & Hagenow-Caprez, 2017, p. 170). Each scenario belongs to one of 11 'domains' (*Handlungsfelder*)[5], i.e., key areas of everyday life and includes series of concrete 'action steps' (*Handlungsschritte*). In the *fide test*, each action step corresponds to a communicative task which must be completed by the test-takers, for example:

Key area: *Transport*

Scenario: *Travelling by train*

Step: *Asking the Swiss Rail employee about train connections to a specific destination*

Task: *Get detailed information about train connections to the desired destination*

---

[4] For more detailed information see: Lenz, Andrey and Lindt-Bangerter (2009, p. 32); Müller and Wertenschlag (2013, p. 28); Schleiss and Hagenow-Caprez (2017, p. 170); SEM (2022, p. 14).

[5] The 11 key areas of everyday life are as follows: *Living*, *Children*, *Work*, *Job-seeking*, *Authorities*, *Media and leisure*, *Transport*, *Shopping*, *Post, banks and insurance*, *Health*, and *Life-long learning*. They correspond with the main key areas defined by *Outline Curriculum* developed in 2007 by Goethe Institute and the German Federal Office for Migration and Refugees (2017).

As the scenarios are not linked to specific CEFR levels, for each task there are descriptors at Levels A1, A2 and B1. The descriptors for the different levels define how active (or reactive) the test-takers act while completing the given communicative task and what linguistic resources they use.

Aim of the present study

Since the vast majority of migrants living in Switzerland are, most probably, familiar with the everyday communicative situations on which the *fide test* tasks are based, it is assumed that the test reflects the same construct for all test-takers and does not advantage or disadvantage any individuals or test-taker groups, i.e., the test is not biased, and is fair in terms of gender and first language diversity (cf. Kunnan, 2004, p. 37).

Accordingly, the aim of the study presented in this paper is to check this hypothesis by answering the following research question: With regard to the overall test results for the oral and written part, is there any difference (i) between male and female test-takers and (ii) across test-taker groups with different first languages?

Methodology

Instrument

fide tests in German and French were considered for this study. The tests being examined were administrated between March and July 2022 in the German and French speaking parts of Switzerland.

Data collected

For the purposes of the present study, the following data were collected, anonymized, and analyzed:

- overall test results for the oral and written part
- test-takers' gender (male or female)
- test-takers first language.

The number of first languages totalled 148. The number of test-takers varied widely from language to language, and was mostly very low, i.e., less than 100 test-takers per language, which would have called into question the reliability of the statistical calculations. It was, therefore, decided to consider languages within the language family or group to which they belong, rather than separately.

Methods of analysis

Since the data included raw-score based percentage figures, which were on an ordinal rather than on an interval scale, and because the distribution of these figures did not approximate normal distribution, nonparametric statistical tests were used. In case of gender, the Mann–Whitney U test was used. The Mann–Whitney U test is the nonparametric alternative to the t-test and compares two population samples to determine if there is a significant difference between them. In case of first or native language, the Kruskal-Wallis H test was applied, which is the nonparametric equivalent of the ANOVA. The Kruskal-Wallis H test helps check whether there are significant differences between two or more groups of an independent variable that do not have a normal distribution (see *Differences between Parametric Test vs. Nonparametric Test*).

Results

A total of eight analyses were performed. The following bar charts summarize the results of these analyses. The bars stand for mean percentage figures; N shows number of test-takers in the analysis. The results will be briefly interpreted in the following section.

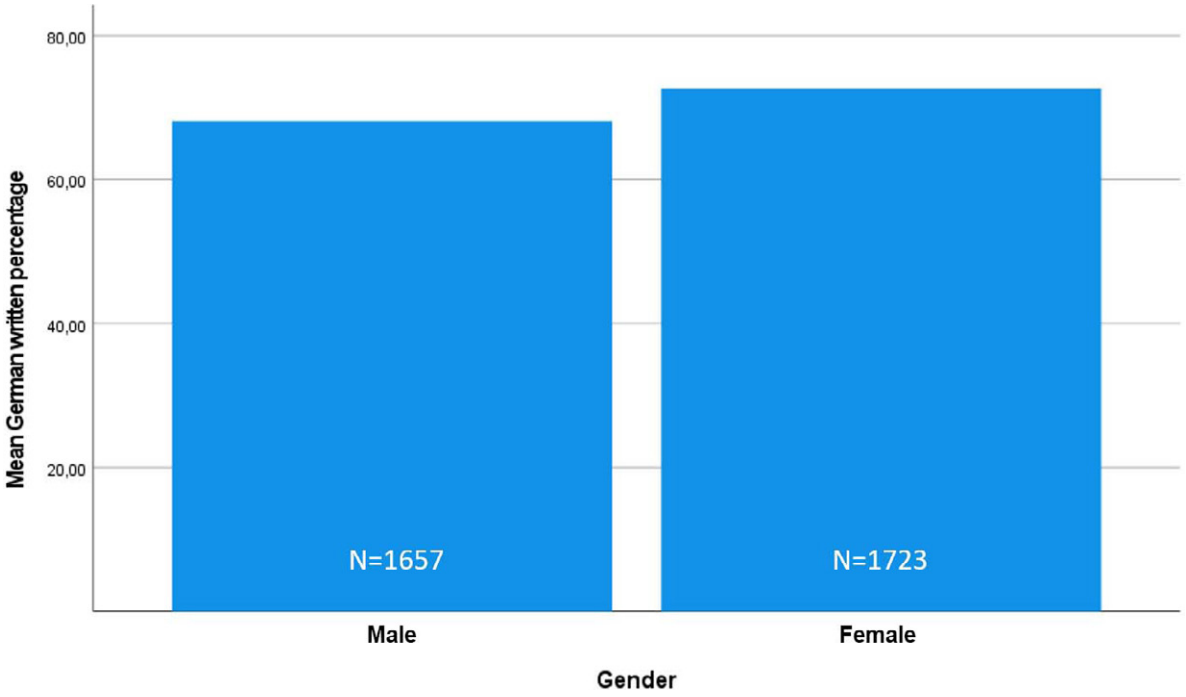


Figure 2 Gender-related overall results for the written part: German language

The difference is significant at p <0.001.

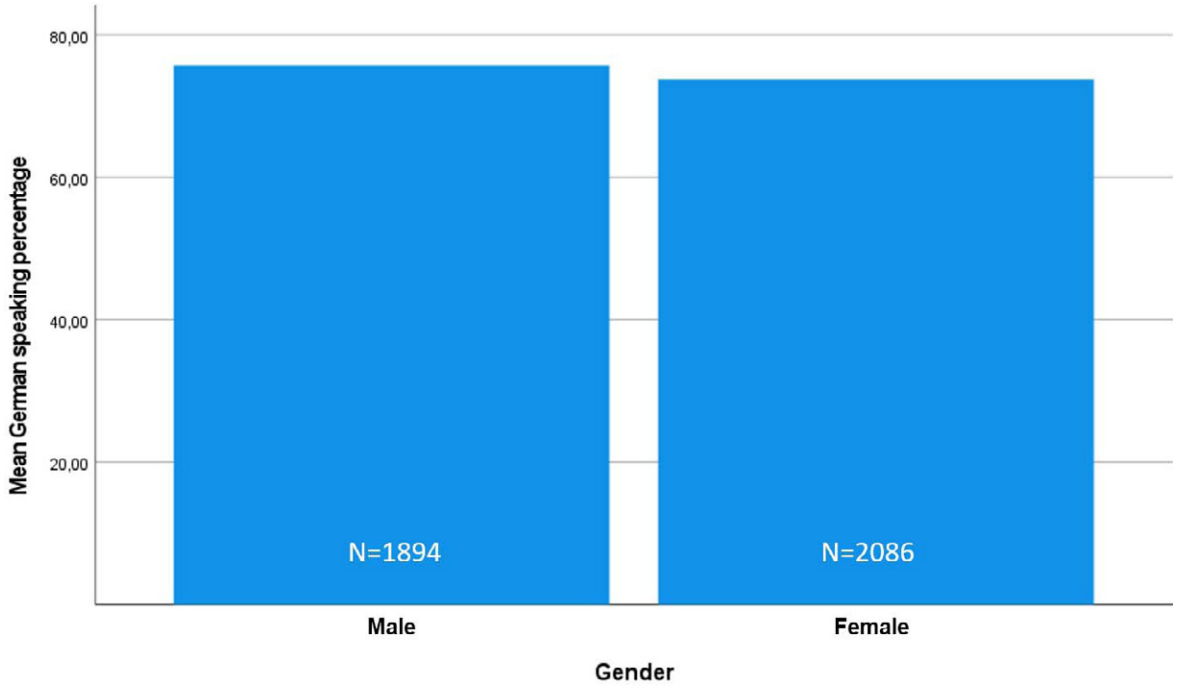


Figure 3 Gender-related overall results for the oral part: German language

The difference is significant at p = 0.007.

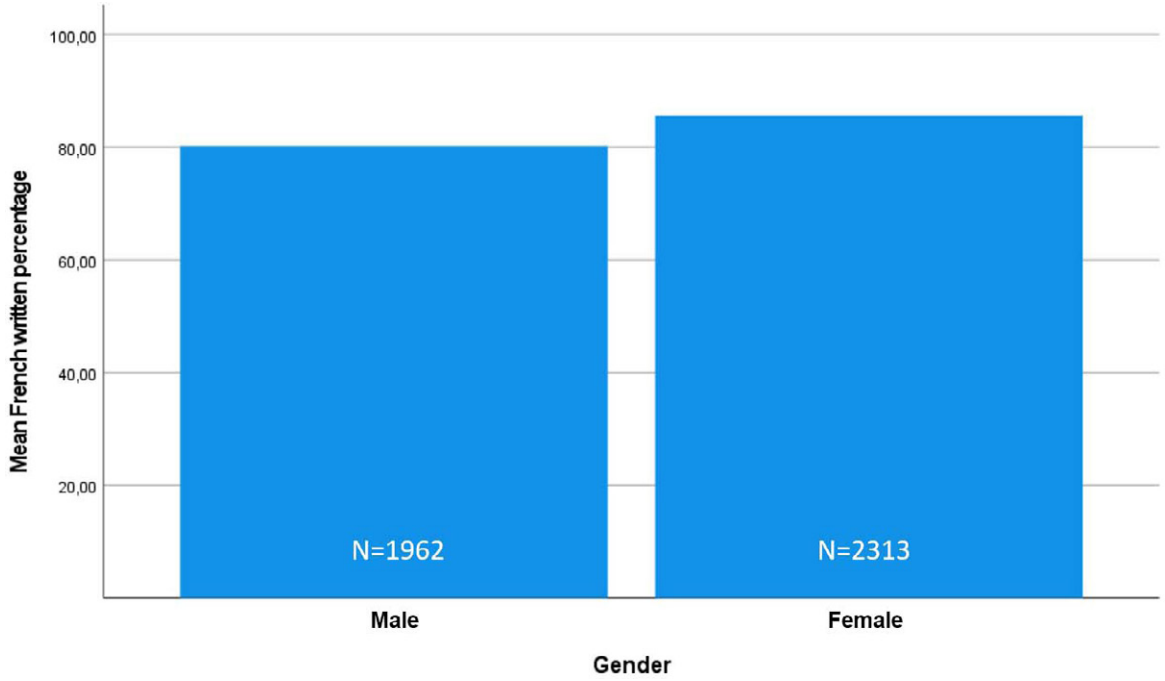


Figure 4 Gender-related overall results for the written part: French language

The difference is significant at p <0.001.

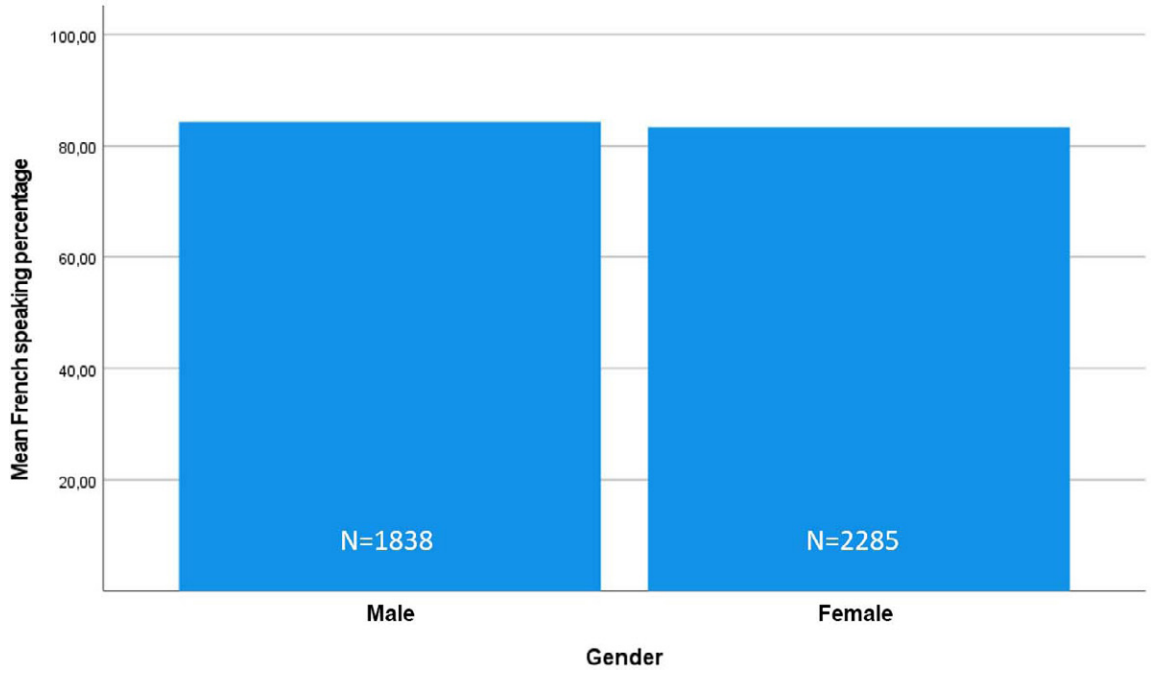


Figure 5 Gender-related overall results for the oral part: French language

The difference is not significant.

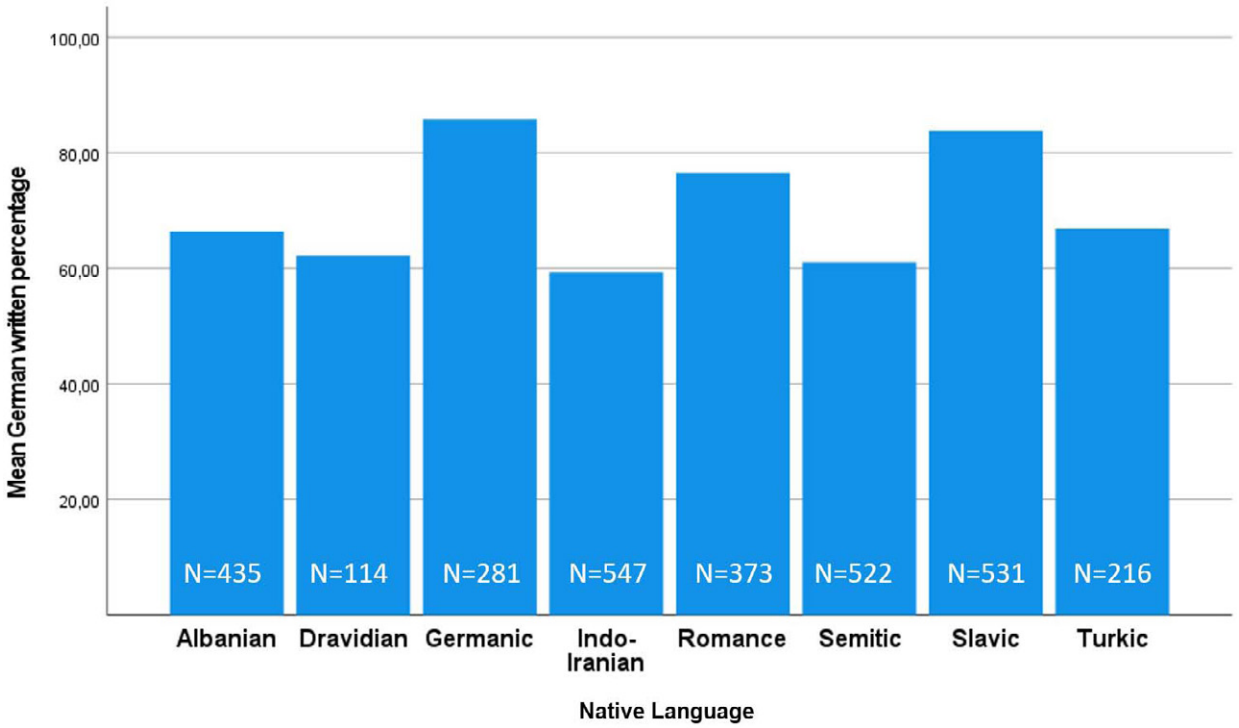


Figure 6 Language-related overall results for the written part: German language

The difference is significant at p <0.001. However, not all language groups' results are significantly different.

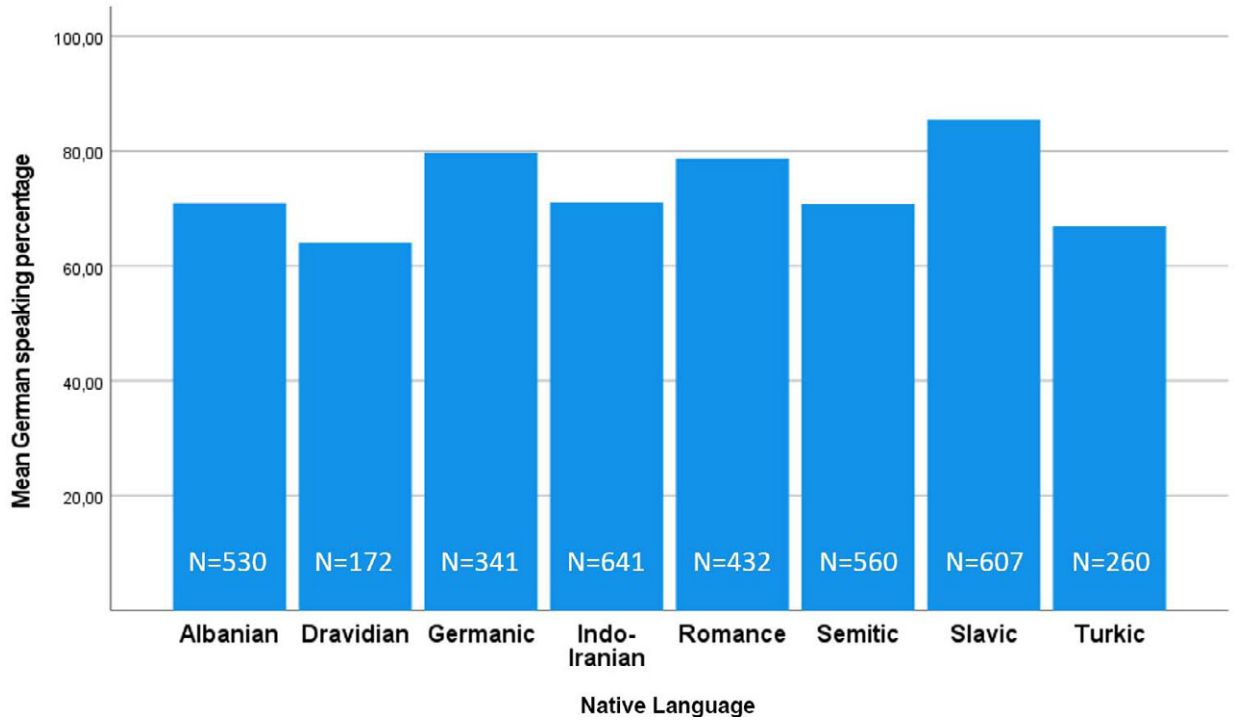


Figure 7 Language-related overall results for the oral part: German language

The difference is significant at p <0.001. However, not all language groups' results are significantly different.

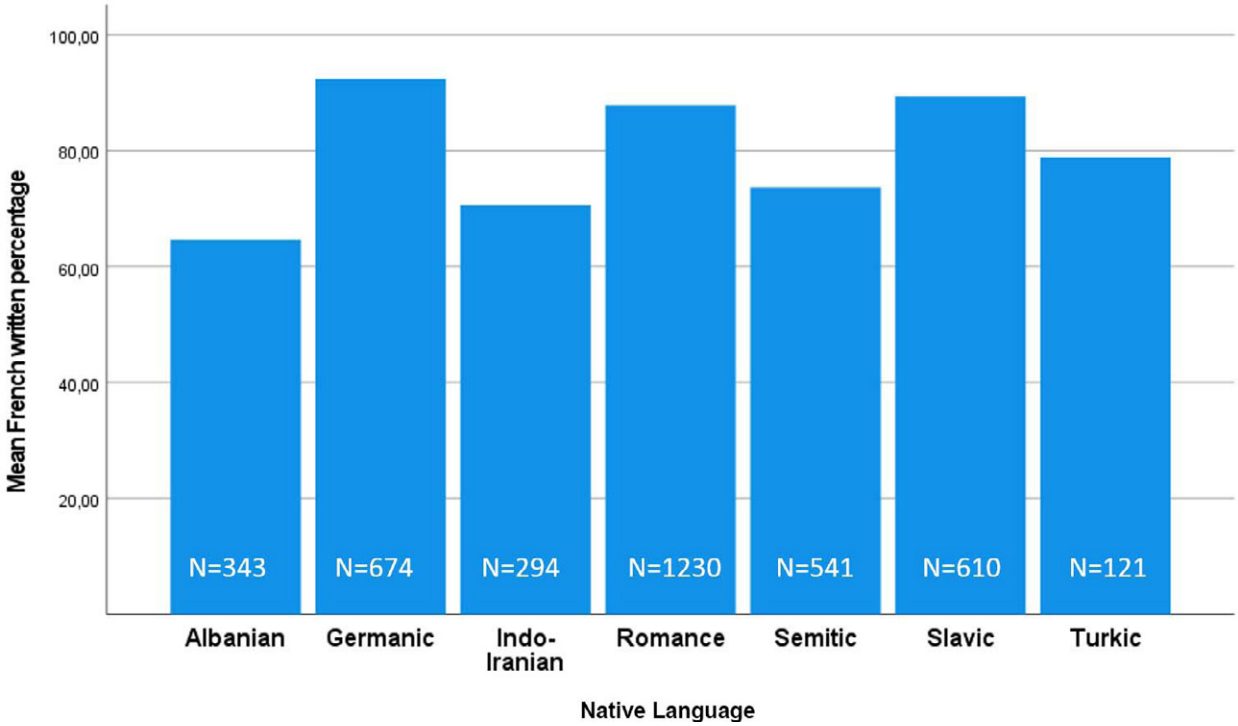


Figure 8 Language-related overall results for the written part: French language

The difference is significant at p <0.001. However, not all language groups' results are significantly different.

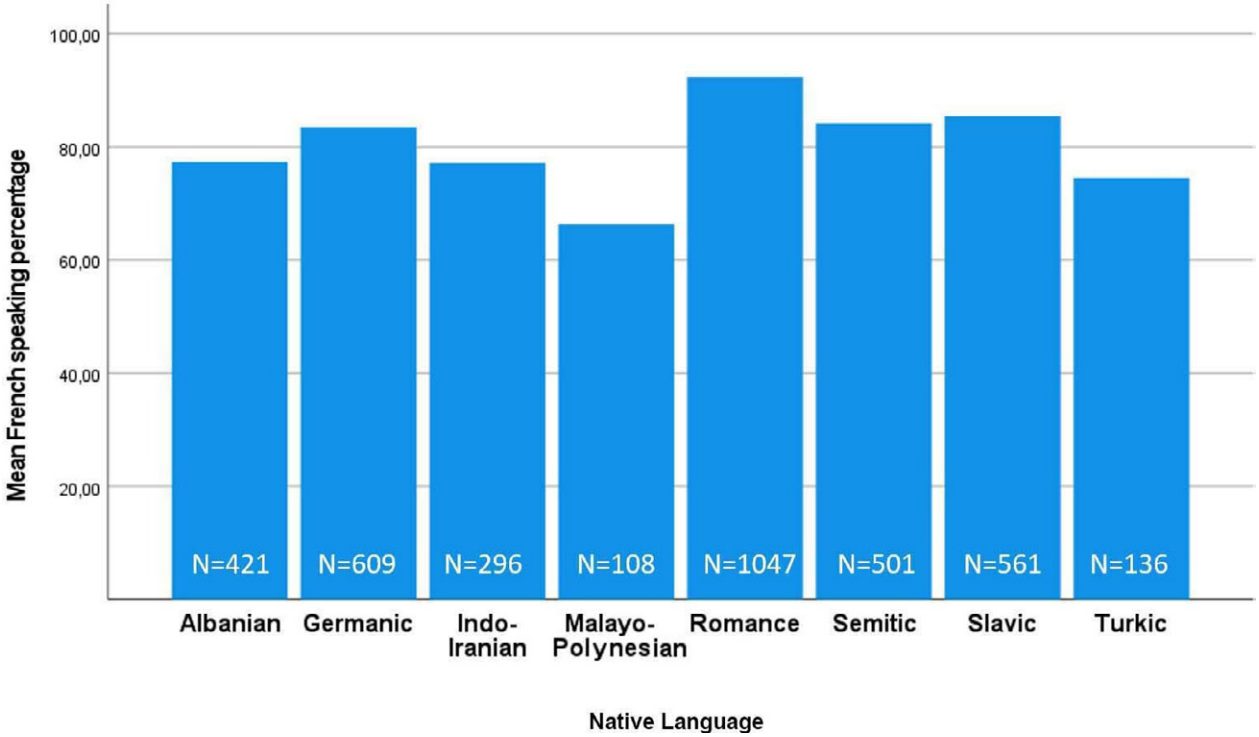


Figure 9 Language-related overall results for the oral part: French language

The difference is significant at p <0.001. However, not all language groups' results are significantly different.

Discussion

This section interprets the results of the quantitative analyses carried out in the course of the study.

As the charts suggest, and as was confirmed by the analyses, results tend to differ significantly both along the lines of gender and native language.

Potential explanations for these results may be:

- The tests may be biased.
  - Although this is a possible explanation for the results, it would need further evidence, as bias seems unlikely owing to the fact that the differences are not always systematic. Bias, however, would, by definition, need to be systematic (cf. Bachman, 2004, p. 149).
- Language-based differences may stem from language groups' degree of relatedness to the target language. Clearly, when a test-taker's native language is more closely related to a target language than some other test-takers', a task may well be easier to perform. Although this may provide an advantage to some test-takers over others, this cannot be considered a facet of the test; rather, it would need to be identified as a candidate-related variable that influences test performance.
- Language-based differences may stem from the degree of socio-cultural congruence. As Ryan and Bachman (1992, p. 22) pointed out, 'native language [..] is most likely a surrogate for a complex of cultural, social, and educational differences'. Thus, candidates for whom certain types of social interaction are more familiar, owing to cultural congruence, may find tasks easier to perform.
- Gender-based differences are largely in line with research into male/female success in second language assessment (cf. Arabski, 1999; Ellis, 1994; Willingham & Cole, 1997). Accordingly, the differences detected may not stem from the tests themselves, but from the generally observable differences in male and female language learners' performances.
- A combination of variables interacting (gender, age, ethnic origin, cultural and language background) may have produced the differences (cf. Ellis, 1994). In other words, it is also possible that, instead of a single variable being responsible for differences, it is the combined effect of some or all of the variables above that produced the different patterns of responses.

Conclusion and outlook

In summary, it can be stated that the differences may have resulted from the effect of several different variables. To check this assumption, further analysis of the data is needed, for instance DIF (differential item functioning) analysis or subgroup analysis (e.g., male/female results within language groups). Additionally, further data need to be collected and analysed on the same exam components. And finally, candidate feedback needs to be collected and analysed qualitatively. Thus, it can be stated that further research is needed in order to determine whether *fide tests* are sufficiently unbiased.

References

ALTE . (2016). *Languages tests for access, integration and citizenship: An outline for policy makers*. Cambridge: ALTE. Available online: https://www.alte.org/resources/Documents/LAMI%20Booklet%20EN.pdf

Arabski, J. (1999). Gender differences in language learning strategy use (A pilot study). In B. Missler & U. Multhaup (Eds.), *The Construction of knowledge, learner autonomy and related issues in foreign language learning. Essays in honour of Dieter Wolff* (pp. 79–90). Tübingen: Stauffenburg.

Bachman, L. F. (2004). *Statistical Analyses for Language Assessment*. Cambridge: Cambridge University Press.

Council of Europe, Language Policy Unit. *Linguistic Integration of Adult Migrants (LIAM)*. Available online: https://www.coe.int/en/web/language-policy/adult-migrants

Differences between Parametric Test vs. Nonparametric Test. https://ca.indeed.com/career-advice/career-development/parametric-test-vs-nonparametric-test

Ellis, R. (1994). *The study of second language acquisition*. Oxford: Oxford University Press.

Federal Act on Foreign Nationals and Integration. Available online: https://www.fedlex.admin.ch/eli/cc/2007/758/en#art_4

Federal Act on Swiss Citizenship. Available online: https://www.fedlex.admin.ch/eli/cc/2016/404/en#art_12

Goethe Institut, & Bundesamt für Migration und Flüchtlinge. (2017). *Rahmencurriculum für Integrationskurse Deutsch als Zweitsprache*. München: Goethe Institut.

Kunnan, A. J. (2004). Test fairness. In M. Milanovic & C. Weir (Eds.), *European Language Testing in a Global Context: Proceedings of the ALTE Barcelona Conference 2001* (pp. 27–48). Studies in Language Testing Volume 18. Cambridge: UCLES/Cambridge University Press.

Lenz, P., Andrey, S. & Lindt-Bangerter, B. (2009). *Rahmencurriculum für die sprachliche Förderung von Migrantinnen und Migranten*. Bern: Bundesamt für Migration BFM.

Müller, M., & Wertenschlag, L. (2013). "Meine Kinder möchten, dass ich auch zum Elternabend gehe". Anmerkungen zum Szenarienansatz und zur Entstehungsgeschichte der fide-Szenarien. *Babylonia, 01*(13), 28–34.

Piccardo, E., & North, B. (2019). *The Action-Oriented Approach: A Dynamic Vision of Language Education*. Bristol: Multilingual Matters.

Ryan, K. E., & Bachman, L. F. (1992). Differential item functioning on two tests of EFL proficiency. *Language Testing, 9*, 12–29.

Schleiss, M., & Hagenow-Caprez, M. (2017). fide – On the way to a coherent framework. In J. Beacco, H. Krumm, D. Little & P. Thalgott (Eds.), *The Linguistic Integration of Adult Migrants* [*L'intégration linguistique des migrants adultes*] (pp. 169–174). Berlin: De Gruyter.

SEM. (Staatssekretariat für Migration). (2022). *Testhandbuch*. Bern: SEM. Available online: https://fide-info.ch/doc/3151/fide_Test_Handbuch_D_Juli_2023.pdf

Willingham, W., & Cole, N. (1997). *Gender and Fair Assessment*. Mahwah: Lawrence Erlbaum Associates, Inc.

The test as an opportunity for less widely tested languages: The case of Romanian

Dina Vîlcu
Babeș-Bolyai University, Romania

Abstract

Language certification has proven to contribute to a larger use of the tested languages and this becomes crucial in the effort of preserving language diversity. In this context, migrants' decision to take a language test in their own language can contribute to its preservation.

This advantage is exemplified here with the case of Romanian Diaspora. Most of the candidates who took the examinations of Romanian as a Foreign Language in Madrid (through the Babeș-Bolyai University – Romanian Cultural Institute Consortium for Testing Romanian as a Foreign Language) had Romanian as heritage language. The certificate they obtained in their mother tongue proved very useful for improving their job prospects in Spain. This case can represent a model for authorities valuing linguistic competence in any language, not only the most used ones, and also for migrants who can be given a very concrete, practical motivation for continuing to use and to cultivate their mother tongue.

Introduction

According to a document elaborated by the UNESCO Ad Hoc Expert Group on Endangered Languages, at least 50% of the world's more than 6,000 languages are losing speakers (UNESCO, 2003, p. 2). It is estimated that 'in most world regions, about 90% of the languages may be replaced by dominant languages by the end of the 21st century' (UNESCO, 2003, p. 2). The vitality profile indicates that almost half of the world languages belong to the levels endangered and extinct (www.ethnologue.com). The risk of language loss has been reported by linguists and addressed by political decision-makers for decades, but still ample and concerted actions are necessary for preserving the world's language diversity, currently under severe threat.

This contribution looks at the loss of speakers which affects Romanian and it presents the test of Romanian as a Foreign Language as an instrument which can contribute to the preservation and promotion of the language. The test will be presented in relation to the target population of the Romanian migrants and will focus on the Diaspora in Spain and the context of labour market.

Romanian culture and language in the context of migration

Economic migration

The phenomenon of economic migration started after the Revolution in 1989 and became dramatic after 2001, when the Romanians could travel without visas and could work in countries of the European Union, and reached a new peak in 2007, when our country joined the European Union (www.insse.ro). On the 1 January 2022 there were more than three million Romanians living in other countries of the EU (ec.europa.eu/eurostat/databrowser/view/migr_pop9ctz/default/table?lang=en). Adding migrants in other parts of the world raises the number to 5.7 million Romanians officially recorded as living abroad (www.mae.ro). Economic migration does not seem to have come to an end; between 100,000 and 200,000 Romanians still leave the country every year. At the end of 2021 more than 3% of the country's population were planning to leave the country (www.rethinkromania.ro), and at the beginning of 2023 more than half (55.1%) of the children in Romania said that they wanted to leave the country and live abroad (www.salvaticopiii.ro/sci-ro/files/89/89904f0b-c151-4b7e-bb34-0ea5ded461d9.pdf).

The cultural dimension

Linguists and sociologists have been studying the changes this phenomenon brings to the migrants' culture and mother tongue. These changes are intrinsically linked to the first generation of migrants' wish to integrate in the host country at various levels (professional, social, cultural, etc.) and to offer their children access to performant educational systems, with good prospects for the local labour market. Many members of the second generation of migrants have done their studies (almost) entirely in the host country and built their social life as part of the local young generation circles. Consequently, while a large part of the first generation of migrants kept their cultural identity and maintained close ties with the country of origin, the second generation developed a 'transcultural identity', through the ability of combining in a creative way elements from the culture in the origin country and in the host country (Arieșan, 2020, p. 198). Many parts of the world have seen, in different periods of time and for different reasons, massive movements of migration. Studies show changes in the original culture patterns of either or both groups, identified as phenomena specific to acculturation (Sam, & Berry, 2010, p. 472). A possible outcome of this process is the one of assimilation: the complete integration in the culture of the host country.

The linguistic dimension

Changes from one generation to the next appear also with reference to language. The first generation of migrants usually keeps the mother tongue, sometimes with linguistic interferences. With the second generation the language of the host country dominates, while the process of language displacement (mother tongue replaced by the language of the host country) is brought to completion by the third generation (see the case of the transitional communities of immigrants in the USA in the early twentieth century – Fishman, 1971).

The displacement of mother tongue from one generation to the next is a process which might occur or vary according to numerous factors. External factors (military, economic, religious, cultural, educational) and internal factors, manifested through the community's negative attitude towards its own language might determine the halt of intergenerational transmission of linguistic and cultural traditions (UNESCO, 2003, p. 2). Occurring often in the context of colonialism (and usually not overcome in the postcolonialism and neocolonialism), this phenomenon can also affect the migrant groups. The tendency is observed, for example, in the communities of Romanian migrants in Spain and Italy. In part, 'the parents (the first generation of immigrants) prefer to abandon their mother tongue and to speak to their children in the language of the host country, in order to avoid possible inconvenience' (Grigoriu, 2014, p. 12). At the same time, the majority of the migrants continue to use Romanian as the language of communication in the family (86% of the migrants in Italy – Grigoriu, 2014).

The process of language attrition seems to be in progress among the representatives of the second generation of migrants. They become part of the educational system in the host countries, diminish the contacts with the country of origin, and integrate in the local community; most of them choose to remain in the country of adoption for higher education and/or for work, and some of them start families in the adoptive country. Consequently, the second-generation migrants use their mother tongue almost only in the family and sometimes only with the older generations (parents and grandparents), the siblings speaking in the language of the host country or code switching very often. When it comes to the second/third generation of migrants, the danger of language loss grows discouragingly. An example in this sense comes from Venice, Italy, where staff from the Romanian Cultural Institute organized an event for the children of the migrant Romanians. A lot of families showed up and brought their kids, most of them at the age of kindergarten. While everyone around them spoke Romanian and all the children came from Romanian migrant families, all of them spoke with each other in Italian.

The test of Romanian as a foreign language: The data

Babeș-Bolyai University – Romanian Cultural Institute Consortium for Testing Romanian as a Foreign Language was founded in 2016 and started to administer examinations for all the levels of the Common European Framework of Reference for Languages (CEFR, Council of Europe, 2001) (A1 to C2). Initially limited to the students in the preparatory year from Babeș-Bolyai University, starting with 2019 the examinations were offered to any candidates who needed to certify their competence in Romanian. Sessions of examination are organized at the Faculty of Letters in Cluj-Napoca, and in headquarters of the Romanian Cultural Institute in Bucharest, Madrid and Venice, with the largest number of candidates so far in Madrid. The exam administration was stopped almost completely during the COVID-19 pandemic. This aspect, together with the novelty of the test and the fact that Romanian is generally a less widely tested language, resulted in a relatively small number of candidates. However, we are confident that by maintaining the quality and validity of our examinations, a purpose to which we are deeply committed, and by promoting them, we will attract a larger candidature in the future. The evolutions in the last year are quite promising.

We collected variables and built a candidate profile in order to adapt our examinations as closely as possible to the needs of our test-takers. A surprising outcome of this process was the fact that 84% of the candidates who took our examination in Madrid

until now were of Romanian origin, with Romanian as a heritage language. We investigated further how a certificate in Romanian as a foreign language can help candidates with Romanian origin.

The test as an opportunity

Language certification is a valuable instrument for the promotion of multilingualism. The context of migration illustrated above with the case of Romanian proves that the language test can also contribute to the preservation of the mother tongue for the migrants, especially for the ones in the second generation (and hopefully in the ones which will follow).

After interviews with candidates in this group, we discovered that most of them used the certificate in Romanian as a foreign language for improving their job prospects for the labour market in Spain. The fact that competence in foreign languages is an asset for a large number of jobs has been long proven (Carvalho, Olim, & Campanella, 2021; Liwiński, 2018; Piri, 2002). It is true that many times certain languages, of very large international use, are preferred and even specifically required for certain jobs or workplaces. From this point of view, the employment policy within the state of Spain can be considered an example of good practice. The certificate of linguistic competence in any language will bring the applicant a significant number of points and improve their prospects for obtaining a job with the Spanish state. An example comes from the official bulletin for the region of Castilla and León in December 2022. For any certificate proving competence in a foreign language the candidate obtains 0.2000 points for temporary jobs and 0.5000 points for permanent employment (https://www.educa.jcyl.es/es/resumenbocyl/orden-edu-1866-2022-19-diciembre-convocan-procedimiento-sel). These adverts are for the domain of teaching, but it is similar for employment in libraries, archives, local administration, positions in universities or research, etc.

The Spanish example can be considered as good practice from the point of view of language diversity preservation for various reasons: 1) competence in all languages, not only the largely used ones is appreciated; this is a crucial point for speakers of languages like Romanian, because it can counterbalance the opinion, perpetuated within their own community, that abandoning your mother tongue in favour of the language of the country of adoption will improve study and job prospects; 2) it can serve as an example for various employers, who might value more foreign language mastering by their job candidates or employees; and 3) it can serve as an example for the cooperation between language specialists and decision makers who can really set the tone in matters related to the relationship between language mastering and employment.

There are numerous other domains which are or can be positively impacted by a higher value added to the language mastering and its certification. The one considered here provides a strong argument not only for foreigners to learn Romanian, but also for Romanian migrants to keep and cultivate their language, for sentimental or cultural reasons, but also for very concrete and lucrative motivations.

Conclusions

This paper focused on the contribution the test of Romanian as a foreign language can bring to the preservation and promotion of the language for its own speakers, more exactly the ones who migrated and the following generations, but it is meant as a possible example for other languages, too. The speakers of less widely spoken/taught/tested languages need to be shown that their own language, like any other language, is important and valuable for the preservation of their identity and cultural values, but also for reasons related to studies, economy and labour market.

References

Arieșan, A. (2020). Limbă și identitate în diaspora românească nouă. Spania și Italia. [Language and identity in the new Romanian Diaspora. Spain and Italy]. In C. Braga (general coord.), *Enciclopedia imaginariilor din România*. [*Encyclopedia of Romanian Imaginaries*], (coord. Elena Platon) *II Patrimoniu și imaginar lingvistic* [*Patrimony and linguistic imaginary*] (pp. 191–209). Iași: Polirom.

Carvalho, A., Olim, L., & Campanella, S. (2021). *The Contribution of Foreign Language Learning to Employability* [Conference presentation]. Seminário Internacional: Educação, Territórios e Desenvolvimento Humano, Porto.

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge: Cambridge University Press.

Fishman, J. A. (1971). The sociology of language: an interdisciplinary approach. In J. A. Fishman (Ed.), *Advances in the Sociology of Language*. *Vol. I* (pp. 214–404). The Hague: Mouton.

Grigoriu, A. (2014). *Românii din Italia. Comunicarea interculturală și păstrarea identității naționale*. [*The Romanians in Italy. Intercultural Communication and National Identity Preservation*]. București: Ars Docendi.

Liwiński, J. (2018). *Does it pay off to learn foreign languages? Evidence from Poland*. Available online: https://caseresearch.medium.com/does-it-pay-off-to-learn-foreign-languages-evidence-from-poland-1-2e465a0a301e

Piri, R. (2002). *Teaching and Learning Less Widely Spoken Languages in Other Countries. Guide for the Development of Language Education Policies in Europe. From Linguistic Diversity to Plurilingual Education*. Strasbourg: Council of Europe.

Sam, D. L., & Berry, J. W. (2010). Acculturation: When Individuals and Groups of Different Cultural Backgrounds Meet. *Perspectives on Psychological Science, 5*, 472–481.

UNESCO Ad Hoc Expert Group on Endangered Languages. (2003). *Language Vitality and Endangerment. Document adopted by the International Expert Meeting on UNESCO Programme Safeguarding of Endangered Languages*. Available online: https://unesdoc.unesco.org/ark:/48223/pf0000183699

HABE C1. Aproximación integral al estudio del DIF

Paula Elosua
Universidad del País Vasco/Euskal Herriko Unibertsitatea, España

Iñaki Villoslada
HABE

Enara Azkue
HABE

Resumen

Una aproximación integral al estudio del funcionamiento diferencial del ítem incorpora, a la aplicación de procedimientos de detección, el estudio de sus causas y el análisis de sus consecuencias. En este trabajo, centrado en la evaluación de la competencia comunicativa en euskera, se ejemplifican estas tres etapas. La investigación se centra en la prueba de comprensión lectora de HABE para el nivel C1, que ha sido respondida por 1,943 candidatos residentes en zonas donde se hablan variedades regionales. Se destaca la importancia de adoptar una perspectiva ecológica en el estudio de validación de las puntuaciones en las pruebas de acreditación lingüística.

Introducción

Estudiar el funcionamiento diferencial del ítem (*Differential item functioning*, DIF) como parte del proceso de validación de puntuaciones aporta evidencias sobre la estructura interna de una prueba y garantías de validez a la interpretación de puntuaciones. Un ítem presenta DIF cuando miembros de diferentes grupos definidos en función de variables como edad, género, idioma, educación, cultura . . ., que presentan el mismo nivel de competencia, tienen diferentes probabilidades de dar una respuesta específica. En las últimas décadas, se ha invertido un considerable esfuerzo en desarrollar y perfeccionar procedimientos para identificar el DIF, y como resultado, contamos con métodos de detección sólidamente establecidos (Holland, y Wainer, 1993; Osterlind, y Everson, 2009; Penfield, y Camilli, 2007). De entre ellos, los procedimientos derivados de la teoría de respuesta al ítem (TRI) y Mantel-Haenszel son los empleados con mayor frecuencia en el ámbito de la evaluación de competencias comunicativas (Li, Hunter, & Bialo, 2021).

Aunque los análisis de DIF generalmente se inician, y en muchas ocasiones concluyen, con la identificación de elementos problemáticos, es conveniente llevar a cabo un análisis adicional para comprender su origen. La literatura reconoce que el DIF, siendo un requisito necesario, no es suficiente para que se aprecie sesgo (Elosua, 2006; Zieky, 1993; Zumbo, 1999); entre otras razones, porque el origen del DIF puede residir en los procedimientos estadísticos utilizados en su detección (Hambleton, 2006). Hasta el momento no se ha formulado ninguna teoría explicativa del DIF y la investigación sobre el tema es escasa (Camilli, y Shepard, 1994; Elosua, 2006; Liu et al., 2016; O'Neill, y McPeek, 1993; Zenisky, Hambleton, y Robin, 2003). Aun así, son varias las causas de DIF recogidas en la literatura; entre ellas, el uso de contenidos con alta relevancia cultural, disparidades en los planes de estudios, o divergencias gramaticales entre formas idiomáticas en términos morfosintácticos o semánticos (Allalouf, Hambleton, y Sireci, 1999; Elosua, y López, 2007; Ercikan, 2002; Gierl y, Khaliq, 2001).

En este sentido, en el abordaje del DIF, sería aconsejable explicitar las variables que pueden generar algún tipo de interacción con el contenido del ítem. Por ejemplo, en contextos sociolingüísticos definidos por la existencia de lenguas minorizadas sometidas a procesos de revitalización lingüística y en los que conviven variaciones dialectales junto a la versión estandarizada, sería adecuado definir grupos en función de las variedades dialectales. El estudio del DIF podría explorar la hipótesis de interferencias entre las expresiones lingüísticas o usos idiomáticos específicos que, no siendo relevantes para el dominio de la competencia en la versión estandarizada, podrían tener un impacto negativo o positivo en comunidades lingüísticas particulares.

Además de ello, y dado que el abordaje del DIF forma parte del proceso de validación de las puntuaciones, resulta recomendable extenderlo hasta ese nivel, y evaluar la presencia de un posible efecto acumulativo conocido como funcionamiento diferencial del test (*Differential Test Functioning*, DTF). La presencia de DTF significaría que las puntuaciones están sistemáticamente sesgadas

y son diferentes entre grupos con el mismo nivel de competencia (Wyse, 2013). Se han desarrollado varios índices derivados de la teoría de respuesta al ítem (TRI) para evaluar el DTF (Chalmers, Counsell, & Flora, 2016; Elosua y, Hambleton, 2018; Raju, van der Linden, & Fleer, 1995). Esos indicadores, básicamente, ofrecen estimaciones del área entre las funciones características del test asociadas a cada uno de los grupos de interés. En caso de observar diferencias condicionales, sería aconsejable adoptar medidas que garantizaran la equidad del proceso evaluativo (Drasgow, Nye, Stark, y Chernyshenko, 2018; Elosua y, Hambleton, 2018; Wyse, 2013).

Propósito del estudio. El euskera

Dentro de este marco, está claro que en contextos lingüísticos donde coexisten varias formas lingüísticas de un idioma minorizado (dialectos y forma estandarizada), los usos dialectales podrían ser un factor generador de DIF. Por ejemplo, un término o expresión podría ser desconocido para grupos de hablantes, o la prueba incluso podría introducir involuntariamente formas dialectales específicas que favorecieran a un grupo de hablantes particular.

Esta es la situación del euskera o vasco, una lengua minorizada hablada en el norte de España y el suroeste de Francia. Además de contar con numerosas variantes dialectales, el euskera tiene una forma estandarizada conocida como Euskara Batua. En el ámbito laboral, para acceder a determinados trabajos, los ciudadanos deben acreditar un nivel de competencia en Euskara Batua. Entre las posibilidades para ello, el examen de competencia HABE nivel C1, gestionado por el organismo público HABE (Helduen Alfabetatze eta Berreuskalduntzerako Erakundea; Instituto de Alfabetización y Reeuskaldunización de Adultos) es uno de los más populares. Su objetivo principal es certificar el dominio de usuario experto del idioma vasco. El examen cumple con los principios de buenas prácticas de ALTE (2001) y está vinculado a los niveles del MCER mediante los procedimientos recomendados por el Consejo de Europa (Council of Europe, 2001). Las habilidades receptivas son evaluadas a través de la comprensión de textos escritos y orales, mientras que las habilidades productivas abarcan la expresión escrita y oral. Este examen está diseñado para candidatos mayores de 16 años.

Preguntas de investigación

Dentro del proceso de validación de las puntuaciones de HABE C1, este trabajo plantea las siguientes interrogantes relacionadas con el DIF:

1. ¿Existe funcionamiento diferencial del ítem entre los grupos residentes en zonas donde se hablan variedades regionales?
2. En caso de detectar DIF, ¿se puede afirmar que existe un sesgo en contra de alguno de los grupos?
3. ¿Se puede concluir la presencia de DTF entre los residentes en la zona donde se habla un dialecto regional u otro?

Para dar respuesta a estas tres cuestiones, se propone abordar el estudio del DIF desde una aproximación integral que comprende tres etapas: (1) detección de DIF, (2) análisis de las fuentes del DIF y (3) evaluación de las consecuencias.

La investigación se centra en la prueba de comprensión lectora HABE C1; la prueba consta de tres tareas diferentes, que en total suman 30 ítems dicotómicos. La primera tarea comprende 10 preguntas de opción múltiple relacionadas con la lectura de textos; la segunda tarea implica completar espacios en blanco y consta de 10 ítems de opción múltiple; por último, en la tercera tarea el candidato tiene que elegir sinónimos para 10 palabras de un texto, para cada una de las cuales se ofrecen cuatro opciones de respuesta. El rango de puntuaciones es 0 - 30, y se establece una puntuación de corte de 15 para aprobar el examen.

Método

Fases del estudio

Primera fase: Detección del DIF. Se aplicaron tres procedimientos de detección: a) estadístico de Wald para la comparación de los parámetros de dificultad del modelo de Rasch, b) Mantel-Haenszel, y c) diferencia de medias estandarizada (STD).

Segunda fase: Evaluación de sus causas. Se constituyó un grupo de discusión con el propósito de analizar el contenido de los ítems mediante la técnica de grupo focal. Se reclutó a seis filólogos (3 mujeres y 3 hombres) con experiencia en la construcción de ítems. Los participantes recibieron una guía para el desarrollo de la reunión y para el fomento de un debate sobre el contenido de los ítems.

Tercera fase: Estimación del funcionamiento diferencial del test. Se estimó el DTF dentro del marco de la TRI.

Muestra

La muestra estuvo compuesta por 1,943 candidatos con una media de edad de 23.08 años (DT = 9.19). Se definieron dos grupos de hablantes en función de los dialectos regionales occidental y central: 1,113 residentes en la zona donde se habla el dialecto occidental y 830 residentes de la zona del dialecto central.

Resultados

La diferencia de medias entre los grupos de hablantes no fue estadísticamente significativa ($F(1,1941) = 1.95$; $p = .16$). Las medias aritméticas obtenidas por los residentes de las zonas donde se hablan el dialecto occidental o central fueron, respectivamente, 18.19 (DT = 3.54) y 18.41 (DT = 3.32).

Fase 1: Detección de DIF entre dialectos regionales

Los tres procedimientos de detección detectaron de forma unánime un único ítem (ítem 12) con DIF significativo y severo, que beneficiaba a los candidatos de la zona del dialecto regional central. Otros dos ítems fueron categorizados como DIF severo según al menos uno de los procedimientos de detección empleados (ítems 23 y 24).

Fase 2. Grupo de discusión

El análisis de contenido concluyó que el ítem 12 aportaba varianza irrelevante con relación a la competencia comunicativa en euskera. El resto de los ítems no mostraron sesgo.

Fase 3. Consecuencias

La estimación de la diferencia entre las curvas características del test arrojó un valor no significativo (DTF = –0.199; IC del 95% [–0.6, 0.18]; $p = 0.324$).

Discusión

Aunque la detección de DIF generalmente se centra en aspectos estadísticos, un análisis integral de DIF requiere examinar tanto sus causas como sus consecuencias. El análisis de causas es una práctica poco común en la evaluación lingüística (Elosua, en prensa; Li et al., 2021), y su impacto sobre las decisiones basadas en las puntuaciones apenas ha sido explorada. Para abordar el análisis de causas, el modelo ecológico propuesto por Bronfenbrenner (1979) proporciona un marco de referencia en el estudio de la validez que algunos autores ya han empleado para formular hipótesis explicativas (Solano-Flores y, Elosua, 2021; Solano-Flores y, Milbourn, 2016; Zumbo et al., 2015). Para evaluar las consecuencias de la presencia de DIF, la teoría de respuesta al ítem proporciona indicadores cuantitativos de funcionamiento diferencial (Drasgow et al., 2018; Raju et al., 1995); además de ellos, es posible estudiar el impacto acumulativo del DIF comparando diferentes escenarios construidos por medio de la exclusión/inclusión de elementos con DIF en las decisiones tomadas en base a las puntuaciones observadas (Elosua, en prensa). Si determinados grupos se vieran afectados negativamente, sería conveniente considerar si se deben mantener o eliminar ítems y/o implementar procedimientos de equiparación compensatorios. En este sentido, Elosua y Hambleton (2018) presentan un método para equiparar puntuaciones en presencia de DIF.

En resumen, este estudio presenta un enfoque holístico para el tratamiento del DIF que se construye sobre una hipótesis contextual. Se incorpora una perspectiva ecológica que, en este ejemplo, se construye sobre la coexistencia de dialectos regionales que conviven junto a una versión estandarizada del idioma vasco. El trabajo ofrece herramientas para analizar las causas del DIF a través de una revisión cualitativa de contenido. Si bien es cierto que esta revisión no siempre determinará el origen del DIF, la implementación sistemática de protocolos similares podría ayudar a acumular evidencia y a formular teorías explicativas sobre su origen. Además, se detallan estrategias para explorar los efectos del DIF a nivel del test. En definitiva, se establece un enfoque integrador que favorece el fortalecimiento de la validez de las puntuaciones (Elosua, 2023), y con ello su correcta interpretación.

Referencias

Allalouf, A., Hambleton, R. K., & Sireci, S. G. (1999). Identifying the causes of DIF in translated verbal items. *Journal of Educational Measurement, 36*(3), 185–198.

ALTE. (2001). *Principles of good practice for ALTE examinations*. Available online: https://www2.testdaf.de/fileadmin/Redakteur/PDF/TestDaF/ALTE/ALTE_good_practice.pdf

Bronfenbrenner, U. (1979). *The Ecology of Human Development*. Cambridge: Harvard University Press.

Camilli, G., & Shepard, L. (1994). *Methods for identifying biased test items*. Thousand Oaks: Sage Publications.

Chalmers, R. P., Counsell, A., & Flora, D. B. (2016). It might not make a big DIF: Improved differential test functioning statistics that account for sampling variability. *Educational and Psychological Measurement*, *76*(1), 114–140.

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge: Cambridge University Press.

Drasgow, F., Nye, C. D., Stark, S., & Chernyshenko, O. S. (2018). Differential item and test functioning. In P. Irwing, T. Booth & D. J. Hughes (Eds.), *The Wiley handbook of psychometric testing* (pp. 885–899). Hoboken: Wiley-Blackwell.

Elosua, P. (2006). Funcionamiento diferencial del ítem en la evaluación internacional PISA. Detección y comprensión. [Differential item functioning in the PISA international assessment. Detection and understanding.] *RELIEVE*, *12*(2). Retrieved from https://ojs.uv.es/index.php/RELIEVE/article/view/4229

Elosua, P. (2023). Hizkuntza-komunikaziorako konpetentziak egiaztatzearen artehunak. [The ins and outs of the linguistic competency certification]. *e-hizpide* (101). Available online: https://doi.org/10.54512/VXMJ6717

Elosua, P. (in press). *A three-step DIF analysis of a reading comprehension test across regional dialects to improve test score validity*.

Elosua, P., & Hambleton, R. H. (2018). Psychological and educational test score comparability across language and cultural groups in the presence of item bias. *Journal of Psychology and Education*, *13*(1), 23–32.

Elosua, P., & López, A. (2007). Potential DIF sources in the adaptation of tests. *International Journal of Testing, 7*(1), 39–52.

Ercikan, K. (2002). Disentangling sources of differential item functioning in multilanguage assessments. *International Journal of Testing*, *2*(3–4), 199–215.

Gierl, M. J., & Khaliq, S.N. (2001). Identifying sources of differential item and bundle functioning on translated achievement tests: a confirmatory analysis. *Journal of Educational Measurement, 38*(2), 164–187.

Hambleton, R. K. (2006). Good practices for identifying differential item functioning. *Medical Care*, *44*(11), 182–188.

Holland, P. W., & Wainer, H. (Eds.). (1993). *Differential item functioning.* Mahwah: Lawrence Erlbaum Associates.

Li, H., Hunter, C.V., & Bialo, J. A. (2021). A revisit of Zumbo's third generation DIF: how are we doing in language testing?. *Language Assessment Quarterly, 19*, 27–53.

Liu, Y., Zumbo, B., Gustafson, P., Huang, Y., Kroc, E., & Wu, A. (2016). Investigating causal DIF via propensity score methods. *Practical Assessment, Research, and Evaluation*, *21*, Article 13.

O'Neill, K., & McPeek, W. (1993). Item and test characteristics that are associated with differential item functioning. In P. W. Holland, & H. Wainer (Eds.), *Differential item functioning* (pp. 255–276). Mahwah: Lawrence Erlbaum Associates.

Osterlind, S. J., & Everson, H. T. (2009). *Differential item functioning* (Second edition). Thousand Oaks: Sage Publications.

Penfield, R. D., & Camilli, G. (2007). Differential item functioning and item bias. In C. R. Rao, & S. Sinharay (Eds.), *Handbook of Statistics Volume 26: Psychometrics* (pp. 125–167). Amsterdam: Elsevier.

Raju, N. S., van der Linden, W. J., & Fleer, P. F. (1995). IRT-based internal measures of differential functioning of items and tests. *Applied Psychological Measurement*, *19*(4), 353–368.

Solano-Flores, G., & Elosua, P. (2021). *Measuring and operationalizing national assessment capacity* [Virtual conference presentation]. *XII International Test Commission Conference*.

Solano-Flores, G., & Milbourn, T. (2016). Assessment capacity, cultural validity and consequential validity in PISA. *RELIEVE*, 22(1).

Wyse, A. E. (2013). DIF cancellation in the Rasch model. *Journal of Applied Measurement*, *14*(2), 118–128.

Zenisky, A. L., Hambleton, R. K., & Robin, F. (2003). Detection of differential item functioning in large-scale state assessments: A study evaluating a two-stage approach. *Educational and Psychological Measurement*, *63*(1), 51–64.

Zieky, M. J. (1993). Practical questions in the use of DIF statistics in test development. In P. W. Holland & H. Wainer (Eds.), *Differential Item FECV*. (pp. 337–347). Mahwah: Lawrence Erlbaum Associates.

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF)*. National Defense Headquarters.

Zumbo, B. D., Liu, Y., Wu, A. D., Shear, B. R., Olvera-Astivia, O. L., & Ark, T. K. (2015). A methodology for Zumbo's third generation DIF analyses and the ecology of item responding. *Language Assessment Quarterly*, *12*(1),136–151.

Describing washback: teachers and students' voices in Jaén (Spain)

Victoria Peña Jaenes, PhD
Cambridge University Press & Assessment, United Kingdom

Abstract

The prevalence of English language as a gateway to better opportunities has led to an increase in the demand for English language courses and accreditation exams in Spain, which has put the focus on the impact that exams may have.

This paper explores teachers' views of the courses and the accreditation exams they were involved in or familiar with. Their perceptions were analysed and contextualised with their students' perspectives and discussed in the light of research.

The study gives evidence of a limited washback on courses and a stronger effect on students' preparation. Learners perceived the B2 First exam – produced by Cambridge University Press & Assessment, which is aligned with the B2 level of the CEFR – mainly as a source of motivation and their views on challenges in the exam were similar regardless of the type of courses they attended.

Introduction

Speaking a foreign language has traditionally been considered an advantage. Although in the past it was perceived more as a privilege, speaking English in particular has become a necessity (Chávez-Zambano, Saltos Vivas, & Saltos Dueñas, 2017, p. 761), regardless of where one lives or what their field of expertise is (Jaimechango, 2009; cited by Chávez Zambano et al., 2017, p. 761). The reason for this has to do with the fact that for the last three or four decades the relevance of English has rocketed, becoming the *lingua franca*, and it is connected with economic success (Graddol, 2006). In this context, where being able to communicate in English is crucial, it is as vital to prove one's communication skills. As a result of the increasing use of exam results to access higher education or for employment, accreditation exams have become high stakes.

Rationale and objectives

In the light of the above, there has been growing interest in the impact of exams among researchers and assessment institutions, aware of the fact that exams results are more and more frequently used to influence life-changing decisions for millions of people around the world (Raban, 2008; University of Cambridge Local Examinations Syndicate, 2016). In the literature, the terms impact and washback are used (as well as backwash) to refer to the effect of exams. This contribution understands washback as the influence of tests which 'leads teachers and learners to do things they would not necessarily otherwise do' as per Alderson & Wall's definition (1993, p. 117) and agrees with Buck's (1988, p. 17) view that the tendency to tailor classroom activities to the demands of the test is especially observed when the test is particularly important for test-takers.

Until fairly recently, the research into washback focused mainly on the teachers and was frequently conducted where a new exam had been introduced. Well-known scholars such as Cheng (2008; cited by Pan, 2014), Green (2013; cited by Sevilla Morales & Chaves Fernández, 2020, p. 207), Booth & Lee (2019, p. 19) had identified a gap in the research into learner washback in particular and in specific contexts of test use. Students are the ultimate stakeholders in any assessment and play a key role in their own learning. Moreover, their perspective towards the test is paramount in the presence or absence of washback (Tsagari, 2007, p. 314). The results presented here were obtained in a research project into the washback that a prestigious proficiency exam may have on learners, which was conducted in Jaén (Spain), where these exams are well-established. The study hoped to offer data about learner washback, and more precisely about perceptions, classroom practice and preparation.

Methodology

The data used in this study was collected using semi-structured questionnaires. There were two versions of the student questionnaire – Entry Questionnaire and End-of-Course Questionnaire – and one for the teachers, which they filled in towards the end of the school year. The research was carried out in two institutions. One of them belonged to the University and the second one was a well-known private language school. In terms of students' profiles, the students enrolled in the Higher Education Institution were adults who attended general English courses or exam-oriented courses. Apart from language tuition, the institution offered a number of accreditation exams. In the private language school context, there were students of all ages attending exam-oriented courses to sit a Cambridge English Qualification. In the study, the 131 students taking part were divided into a control group, made up of learners attending general English courses in the higher education context, and into an experimental group, with those attending exam preparation courses in the private language school. The views of the eight professionals teaching these students in both institutions were also gathered.

The first part of the Entry Questionnaire for students included a number of questions which provided data about students' objectives as well as perspectives about the exam and about English, which could have an impact on the quality and intensity of the washback. There was an equivalent section in the teacher questionnaire to learn about teachers' previous experience and degree of familiarity with the exams. Responses showed that the majority (58%) of the students who took part in the study were between 12 and 17 years old and 70% had prior experience sitting an official accreditation exam. In terms of their objectives and perceptions towards English, for 78% of them English was perceived as a gateway to make progress either in Spain or abroad, confirming the prevalence of English described above. 67% of the participants aimed to pass B2 and 62% felt confident they would be able to pass B2 First after 18 months of instruction. In terms of preparation, 87% of the learners identified working on production, interaction and reception as key for success in a B2 exam. A total of 84% of the learners considered it (very) important to pass an accreditation exam and 55% decided what exam to sit on the basis of its prestige and recognition. As for teachers, the professionals taking part were all qualified English teachers who were familiar with Cambridge exams and had experience teaching different levels and courses modalities. The majority (86%) prepared for external examinations, most of them for B2 First (71%). Based on their experience and knowledge of the exam, all the teachers agreed that B2 First was a reliable measure of the B2 level of the Common European Framework of Reference for Languages (CEFR, Council of Europe, 2001).

Results

Washback on classroom practice

The analysis of the results in terms of teachers' perceptions of the washback that the exam had on their lessons showed that professionals saw a difference between exam-oriented and general English courses based on the focus of the former on the exam. In fact, 60% of them expressed that B2 First was the most important factor influencing their lessons. To have a better understanding of the influence of the exam, teachers were asked about the activities carried out in class, which showed that exam practice and format were included by all the teachers in exam-oriented courses and by 50% of the teachers in general English courses. It was perhaps slightly surprising that while all the teachers in exam-oriented courses reported including feedback in their classes none of the teachers in general English courses reported that – perhaps because they did not see it as an activity on its own but rather as part of other activities. Finally, 80% of the teachers in exam-oriented courses reported using authentic material in their lessons compared to 0% in general courses. All the teachers, regardless of the course modality, reported working on exam strategies, and all four skills as well as grammar and vocabulary. Teachers answered one more question about activities, in this case about how often they worked on speaking, writing, listening, reading, grammar and vocabulary, exam format, course book, resource book, exam practice and authentic material. The main differences between course modalities were found in writing, which was more frequently covered in exam-oriented courses, the same as exam practice. It was interesting that the main factor affecting how often exam practice was included in the lessons in exam-oriented courses was the teacher, as each professional reported working on exam practice with different frequencies. Students were also asked about activities; grammar and vocabulary, four skills and exam practice were the top three options for students regardless of the course modality they attended.

Washback on students' preparation and feelings

Another key area of the research was the effect of the exam on students' preparation for a B2 exam. A similar percentage of students in exam-oriented courses (61%) and in general English courses (57%) chose course exams and homework as the main activity they carried out to prepare for the exam. In terms of the differences, the experimental group had a preference for exam-related activities, while the control group focused on studying English in general and working on the grammar and vocabulary they thought may be included in the exam. In terms of the exam effect on learners, teachers reported that taking an official exam encouraged their students to work harder and that B2 First in particular was a source of motivation to put in additional effort

although it could also increase students' stress and workload. When asked about how they felt when they received low results in the mock exams, most students reported that it helped them understand how they could improve (57% of students in general English courses and 56% in exam-oriented courses), the second most frequent option being that it was a motivation to study harder (14% control group, 16% experimental group).

Washback on students and teachers' perceptions towards the exam

Students and teachers were also asked about the challenges students faced in the different exam parts. In the listening paper, teachers identified accent as being the most problematic aspect for their students. On their part, the highest percentage of students in the control group (41%) and in the experimental group (47%) pointed out that what they found most difficult was the speed of speech. The main difficulty from the teachers' perspective in the other receptive skill was time management and using suitable reading subskills while students singled out vocabulary as their main concern (58% experimental group and 75% in the control group). Unknown lexis also appeared as the main problem for students in the Use of English component and was also identified by teachers after grammar, which came second for students. When it came to productive skills, teachers summarised students' concerns when 60% of them explained that students found it difficult to show what they knew. Students in the control group mentioned being able to use B2 language and their peers in the experimental group pointed at overcoming nerves to make the most of their abilities. Language was the main concern for students in writing regardless of the course modality they attended, while teachers identified lack of planning as the main obstacle.

Conclusions

The results showed that exam preparation is present in both general and exam-oriented courses classroom practice, although it is more frequent in the latter. However, results do not suggest there is a narrowing of the curriculum as exam-oriented courses include a wider range of activities than general courses. In terms of students' preparation, washback could be observed, as students in the experimental group tended to focus more on test-related activities, compared to their peers in the control group. Finally, the washback on students' perceptions was limited because although preparing for B2 First was mainly perceived as a source of motivation, the views on the challenges in the different exam papers were similar in the control and the experimental group.

References

Alderson, J. C., & Wall, D. (1993). Does Washback Exist?, *Applied Linguistics, 14*(2), 115–129.

Booth, D. K., & Lee, N. D. (2019). Learner Perceptions and Washback of the Paper-Based TOEFL Test on Student Affect at one Japanese University. *KSAALT-TESOL Academic Journal, 1*(1), 9–23.

Buck, G. (1988). Testing listening comprehension in Japanese university entrance examinations. *JALT Journal, 10*, 12–42.

Chávez-Zambano, M., Saltos-Vivas, M. A. & Saltos-Dueñas, C. M. (2017). La importancia del aprendizaje y conocimiento del idioma inglés en la enseñanza superior. *Dominio de las Ciencias, 3*, 759–771.

Cheng, L. (2008). Washback, impact and consequences. In E. Shohamy & N. H. Hornberger (Eds.), *Encyclopedia of Language and Education. Language Testing and Assessment Volume 7*, (2nd edition) (pp. 349–364). New York: Springer.

Council of Europe. (2001). *Common European Framework of Reference for Languages. Learning, Teaching, Assessment*. Cambridge: Cambridge University Press.

Graddol, D. (2006). *English Next*. London: British Council.

Green, A. (2013). Washback in Language Assessment, Achieving Beneficial Backwash. *International Journal of English Studies, 13*(2), 39–51.

Jaimechango. (2009). *Importancia del inglés en la educación* [PowerPoint slides]. Available online: https://es.slideshare.net/jaimechango/importancia-del-ingles-en-la-educacon

Pan, Y. C. (2014). Learner Washback Variability in Standardized Exit Tests. *Teaching English as a Second or Foreign Language, 18*(2).

Raban, S. (2008). *Examining the World. A History of the University of Cambridge Local Examinations Syndicate*. Cambridge: Cambridge University Press.

Sevilla Morales, H., & Chaves Fernández, L. (2020). Washback Effects of Board-Based Speaking Tests. *Letras*, *68*, 199–238.

Tsagari, D. (2007). *Review of washback in language testing: How has been done? What more needs doing?*. Available online: https://files.eric.ed.gov/fulltext/ED497709.pdf

University of Cambridge Local Examinations Syndicate. (2016). *Principles of Good Practice. Research and Innovation in Language Learning and Assessment*. Cambridge: Cambridge English Language Assessment.

Testing aptitude with the MLAT-EC in young learners: The role of age and beyond

Maria-del-Mar Suárez
Universitat de Barcelona, Spain

Abstract

Young learners' language aptitude is understudied due to a lack of tests covering this period of life. Also, young learners are still acquiring their L1 while developing their literacy skills. Consequently, a language-dependent aptitude test for them should be carefully designed. An additional challenge is found when the testees are bilingual, as is the case of the Catalan/Spanish community in Catalonia. This was the scenario for the adaptation of the Modern Language Aptitude Test –Elementary in Spanish (MLAT-ES – Stansfield, Reed, & Velasco, 2005) into Catalan (MLAT-EC – Suárez, 2010). Despite the linguistic proximity between these languages, several issues had to be considered across grades, such as the use of certain words, the test font type and size as well as the distractors. The resulting MLAT-EC test has proven to tap into the same factors as the original in several contexts and, therefore, is a valid and reliable measure despite the challenges in its adaptation.

Introduction

Aptitude in young learners has been neglected in the literature for its difficulty to measure it. One of the reasons is the fact that young learners are, first, still acquiring both their literacy skills and their L1, and second, they are still developing cognitively. This means that their aptitude is dynamic. Its measurement becomes even more complex when young learners are bilingual from birth, as simultaneous bilingualism implies the development of two linguistic systems which, of course, can complement each other, but they can also cause linguistic interferences and add an extra element of difficulty in one's L1 acquisition process. Consequently, the validity of a language-dependent aptitude test in any of the two languages spoken by bilinguals should not be taken for granted as some apparently optimal items might carry hidden difficulty or flaws when presented to such testees. This is the case of the MLAT-EC (Suárez, 2010), the Catalan version of the MLAT-ES (Stansfield, Reed, & Velasco, 2005). Therefore, our aim is to present the main challenges when adapting the four parts of the MLAT-ES into Catalan, as both linguistic and cognitively-dependent aspects played a role in the functioning of certain items when answered by bilingual Catalan/Spanish young learners of ages 8 to 14.

The MLAT-ES and the MLAT-EC

These aptitude tests for young learners are probably the most widely used tests to measure aptitude in young learners. They both stem from the Modern Language Aptitude Test-Elementary (Carroll & Sapon, 1967), which at the same time stems from the Modern Language Aptitude Test (Carroll & Sapon, 1959). The Elementary versions consist of four parts.

Part 1: Hidden Words (*Paraules ocultes*). Based on the MLAT's Spelling Clues, this test presents easier vocabulary. It is believed to measure L1 vocabulary and sound-symbol association. (30 items)

Part 2: Words in Sentences (*Paraules que es corresponen*). This part measures grammatical sensitivity without using formal grammatical terms. Learners are to find the word in a sentence performing the same function as the capitalized word in the item's stem. (29 items)

Part 3. Rhyming Words (*Paraules que rimen*). This test has no counterpart in the MLAT. It measures the ability to hear speech sounds while selecting words that rhyme. (38 items).

Part 4. Learning Numbers (*Números en un altre idioma*). In this test, test-takers learn six numbers (units and tens) and how to combine them in an artificial language. This part taps into rote memory learning as well as vocabulary learning and the ability to form and remember associations between speech sounds. (25 items)

While the MLAT-ES consists of 123 items, the MLAT-EC, after a careful qualitative and quantitative item analysis, ended up with 122 items. It could be said that both tests are equivalent or exchangeable in Catalan/Spanish bilinguals, as it has been widely

demonstrated that the testees' language preference of one language over the other, or their lack of preference does not interfere with their test performance across grades (Suárez & Stansfield, 2023). Also, both tests present the same patterns across grades, with only anecdotical exceptions when considering the testees' performance and their cognitive development (Suárez & Muñoz, 2011).

This almost perfect equivalence between tests, though, was the product of a careful item creation and analysis of both the MLAT-EC and the MLAT-ES, the latter being validated using a wider, eminently monolingual population.

The following is an account of the main issues encountered when adapting the MLAT-ES into Catalan and the possible rationale behind them considering the testees' age.

The present study

Participants and their bilingual status

A convenience sample of 629 Catalan/Spanish bilingual students between Grades 3 to 7 (age range 8.3–14.9 years) participated in the main validation study (Suárez, 2010). They all took both the MLAT-ES and the MLAT-EC in counterbalanced order to eliminate the test-training effects found. Here, though, only the results of the students who took the MLAT-EC first (N = 304) are analyzed. They were all schooled in immersion schools, where Catalan is meant to be the vehicular language, except for both the Spanish and the English language subjects. They were also asked about their language preference, but as it did not yield significant differences (Suárez & Stansfield, 2023), they are all considered one group despite their bilingual status.

Qualitative decisions in adapting the MLAT-ES into Catalan

Despite the similarity between Spanish and Catalan, both Romance languages, translation between them oftentimes is challenging. In Part 1, *Paraules ocultes*, the cognitive challenges in the MLAT-ES were preferred over a direct translation of items. This implied changes in the meaning of the word of the item stem. One such example is MLAT-ES item 7 <ddo> – *está en la mano*, which refers to *'dedo'* (a finger) became <ungl>, meaning *'ungla'* (a nail), because it was very difficult to play with the spelling of the equivalent Catalan word, *'dit'*. With this change, the distractor *'para jugar'* (to play) was pointless (which would correspond to the Spanish distractor *'dado'* (dice)). The solution was to opt for *'part d'un triangle'* (part of a triangle), as <ungl> could lead test-takers to interpret it as *'angle'* (an angle), which belongs to the semantic field of geometry.

In Part 2, *Paraules que es corresponen,* the items were also literally translated from the Spanish version whenever possible, but some isolated words were deliberately changed as they got the participants confused due to cultural and usage reasons. Outdated proper nouns in the Peninsular variety of Spanish 'Leila' and 'Perla' became 'Laura' and 'Paula', for instance. Also, Catalan uses more words than Spanish in the construction of some verb tenses, possessive adjectives, and proper nouns. This implied choosing where to put the checkboxes for students to choose the 'corresponding word'.

Part 3, *Paraules que rimen*, was the most challenging one. First, while there is almost a one-to-one correspondence between phoneme and grapheme in Spanish, this is not the case of Catalan, where there are up to eight vocalic sounds (standard Spanish has five), and differences in alveolar fricative (voiced vs voiceless <s> and a non-existing letter in Spanish, that is, <ç>) as well as a wider variance in similar consonants as it happens in the bilabial plosive /b/ versus its approximant and labiodental counterparts (/β/ and /v/, respectively).

The items of Part 4, *Números en un altre idioma,* were kept as they were in the MLAT-ES although they were re-recorded following the Catalan phonetic system. This decision was made because the invented numbers in Spanish and Catalan had different distractor targets. For example, *'vein'* resembles *'veinte'* (twenty) in Spanish, but not so much Catalan *'vint'*, which, instead, shares the same starting syllable with *'vinca'*.

Quantitative decisions in adapting the MLAT-ES into Catalan

As in old-school item analysis tradition, all items were inspected for content validity, difficulty, and reliability. The index of the facility was calculated following these criteria: IF >0.74: very easy - <0.25: very difficult. The coefficient of discrimination power of items (Di) ranged from –1 to +1, with positive numbers over 0.2 reliably implying that carried a positive Di. The corrected item-total correlations were considered, like other types of correlations, low if they were <.300 and high if they were >.600. All tests had excellent Cronbach alphas across grades and parts, but individual item performances were nevertheless inspected to see if overall reliability was affected when removing the red-flagged items.

Test speed was an influential factor in Parts 1 to 3 in the lower grades. The percentage of participants who left Parts 1 and 3 unfinished is rather larger than those in Part 2 in the lower grades in both tests. From Grades 4 to 7, most of the students managed to finish the test. Leaving items blank in the lower grades could be attributed to L1 literacy development, but in the upper grades these are probably due to the test answering strategy adopted by the testees, regardless of their age/grade.

Those items presenting issues in either content validity, difficulty and reliability aspects were further examined and several decisions (rewriting or removal) were taken depending on the case (see Suárez, 2010). What is most remarkable is that most issues appeared in both versions of the tests in the same grades, which could mostly be attributed to the testee's cognitive developmental stage at the moment of taking the tests.

Conclusions and lessons learned

Language-dependent aptitude tests like the MLAT-E, MLAT-ES or MLAT-EC might be an easy target for criticism due to their clinginess to L1 and its possible drawbacks. Catalan and Spanish are certainly close Romance languages, but that does not mean that a light translation of the MLAT-ES into Catalan could guarantee a perfect fitting test in the minoritarian language for several reasons. Therefore, we claim that decisions when adapting tests, be them aptitude or performance tests, are taken despite the proximity between the languages at work. There might be cognates that facilitate test adaptations, but non-cognates must also be carefully chosen and examined as they might imply collateral decisions. Indeed, despite the careful item analysis that the MLAT-EC went through, the MLAT-EC is slightly more difficult than the MLAT-ES, though not significantly so, across ages.

Also, when dealing with young learners, especially Grade 3 – and possibly younger ones – it must be taken into account that these are still in the process of acquiring their L1 (or L1s, if bilingual), besides their young cognitive age if compared to pre-adolescents. Therefore, not only aptitude tests but any type of linguistic performance test should consider not only test fatigue, but also attention span and alternative coding systems (pictograms, oral input and output) so as to have wider evidence to measure any trait in young participants.

Finally, and focusing on more local testing issues, while the validity and reliability of both the MLAT-ES and the MLAT-EC have been proven in a Catalan/Spanish population across ages (Suárez, 2010), the MLAT-EC solves some specific linguistic and cultural problematic items in the MLAT-ES, especially for Grade 3. Tests, unfortunately, age, so items should also be revised so they can be adapted to the current times.

References

Carroll, J. B., & Sapon, S. (1959). *Modern Language Aptitude Test (MLAT): Manual.* New York: Psychological Corporation.

Carroll, J. B., & Sapon, S. (1967). *Modern Language Aptitude Test-Elementary.* New York: Psychological Corporation.

Stansfield, C. W., Reed, D. J., & Velasco, A. M. (2005). *Modern Language Aptitude Test – Elementary: Spanish Version: Manual 2005 Edition*. Rockville: Second Language Testing Foundation.

Suárez, M. M. (2010). *Language aptitude in young learners: The Elementary Modern Language Aptitude Test in Spanish and Catalan*. [PhD dissertation]. Universitat de Barcelona.

Suárez, M. M., & Muñoz, C. (2011). Aptitude, age and cognitive development: The MLAT-E in Spanish and Catalan. *EUROSLA Yearbook, 11*, 5–29.

Suárez, M. M., & Stansfield, C. (2023). Young learners' bilingual status and cognitive development in foreign language aptitude testing. *ITL – International Journal of Applied Linguistics, 174* (2), 291–315.

Assessment in the early years: Mapping concepts and practices in four Brazilian states

Juliana Reichert Assunção Tonelli
State University of Londrina, Brazil

Gladys Quevedo-Camargo
University of Brasília, Brazil

Abstract

There has been considerable growth in the offer of English in the curriculum in the early years of elementary school (first to fifth grades) in many Brazilian municipalities. With no official guidelines for this educational phase, each municipality organises its curriculum. A one-year study, supported by the British Council Brazilian office, aimed to map the municipalities in four Brazilian states that offer English in the early years of elementary school and to write guidelines to support the development of public policies for English language teaching in that context. This paper reports the data collected with respect to assessment identified in the curricula of the municipalities. It also briefly presents the guidelines produced in the project and published by the British Council Brazilian office with the aim of supporting the development of public policies for the teaching of English to young learners.

Introduction

All around the world, research in the area of teaching additional languages to children has seen a steady growth (Johnstone, 2019). This scenario is not different in Brazil (Seccato, Tonelli, & Selbach, 2022). Although, in our country, the offer of English is compulsory only when students start their sixth year (which corresponds to the second phase of elementary education), the teaching of this language in kindergarten and pre-primary schooling has become more and more common.

The inclusion of additional languages in the early years of education demands, among other changes, the need to redesign language teacher education courses (Tonelli, Ferreira, & Belo-Cordeiro, 2017) and improve teachers' assessment knowledge (Bueno, 2020; Tonelli, & Quevedo-Camargo, 2019) since assessment is a central part of the teaching process. In addition, several studies highlight the importance of developing knowledge about additional language assessment and the urgency to develop skills to enable well-informed work with assessment (Quevedo-Camargo, 2020).

This paper aims at presenting partial results of a broader project carried out in 2020–2021, supported by the British Council Brazilian office. One of its objectives was to identify the municipalities in four Brazilian states that offer English in the early years of elementary school. Hence, this paper presents: 1) the mapping of the municipalities in the Brazilian states that offer English in the early years of elementary school, have local curricula to guide the teaching of that language, and, mainly, mention assessment types in their local curricula; and 2) an overview of the guidelines, written by the group of teachers, educators and researchers involved in the project and published by the British Council Brazilian office, aiming to support the development of public policies for the teaching of English to young learners in Brazil.

The project and some of its findings

In order to identify the municipalities that offer English, an online questionnaire was sent to 1.369 municipalities located in four Brazilian states: São Paulo, Paraná, Goiás, and Espírito Santo. Basically, the questionnaire focused on whether or not the municipalities offered English in the early years, had specific material to teach young learners, and offered teacher development courses for those teaching young learners.

In São Paulo, out of 645 municipalities (100% – total number), only 155 (24%) answered the questionnaire. Among those, 115 (74%) stated that they offered English in the early years; 63 (54%) of them affirmed that they had specific material to teach English.

In the State of Paraná, there are 399 (100%) municipalities. There were 147 (36%) respondents, out of which 61 (41%) stated offering English in the early years and, out of the latter, 31 (50%) said they had specific material to teach English.

In the State of Goiás there are 247 (100%) municipalities. Respondents to the questionnaire were 124 (50%), out of which 68 (54%) offered the English language in the early years. However, only 19 (27%) of them mentioned they had specific material.

The fourth state was Espírito Santo, with 78 (100%) municipalities. The questionnaire was answered by 62 (79%) of them. Out of this number, 22 (35%) offered the English language, seven of whom (31%) declared they had specific teaching material.

To sum up, the questionnaire was sent to the education secretariats of 1,369 (100%) municipalities, but only 378 (27%) of them replied. Out of the respondents, 266 (70%) stated they offer English in the early years of Elementary school, although only 120 (45%) of them state they have specific material to teach English at this level.

These numbers are important because, although they represent only 25% of the overall number of municipalities in the country[1], they confirm the presence of the English language in Brazilian public classrooms. In addition, the data indicate that 45% of the municipalities have proper material to teach English at that level. However, because there are no official guidelines for the teaching of additional languages in the first years of schooling, each secretariat of education designs their local curriculum, which varies according to the understanding of those in charge of it, which may cause serious disparities in terms of education and enhance social inequalities. Also, there are cases in which no documents are available to guide the pedagogical practices.

When further analyzing the curricula provided by the education secretariats, we carried out a categorization of the information by using a colour code. One of the categories found, which is directly relevant for this paper, was related to assessment. The next section is dedicated to this matter.

Concerning assessment

In some of the curricula provided by the education secretariats of the municipalities, it was possible to identify information about the assessment system, the assessment criteria, and the assessment mode used.

In the State of São Paulo, 99 (100%) municipalities stated they had a curriculum to guide the teaching of English in the early years of regular education. However, only 28 (29%) made their curricula available. Out of these 28 curricula, 12 (42%) mentioned assessment.

In Paraná, 58 (100%) affirmed having a curriculum, and 44 (76%) made them available to the researchers. The number of these available curricula mentioning assessment was 29 (66%).

In Goiás, only two municipalities made their curricula available out of the 44 (100%) municipalities in the state. However, only one of the curricula mentioned assessment.

Finally, in Espírito Santo, out of the 19 (100%) municipalities in the state, 14 (74%) delivered their curricula to the researchers, and only three (21%) of them mentioned assessment.

Summing up, out of the 220 (100%) municipalities of the four states that declared having curricula, 88 (40%) made their curricula available for analysis, and 45 (51%) of them mentioned assessment.

Such data are very concerning, as they suggest that there are not many municipalities in the four states that have curricula to guide teachers in their work with young learners. In addition, the data evidence that the number of curricula that include assessment guidelines is very small.

The guidelines framework

As mentioned, considering that in Brazil there are no official documents regarding the teaching of English in the first years of schooling, the researchers involved in the British Council project, based on their experience as teacher educators and on the data collected in the project, wrote a set of guidelines to support the development of public policies for the teaching of English to young learners in Brazil.

The guidelines, entitled 'Framework for a National Curriculum for the English Language in Primary Education', 1) present a brief history of the teaching of English to young learners in Brazil; 2) justify the importance of teaching English to children; and 3) highlight the need for public policies at that school level.

[1] Brazil has 5.568 municipalities (Belandi, 2023).

In addition, the framework approaches four important principles underlying the teaching of English at early ages, which were specially investigated in order to subsidise the proposal made in the Framework: conceptions of language in the English teaching to children; the role of English in the world; the place of English in the curriculum; and the assessment system (whose data were presented in the previous section).

With the aim of providing some practical information for those who decide to use the guidelines, some discursive genres were also included as suggestions in order to support the work of the municipalities that either do not have any curricula to guide the work with young learners or would like to improve the existing ones.

The Framework was built on the understanding of language as social practice and is based on the following six premises: 1) lucidity; 2) linguistic education through cross-disciplinary projects; 3) multiple literacies; performative means that are part of the child's play universe; 4) interculturality and linguistic sensitivity; 5) development of the whole child and building citizenship; and 6) genre-based communication.

Taking into consideration that few curricula analysed mention how to assess children learning English as an additional language, the assessment system recommended for primary school English language is the formative assessment, which has two distinct approaches: assessment for learning and assessment as learning (Earl, 2013). With the first approach, both the learning process and the student progress are assessed regularly on the basis of the learning objectives set in the teaching plan. The assessment is carried out using a variety of instruments associated with formative assessment: student portfolios, class observation and educational games. With the second approach (assessment as learning), the children are gradually encouraged to reflect on their own learning, to self- and peer-assess, thus experiencing assessment as a way of achieving autonomy, learning about themselves and their colleagues, and of self-regulating.

Summary and conclusion

Based on the fact that there has been considerable growth in the offer of English in the early years of elementary school in many Brazilian municipalities, and that there are no official guidelines for this educational phase, a one-year study, supported by the British Council, was designed to map the municipalities in four Brazilian states that offer English in the early years of elementary school. Based on the data collected from the education secretariats of municipalities in the four states, the project allowed the production of guidelines aimed to support the development of public policies for the teaching of English to children in Brazil. In this article, we reported partial results of that project, as we focused on the data collected with respect to assessment in the curricula of the municipalities. We also briefly presented the guidelines written with the aim of supporting the development of public policies for the teaching of English to young learners.

Overall, the British Council project developed in the four Brazilian states, although representing only one quarter of the municipalities in the country, draws our attention, firstly, to the urgent need to have official guidelines for the work with English in the early years of public schools as it is an undeniable reality; secondly, to the importance of redesigning language teacher education courses so as to prepare teachers to teach English to children; and thirdly, to the fundamental importance of improving the language assessment literacy of different stakeholders, starting with (future) teachers.

References

Belandi, C. (2023). *IBGE atualiza dados geográficos de estados e municípios brasileiros*. Available online: https://encurtador.com.br/BFJM7

British Council. (2022). *Documento-base para a elaboração de diretrizes curriculares nacionais para a língua inglesa nos anos iniciais do ensino fundamental*. Available online: https://www.inglesnasescolas.org/wp-content/uploads/2022/05/Diretrizes_Ingles_Anos-Inciais-Molic-BritishCouncil-2022.pdf

Bueno, B. A. G. (2020). *Chameleon: o jogo de tabuleiro como instrumento de avaliação para a aprendizagem de língua inglesa por crianças* [Final paper - Professional Masters]. Centro de Letras e Ciências Humanas, Universidade Estadual de Londrina.

Earl, L. (2013). *Assessment as Learning: Using Classroom Assessment to Maximize Student Learning* (Second edition). Thousand Oaks: Corwin Press.

Johnstone, R. (2019). Languages Policy and English for Young Learners in early education. In S. Garton & F. Copland (Eds.), *The Routledge Handbook of Teaching English to Young Learners* (pp. 13–29). New York: Routledge.

Quevedo-Camargo, G. (2020). Formação de professores de línguas adicionais e letramento em avaliação: breve panorama e desafios para os cursos de licenciatura em LEM no Brasil. *Calidoscópio*, *18*(2), 435–459.

Seccato, M. G., Tonelli, J. R. A., & Selbach, H. V. (2022). A panorama of the teaching of additional languages to children in Brazil. *Revista Letra Magna*, *18*(29), 34–46.

Tonelli, J. R. A., & Quevedo-Camargo, G. (2019). Saberes necessários ao professor para avaliar a aprendizagem de crianças na sala de aula de línguas estrangeiras. *fólio - Revista De Letras*, *11*(1), 583–607.

Tonelli, J. R. A., Ferreira, O. H. S., & Belo-Cordeiro, A. E. (2017). Remendo novo em vestido velho: uma reflexão sobre os cursos de letras-inglês. *REVELLI - Revista de Educação, Língua e Literatura*, *9*, 124–141.

Embrace the future of minority language testing: Insights from Zhuang Language Proficiency Test in China

Andy Jiahao Liu
International Foundations Writing Program, Department of English, University of Arizona, USA

Abstract

Against the trending English-medium knowledge construction in the language testing field (Weir & Saville, 2016), this paper introduces and highlights the importance of minority language testing in enhancing linguistic diversity. Drawing on the Assessment Use Argument framework (Bachman & Palmer, 2010), it presents and evaluates a somewhat less represented test—the Zhuang Language Proficiency Test—as an example. By doing so, it provides language testers worldwide with welcoming developments in minority language testing. The paper concludes with a call for more studies and discussions on languages other than English and multilingual and multicultural assessment, hence enriching the diversity in language testing.

Introduction: Minority language testing matters

Over the past 60 years or so, various English assessments, as a result of the globalization of English, have emerged and have been well-researched in the field of language testing and assessment. In particular, I have read a great deal about the International English Language Testing System (IELTS), the Test of English as a Foreign Language (TOEFL), the Duolingo English Test (DET), and many others. Languages other than English (LOTE) have been consequently underrepresented in the accessible literature. Or, put it in another way, 'We rarely read about what goes on elsewhere in assessing other languages' (Weir & Saville, 2016, p. ix). The situation since these years has changed to some extent, as researchers develop an increasing interest in LOTE assessments.

Perhaps one of the important reasons to consider in relation to the growing interest in LOTE assessment is the continuous advocation of language assessments and multilingualism from the Association of Language Testers in Europe (ALTE). In recent years, ALTE has thematized multilingualism at its conferences at least twice: *Multilingualism and assessments* (ALTE 2005, Berlin) and *Language assessment for multilingualism* (ALTE 2014, Paris). Not surprisingly, this year's *Diversity and Inclusion in Language Assessment* strand is also a partial reflection of ALTE's efforts in promoting multilingualism. Another essential reason behind this welcome change may come from the LOTE policy level. For example, the British Council (2013) launched the Languages for the Future initiative to identify the languages the UK needs. In a similar vein, China's Belt and Road Initiative (BRI) also plays a positive role in promoting the teaching and learning of languages from the concerned countries. Additionally, the 'sensitive' language topic 'touches on issues of culture and personal identity' (Taylor, 2008, p. 279) since all languages arguably represent the historical and cultural trajectories of their users. In this sense, minority languages are and will be a continual enhancer of the linguistic diversity of world languages.

Language assessment, as part of language planning and policy, plays an indispensable role in promoting the status of a given language. Though the said efforts are important steps toward representing minority languages in the language testing and assessment field, much work remains to be done if we intend to close that gap. This being the case, I, in what follows, use the Zhuang Language Proficiency Test developed in China as a vehicle to expand the research landscape in multilingual assessment.

The Zhuang Language Proficiency Test in China

The Zhuang Language Proficiency Test, officially titled *Vahcueng Sawcuengh Suijbingz Gaujsi* (VSSG), serves as a pioneer in standardizing minority language learning in China. As a '*de facto* local test' (Wu, Silver, & Hu, 2022, p. 5), the VSSG has been administered annually for free since 2012 within the Guangxing Zhuang Autonomous Region, where the highest number of

Zhuang ethnic people reside. According to the Zhuang Language Test Syllabus (Trial version, 2021), the three-level VSSG is developed to fit for the following purposes:

- Providing a reference for the implementation of the *Zhuang Language Script Scheme*.
- Promoting the Zhuang language and its culture.
- Certifying Zhuang language proficiency levels (i.e., Basic (Cogaep), Intermediate (Cunggaep), and Advanced (Gaugaep)) for test-takers and employers.
- Certifying the training progress for the Zhuang language institutions.

The VSSG is designated as an exclusively written examination, including but not limited to the following item formats: multiple choice, cloze, fill-in blanks, translation, and writing. Interested readers can read more about the test content of VSSG in Wu et al. (2022). The length of VSSG varies from 120 minutes to 150 minutes, based on the degree of linguistic complexity across levels. The VSSG sets 60 as the cut-off score, and the results are reported in three ranks: Level A (80–100), Level B (60–79), and Fail (0–59).

Theoretical framework, data, and analysis

My evaluation of the VSSG is guided by Bachman and Palmer's (2010) Assessment Use Argument (AUA) framework, which specifies the connections from assessment tasks to the consequence of test use. In brief, the AUA includes five main chains: assessment tasks, assessment records, interpretations, decisions, and consequences. Informed by the AUA, I conducted content analysis on the VSGG-related documents, such as test syllabus, question papers, and test announcements. Even with the limited data source, there are interesting preliminary findings for worldwide language testers' reference. The findings are presented and integrated under the five main AUA chains in the following section.

Preliminary findings

Assessment tasks

As mentioned above, the VSSG is a written examination, primarily assessing reading and writing skills. All items in the *reading* skill dimension are selective response tasks (e.g., multiple-choice questions (MCQs) and cloze). The *writing* skill dimension mainly includes translation tasks and prompt-based essays. Generally speaking, item variety develops on a continuum of difficulty and proficiency level. For example, the Basic level includes 30 MCQs on vocabulary and language use and a 200-syllable-long essay task, and the Advanced level includes 10 MCQs and an 800-syllable-long essay task. In terms of authenticity, no technical report is publicly available online, though the test appears to be developed with a corpus of real-life Zhuang language materials (Zhuang Language Test Syllabus (Trial), 2021).

Assessment records

The total score for each proficiency level is 100, and the cut-off score is 60. Results are available to candidates after 10 working days. Although the testing agency subdivides the results into three categories—pass with Level A, pass with Level B, and fail—there is no information available on the format of score reports or certificates. In the meantime, no research or technical reports concerning the assessment consistency can be found, though the test developer claims the test to be consistent in a released news report in 2020. This claim seems to be supported by the VSSG results (2012–2020) in Wu et al. (2022), as the mean scores of the advanced level in 2013, 2014, and 2020 were 66.6, 64.2, and 65, respectively. It should however be noted that the unavailability of other yearly reports still casts doubts on the consistency claim.

Interpretations

The test developer states that the VSSG aims to measure the general language proficiency of Zhuang language and is criterion-referenced (Zhuang Language Test Syllabus (Trial), 2021). Specifically, the ability descriptors of each level clearly specify the required vocabulary size and 'can do' statements. For instance, the Basic level assesses test-takers' ability to know basic sentence structure in the Zhuang language, the Intermediate level measures test-takers' ability to use relatively complex grammar, and the Advanced level measures test-takers' ability to use grammar easily and fluently. However, greater transparency is needed, as there is no publicly available evidence on construct validity and the alignment between claimed abilities and real-life using practices.

Decisions

Influenced by the missing evidence described above under *Interpretations*, limited decisions can be made with the VSSG, which will be elaborated on in the following section. Interestingly, when I performed the search online with the keywords 'Zhuang language recruitment requirement', I did notice that Guangxi-based employers, especially government-based ones, would explicitly state the requirement of VSSG (though no level-specific information) in their recruitment notice. Though this may be a good sign regarding decision making, no technical report supporting the certified Zhuang language users is accessible.

Consequences

The test consequence is intended to be beneficial. As mentioned in *The Zhuang Language Proficiency Test in China* section, the VSSG bears four purposes. Overall, some of the intended purposes of VSSG have been achieved but only bring limited positive consequences. The first two are that VSSG has limited influence on providing a useful reference for the implementation of the *Zhuang Language Script Scheme* and promoting the Zhuang language. The VSSG is surprisingly only popular among a relatively small number of test-takers, and the majority of test-takers are still Zhuang people. According to the official yearly news report, the registered candidates from 2016 to 2021 were 430, *unknown*, 371, 538, 299, and 686, respectively. Meanwhile, no information about the actual testee number can be found.

The third purpose of the VSGG is to certify Zhuang language proficiency level for test-takers and employers. Unfortunately, the VSSG could only reflect the proficiency levels of the Zhuang language in a partial manner. One potential reason is that the exclusive written examination of the current VSSG provides no proficiency information on the oral abilities of test-takers. Such an underrepresentation thus hinders test-takers and employers from making appropriate references to their Zhuang language proficiency. This is also supported by Wu et al.'s (2022) finding that the VSSG fails to capture the proficiency level of test-takers. In terms of the VSSG's impact on training institutions (i.e., the fourth purpose), little is known due to the lack of available evidence.

Conclusion and implication

Based on the above analysis using the AUA framework, the VSSG is indeed a 'de facto local test' (Wu et al., 2022, p. 5) and is still at an early stage, as its implications and social consequences are limited to the Guangxi Zhuang Autonomous Region. It is however understandable because the VSSG itself is the first minority language testing in China, and there are limited language testers proficient in Zhuang language. To conclude, the pioneering VSSG, in its current form, does bring limited positive consequences to the promotion of the *Zhuang Language Script Scheme* and the use of Zhuang language. Nevertheless, more research and transparency on the VSSG are needed for its healthy development to fit for the multipurpose. Looking ahead, I anticipate that in another 5 to 10 years, a revised VSSG will address the concerns raised here and bring the necessary changes.

Though the main thrust of this paper has revealed the deficiency of the current VSSG, I still see welcoming developments in minority language testing, such as the efforts in maintaining consistency and validity and bringing positive social and personal consequences. As the language testing field continues to put English assessment in focus, I hope that the preliminary findings presented here, along with other similar research (e.g., Portuguese: Zhao & Liu, 2019; Ukrainian language: Ozernyi & Suvorov, 2023), can spark more assessment-related discussions on the languages other than English in this multilingual and multicultural era, hence enriching the linguistic diversity in language testing and assessment.

References

Bachman, L. F., & Palmer, A. F. (2010). *Language Assessment in Practice*. Oxford University Press.

British Council. (2013). *Languages for the future: Which languages the UK needs most and why*. Available online: https://www.britishcouncil.org/sites/default/files/languages-for-the-future-report.pdf

Ozernyi, D. M., & Suvorov, R. (2023). Ukrainian language proficiency test review. *Language Testing*, *40*(3), 828–839.

Taylor, L. (2008). Language varieties and their implications for testing and assessment. In L. Taylor & C. J. Weir (Eds.), *Multilingualism and Assessment: Achieving Transparency, Assuring Quality, Sustaining Diversity – Proceedings of the ALTE Berlin Conference May 2005* (pp. 276–295). Studies in Language Testing Volume 27. Cambridge: UCLES/Cambridge University Press.

Weir, C. J., & Saville, N. (2016). Series Editor's Note. In C. Docherty & F. Barker (Eds.), *Language Assessment for Multilingualism: Proceedings of the ALTE Paris Conference, April 2014* (pp. ix–xi). Studies in Language Testing Volume 44. Cambridge: UCLES/Cambridge University Press.

Wu, Y., Silver, R. E., & Hu, G. (2022). Minority language testing: the social impact of the Zhuang language proficiency test in China. *Journal of Multilingual and Multicultural Development*. Advance online publication. Available online: https://doi.org/10.1080/01434632.2022.2097249

Zhao, C. G., & Liu, C. J. (2019). An evidence-based review of Celpe-Bras: The exam for certification of proficiency in Portuguese as a foreign language. *Language Testing*, *36*(4), 617–627. Available online: https://doi.org/10.1177/0265532219849000

Zhuang Language Test Syllabus (Trial). (2021). Available online: http://mzw.gxzf.gov.cn/gzyw/tzgg/P020210624318083496046.docx

Italian language testing regime: Alternative perspectives

Paola Masillo
University for Foreigners of Siena, Italy

Giulia Peri
University for Foreigners of Siena, Italy

Sabrina Machetti
University for Foreigners of Siena, Italy

Abstract

Recently a significant number of European countries have introduced linguistic requirements for the purposes of migration (Rocca, Carlsen, & Deygers, 2020). In Italy, to obtain a long-term residency permit, non-EU citizens must demonstrate a certain level of proficiency in L2 Italian (Law no. 94/2009).

The nature of competencies and tools required in the field of integration and language policies needs to be investigated to understand whether they are adequate to reflect the actual behaviour expected of migrants in the host country.

The paper aims at (a) reporting the development of policies and practices designed to support the linguistic integration of adult migrants in Italy and (b) exploring new perspectives for this purpose.

Introduction

The study illustrates language testing practices carried out at national level within the framework of current Italian legislation, and focuses on the use of language requirements in order to issue a long stay permit for non-EU citizens. The impact of those language policies on migrants is analysed and it is linked to research investigating the use of language tests as a power tool (Shohamy, 2001).

Therefore, the study aims at (a) reporting the development of policies and practices designed to support the linguistic integration of adult migrants in Italy (Machetti, Barni, & Bagna, 2018) and (b) exploring new perspectives for this purpose. We advocate the adoption of a Learning-Oriented Approach (LOA) to Assessment (Purpura & Turner, 2018) by proposing the development of a scenario-based test for L2 Italian (Purpura, 2021).

Research context and hypotheses

Over the past two decades, a significant number of European countries have decided to introduce linguistic requirements for the purposes of migration, such as first entry, permanent residency, and citizenship (ALTE, 2016; Carlsen & Rocca, 2021; Machetti et al., 2018; Rocca et al., 2020).

As underlined by reports promoted by the Council of Europe, an increasing number of countries have imposed policies of control by measuring proficiency in the host country's language (Barni, 2012; Extramiana, Pulinx, & Van Avermaet, 2014).

Since 2009, Italy has provided that the issuance of a long-term residency permit is conditional on passing an Italian language proficiency test. Migrants must be officially recognized as having an A2 level in the Italian language (Ministerial Decree of June 4, 2010).

According to the agreement signed in November 2010 between the Ministry of the Interior and the Ministry of Education, the latter has drawn up official *Guidelines* (MIUR, 2010) for test design and assessment procedures. The test consists of three components: Listening, Reading, and Written Interaction. The decision of the Italian government is neither to use a centralized, national test, nor to involve the L2 Italian Certification Centres in the assessment procedures. Italy assigned test development, administration

and rating to the Public Adult Educational Centres distributed throughout the country (the so-called CPIA – Centro Provinciale Istruzione Adulti).

This move has raised a significant number of critical issues, in terms of fairness, justice, validity, reliability (Carlsen & Rocca, 2021). Each CPIA has in fact developed its own tests, following the official *Guidelines*. Consequently, the results obtained were very different from region to region, also due to the different difficulty level of the items. The tests appeared to be subject to geographical factors and socio-political implications (Masillo, 2021).

Our main research hypothesis, arising from the decentralized test management, questions the fairness, justice, validity, and reliability of the test. The issue on which we would like to reflect is the weight that the specificity of local realities may have in the test design and administration. Given the observed inhomogeneity in the test pass rates, we investigated the degree of comparability of the tests designed in the various CPIAs, and thus whether they are correlated and equivalent.

A second critical issue concerns the central role played in a language test by the theoretical framework of reference for the definition of competences. In the last two decades, the social, cultural, and economic context has undergone global changes in technologies, cultures, and languages. Therefore, the competencies required of anyone living integrated in the dynamics of a globalised and increasingly technological world have changed (Purpura, 2021).

Research procedures

First study

In the first study, we monitored the assessment procedures carried out within the CPIAs. We collected examples of tests used (no. 83) and implemented a database of tests administered.

Focusing on test results, according to the official data presented by the Ministry of the Interior, the range of the pass rate goes from a maximum average of 90.2% (in north-western Italy) to a minimum average of 71.6% (in north-eastern Italy).

To search for useful evidence to confirm the hypothesis of test unfairness, we selected two test samples. We administered them to a group of candidates as representative as possible of the original target public for which the tests were designed (157 candidates, non-EU citizens attending Italian classes at A2 level).

Then the data analysis was divided into two stages. The first was a qualitative analysis, focusing on the test content in relation to the construct, in line with the standard-setting procedures (Council of Europe, 2009). The second was a quantitative analysis to measure the degree of comparability of the two tests.

Second study

The second study was conducted between 2017 and 2021. The study was supported by European funds and conducted by the CLIQ Association (https://www.associazionecliq.it/progetto-fami-1603-2017-2021/)[1].

One of the aims of the study was the validation of the assessment procedures of the CPIAs, implemented within the framework of the current legislation. The study involved 27 CPIAs. Survey instruments included interviews with CPIAs principals, online questionnaires addressed to teachers, and paper questionnaires addressed to migrants attending language courses, and/or participants in the civic training session, and/or candidates for the Italian test.

The monitoring was conducted based on a list of indicators aimed primarily at detecting good practices concerning the assessment procedures applied in the CPIAs, and also at subsequently providing guidelines related to the standardization of assessment procedures and tools, with the aim of increasing their validity and reliability.

Table 1: Frequency distribution

	Test 1	*Test 2*
Listening	63.7%	28.7%
Reading	28.7%	17.8%

Findings

The unfairness of the two tests has been confirmed in terms of the 'unbalance' of the test difficulty level from the very first item analyses. Test 1 appears to be easier than Test 2.

[1] Certificazione Lingua Italiana di Qualità

The data emerging at the level of descriptive statistical analysis confirms the picture described above; the maximum score for the Listening test was obtained by 63.7% of the candidates in Test 1, and only 28.7% of the same sample of candidates in Test 2.

Paying particular attention to the mean, a relatively higher value was found in the two sections of Test 1, namely 9.31 (Listening) and 8.20 (Reading); while for Test 2 the values were 8.46 for Listening and 7.13 for Reading.

Table 2: Descriptive statistics (mean)

	Test 1	*Test 2*
Listening	9.31	8.46
Reading	8.20	7.13

To demonstrate the lack of comparability between the two tests, we measured reliability on two levels: a) internal consistency; b) correlation between the two tests.

The hypothesis of a questionable level of test reliability found an initial confirmation in the values emerging from the calculation of Cronbach's alpha (see Table 3).

Table 3: Internal consistency

	Listening	*Reading*
Test 1		
Cronbach's alpha	.39	.57
Test 2		
Cronbach's alpha	.51	.63

The level of test reliability does not reach the minimum threshold of acceptability (.7) in either test. This finding may find an initial overall justification in the number of items that compose the individual skills (no. 10), which is the minimum number required for a reliable calculation of the Cronbach's alpha coefficient.

Examining the correlation analysis carried out between the total scores obtained for Test 1 and Test 2, we found a correlation coefficient that indicates a moderate relationship, insofar as it is not possible to predict perfectly the performance in one test from the results obtained in the other.

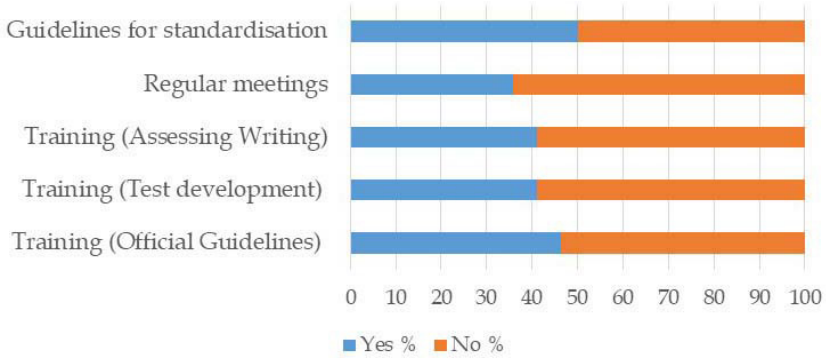


Figure 1 Areas for improvement in assessment procedures (CLIQ, 2021, p. 43)

While migrants recognise a general adequacy of the contents and themes proposed (78.60%), they identified an area for improvement in the need to provide more pictures as a facilitating element aimed at contextualising the contents addressed (48.20%).

Figure 2, related to the component considered by candidates most challenging, is not surprising: in 64.4% of the cases, Writing worries the candidates the most, since it concerns the linguistic activity less connected to everyday life in Italy.

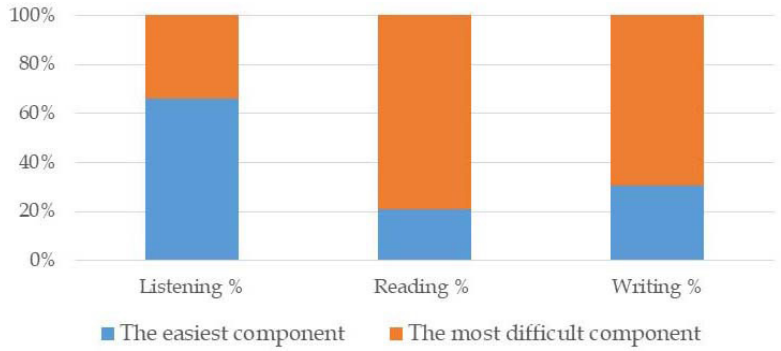


Figure 2 Perception of test difficulty – Candidate feedback (CLIQ, 2021, p. 45)

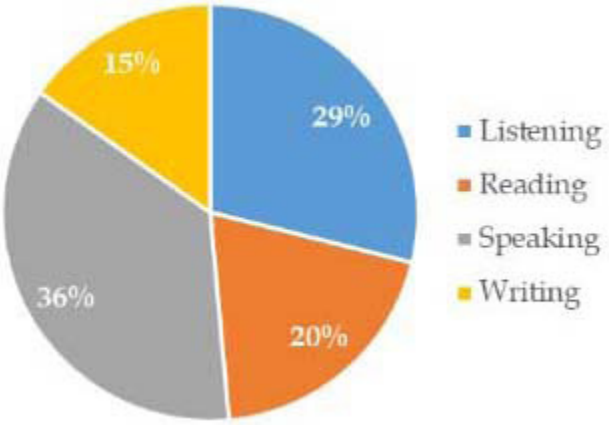


Figure 3 Use of Italian in real life (CLIQ, 2021, p. 46)

This figure can be correlated with the detection of the habitual use of the target language; as Figure 3 shows, oral skills clearly exceed written ones.

Conclusion: A new perspective

As Carlsen and Rocca (2021) point out, although the Common European Framework of Reference for Languages (CEFR, Council of Europe, 2001) recommends a profiled approach to both teaching and language assessment – due to the diversity of L2 learner profiles, which is particularly evident in relation to migrants – the tests for obtaining a long-term residence permit in Italy developed since 2010 by CPIAs merely correspond to the 'standard' A2.

Considering our data and results, we advocate the adoption of a LOA to Assessment (Purpura & Turner, 2018; Turner & Purpura, 2016) by proposing the development of a scenario-based test for L2 Italian (Purpura, 2021).

Grounding on the basic principles and action strategies of European regulations concerning integration, we understand that 'competencies in being able to contribute effectively to society' (Purpura, 2021, p. 3) are what is actually needed to build a new and global society as the product of a bilateral process of mutual accommodation. Moreover, as we know, these competencies are necessary to navigate successfully within different domains and situations.

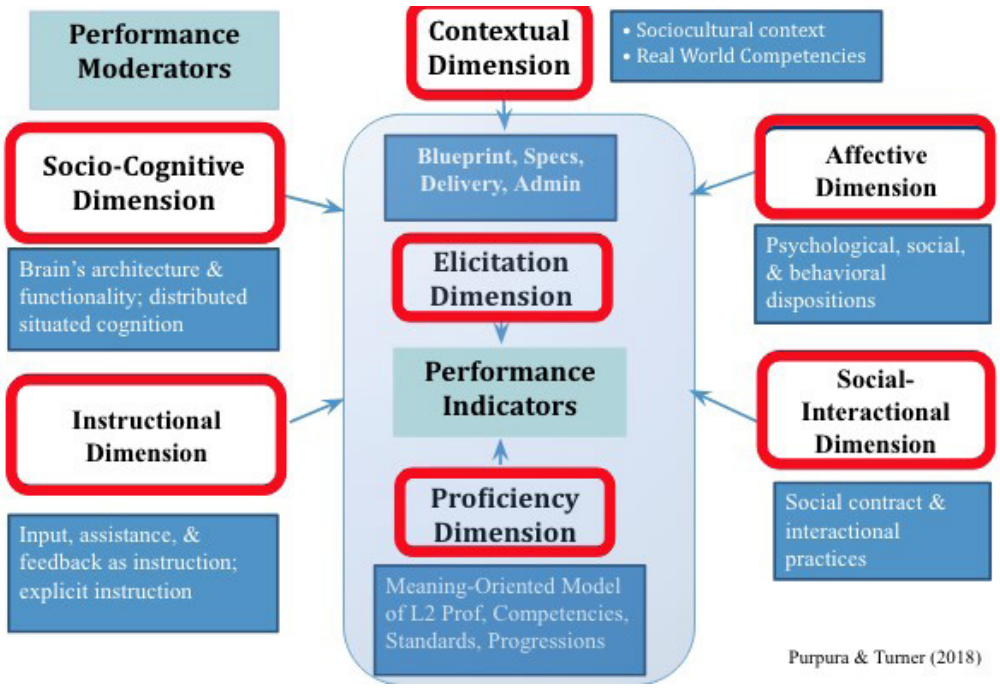


Figure 4 LOA framework

To do this, as Purpura (2021, p. 6) points out, users need content resources and the ability to coordinate these resources to accomplish different tasks in different contexts. At the same time, they need a repertoire of socio-cognitive skills, positive affective dispositions and, in today's world, technological skills (ibid.)

Scenarios mirror as much as possible what can be encountered in real life, which is why they are considered proxies for real-life situations (Purpura, 2021). The learner is therefore presented with a coherent imaginary sociocultural context where they act as an active (emotionally, linguistically, and topically) participant member of a team online to work in collaboration with peers to achieve a goal. The scenario represents the context of this goal and is 'engineered' through different scenes thanks to an architecture of coherent and sequenced tasks (both integrated and independent). This 'scenario narrative' is naturally flexible and can be modeled according to different needs (Purpura, 2021).

At the same time, the process culminating with the problem that needs to be solved is intended as a worthwhile educational experience.

To systematically identify and describe the factors that users need to know how to handle in real-life (whether that be in the context of assessment, instruction, or naturalistic interaction), Turner and Purpura (2016) and later Purpura and Turner (2018) proposed the following framework of LOA.

These seven dimensions (eight also considering the technological dimension) are highly interrelated and can be operationalized in the scenario-based assessment (SBA) through scenario narratives. Operationalizing the characterizing elements of these dimensions could support test construction and validation in reflecting the complexities of real-life second foreign language assessment, turning the test into an assessment process but also a formative one (instructional dimension). By sharing and consolidating knowledge through collaborative problem solving, in interaction, thus tackling the social-interactional dimension, we can operationalize in the test the 'bilateral process' necessary for integration. This aspect also allows us to work directly on the affective dimension, for example, or the socio-cognitive dimension.

As we said, the proficiency dimension, the elicitation dimension, and the contextual dimension alone, as generally considered in traditional testing may not be sufficient to truly reflect the test criterion, especially in today's world and in a normative situation such as the one illustrated so far.

Therefore, LOA and SBA, in their flexible, adaptive and comprehensive nature could be viable assessment alternatives supporting the view of foreign immigration as an (educational) resource and an opportunity which can encourage a project of *linguistic planning and education* aimed at a broad plurilingual perspective. This perspective should not only encourage and support language training courses in the language of the host country, but should also give due recognition, at institutional but also at social level, to the identities and cultures of which citizens of foreign origin are bearers.

References

ALTE. (2016). *Language tests for access, integration and citizenship: An outline for policy makers*. Available online: Available online: https://www.alte.org/resources/Documents/LAMI%20Booklet%20EN.pdf

Barni, M. (2012). Diritti linguistici, diritti di cittadinanza: l'educazione linguistica come strumento contro le barriere linguistiche. In S. Ferreri (Eds.), *Linguistica educativa: atti del XLIV Congresso Internazionale di Studi della Società di Linguistica Italiana (SLI). Viterbo, 27–29 settembre 2010* (pp. 213–223). Rome: Bulzoni.

Carlsen, C., & Rocca L. (2021). Language Test Misuse. *Language Assessment Quarterly*, 18 (5), 477–491.

CLIQ. (2021). *Progetto FAMI 1603 Studio e analisi dell'impatto dei percorsi formativi e valutativi*. Documento conclusivo di Progetto.

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge: Cambridge University Press.

Council of Europe. (2009). *Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR). A Manual*. Strasbourg: Language Policy Division.

Extramiana, C., Pulinx, R., & Van Avermaet, P. (2014). *Linguistic integration of adult migrants: policy and practice. Draft report on the 3rd Council of Europe survey*. Strasbourg: Council of Europe.

Machetti, S., Barni, M., & Bagna, C. (2018). Language policies for migrants in Italy: The tension between democracy, decision-making, and linguistic diversity. In M. Gazzola, T. Templin, & BA. Wickström (Eds.), *Language policy and linguistic* justice (pp. 477–498). New York: Springer.

Masillo, P. (2019). *La valutazione linguistica in contesto migratorio: il test A2*. Pisa: Pacini Editore.

Masillo, P. (2021). Lingua e cittadinanza italiana: uno studio sulla validità della valutazione linguistica per la cittadinanza. *Studi Italiani di Linguistica Teorica e Applicata*, 1(2021), 169–193.

MIUR (Ministero dell'Istruzione, dell'Università e della Ricerca). (2010). *Direzione generale dell'istruzione e formazione tecnica superiore e per i rapporti con i sistemi formativi delle regioni - Ufficio IV. Vademecum - Indicazioni tecnico-operative per la definizione dei contenuti delle prove che compongono il test, criteri di assegnazione del punteggio e durata del test*.

Purpura, J. E. (2021). A rationale for using a scenario-based assessment to measure competency-based, situated second and foreign language proficiency. In M. Masperi, C. Cervini & Y. Bardière (Eds.), *Évaluation des acquisitions langagières: Du formatif au certificatif. MediAzioni*, 32, A54–A96, Available online: https://mediazioni.sitlec.unibo.it/index.php/no-32-2021.html

Purpura, J. E., & C. E. Turner (2018). *Using Learning-Oriented Assessment in Test Development* [Workshop]. Language Testing Research Colloquium, Auckland, New Zealand.

Rocca, L., Carlsen, C., & Deygers, B. (2020). *Linguistic integration of adult migrants: requirements and learning opportunities*. Strasbourg: Council of Europe.

Shohamy, E. (2001). *The power of tests. A critical perspective on the use of language tests*. New York: Pearson.

Turner, C. E., & Purpura, J. E. (2016). Learning-oriented assessment in second and foreign language classrooms. In D. Tsagari & J. Baneerjee (Eds.), *Handbook of Second Language Assessment* (pp. 255–272). Boston: De Gruyter, Inc.

Language needs of adult refugees and migrants and the context of language use in Greece and Italy: Domains, communication themes, and language use situations in L2 Greek and L2 Italian

Anna Mouti
Aristotle University of Thessaloniki and Hellenic Open University, Greece

Abstract

Greece and Italy share a double role as host and transition countries, two of the main EU entry points for refugees and migrants since the 2015 refugee crisis. In this study, individual differences related to language will be examined in Greece and Italy. Linguistic repertoires and multilingual needs of adult refugees and migrants will be explored and identified through the perspectives of their language teachers, with a particular focus on L2 Greek and L2 Italian. The primary purpose of this study is to investigate the domains, communication themes, and language use situations regarding the acquisition of L2 Greek and L2 Italian.

Introduction

The migrant's journey, the migration path, and the migration in general have almost always been accompanied by language-related processes and concerns. Saville (2009) says that 'language has always been a critical factor for migrants' and discusses both the *physical* and the *metaphorical journeys* of the migrants, from one place to another, from one society and culture to another, from one language community to another, even from one identity to another. These journey dimensions interact with numerous language-related issues, including language learning, teaching, and assessment, and Saville (2009, p. 22) attempts to illustrate the 'journey' from arrival in a country to application for citizenship (through the stages of the newcomer, the oldcomer, the settled migrant, and the citizen) and the ways in which government policy interacts with the social, economic, and linguistic factors. According to Beacco, Little and Hedges (2014), account will also have to be taken of the 'timing' relative to migration, e.g., during the pre-migration phase or upon arrival in the host country (when the need is urgent), and form of settlement (sojourn, settlement involving alternation between countries, long-term settlement, settlement with a fixed plan of return, permanent settlement, etc.). Certainly, the migration journey and the migration path may last up to several years, consisting of a rather long migration experience for refugees and asylum seekers as forced migrants, who often 'cross various countries, staying in one or more of them for a longer period than in others, entering in contact with the local population and learning their language' (Bianco & Ortiz Cobo, 2018, p. 2). Migration and relocation plans may define the multilingual and language needs of the refugees and migrants and every story can be different as '. . . there is no such thing as a typical migrant' (Beacco, Krumm, & Little, 2017, p. 4) and there is no such a thing as a typical migration path.

Assessing adult refugees' and migrants' multilingual needs and specific subjective language needs should be a prerequisite for designing and developing tailor-made language courses. 'A description of domains and language use situations must be further refined to be convertible into a workable tool for curriculum, syllabus, or assessment design. After all, domains and situations only describe contexts in which language is used, but do not specify what particular things a language learner should do with language to function efficiently in these situations' (van Avermaet & Gysen, 2008).

Language education for adult refugees and migrants in the Greek and in the Italian context has been examined through a variety of studies, providing information about teachers' and learners' profiles, language courses and materials, etc. (e.g., Androusou & Iakovou, 2020; Chatzidaki & Tsokalidou 2021; Kantzou, Manoli, Mouti, & Papadopoulou, 2017; Karavas, Iakovou, & Mitsikopoulou, 2021; Machetti & Rocca, 2017; Machetti, Barni, & Bagna, 2018; Mattheoudakis, Griva, & Moumtzi, 2021; Rocca, 2017), but also

raising concerns and actions regarding other social or educational needs e.g., alphabetization to adults, and adapting the Italian language curricula and syllabi to low-literate language learners (Borri, Minuz, Rocca, & Chiara, 2014; Minuz & Borri, 2017). Other studies (e.g., Barn, Di Rosa, & Kallinikaki, 2021; Bertotti, Di Rosa, & Asimopoulos, 2023; Brändle, Eisele, & Trenz, 2019; Samek Lodovici et al., 2017) attempt to provide a comparative approach to the two countries in the migration context but none of them focuses on the linguistic integration aspects of adult refugees and migrants in the two countries. In these comparative studies Greece and Italy were chosen as the 'first arrival countries for people arriving on dangerous routes via the Mediterranean Sea', as the 'first countries of entry to the EU', as 'the two main gateways to the European continent', as 'transit countries' and as 'main entry points to the EU'.

This study is part of a larger project, funded by the Research Committee of Aristotle University of Thessaloniki, Greece, examining individual differences related to language in Greece and Italy and, more specifically, plurilingual profiles and multilingual needs of adult refugees and migrants, with a particular focus on L2 Greek and L2 Italian in the two contexts in a comparative way. The purpose of the paper in the broader field of linguistic integration of adult refugees and migrants in Greece and Italy is to provide more detailed information based on empirical data regarding the two countries and to examine an even more concrete field: multilingual needs and L2 Greek and L2 Italian language needs. In this paper, we will focus on the language teachers' perspectives through an online open-ended questionnaire, and we will try to explore the plurilingual profiles and needs of the learners by considering the context(s) of language use, establishing relevant domains, and specifying the communication themes and language use situations or communicative functions that the target group should be able to cope with.

Similar studies have already been conducted in Greece and Italy but not in a comparative way. Mouti, Maligkoudi and Gogonas (2021; 2022), Kyrligkitsi and Mouti (2023), and Androulakis, Gkaintartzi, Kitsiou and Tsioli (2017) have examined the language needs regarding L2 Greek in a variety of different methods. Interestingly Bianco and Ortiz Cobo (2019), examining the real needs of refugees in the Italian context, have demonstrated that 'language competence is not only a means of assisting with integration, but also the result of the integration itself' and that Italian is not the only language spoken in Italy by refugees and migrants, especially in social contexts such as the workplace.

Method

In our study, a total of 89 language teachers completed an online open-ended questionnaire (38 L2 Greek language teachers, and 51 L2 Italian language teachers), offering language education/support to adult refugees and migrants in a variety of settings. All Greek participants had extensive teaching experience, with an average of 9.25 years of language teaching to adult learners in various formal and non-formal education settings, non-governmental organizations (NGOs), municipalities, solidarity schools, etc. Their current position is mainly in state settings and language courses organized by NGOs. Italian participants had an average of 13.25 years of teaching experience to adults, while the majority of them are currently employed in CPIAs (Centri Provinciali per l'Istruzione degli Adulti/Provincial Centers for Adult Education), and the rest in associations, emergency reception centers, or other migrants' integration projects. Notably, all participants are plurilingual. All the Greek participants have L1 Greek and L2 English at B2 level and above. French is the second most frequent L2 (17 cases), followed by German (11), Spanish (11), and Italian (10). Few participants have mentioned Russian (4) and Arabic (3), and Ukrainian, Portuguese, Farsi, Dutch, Danish, Turkish, Czech, Romanian, and Catalan were each only reported once. All Italian participants have L1 Italian and L2 English in the majority of the cases (44). French is also the most frequent L2 (25), followed by Spanish (13), German (7), and Russian (5). Only two participants mentioned Chinese, and Arabic, Ukrainian, Albanian, Polish, Greek, and Esperanto each got one mention.

The tool implemented in this study was an open-ended questionnaire distributed online to a large number of language teachers of adult refugees and migrants in Greece and Italy. The anonymous questionnaire, including closed and open-ended questions, was provided in Greek and Italian. In particular, it aimed to collect demographic data, information on the learners' and classrooms' profiles, and information regarding plurilingualism and Greek/Italian everyday language use. Finally, it contained items on general language needs but also more specific language needs of the learners.

Results

Mouti and Rocca (under review) have examined the plurilingual profiles, everyday language use, and relocation plans of adult refugees and migrants based on their language teachers' perceptions. The relocation plans of adult refugee and migrant learners, as explored by Mouti and Rocca (under review), can be closely linked to the motivation and needs for learning L2 Greek and L2 Italian. In both contexts, there were three main categories of learners: the ones who want to stay in Greece and Italy permanently, those who want to relocate to another European country, and those who want to return to their homelands. This reality certainly influences the multilingual needs of the learners, as L2 Greek and L2 Italian do not seem to be the main and only needs. English was found to be a language that many of the students would like to learn for work purposes either in Greece or Italy or in any other destination country as a *lingua franca*. German was a strong reference by the teachers, and other languages

mentioned were Swedish and French in the Greek context, while in the Italian context, Spanish, French, and Chinese. The countries that were mainly mentioned regarding the relocation plans were Germany, France, Scandinavian countries, and the UK, together with homelands like Ukraine.

In this study, more specific language needs were observed regarding L2 Greek and L2 Italian. When language teachers were asked why their learners wanted to learn L2 Greek or L2 Italian, their motives were ranked similarly in both contexts. Both Italian (IT) and Greek teachers (GT) ranked language use situations, as shown in Figure 1. The most important situation mentioned by the Greek teachers was the desire to communicate more effectively (4.72) and to enhance professional prospects (4.44), as indicated by Italian teachers. Speaking a second language at home was considered less important by both groups (2.70 and 2.76). Other highly ranked language use situations in both contexts – above 4.0 – were: to integrate better into the society, to communicate with doctors, and in public services.

When participants were asked to identify why their learners felt that they have to learn L2 Greek and L2 Italian for formal social integration, formal language requirements were identified as follows: for professional purposes (82.35% for IT and 71% for GT), residence permit (78.43% for IT and 50% for GT), citizenship (66.67% for IT and 47.37% for GT), university studies (82.35% for IT and 71% for GT).

The next aspect to be investigated was the importance of specific communication themes and language use situations, and what the teachers themselves perceived as more important for their students to learn. Less important is to converse with neighbors (especially for the IT – 3.38), and engage in sports (especially for the GT: 3.05). Most crucial is communicating with public services (GT:4.58) and reading and writing effectively at work (IT:4.60). Other cases reaching an average level above 4.40, are: conversations with doctors, conversations with people at shops, and interactions with other individuals at their children's school.

Regarding the perceived importance of macro-skills and competences among language teachers, both for L2 Greek and L2 Italian, it is evident that oral skills are considered the most crucial. This finding aligns with the learners' uneven language proficiency profiles estimated in L2 Greek and L2 Italian by the teachers, where oral skills were better developed. Written reception also ranks high. In contrast, receptive skills and especially written reception came first in other cases. It appears that oral skills and receptive skills are more advanced but also most urgent to learn, with written production being the last to develop or the least important to be developed.

Language use domains were similarly ranked by language teachers in both contexts. Health, Workplace, and Filling Forms took the top positions, while Food ordering was rated the least important. Communicative language needs in the public and occupational domains were considered more crucial, followed by the educational and personal domains.

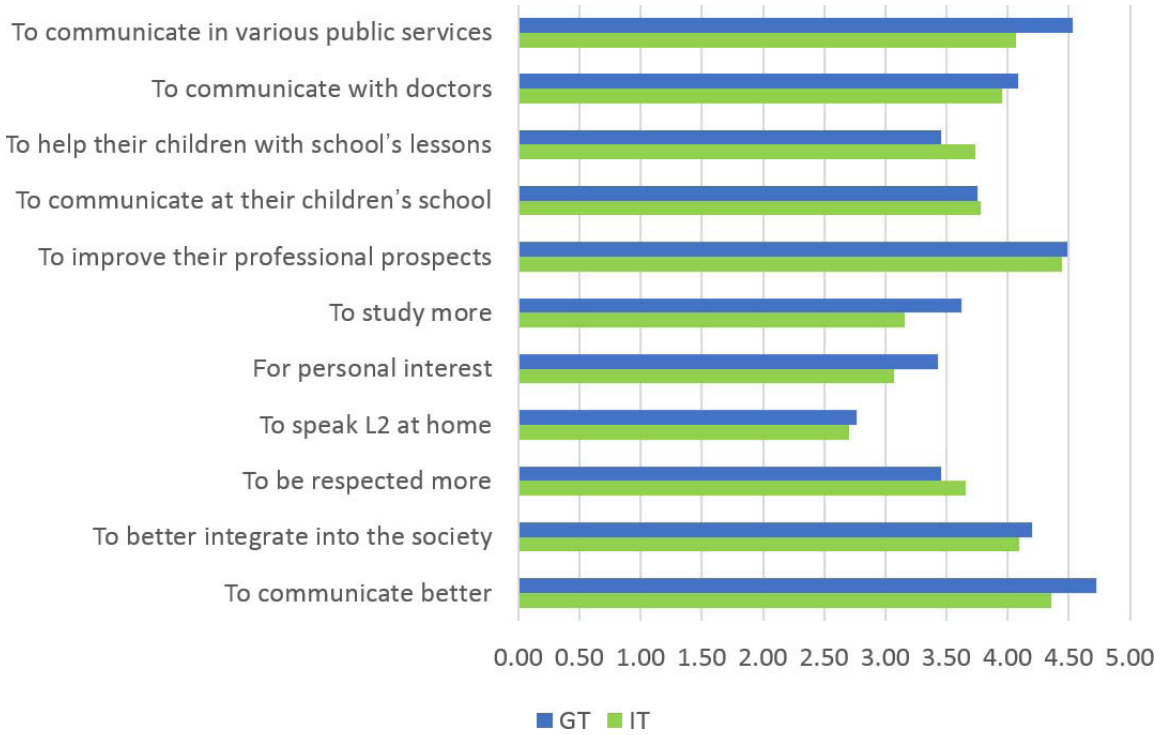


Figure 1 Why learning L2 Greek/L2 Italian?

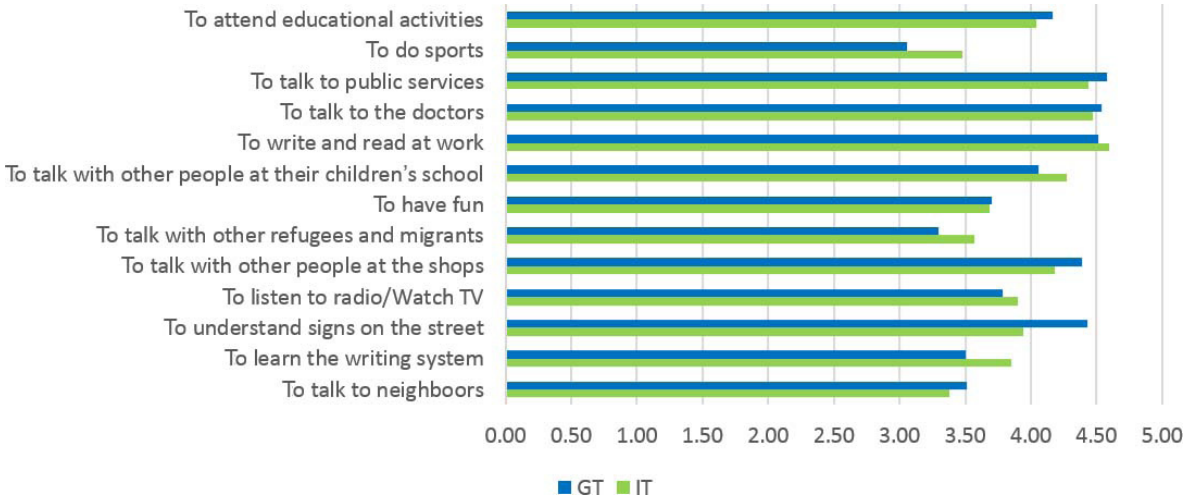


Figure 2 Language use situations: Communication themes

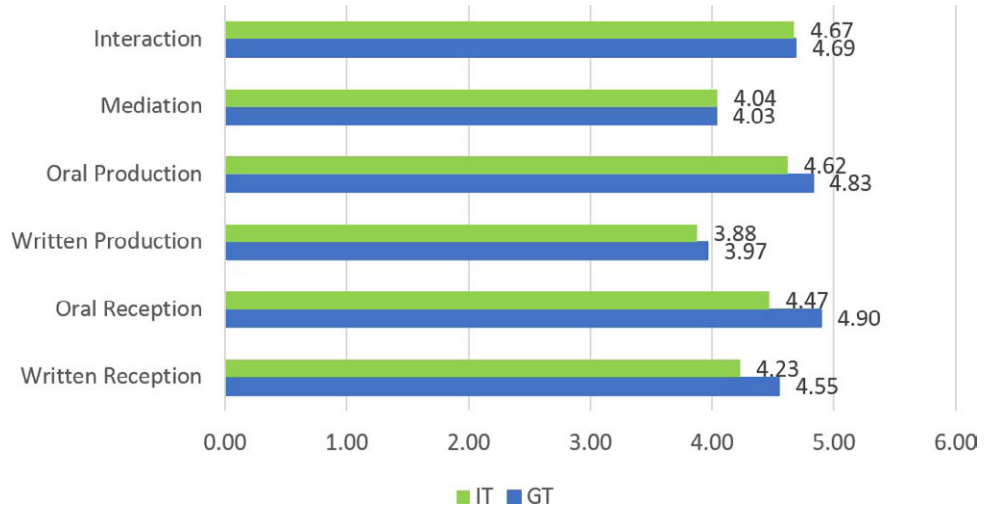


Figure 3 Language skills

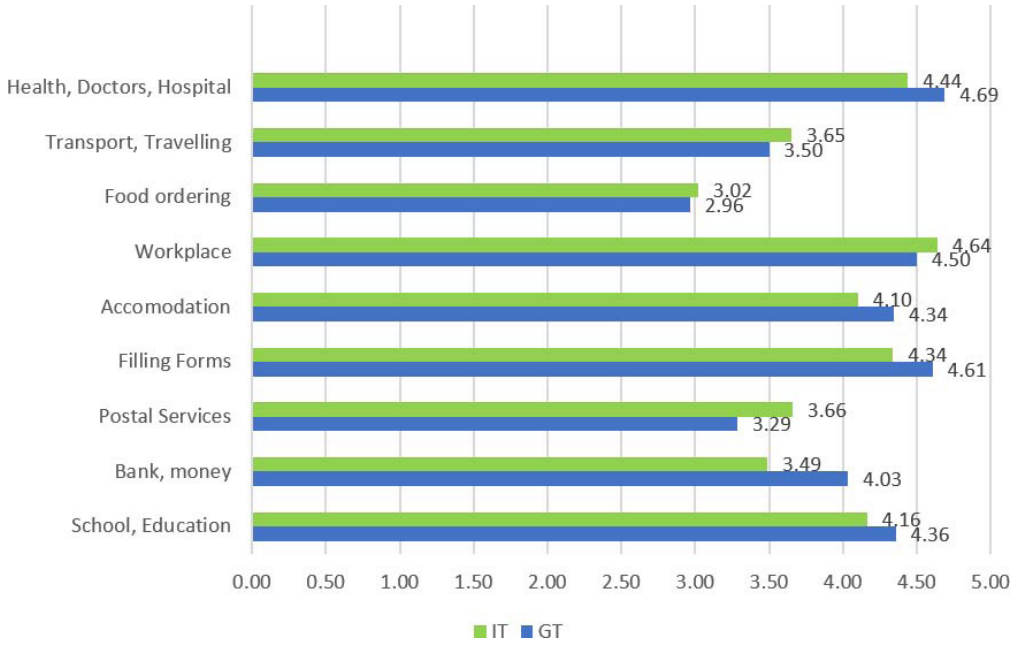


Figure 4 Language use sub-domains

Discussion and concluding remarks

This paper provides information on specific language needs regarding L2 Greek and L2 Italian. These findings align with prior research by van Avermaet and Gysen (2008), Androulakis et al. (2017), as well as Mouti et al. (2021; 2022), Kyrligkitsi and Mouti (2023) and Bianco and Ortiz Cobo (2018; 2019), particularly in specific sub-domains and communication themes, as Work/ Business/Access to the labor market, Communication in public services, Communication with doctors, and Education/training/ children's education consistently ranked highest.

There is a need to visualize the population of adult refugees and migrants as a language learning group by examining background variables, individual differences, profiles, and repertoires. Greece and Italy share a double role, both as host countries but also as transition countries, and as Bianco & Ortiz Cobo (2019) mention ' the migration path of some refugees does not conclude in Italy,' and we may add that it certainly does not conclude in Greece. Therefore, refugees and migrants may also be interested in other languages, especially those needed in a possible future relocation. But it is not only the future but also their past migration path and this migration experience that have shaped their existing linguistic repertoires, as many of them could learn the languages met during their migration journey in various countries (Bianco & Ortiz Cobo, 2018).

There are social justice and fairness issues related to language learning opportunities offered to and provided for adult refugees and migrants that should be taken into consideration when designing and implementing language courses. Questions arise regarding the languages for which courses are developed, the language proficiency levels targeted, and whether tailor-made courses are created to address specific language needs. Furthermore, exploring how diversity issues are managed within diverse and heterogeneous working groups is essential. Specific challenges and social needs of language learners, such as seasonal workers or mothers with children, should be considered. Lastly, there is an urgent need to examine other literacy limitations and needs for low-literate migrants (Minuz & Kurvers, 2021; Minuz, Kurvers, Schramm, Rocca, & Naeb, 2022) and possible exceptions regarding language requirements for social integration (see Carlsen, Deygers, Rocca, & Van Oers, 2023). Further research is required in this direction.

References

Androulakis, G., A. Gkaintartzi, R. Kitsiou, & Tsioli, S. (2017). Research-driven task-based L2 learning for adult immigrants in times of humanitarian crisis: results from two nationwide projects in Greece. In J-C. Beacco, H-J. Krumm, D. Little & P. Thalgott (Eds.), *The Linguistic Integration of Adult Migrants. Some Lessons from Research* (pp. 181–186.). Berlin/Boston: Walter de Gruyter GmbH.

Androusou, A., & Iakovou, M. (2020). Refugee children's integration in Greece: training future teachers to face new educational challenges. *International Journal of Early Years Education*, *28*(2), 162–175.

Barn, R., Di Rosa, R. T., & Kallinikaki, T. (2021). Unaccompanied Minors in Greece and Italy: An Exploration of the Challenges for Social Work Within Tighter Immigration and Resource Constraints in Pandemic Times. *Social Sciences 10*(4). Available online: https://www.mdpi.com/2076-0760/10/4/134

Beacco, J-C., Little, D., & Hedges, C. (2014). *Linguistic integration of adult migrants: Guide to policy development and implementation*. Strasbourg: Council of Europe.

Beacco J-C., Krumm, H-J., & Little, D. (2017). Introduction. In J.C. Beacco, H.J Krumm, D. Little & P. Thalgott (Eds.), *The Linguistic Integration of Adult Migrants. Some Lessons from Research* (pp. 1–5). Berlin/Boston: De Gruyter.

Bertotti, T., Di Rosa, R. T., & Asimopoulos, C. (2023). Child Protection in Mediterranean Countries: Italy and Greece. In J. D. Berrick, N. Gilbert & M. Skivenes (Eds.), *The Oxford Handbook of Child Protection Systems* (pp. 261–288). Oxford: Oxford University Press.

Bianco, R., & Ortiz Cobo, M. (2018). The migration experience and the informal language learning of refugees. In *Conference Proceedings. Innovation in Language Learning International Conference 2018*, *Bologna*. Available online: https://conference.pixel-online.net/files/ict4ll/ed0011/FP/5097-LSM3411-FP-ICT4LL11.pdf

Bianco, R., & Ortiz Cobo, M. (2019). The linguistic integration of refugees in Italy. *Social Sciences 8*(10), 284. Available online: https://www.mdpi.com/2076-0760/8/10/284

Borri, A., Minuz, F., Rocca, L., & Chiara, S. (2014). *Italiano L2 in Contesti Migratori*. Torino: Loesher.

Brändle V. K., Eisele, O., & Trenz, H-J. (2019). Contesting European Solidarity During the "Refugee Crisis": A Comparative Investigation of Media Claims in Denmark, Germany, Greece and Italy. *Mass Communication and Society*, *22*(6), 708–732.

Carlsen, C., Deygers, B., Rocca, L., & Van Oers, R. (2023). The consequences of migration tests on low-literate adult migrants : a survey of teacher opinions in 20 European countries. In AERA (Ed.), *2023 AERA Annual Meeting, Proceedings* (pp. 1–13). Washington, D.C.: AERA.

Chatzidaki, A., & Tsokalidou, R. (Eds.). (2021). *Challenges and initiatives in refugee education: The case of Greece.* Newcastle: Cambridge Scholars Publishing.

Kantzou, V., Manoli, P., Mouti, A., & Papadopoulou, M. (2017). Γλωσσική εκπαίδευση προσφύγων και μεταναστών/ριών: Πολλαπλές μελέτες περίπτωσης στον Ελλαδικό χώρο. *Διάλογοι! Θεωρία και πράξη στις επιστήμες αγωγής και εκπαίδευσης*, *3*, 18–34.

Karavas, E., Iakovou, M., & Mitsikopoulou, B. (2021). Responding to the Challenges of Adult Refugee Language Education through Action Research. *International Journal of Learner Diversity & Identities*, *28*(2).

Kyrligkitsi, A., & Mouti, A. (2023). A Multi-Method Profiling of Adult Refugees and Migrants in an L2 Non-Formal Educational Setting: Language Needs Analysis, Linguistic Portraits, and Identity Texts. *Societies 2023*, *13*. Available online: https://www.mdpi.com/2075-4698/13/8/186

Machetti, S., & Rocca, L. (2017). Integration of migrants, from language proficiency to knowledge of society: The Italian case. In J-C. Beacco, H-J. Krumm, D. Little & P. Thalgott (Eds.), *The Linguistic Integration of Adult Migrants. Some Lessons from Research* (pp. 213–218). Berlin/Boston: Walter de Gruyter GmbH.

Machetti, S., Barni, M., & Bagna, C. (2018). Language policies for migrants in Italy: the tension between democracy, decision-making, and linguistic diversity. In M. Gazzola, T. Templin, & B-A. Wickström (Eds.), *Language Policy and Linguistic Justice: Economic, Philosophical and Sociolinguistic Approaches* (pp. 477–498). Cham: Springer International Publishing.

Mattheoudakis, M., Griva, E., & Moumtzi, M. (2021). *Migration and Language Education in Southern Europe. Practices and Challenges.* Newcastle: Cambridge Scholars Publishing.

Minuz, F., & Borri, A. (2017). Literacy and language teaching: tools, implementation and impact. In J-C. Beacco, H-J. Krumm, D. Little & P. Thalgott (Eds.), *The Linguistic Integration of Adult Migrants. Some Lessons from Research* (pp. 357–363). Berlin/Boston: Walter de Gruyter GmbH.

Minuz, F., & Kurvers, J. (2021). LASLLIAM. A European Reference Guide for LESLLA Learners. *LESLLA Symposium Proceedings, 14*(1), 453–470.

Minuz, F., Kurvers, J., Schramm, K., Rocca, L. & Naeb, R. (2022). *Literacy and Second Language Learning for the Linguistic Integration of Adult Migrants*. Strasbourg: Council of Europe.

Mouti, A., & Rocca, L. (under review). *Linguistic integration of adult migrants in Greece and Italy: Learners' profiles, multilingual needs, and relocation plans*.

Mouti, A., Maligkoudi, C., & Gogonas, N. (2021). Assessing language needs of adult refugees and migrants in the Greek context. In ALTE. (Ed.), *Collated Papers for the ALTE 7th International Conference*, Madrid (pp. 229–232). Available online: https://www.alte.org/IntConfProceedings

Mouti, A., Maligkoudi, C., & Gogonas, N. (2022). Language needs analysis of adult refugees and migrants through the CoE-LIAM Toolkit: The context of language use in tailor-made L2 material design. *Selected Papers on Theoretical and Applied Linguistics*, *24*, 600–617.

Rocca, L. (2017). *Language support for adult refugees: A Council of Europe toolkit. Report on piloting carried out in Italy from February to April 2017*. Available: https://rm.coe.int/language-support-for-adult-refugees-a-council-of-europe-toolkit-report/168073cf1d

Saville, N. (2009). Language Assessment in the Management of International Migration: A Framework for Considering the Issues. *Language Assessment Quarterly*, *6*(1), 17–29.

Samek Lodovici, M., Drufuca, S. M., Orlando, N., Crepaldi, C., Pesce, F., Koulocheris, S., & Borbély, S. (2017). *Integration of Refugees in Greece, Hungary and Italy: Comparative Analysis*. Brussels: European Union.

van Avermaet, P., & Gysen, S. (2008). *Language learning, teaching and assessment and the integration of adult immigrants. The importance of needs analysis*. Available online: https://rm.coe.int/16802fc1d5

Defining alternative constructs of multilingual assessment in higher education: The case of English in contact with other languages in mainland US and Puerto Rico

Eva Rodríguez-González
University of New Mexico

Rosita L. Rivera
University of Puerto Rico – Mayagüez Campus

Abstract

This study is part of a larger book publication (Rodríguez-González, & Rivera, 2022) that provides practitioners within the field of heritage languages contexts in the Americas with examples of challenges faced by these academic communities in the design and implementation of effective assessment practices. Recent research in the field of applied linguistics has addressed the complex and contextual realities of multilingual language learners. These perspectives include different language learner profiles and different learning contexts in higher education. The challenge of assessment in these communities requires that educators contest more traditional and prescriptive notions of assessment to better serve their communities of learners. One example described in our study examines Spanish learners' self-efficacies in speaking and writing tasks at a southwestern university in mainland US. Another example discusses the bilingual context of Puerto Rico and the teaching of an L3 as a foreign language at a university on the west coast of the island.

Introduction[1]

As the field of language learning becomes more interdisciplinary due to the complexities of the contexts educators tackle in Higher Education classrooms, assessment methodologies continue to evolve based on shifting paradigms. Approaches are context-based and eclectic in nature. This ever-evolving field has been transformed by the dynamics of students and their individual and collective experiences as part of their language learning communities of practice. Research and practice are also more intertwined in university settings where students bring with them not only their content knowledge, but also their linguistic ability and (inter) cultural competence to communicate in multiple discourses. The challenge of assessment in these communities requires that educators contest more traditional and prescriptive notions of assessment to better serve their communities of learners. Some of these diverse contexts in higher education include different languages and different learner profiles. Research in assessment and textbooks dealing with methodologies for assessing language learners have approached the study of assessment from a more one-dimensional perspective. These perspectives end with a recommendation for one specific type of assessment. Although very valid and useful for language educators, this approach to study assessment does not provide a more holistic view of assessment for practitioners. We share two case studies as examples of the incorporation of a multidimensional perspective in which different communities of learners are represented when using alternative assessments in language classrooms in mainland US and Puerto Rico.

[1] We would like to thank and acknowledge the collaboration of Marián Giráldez-Elizo, Sarah Schulman and Mildred M. Vargas in the two case studies we are highlighting here.

Example 1: Alternative assessments in mainland US Spanish language programs

Student self-assessment in language classrooms occurs when learners assess their own performance. Self-assessment via self-efficacy gives learners a greater amount of agency regarding assessment, thus enriching their learning. *LinguaFolio*, a self-monitoring learner portfolio tool that enables goal setting and collection of evidence of language achievement, was specifically created for measurement of progress and growth in second languages other than English in the context of the US. As such, *LinguaFolio* can serve as a type of assessment as it contains a set of multiple language learning standards that have been adapted into classroom goals as 'Can Do' statements that follow the *American Council of Teaching Foreign Languages* (ACTFL) proficiency guidelines. 'Can Do' statements have been shown to increase learner motivation, language proficiency, and academic achievement (Moeller, Theiler, & Wu, 2012). Although 'Can Do' statements were originally designed to enhance the learning of second language learners, the same guiding principle can also be applied to heritage learners (Cox, Malone, & Winke, 2018, p. 106). By identifying learners' perceived learning abilities, language teachers can better target their instruction to support learners' developing linguistic proficiency (Hlas, 2018, p. 49). Given that speaking and writing are often identified by language learners as the most difficult skills to learn, the study focused on learners' self-assessment of their capabilities in these two skills.

When referencing home and community speakers of the target language, the term *heritage language learner* (HLL) is often used. Valdés (2001) provides the most frequently referenced description of a HLL as 'a language student who is raised in a home where a non-English language is spoken, who speaks or at least understands the language, and who is to some degree bilingual in that language and in English' (p. 38). Given that HLLs are typically exposed to the target language at a young age, language educators often assume they will perform equally to or better than L2s on communicative tasks. When HLLs are unable to meet academic register or demonstrate metalinguistic knowledge, this experience may lead to lower ratings of self-efficacy on specific language learning tasks that are also experienced by L2 learners.

The data set of this study included self-assessment survey responses from a total of 133 Spanish language learners enrolled in first and second year Spanish college courses. Participants were students enrolled in two different language programs based on their academic or home and/or community exposure to the Spanish language. Participants therefore included Spanish as Second Language learners (SSL, N = 67) and Spanish as Heritage Language Learners (SHL, N = 66). Participants ranged in proficiency from Novice High to Advanced Low and responded to a Can Do Statement questionnaire (ACTFL, 2017) that was directly aligned to course objectives. The study aimed to determine to what extent learners perceive themselves as more proficient in Spanish speaking or writing as they progress through different periods of coursework and to see whether differences in self-efficacy exist between L2 or HL.

Results for the interpersonal speaking and presentational writing domains suggest that participants differed in their self-efficacy based on their language program. Learners in the SSL program predominantly placed themselves within the Novice-Mid continuum. SHL learners, however, expressed greater confidence than SSL learners in their interpersonal speaking abilities, in the Novice-High range. In second year courses, SSL learners mostly identified themselves as feeling capable of engaging in interpersonal speaking tasks at the Intermediate-Low level, with some considerable ratings also reported in Intermediate Mid-High. SHL learners also identified self-efficacy in interpersonal speaking at the Intermediate-Low range, while some SHL learners also placed themselves within the Intermediate-High level.

In terms of presentational writing, findings indicated that learners in SHL first year have higher ratings of self-efficacy in presentational writing, as compared to their SSL peers. In second year courses, there was an overall effect of similarity between SSL and SHL in second-year coursework with most learners, independently of language program, rating themselves as capable of performing at the Intermediate-Low sub-level of proficiency.

In addition to examining the nature of self-efficacies in speaking and writing by Spanish L2 and HLLs, a comparison was made between learners' self-perceptions with programmatic student learning outcomes identified by instructors in first year coursework of both SSL and SHL programs in speaking and writing. The results confirm a direct alignment of course objectives and learners' perceptions of self-efficacies in the first year courses in both programs. The second year coursework expectations and self-perceived abilities of language proficiency in interpersonal speaking and presentational writing did not match in the SHL data set. More specifically, the distance between an expected outcome of language proficiency and self-efficacy ratings is considerable in the case of learners in the second year Spanish coursework in the SHL program (Advanced-Low vs. Intermediate-Low).

In conclusion, the findings in the study call for a more contextualized approach to language instruction and planning that takes into consideration the learning outcomes that the students themselves identify as goals to their success. By integrating learners' voices on self-perceived capabilities into language coursework, instructors may draw on this kind of data as a baseline for the development of a more reliable set of course learning outcomes in both cross-sectional and vertical curriculum alignment.

Example 2: Alternative assessments in Puerto Rico English language programs with multiple L1 student language profiles

The linguistic landscape of Puerto Rico incorporates various contexts in which English and Spanish are used in both formal and informal settings. These contexts transcend domains when dealing with the use of English. Dayton and Blau (1997) examined Puerto Rico as an English-using society. In the case of higher education, studies have dealt with the use of English and Spanish in higher education classrooms (Carroll, Rivera, & Santiago, 2015; Mazak & Herbas-Donoso, 2015). English as a *lingua franca* is evident in the use of English in government, technology, business, media, and education domains. A great number of English language borrowings and anglicisms are also part of the Puerto Rican discourse (Dayton & Blau, 1997). Thus, code-switching is a common practice in formal settings, such as schools and universities, and in informal settings, such as social media and video-game communities.

This study explored the role of English and Spanish in foreign language teaching and assessment at a university in Puerto Rico. Our research site were two foreign language classrooms at the University of Puerto Rico in Mayagüez (UPRM) where students' native language is Spanish and the use of English as either a first or second language varies depending on geographic region and social background. Foreign languages are not mandatory in grade school. As a result, foreign language instruction is more evident in higher education contexts. We focused on how professors incorporate multilingual practices in their L3 classrooms. Through ethnographic observations and analysis of classroom materials, we examined the use of language(s) in L3 classrooms at UPRM.

Most of the research dealing with language teaching and assessment in Puerto Rico's higher education system focuses on bilingualism (English – Spanish) due to the status and use of languages in Puerto Rico. Alvarez Aguirre (2000) studied the attitudes and experiences of Puerto Rican college students towards learning French and English. She found that at the time the study was conducted, the instruction of French was mostly based on a 'French Only' policy and students learned conversational skills that would help them to face a real situation. The French course in the study was adapted to fit the Puerto Rican culture and students noticed the similarities between French and Spanish. Alvarez Aguirre (2000) argues that students strongly believe in the power and the direct connection between knowing multiple languages and social mobility.

The data analysis provided evidence of multilingual use and *translanguaging* practices in the two classrooms involved in the study. Both instructors involved in the study used a multilingualism approach when teaching and assessing a foreign language. Materials and assessment provide evidence for how English, Spanish, and the target language were used by instructors to help students learn the language. Participants used multilingualism as a tool to communicate in the classroom and as a learning tool to scaffold and mediate language learning. All the participants code-switched at some point in their class between English, Spanish and the target language. These instances of language use can be considered multilingual practices in the formal setting of classroom instruction. Further research in these areas can help shed light on this particular question regarding multilingualism and assessment in a foreign language course at university level.

Conclusions

Whether the student is a second/third language learner, a heritage language learner, a multilingual language speaker, we provided examples of assessment that do not follow a single universal or standardized design but an applicable one based on the needs and context of a given community. The examples described above follow Justice, Equity, Diversity and Inclusion (JEDI) assessment practices that are based on observation, examination and integrative notions of diverse language scenarios. By involving instructors and students in reflection, dialogue and decision making, we will be promoting assessment FOR learning (Leung & Rea-Dickins (2007); see also Green's (2014) PRICE principles for promoting effective classroom assessment: Planning, Reflection, Improvement, Cooperation, and Evidence). Additionally, we also call for an ecological approach to language assessment: despite multiple shared characteristics of learners' profiles, assessment tools are unique and different depending on the *habitat*. To avoid habitat fragmentation that produces isolated patches (minority language profiles such as heritage language learners for instance), assessment needs to be accessible and inclusive to all in an equitable manner and should keep evolving in varied patterns.

References

ACTFL. (2017). *NCSSFL-ACTFL Can-Do Statements*. Available online: https://www.actfl.org/educator-resources/ncssfl-actfl-can-do-statements

Alvarez Aguirre, M. (2000). *Experiences of Puerto Rican Students in Learning English as a Second Language and French as a Foreign Language*. New York: New York University.

Carroll, K. S., Rivera, R. L., & Santiago, K. (2015). Questioning Linguistic Imperialism: Language Negotiation in an Agriculture Classroom. In A. Fabricius & B. Preisler (Eds.), *Transcultural Interaction and Linguistic Diversity in Higher Education* (pp. 164–187). New York: Palgrave Macmillan.

Cox, T. L., Malone, M. E., & Winke, P. (2018). Future directions in assessment: Influences of standards and implications for language learning. *Foreign Language Annals*, *51*(1), 104–115.

Dayton, E., & Blau, E. (1999). Puerto Rican English: An acceptable non-native variety?. *Milenio*, *3*, 176–193.

Green, A. (2014). *Exploring Language Assessment and Testing: Language in Action*. New York: Routledge.

Hlas, A. C. (2018). Grand challenges and great potential in foreign language teaching and learning. *Foreign Language Annals*, *51*(1), 46–54.

Leung, C., & Rea-Dickins, P. M. (2007). Teacher assessment as policy instrument: contradictions and capacities. *Language Assessment Quarterly*, *4(1),* 16–36.

Mazak, C. M., & Herbas-Donoso, C. (2015). Translanguaging practices at a bilingual university: a case study of a science classroom. *International Journal of Bilingual Education and Bilingualism*, *18*, 698–714.

Moeller, A., Theiler, J., & Wu, C. (2012). Goal setting and student achievement: A longitudinal study. *Modern Language Journal*, *96*, 153–169.

Rodríguez-González, E. & Rivera, R. (Eds.). (2022). *Integrating Context-based Approaches to Language Assessment in Multilingual Settings*. Current Issues in Bilingualism Series, Language Science Press [Open-Access].

Valdés, G. (2001). Heritage language students: Profiles and possibilities. In J. K. Peyton, D. Randard & S. McGinnis (Eds.), *Heritage Languages in America: Preserving a national resource* (pp. 37–80). Washington, DC: Center for Applied Linguistics.

A pilot material for a fair and accessible A2 listening test for adult immigrants with diverse educational backgrounds

Elina Stordell
Testipiste Language Assessment Centre for Adult Migrants, Finland

Abstract

Teachers often give us test developers feedback that results from listening tests that do not correlate with their classroom assessment. Furthermore, in Finnish integration training for immigrants, test results in listening are often weaker than in speaking, when quite similar proficiency levels could be expected.

For immigrants with diverse educational backgrounds, formal listening tests requiring reading and writing can be a challenge. However, students with a lower literacy level and possibly with some literacy training background deserve a fair opportunity to show their oral skills, which are usually their strongest ones.

Thus, an attempt has been made to design a more accessible listening test for A2 level. The first part is completely auditive and simulates a kind of interview or informal discussion; the second part includes simple multiple-choice and short-answer questions. First test versions have been piloted in the integration training including students with various educational backgrounds.

Introduction

Testipiste, a language assessment centre for adult migrants in Finland (www.testipiste.eu), has designed placement and proficiency tests for Finnish integration training since 2010. Widely used tests have been warmly welcomed with one exception. According to teachers' experience, results in listening tests do not correlate with their classroom assessment. Many students' results in listening are weaker than expected. Furthermore, tests include too much reading and writing (multiple-choice and short-answer questions), and that is why less educated students with lower literacy levels often just guess.

To further investigate teachers' feedback, a sample of 499 students taking proficiency tests at the end of the official integration training was collected. The data show that when speaking is at B1.1 level (target level of the integration training), 96% of test-takers get lower results in the listening test, when a more even proficiency profile could be expected. Possible explanations exist, such as too strict cut scores in the listening tests, and more lenient assessment due to a human factor in face-to-face speaking tests, but weaker study skills might play a remarkable role as well.

The same kind of observations had been made in the placement testing at Testipiste. Many beginners were able to manage a simple face-to-face interview and a speaking test with an examiner, but struggled in a traditional listening test with multiple-choice questions. Recently, a new oral listening test was introduced, and those who got 0, pre-A1 or A1 level in the traditional listening test, took the new one as well. A small sample of 101 test-takers shows that almost 50% of beginners do better in the oral test, and only 10 percent do better in the traditional one.

Thus, a more accessible listening test for less educated test-takers or test-takers with a literacy training background was piloted when new proficiency tests for the integration training were designed.

Test development

Test tasks

A new proficiency test to be used in the middle of the integration training was needed, and a more accessible listening test was set as a goal. The target level is at A2.1 but because of a heterogeneous population, the test is designed to cover both levels A2.1 and A2.2. Some A1 warm-up questions are included.

The test consists of two parts. The written part includes a simple multiple-choice task (short radio news) and a short answer task (an informal conversation), where students need to read and write as little as possible. These task types were kept for comparative reasons, to see if there is any difference between traditional and new oral questions. The written part is done with pen and paper. The oral part is completely auditive and simulates a kind of interview or informal discussion on four topics. Each topic includes several questions at different levels. Here are example questions from one topic translated into English:

First, let's talk about your home. Where do you live?

What do you do at home?

What's near your home?

The area where you live in, is it good? Why?

The oral part is delivered as an audio file and requires recording premises.

Two test versions were created and the versions were linked together by some shared oral and multiple-choice questions.

Piloting, statistical analyses and standard setting

The two test versions were piloted in the integration training with 435 and 356 test-takers, respectively. The target level at the end of the first phase of the training is A2.1, in a more intensive study path A2.2.

The answers were then marked by test developers. Multiple-choice questions were scored with an answer key, and short-answer questions and oral questions with examples of accepted and failed answers. In the oral task very different answers at different levels could be scored with one point. The answer had to show that the question had been understood, but otherwise speaking itself was not assessed. Thus, in some cases, an answer with just one word, a short expression and a grammatically correct more complex sentence all resulted in one point. Here are some examples of the first question mentioned earlier, and all the answers were scored with one point, because they show that the test-taker has understood the question:

Helsinki.

At Helsinki.

I live Kontula, in Helsinki.

I have lived in Helsinki for four years now.

Data was statistically analysed with Winsteps software, and some inconsistent items were deleted or edited. Cut scores were set in a standard setting using the Angoff yes/no method. The jury assessed some items at B1 level, and in the future these items should be deleted or reviewed as well.

Results

Comparing the results of oral and written tasks

Later, a small follow-up study was made. Fresh data with some background information and feedback from 155 voluntary test-takers was collected. Students had a higher educational background than expected. They were asked to indicate their education in years, and it is possible that some included pre-school in their education as well. Teachers could give feedback as well.

Statistical analyses run by Winsteps show high item and person reliability, 0.98 and 0.86. However, the multiple-choice questions show more guessing and lower discrimination while the short answer questions look better. Most of the oral questions work well, although warm-up items look quite predictable and easy with less variation, and some items are surprising outliers.

Other analyses show a difference between oral and written parts. Test-takers achieve higher levels in the oral part: 52% achieve the highest possible level A2.2 or above, 40% get A2.1, and only 8% A1 or below. In the written part results are clearly weaker: only 22% are at A2.2 level, 53% at A2.1 and 25% at A1 or below. A related-samples Wilcoxon signed rank test for non-parametric related-samples suggests that the null hypothesis *The median of differences between written and oral tasks equals 0* can be rejected, and the result is statistically significant (sig. 0.001). There is a moderate correlation (r_s 0.672, sig. <0.001) as expected: both parts of the test are tapping into the listening skill.

Comparing the results of students with diverse educational backgrounds

When the test-takers are split into several groups according to their educational background, the difference between the oral and written tasks becomes even more obvious. Also test-takers with low education or no education at all (0–6 years) can reach the highest level in the oral part of the test. Each level covers 33% of the least educated test-takers. On the other hand, in the written part 75% of the less educated stay at A1 level (or below it, the test does not separate any lower levels), only 25% achieve A2.1, and no one is at A2.2 level.

The correlation between performance and educational background is moderate for the writing tasks (r_S 0.419, sig. <0.001), and weak for the oral task (r_S 0.202, sig. <0.001). So, no matter what kind of education one has, it is possible to achieve high results in the oral listening test. Also, an independent-samples t-test shows that the null hypothesis *There will be no difference in the results of less vs. more educated test-takers on the written tasks* can be rejected (sig. <0.001). For the oral task, Mann-Whitney U test for non-parametric independent samples was carried out. This time the null hypothesis *The distribution of scores in the oral task is the same across categories of education level* should be retained, so there is no difference between the less and more educated test-takers in the oral task (sig. 0.094).

Feedback from the test takers and the teachers

The students (N = 155) took a short questionnaire before and after the listening test. The teachers were asked to assist their students with the questionnaire. Nevertheless, quite a few questions were left empty. Beforehand, 44% of those who answered thought that listening is the most difficult language skill. After the test 58% found the short-answer questions the most difficult task. The multiple-choice questions and oral questions were considered easy: 40% thought multiple-choice and 40% oral questions were the easiest test tasks.

Only six teachers gave feedback on the test. Five of them found the new kind of listening test reliable or reliable enough, and even the one doubting the reliability of the test found the new oral listening task reliable. No one saw the multiple-choice questions as a reliable tool for assessment, and guessing and writing challenges were mentioned as weaknesses in the written part of the test. All teachers agreed that the oral listening task is well-suited for students with diverse educational backgrounds, and no one complained that assessing the new task was remarkably more time-consuming than other task types.

The teachers made some suggestions for future test development. They wished for slower speaking tempo in all audios, including input and oral questions, as well as for practice material for their students.

Conclusions

For students with diverse educational backgrounds, formal listening tests requiring reading and very often writing can be a challenge. However, test-takers with a lower literacy level and possibly with a literacy training background deserve a fair opportunity to show their oral skills, which are usually their strongest ones. Field trials and a small-scale follow-up study suggest that an oral listening task increases the accessibility of a listening test. A more accessible listening test for A2 level consists of a completely auditive task simulating a kind of interview or informal discussion, and simple multiple-choice and short-answer questions.

According to the statistical analyses, oral items work well: there is good discrimination and less guessing than in multiple-choice questions. Students with lower educational levels and/or lower literacy levels can get the highest scores at least in a test covering A2.1 and A2.2 levels. Test results in the oral part of the test correlate only weakly with educational background.

The adult students in the integration training in Finnish find the oral task type easy. Moreover, the teachers trust the task type: they find it reliable enough and suitable for students with diverse educational backgrounds. Even though assessing spoken answers takes longer, the teachers do not complain about this.

Thus, an oral listening task instead of multiple-choice questions can be recommended at least for lower language levels. In these first test versions, questions simulated an interview or discussion and dealt with test-takers' background, everyday life and concrete plans. However, there could be more variation in question types, such as questions about familiar topics requiring no special knowledge or responding to everyday situations. Test developers could consider the possibility of allowing test-takers to listen to the questions twice, as the input is often played twice in many listening tests. Furthermore, students should be familiarised with the task type through learning activities during the training.

Balancing the need for native and non-native speakers in ELF listening tasks: to what extent do accents affect comprehension?

Anna Maria De Bartolo
University of Calabria, Italy

Jean Marguerite Jimenez
University of Calabria, Italy

Ian Michael Robinson
University of Calabria, Italy

Abstract

The aim of this study is to contribute to research on the feasibility of including a variety of accents in aural comprehension tasks, in a move away from standard English accents towards an inclusive approach which focuses on the use of both native and non-native speaker accents as a response to the increasing use of English as a *lingua franca* around the world. In particular, the study investigates the extent to which a speaker's accent may affect English L2 learners' listening comprehension. Specifically, 120 Italian L1 undergraduate students from five different degree courses were assessed on their comprehension of five extracts of a lecture delivered by fluent speakers of English with different L1s (American English, Arabic, British English, Hungarian, Italian). Results indicate that accent diversity may have an effect on students' performance on listening tasks. This would suggest that including a variety of accents in listening activities in the classroom as well as in listening assessment tasks may better prepare learners to interact in multilingual contexts where English is the main means of communication.

Introduction

In a rapidly changing world in which English has achieved a status of an international language and is spoken as a *lingua franca* (ELF) by growing numbers of non-native speakers in academic, cultural, educational and professional domains, the way it is assessed as an L2 has become a topic of heated debate. This brings into question whether native speaker norms should still be considered the standard in language assessment, especially when more attention is being placed on the intercultural and plurilingual aspects of language learning (Council of Europe, 2020). Although calls for the development of tests which reflect the communicative and contingent variability nature of ELF have been largely unheeded by international language examination certifiers (Jenkins & Leung, 2017), a careful rethinking of a test construct which takes this variability into account warrants further investigation.

Turning specifically to the assessment of listening, this would imply a move away from standard English accents towards an inclusive approach which focuses on the use of both native and non-native speaker accents (see Abeywickrama, 2013; Harding, 2008, 2011, 2012; Kang, Thomson, & Moran, 2018; Major, Bunta, Fitzmaurice, & Balasubramanian, 2002; Newbold, 2017). This may be particularly important in tertiary education, where students are likely to encounter a variety of accents throughout their course of study. Multilingual universities increasingly welcome non-native lecturers who speak English with different accents, and there is a growing international student population who uses English as a means of communication. Considering this, the present study focuses on accent diversity and investigates the extent to which B1 level Italian L1 speakers' performance on listening tasks might be affected by the speaker's accent. Specifically, 120 undergraduate students from different degree courses were assessed on their comprehension of lecture extracts delivered by fluent speakers of English with different L1s (American English, Arabic, British English, Hungarian, Italian). The test tasks were similar regarding topic, speech rate, and syntactic complexity, and were delivered to the participants in a random order.

Brief state of the art

> Accent is potentially a very important variable in listening comprehension. When listeners hear an unfamiliar accent . . . this can cause problems and may disrupt the whole comprehension process.
>
> (Buck, 2001, p. 35)

This issue of accent is also seen as important in the Common European Framework of Reference for Languages (CEFR), where the C1 descriptor for listening includes the following statement:

> I have no difficulty in understanding any kind of spoken language, whether live or broadcast, even when delivered at fast native speed, provided I have some time to get familiar with the accent.
>
> (Council of Europe, 2001, p. 27)

These examples highlight how accent is considered a valid factor of potential difficulty in general language comprehension, even at high levels of competence. This is true with regards to general 'standard' English, but even more so when considering ELF, which is used internationally in academic settings and should therefore be addressed both in the classroom as well as in assessment (Jenkins, 2018). Indeed, various scholars (e.g., Abeywickrama, 2013; Canagarajah, 2006; Harding, 2008; Harding & McNamara, 2018; Jenkins, 2006; Jenkins & Leung, 2017; Newbold, 2015a, 2015b) have enquired into the use of different accents, especially those of speakers who do not have English as their first language (the so-called non-native speakers). These authors raise many important issues concerning this topic, namely how to assess ELF when it is still relatively new and in evolution. Although the inclusion of a broad range of accents in listening assessment may be viewed as beneficial, it also raises several problematic issues. For example, it may lead to a breakdown in comprehension if listeners find particular accents difficult to understand. It may also lead to test bias as listeners who share a speaker's L1 may be advantaged over others when listening to that speaker (Harding & McNamara, 2018). However, if world use of English is more often that of non-native speakers who communicate in English with different accents (see, for example, Kachru, 1985), why should we continue to do listening tasks with only native speakers? This is an area that demands further investigation.

The research project

The aim of the study was to investigate the extent to which B1 English level students' performance on listening tasks might be affected by the speaker's accent. Based on our experience as English teachers as well as on findings from previous studies, we would expect that, when listening to an extract of a lecture in English, students would find it easier to understand a speaker who shares their same L1 or who has an L1 they are familiar with. Conversely, they may have difficulty understanding a speaker who has an L1 they are not familiar with.

Participants

120 undergraduate students enrolled in five different degree courses participated in the study. Specifically, 50 participants were studying Linguistic Mediation, 16 Political Science, 14 Administration Sciences, 21 Media and Digital Society, and 19 Statistics. The students had just completed a B1 level English for basic academic purposes course and were starting an English course specific for their degree course. The data were collected during the third week of this second module.

The participants completed an informed consent form before taking part in the study, but they were not initially told the specific nature of the project.

Data collection tools

Assessment tools

The data were collected through the administration of five listening tasks which consisted in a short audio (approx. 530 words lasting about four minutes), followed by 10 multiple-choice comprehension questions. Five fluent speakers of English with different L1s were involved in the recording of the audios.

Each speaker recorded a different script, which was based on a general topic, namely: A: Educating girls; B: Interracial marriage; C: Advertising and children; D: The Mozart effect; E: Stress and pets. The scripts were prepared by the authors, after which an online programme (Text Analyzer: https://www.online-utility.org/text/analyzer.jsp) as well as expert reviews were used to ensure that the texts were of similar difficulty. For the piloting phase, the audios were all recorded by a British male speaker with a standard accent and a listening comprehension task was then devised for each text. The pilot study involved 28 undergraduate

students from one degree course (Social Services), although only 14 completed all five tasks. The results were analysed and adjustments made to ensure that the difficulty level of all five listening tasks was the same.

The texts were then recorded with L1 speakers with different accents, specifically: A: Educating girls (Arabic, female); B: Interracial marriage (United States, female); C: Advertising and children (Italian, male); D: The Mozart effect (Hungarian, male); E: Stress and pets (British, male).

It was expected that most participants would be familiar with the British and US accents, as well as with English spoken with an Italian accent, which was their L1. Hungarian and Arabic accents were chosen for being, most likely, unfamiliar to the students.

Students' feedback

In addition to the quantitative data, it was important to collect feedback from the students. During a follow-up session the participants were asked if they could guess what the purpose of the study was. Then, after playing short clips of the five audios, participants were asked if they recognized the different accents and if they thought that any of the speakers were easier/more difficult to understand.

Data collection schedule

The experimental schedule is illustrated in Table 1.

For the main study, six groups were formed: Linguistic Mediation (two groups); Political Science; Administration; Media and Digital Society; Statistics. The participants listened to all five audios and completed a test after each one. There was a studied randomization of the tasks so that the members of the different groups would not hear the audios in the same order. The recordings were played twice.

The independent variable is the speaker's accent, while the dependent variable is the test score. A possible intervening variable may be familiarity with the topic.

Table 1: Data collection schedule

Session 1: Three listening tasks
↓ *2/5 days later*
Session 2: Two listening tasks
↓ *2/3 days later*
Follow-up session: Students' feedback

Results and discussion

Descriptive statistics are illustrated in Table 2. As can be seen, listening E (British accent) has the highest mean score, followed by C (Italian accent). It is interesting to note that this is followed by listening A, which is an accent that the students are not familiar with, namely Arabic. Listening D (Hungarian) has the lowest mean score.

In order to understand whether these differences are significant, a one-way repeated measures ANOVA was conducted to compare listening comprehension scores on the five tasks. There was a significant effect for accent ($p < .001$), partial eta square was .438 (large effect size). Table 3 illustrates significant differences between accents (√).

As can be noted, there is a significant difference between the task involving the Hungarian accent and all the other accents, which was not surprising as the students were not familiar with this accent and hence might have had problems understanding it. It was more surprising that there was a significant difference in comprehension between American English and British English, and between American English and Italian.

In the follow-up session, the participants reported being familiar with British and American accents and were able to identify them. In fact, it had been expected that this would be so since Italian undergraduates have been exposed to both British and American accents of English through coursebooks they have previously used as well as television programmes, films, and popular

Table 2: Descriptive statistics

	Mean	*Standard deviation*	*N*
Test score A	7,3250	2,25352	120
Test score B	7,2500	1,83454	120
Test score C	7,7500	1,85277	120
Test score D	6,4500	1,97399	120
Test score E	7,9083	1,93159	120

Table 3: Significant differences

Accent	*Arabic*	*US*	*Italian*	*Hungarian*	*British*
Arabic				√	√
US			√	√	√
Italian		√		√	
Hungarian	√	√	√		√
British	√	√		√	

songs. However, this familiarity did not translate into these two being the easiest for them to understand during the listening comprehension tasks. In fact, comprehension of American English came behind that of L1 speakers of Italian and Arabic.

The participants were also able to immediately identify the Italian accent, but they reported difficulty recognising Arabic and Hungarian accents, with some stating that they had problems understanding the latter. Perhaps surprisingly, some participants said that the Italian accent was not always an aid to understanding, as we had envisaged it might be since the speaker shared their L1. Some participants found the Arabic accent easier to understand than the Italian one, even if they could not identify it as Arabic.

The results would seem to indicate that accents may affect listening comprehension. However, it needs to be pointed out that, although great care was taken to ensure that the difficulty level of the listening tasks was equal and thus the only variable was the accent used, the content itself may have played a role in the comprehension of the audios. This aspect presents the main limitation of this study. Further research could involve comparable groups that listen to the same text, but which is delivered by speakers with different L1s. An investigation could also be conducted to compare groups of learners with different levels of language competence to further examine the effect of proficiency on understanding accents.

Reflections

Treating English as a *lingua franca* opens up new possibilities, but also new challenges. The challenge investigated in this study regards how moving away from standard native speaker accents in listening tasks could affect assessment of the learners' listening ability.

The study indicates that students may have difficulties in recognising and understanding L2 speakers of English who have an accent that learners are not familiar with. This does not mean that such accents cannot be used, but that any comparisons would have to take this variable into account. It might be desirable to include an assortment of accents in a listening assessment task, especially as the backwash effect of this could be the use of a greater variety of accents in the classroom. Indeed, there is a growing awareness of the potential validity of an ELF construct (Harding & McNamara, 2018), including the relevance of designing tasks which reflect accent variety. Certainly, ELF communication represents a significant challenge within language testing and assessment, and as language professionals we need to take this aspect into serious consideration.

References

Abeywickrama, P. (2013). Why Not Non-native Varieties of English as Listening Comprehension Test Input?. *RELC Journal*, 44(1), 59–74.

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

Canagarajah, A. S. (2006). Changing communicative needs, revised assessment objectives: Testing English as an international language?. *Language Assessment Quarterly*, *3*(3), 229–242.

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment.* Cambridge: Cambridge University Press.

Council of Europe. (2020). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Companion Volume*. Strasbourg: Council of Europe Publishing.

Harding, L. (2008). Accent and academic listening assessment: A study of test-taker perceptions. *Melbourne Papers in Language Testing*, *13*(1), 1–33.

Harding, L. (2011). *Accent and Listening Assessment: A Validation Study of the Use of Speakers with L2 Accents on an Academic English Listening Test.* Language Testing and Evaluation. Frankfurt: Peter Lang.

Harding, L. (2012). Accent, listening assessment and the potential for a shared-L1 advantage: A DIF perspective. *Language Testing*, *29*(2), 163–180.

Harding, L., & McNamara, T. F. (2018). Language assessment: The challenge of ELF. In J. Jenkins, M. J. Dewey & W. Baker (Eds.), *Routledge Handbook of English as a Lingua Franca* (pp. 570–582). Abingdon: Routledge.

Jenkins, J. (2006). Current Perspectives on Teaching World Englishes and English as a Lingua Franca. *TESOL Quarterly*, *40*(1), 157–181.

Jenkins, J. (2018). The future of English as a lingua franca?. In J. Jenkins, M. J. Dewey & W. Baker (Eds.), *Routledge Handbook of English as a Lingua Franca* (pp. 594–605). Abingdon: Routledge.

Jenkins, J. & Leung, C. (2017). Assessing English as a Lingua Franca. In E. Shohamy (Ed.), *Language Testing and Assessment* (pp, 1–15). Encyclopedia of Language and Education. New York: Springer.

Kachru, B. B. (1985). Standards, codification and sociolinguistic realism: the English language in the outer circle. In R. Quirk & H. G. Widdowson (Eds.), *English in the World: teaching and Learning the Languages and Literatures* (11–30). Cambridge: Cambridge University Press.

Kang, O., Thomson, R., & Moran, M. (2018). Empirical approaches to measuring intelligibility of different varieties of English in predicting listener comprehension. *Language Learning*, *68*(1), 115–146.

Major, R. C., Bunta, F., Fitzmaurice S. F., & Balasubramanian, C. (2002). The effects of non-native accents on listening comprehension: implications for ESL assessment. *TESOL Quarterly*, *36*, 173–190.

Newbold, D. (2015a). Assessing ELF in European Universities. In Vettorel, P. (Ed.), N*ew Frontiers in Teaching and Learning English* (pp. 205–206). Newcastle: Cambridge Scholars Publishing.

Newbold, D. (2015b). Engaging with ELF in an entrance test for European university students. In Y. Bayyurt & S. Akcan (Eds.), *Current Perspectives on Pedagogy for English as a Lingua Franca* (pp. 205–222), Berlin: Walter de Gruyter and Company.

Newbold, D. (2017). Co-certification: a close encounter with ELF for an international examining board. *Journal of English as a Lingua Franca*, *6*(2), 367–388.

Citizenship tests as a means of inclusion. How far have we gone till now?

Thomais Rousoulioti
Aristotle University of Thessaloniki & Democritus University of Thrace, Greece

Anna Kokkinidou
Aristotle University of Thessaloniki & Democritus University of Thrace, Greece

Abstract

The process of citizenship concerns multiple criteria which touch upon linguistic, cultural, historical, geographical and political knowledge of the host country that the applicant must display. The main naturalization process can be different among countries in the European Union.

The main goal of this paper is to present the citizenship exams in Europe and highlight the points that have been amended within the former framework for the citizenship exams in Greece (Kokkinidou, Rousoulioti, Pasia, Antonopoulou, & Zervou, 2021). It still remains a duty of the host country to respect the rights of the migrant test-takers. Under these conditions it becomes clear that these are high-stakes exams in terms of the lives of migrants, and the examination process should be fair, accurate, reliable and most of all fit for purpose (ALTE, 2016), considering their inclusion in the European Union.

Introduction

The 2030 Agenda for Sustainable Development was adopted by the United Nations (2015). It is a broad and universal policy agenda, with 17 Sustainable Development Goals (SDGs), which seeks to guide Member States to transform their approach to achieving inclusive, people-centered, and sustainable growth, leaving no one behind. Under this light, citizenship exams' issues around the world, with emphasis on the Greek context, are investigated.

The most prevalent definition of citizenship in use today, which has to do with a person's legal relationship with the state, reflects the idea of citizenship. Many people on Earth are legitimate citizens of one or more nation states, and as such, they are entitled to certain benefits or rights. Additionally, having citizenship imposes obligations in terms of what the state requires of people living under its control. As a result, citizens fulfill certain duties to their government and are entitled to the protection of their fundamental interests (Brander et al., 2020). In Europe alone, '[I]n 2021, EU Member States granted citizenship to 827.300 persons having their usual residence on EU territory, an increase of around 14 % compared with 2020' (Eurostat, 2023).

The following two research questions were formulated during this research:

1. What do citizenship tests in Europe examine?
2. What changes have been made to citizenship exams in Greece after 2020?

Acquiring citizenship in the European Union

One can obtain EU citizenship in one of three ways as listed below (Schengen visa info, n.d.):

a) By descent – if someone has a family member to whom they can pass on EU citizenship.

b) By investment – if someone has the required money to invest in an EU country and receive citizenship.

c) By naturalization – if someone lived and worked long enough in an EU country to qualify for citizenship by taking part in citizenship exams. In this case, the citizenship decision is a complex decision that entails various factors, and which depends on the degree of acculturation of the candidate (DeSipio, 1987, p. 390).

The citizenship exams refer usually to the process of citizenship by naturalization. Naturalization confers upon the person concerned all the civil and political rights associated with being a national of a member state. These rights do not have any

retroactive effect (Kokkinidou et al., 2021). A brief overview of the testing requirements in certain countries follows, which depicts the current situation in Europe:

- In France, TCF IRN (*Test de Connaissance du Français – Integration, residence and citizenship*) has been mandatory since 2022. The TCF IRN consists of four compulsory parts, which concern the receptive (listening and reading) and productive (speaking and writing) skills at language Level B1, according to the Common European Framework of Reference for Languages (CEFR) (Council of Europe, 2001).
- In Federal Republic of Germany, language proficiency at B1 level is required as a prerequisite accompanied by the *Einbürgerungstest*. The 33 questions are divided into three broad categories: Life in Democracy, History and Responsibility, as well as Man and Society. There is also a number of questions specific to German states.
- Italy does not require a citizenship test by law. However, candidates must prove knowledge and understanding of the Italian language (minimum Level B1) by presenting a certificate (*Certificazione Lingua Italiana di Qualità* – CLIQ).
- Since 2015, Spain has offered exams of knowledge of the culture and history of Spain and knowledge of the Spanish language as a prerequisite for migrants wishing to become citizens (Bruzos, Erdocia, & Khan, 2018). The Spanish test is entitled *Conocimientos Constitucionales y Socioculturales de Espana-CCSE TEST* (Constitutional and Sociocultural Knowledge of Spain). This is an A2 level CEFR language level test in Spanish that includes all basic communication skills (reading, listening, writing and speaking) along with 25 questions concerning two main subject areas: government and law (60%), culture/history and society (40%).
- In the UK, candidates can meet UK language and life requirements if they have passed the Life in the UK exam by answering 24 questions and have a certificate of speaking and listening in English at language Level B1 according to the CEFR (2001) or higher, i.e., one from the list of recognized exams at an approved test center. The language requirement is also met when the applicant has obtained an academic qualification considered by the UK NARIC[1] to meet the recognized standard of a Bachelor's, Master's or doctoral degree in the UK and the qualification was taught or researched in the UK or a majority English-speaking country outside Canada, or they are a national of a majority English-speaking country (Kokkinidou et al., 2021).

Citizenship tests in Greece

Citizenship tests in Greece until 2020

By 2020, following the law amending the Citizenship Code, a circular has been issued by the competent Ministry of Interior setting out the substantive conditions for naturalization in Greece. The successful candidate should possess a minimum level of competence in each of the following three areas of substantive requirements:

- Knowledge of Greek
- Adequate degree of inclusion in the socioeconomic life of the country
- Opportunities for active and meaningful participation in the political life of the country

The target audience for naturalization is foreign citizens who may be long-term residents who have already been tested in Greek. For the first-time acquisition of a long-term residence permit, A2+ level plus elements of history and culture/B1 level are required (Law 4251/2014, Art. 107).

Participants in the examination procedure must take part in an interview in which they will be examined on language, history, culture and geography. Candidates can download from the website for the naturalization exams the material for study, an e-book entitled *Greece. A Second Homeland* (Vasiliou & Giavi, 2001). The book deals with subjects related to Greek history, culture and geography, but there is a note on the website informing those interested that the specified material was produced under the previous legislative framework and for this reason candidates are encouraged to use other sources of information.

It is obvious that, until 2020, during the citizenship exams in Greece, the language component was not examined. Knowledge of the Greek language was examined through the preparation material for those who had to take the exams and participate in an interview. Under these circumstances, the linguistic level of the texts in the preparatory material was checked, because a text can be a stimulus to start a discussion in a specific context during the interview (Kokkinidou et al., 2021). The texts were checked with

[1] UK NARIC (National Academic Recognition Information Centre) is the national agency responsible for providing information and expert advice on international qualifications and skills in the UK.

Table 1: Changes implemented to the typology of the citizenship exams in Greece

Year	*Until 2019*	*From 2020 and on*
Exams material	A book candidates learn by heart	Item Bank
Typology of item	Questions and answers requiring simple reproduction of the study material	Closed and open-ended questions
Subjects	Language tested indirectly: Geography Culture History Institutions of the State Targeted questions regarding the KoS[2] in practice not included	Language tested directly: Geography Culture History Institutions of the State Targeted questions regarding the KoS–LIAM project (Rocca, Hamnes Carlsen, & Deygers, 2020) included
Type of exams	Interview	Written exams

the readability software designed by the Centre for the Greek Language (Ventouris & Rousoulioti, 2020) and were found to range from A2 to C1, according to the CEFR.

Little (2008) argued that the CEFR should always be adapted to the national context in which migrants work and live. In this light, the proposed language level for texts intended for citizenship exams depending on the type of naturalization should range from A2 to B1+ (Kokkinidou, Markou, Rousoulioti, & Antonopoulou, 2014). Language test exercises should focus on understanding the text and meaning of words, rather than their form, for candidates to give adequate answers to relevant questions, considering the realistic aspects of a question rather than the grammatical ones. On the other hand, questions and discussion during the interview should successfully meet migrants' daily needs, including their rights and obligations.

Citizenship tests in Greece from 2020 on

Citizenship exams in Greece changed since 2020 (Table 1). The Ministry of Interior of Greece, in implementation of law 4735/2020, introduced a substantial reform in the naturalization process of foreigners. The Certificate of Knowledge Adequacy for Naturalization was established, the acquisition of which is a prerequisite for the submission of the naturalization application. It is an examination process based on the model of high-stake exams that take place every year in Greece for those interested in studying at Greek public universities, to ensure the objectivity, universality, and integrity of the process. Participants in this process must take written exams in which they will be examined in the Greek language, Greek history and geography, Greek culture as well as the institutions of the country's regime. This is a B1 CEFR language test in Greek that includes all basic communication skills (listening, reading, speaking, and writing) along with 20 questions related to the other four main subject areas: geography, culture, history and state institutions.

Test items are selected randomly on the day of the exams from a thematic bank that is kept electronically under the care and responsibility of the General Secretariat for Citizenship of the Greek Ministry of Interior and is open to anyone interested. Success in the exams occurs when the candidate has 70% from 100% of the maximum possible score in all the examined subjects, provided that they have achieved at least 40% out of 60% in the exams of the Greek language and at least 20% out of 40% in the other subjects.

Conclusion

The citizenship decision is a complex decision, a long journey (Saville, 2009) at the end of which a migrant must take citizenship exams. From 2020 (Law 4735/2020) the exam process for the acquisition of citizenship in Greece changed. This article tried to shed light on what has changed and how, as well as what still needs to be done regarding citizenship exams in Greece. Citizenship tests in the Greek context could be improved in terms of:

[2] Knowledge of Society

- *transparency*: since authentic texts are used, their source should be cited;
- *practicality*: the time required for the test and the marking of each item must be written next to it;
- *design of the tests*: there must be an example at the beginning (Tsagari et al., 2018) for each task, as participants in citizenship tests are often unfamiliar with the typology of tests' tasks;
- *topics*: the topics of the texts should be closely linked to the daily life and personal experiences of candidates in Greece, to demonstrate their familiarity with the country and correspond to their real interests and the purpose of the specific exams (Tsagari et al., 2018), because, when 'the tests are used for immigration purposes, they should meet the highest standards of quality and fairness' (Rocca et al., 2020).

References

ALTE. (2016). *Language Tests for Access, Integration and Citizenship: An Outline for Policy Makers.* Strasbourg: Council of Europe. Available online: https://www.alte.org/resources/Documents/LAMI%20Booklet%20EN.pdf

Brander, P., De Witte, L., Ghanea, N. Gomes, R., Keen, E., & Justina, A. N. (2020). *Compass. Manual for Human Rights Education with Young People* (Second edition). Strasbourg: Council of Europe.

Bruzos, A., Erdocia, I., & Khan, K. (2018). The path to naturalization in Spain: Old ideologies, new language testing regimes and the problem of test use. *Lang Policy*, *17*, 419–441.

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge: Cambridge University Press.

DeSipio, L. (1987). Social science literature and the naturalization process. *International Migration Review*, *21*(2), 390–405.

Eurostat (2023). *Migration and migrant population statistics*. Available online: https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Migration_and_migrant_population_statistics#Acquisitions_of_citizenship:_EU_Member_States_granted_citizenship_to_827_300_persons_in_2021

Greek Ministry of Education. (n.d.). *Examinations for the Certificate of Knowledge Adequacy for Naturalization.* Available online: https://exetaseis-ithageneia.ypes.gr

Kokkinidou, A., Markou, B., Rousoulioti, T., & Antonopoulou, N. (2014). The contribution of the Common European Framework of Reference for Languages to teaching and assessment. In N. Lavidas, Th. Alexiou & A-M Sougari (Eds.), *Major Trends in Theoretical and Applied Linguistics 3* (pp. 163–182). Versita: Great Britain/De Gruyter Open: Poland.

Kokkinidou A., Rousoulioti T., Pasia A., Antonopoulou S. & Zervou, A. (2021). Naturalisation and the acquisition of citizenship: An overview, aspects and proposals. In M. Mathaioudakis, E. Griva & M. Moumtzi (Eds.), *Migration and Language Education in Southern Europe: Practices and Challenges* (pp. 54–73). Newcastle: Cambridge Scholars Publishing.

Law 4251/2014, Art. 107. Government Gazette 1.4.2014 & Joint Ministerial Decision No 30825/4.6.2014.

Law 4735/2020. Government Gazette A' 197/12.10.2020.

Little, D. (2008). *The Common European Framework of Reference for Languages and the Development of Policies for the Integration of Adult Migrants*. Available online: https://rm.coe.int/16802fc0b1

Rocca, L., Hamnes Carlsen, C., & Deygers, B. (2020). *Linguistic integration of adult migrants: Requirements and learning opportunities. Report on the 2018 Council of Europe and ALTE survey on language and knowledge of society policies for migrants*. Strasbourg: Council of Europe.

Saville, N. (2009). Language Assessment in The Management of International Migration: A Framework for Considering The Issues. *Language Assessment Quarterly*, *6*(1), 17–29.

Schengen visa info. (n.d.). *Schengen visa information*. Available online: https://www.schengenvisainfo.com/

Tsagari, D., Vogt, K., Froelich, V., Csépes, I., Fekete, A., Green A., Hamp-Lyons, L., Sifakis, N., & Kordia, S. (2018). *Handbook of Assessment for Language Teachers*. Available online: http://taleproject.eu/

Vasiliou, T., & Giavi, V. (Eds). (2001). *Greece. A Second Homeland*. Athens: Greek Ministry of Interior.

Ventouris, A., & Rousoulioti, T. (2020) Measuring the readability of texts for Reading Comprehension: the readability software of the Centre for the Greek Language. *Kathedra*, *6*(1), 111–133.

Inclusive formative assessment practices (IFAP) in Higher Education: Promoting education for social justice

Eleni Meletiadou
London Metropolitan University, United Kingdom

Abstract

The aim of the current study was to enhance students' motivation and writing performance and ensure that no student is left behind irrespective of their background. The project developed and piloted the Inclusive Formative Assessment Practices (IFAP) scheme in Higher Education (HE), taking into consideration the scarcity of research in HE and implementing more than one formative assessment method using a mixed-methods approach. This study was funded by London Metropolitan University, promoted its Education for Social Justice Framework, and explored the beneficial impact of inclusive modern educational assessment practices on student outcomes, overall experience, and continuous professional development. This project wished to inform scholarly debate around inclusive assessment practices that can enhance students' learning and motivation and cater for their diverse needs.

Introduction

The current study examined the use of peer assessment (PA) and digital portfolios as inclusive assessment methods that enhance undergraduate students' writing performances and willingness to write and learn in Business and Management Education. Inclusive assessment refers to a stance towards assessment in terms of which individual students' needs, disparities and perceptions are catered for, as much as possible, to ascertain that all learners have an opportunity to succeed by targeting their strengths rather than their shortcomings with the intention of revealing areas for development and helping them as they try to learn (Meletiadou, 2022). As educators in Higher Education (HE) have been welcoming increasingly diverse cohorts recently, inclusive assessment does not necessarily refer to students with special educational needs or disabled learners. Higher Education institutions (HEI) are particularly concerned about using inclusive teaching, learning and assessment practices to respond to the requirements of multilingual and multicultural students who need to work together and succeed first in their academic contexts and later in an increasingly complex workplace, which has high expectations from individuals who wish to find career-enhancing positions and succeed in their professional lives.

Literature review

Assessment as learning (AaL) has shifted the responsibility for learning from educators to students and is currently regarded as a significant alternative assessment approach that may increase student learning and engagement (Fung, Su, Perry, & Garcia, 2022). Peer assessment (PA), also referred to as peer review, is a ground-breaking AaL method which empowers learners as it invites them to reflect on and negotiate their learning process with their peers, allowing them to increase their academic performance as they are asked to take responsibility for their own learning by relying on themselves and their peers rather than their lecturer (Meletiadou, 2023; Yu & Liu, 2021). It is also described as 'a communication process through which learners enter into dialogues related to performance and standards' (Liu & Carless, 2006, p. 280). When involved in PA tasks, students can socially construct knowledge through the exchange of peer feedback which allows learners to detect problems in their texts. Subsequently, they are guided to take action to rectify their mistakes and resolve their cognitive conflict (Zhao, 2018). As Universities and tertiary education increasingly focus on self-reliance and collaborative learning (Voogt, Erstad, Dede, Mishra, 2013), educators experiment even more with the use of collaborative tasks that urge learners to become more active as they engage in learning to write (Loh & Ang, 2020).

There are six theories that support the use of PA and portfolio activities in the English as a Second Language (ESL) writing classroom from both cognitive and psycholinguistic perspectives: a) process writing theory, b) collaborative learning theory,

c) social cognitive theory, d) interaction and second language acquisition (SLA), e) cognitive constructivist theory, and f) self-regulation theory. These in fact complement and to some extent overlap each other. Research based on these theoretical stances has provided substantial evidence that PA and portfolio tasks help learners develop their writing, collaborative, and self-regulation skills through the negotiation of meaning that normally takes place during these activities (Lam, 2022; Topping, 2009).

Lately, an increasingly larger number of researchers and educators has been experimenting with portfolio assessment as the literature indicates that it may improve student learning, facilitate lecturers' work by decreasing their workload, improve the learning and assessment process (Yang, Tai, & Lim, 2016), promote autonomous learning (Tur, Urbina, & Forteza, 2019), increase students' attitudes toward learning (Beckers, Dolmans, & Van Merriënboer, 2016), and promote reflection and the development of metacognitive skills (Weber & Myrick, 2018). Evans, Hawes and Shain (1999) refer to portfolio assessment as 'an evolving collection of carefully selected or composed professional thoughts, goals, and experiences that are threaded with reflection and self-assessment. It represents who you are, what you do, why you do it, where you have been, where you are, where you want to go, and how you planned to getting there' (p. 147). Therefore, portfolios can show how individual students' learning evolves, their most significant milestones and challenges along their learning journey, and can be used as a reference to showcase their achievements in their future professional life.

Handwritten portfolios have now been replaced by digital portfolios as they are easier to use (Sanders, 2000) and students can also be more creative and employ additional resources to unravel their digital and even artistic skills by using, e.g., video clips or interactive elements. Digital portfolios are easier to store, more environmentally friendly and easier to share with the educators, friends, and possible future employers as they can be integrated in blogs, websites and shared through social media. This enables learners and future professionals to exchange ideas and artefacts among members of a learning and/or professional community and to become more innovative and creative as they can cooperate with diverse teams.

Method

The current study explored the impact of group PA and digital portfolios (DP) on 200 undergraduate first-year students' writing performances, development of professional skills and motivation towards learning. Its main goal was to provide an insight into students' viewpoints regarding the implementation of PA and portfolios in HEI classrooms with the aim of enhancing student academic achievement and willingness to engage in academic writing.

This semi-experimental study used a pre-test post-test design to explore the impact of PA and DP on students' writing skills. Students were asked to write a short report as a pre-test during the first week of the semester and then a short report at the end of the second semester as a post-test. After the pre-test students received training in PA and were then involved in one round of anonymous group PA which was followed by a second round of lecturer feedback, as PA was complementary to lecturer feedback. In the second semester, students were asked to create digital portfolios individually and were then also involved in anonymous group PA. The lecturer again provided feedback which was complementary to students' comments. The aim was to familiarise students initially with PA and then allow them to combine PA with digital portfolios as these allow students to be more creative. Students were invited to use these portfolios to showcase their achievements. They were gradually introduced first to PA and then to portfolio assessment in order not to intimidate them as they had not used any form of alternative assessment before.

Six lecturers implemented the scheme after receiving relevant training in implementing alternative assessment methods in their classes. They kept a diary and made notes during the implementation regarding the benefits and challenges they and their students encountered. Students were also invited to provide feedback regarding the implementation by writing a short report about their learning experience twice, first at the end of semester one and then at the end of semester two. Descriptive statistics were used to analyse the quantitative findings of the study and thematic analysis was used to analyse the findings from the lecturers' diaries and students' reports.

Summary of findings and discussion

Findings indicated that students increased their writing performance by almost 30% in two semesters. The researcher undertook this implementation as students complained about this module, did not attend the lectures, and submitted assignments of low quality. The current study indicated that when this specific scheme is implemented in large mixed-ability classes with multilingual and multicultural students, PA may help cover students' knowledge gaps, expand students' resources, and increase students' self-reliance, helping them improve their academic performance (Pintrich & Zusho, 2007). Moreover, taking into consideration participants' feedback and lecturers' observations, the combined use of PA and DP increased students' internal motivation as they were gradually trained to effectively plan, present, and assess assignments within a short period (Syzdykova, Koblandin, Mikhaylova, & Akinina, 2021) and use their creativity to design their own portfolios and enrich them with many interactive elements. This increased student retention, attendance, involvement in the module, co-creation of the module so that

they could eliminate elements that decreased their performance and attitude to write and learn. Asking students to read each other's work and create their own digital artefacts teaches them a range of learning skills and fosters more self-reliant thinking and reflection on a deeper level (Weaver & Esposto, 2012). PA enables students to develop their academic and professional skills by focusing on feedback based on comparison and contrast with their own work (Topping, 2017). Moreover, PA and DP can be used as inclusive assessment strategies as they help low-achieving students develop their reflective skills and detect their strengths and weaknesses. They both stimulate higher-order skills and promote critical thinking, increasing student engagement, interaction, and interest in learning, and have numerous affective benefits such as ownership and confidence building (Topping, 2017), social and transferrable skills which will be helpful in future studies, and work-learned skills which include teamwork, verbal and written communication, problem-solving, constructive criticism, mindfulness, and diplomacy (Nortcliffe, 2012). Finally, students confessed that when PA and DP are used alongside lecturer's feedback, they can assist them in enhancing their academic achievement.

However, the lecturers who participated in this study revealed that students can be reluctant to accept PA to improve their learning products because they often doubt its accuracy and the proficiency of the provider (Panadero, 2016). Therefore, training and instructional scaffolds e.g., rubrics should be used to support learners' engagement in PA and support their digital and creative skills while preparing their digital portfolios. Learners, especially international multilingual and multicultural students, also confessed that the use of PA and DP helped them shape good writing and reflective habits so that they could complete their written tasks more effectively.

Lecturers also detected that multi-PA can provide more total feedback than from an increasingly busy lecturer supporting a large mixed-ability class, more convincing feedback when several reviewers identify the same problems, and feedback reflecting more varied audience perspectives. Students also recognised the value of digital portfolios and PA in developing their organization, meta-cognition and the role of lecturer– students' partnership as learners can work in a non-threatening environment.

Implications and conclusion

Despite increasing interest, PA and DP still remain marginalized as assessment methods in HEI (Nicol, Thomson, & Breslin, 2014) as lecturers still control the learning and assessment process and prevent students' creativity and involvement in and development through assessment (Spiller, 2012). There is growing literature about the impact of PA and DP on learners' attitudes and writing performance (Barbera, 2009), but there is a need for more studies to show how the design and implementation of PA can be made more effective. This paper aspired to contribute to this growing literature by focusing on what works and what does not in HEI for lecturers who implement PA and DP to improve their undergraduate students' learning experience.

References

Barbera, E. (2009). Mutual feedback in e-portfolio assessment: an approach to the netfolio system. *British Journal of Educational Technology*, *40*(2), 342–357.

Beckers, J., Dolmans, D., & Van Merriënboer, J. (2016). e-Portfolios enhancing students' self-directed learning: A systematic review of influencing factors. *Australasian Journal of Educational Technology*, *32*(2).

Evans, M., Hawes, R. H., & Shain, C. (1999). Does portfolio assessment have a place in history and social studies programs?. *Canadian Social Studies*, *34*(1), 146–149.

Fung, C. Y., Su, S. I., Perry, E. J., & Garcia, M. B. (2022). Development of a socioeconomic inclusive assessment framework for online learning in higher education. In M. B. Garcia (Ed.), *Socioeconomic Inclusion During an Era of Online Education* (pp. 23–46). Pennsylvania: IGI Global.

Lam, R. (2022). E-Portfolios for self-regulated and co-regulated learning: A review. *Frontiers in Psychology*, *13*, 1079385.

Liu, N. F., & Carless, D. (2006). Peer feedback: The learning element of peer assessment. *Teaching in Higher Education*, *11*(3), 279–290.

Loh, R. C. Y., & Ang, C. S. (2020). Unravelling cooperative learning in higher education. *Research in Social Sciences and Technology*, *5*(2), 22–39.

Meletiadou, E. (2022). The Use of Peer Assessment as an Inclusive Learning Strategy in Higher Education Institutions: Enhancing Student Writing Skills and Motivation. In E. Meletiadou (Ed.), *Handbook of Research on Policies and Practices for Assessing Inclusive Teaching and Learning* (pp. 1–26). Pennsylvania: IGI Global.

Meletiadou, E. (2023). Transforming multilingual students' learning experience through the use of Lego Serious Play. *IAFOR Journal of Education*, *11*(1), 1–24.

Nicol, D., Thomson, A., & Breslin, C. (2014). Rethinking feedback practices in higher education: a peer review perspective. *Assessment & Evaluation in Higher Education*, *39*(1), 102–122.

Nortcliffe, A. (2012). Can students assess themselves and their peers?: A five year study. *Student Engagement and Experience Journal*, *1*(2).

Panadero, E. (2016). Is it safe? Social, interpersonal, and human effects of peer assessment: A review and future directions. In G. T. L. Brown & L. R. Harris (Eds.), *Handbook of Human and Social Conditions in Assessment* (pp. 247–266). New York: Routledge.

Pintrich, P. R., & Zusho, A. (2007). Student motivation and self-regulated learning in the college classroom. *The Scholarship of Teaching and Learning in Higher Education: An Evidence-based Perspective*, 731–810.

Sanders, M. (2000). Web-based Portfolios for Technology Education: A Personal Case Study. *Journal of Technology Studies*, *26*(1), 11–18.

Spiller, D. (2012). *Assessment Matters: Self-assessment and Peer Assessment*. Hamilton: The University of Waikato.

Syzdykova, Z., Koblandin, K., Mikhaylova, N., & Akinina, O. (2021). Assessment of E-portfolio in higher education. *International Journal of Emerging Technologies in Learning (iJET)*, *16*(2), 120–134.

Topping, K. J. (2009). Peer assessment. *Theory into Practice*, *48*(1), 20–27.

Topping, K. J. (2017). Peer assessment: Learning by judging and discussing the work of other learners. *Interdisciplinary Education and Psychology*, *1*(1), 1–17.

Tur, G., Urbina, S., & Forteza, D. (2019). Rubric-Based Formative Assessment in Process Eportfolio: Towards Self-Regulated Learning. *Digital Education Review*, *35*, 18–35.

Weaver, D., & Esposto, A. (2012). Peer assessment as a method of improving student engagement. *Assessment & Evaluation in Higher Education*, *37*(7), 805–816.

Weber, K., & Myrick, K. (2018). Reflecting on Reflecting: Summer Undergraduate Research Students' Experiences in Developing Electronic Portfolios, a Meta-High Impact Practice. *International Journal of ePortfolio*, *8*(1), 13–25.

Voogt, J., Erstad, O., Dede, C., & Mishra, P. (2013). Challenges to learning and schooling in the digital networked world of the 21st century. *Journal of Computer-assisted Learning*, *29*(5), 403–413.

Yang, M., Tai, M., & Lim, C. P. (2016). The role of e-portfolios in supporting productive learning. *British Journal of Educational Technology*, *47*(6), 1,276–1,286.

Yu, S., & Liu, C. (2021). Improving student feedback literacy in academic writing: An evidence-based framework. *Assessing Writing*, *48*, 100525.

Zhao, H. (2018). Exploring tertiary English as a Foreign Language writing tutors' perceptions of the appropriateness of peer assessment for writing. *Assessment & Evaluation in Higher Education*, *43*(7), 1,133–1,145.

An education action plan to improve assistance to Autistic Spectrum Disorder (ASD) test-takers in written large-scale exams

Anarcisa de Freitas Nascimento
Brazilian National Institute of Educational Studies and Statistics

Gladys Quevedo Camargo
University of Brasília

Abstract

The growing presence of test-takers with Autistic Spectrum Disorder (ASD) in the Brazilian National High School Exam (Exame Nacional do Ensino Médio; Enem) and the need of specialized assistance in the writing component of the exam inspired the present study. This article shows analyses of accommodation procedures offered to the demands for specialized attendance in writing for students with ASD in Canada, the United States and the United Kingdom, and highlights the linguistic characteristics of such candidates based on the analyses of texts produced by Enem 2017 ASD participants. The two types of analyses resulted in an Educational Action Plan[1] for Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep) management teams to improve the training plans for readers and transcribers of the exam for students with autism, the logistical support of the exam, the application of a written test on a computer and the inclusion of linguistic resources.

Introduction

Every year, the Brazilian National Institute of Educational Studies and Research (Instituto Nacional de Estudos e Pesquisas Anísio Teixeira – Inep) applies the National High School Exam (Exame Nacional do Ensino Médio – Enem). This exam has been administered since 2000. At first, its aim was to evaluate secondary school students' knowledge after they completed the course, but since 2009 it has been used to allow access to higher education for young adult learners in many Brazilian universities and colleges by means of a unified entrance system.

Since the beginning of the administration of Enem, specialized attendance has been a challenge, mainly for those participants who have Autistic Spectrum Disorder (ASD), with special difficulties in written tests. The number of participants with ASD grew from 45 in 2012 to 3,031 participants in 2022, according to Nascimento (2020) and Inep's information available in the Brazilian government website (2023)[2]. Therefore, in addition to facing problems in written tests, ASD participants also have problems with the accommodations that are offered to assist them in the moment of the exam.

Taking this scenario into account, the aim of this article is to briefly report an MA study made by Nascimento (2020), whose main objective was to enhance accommodations for ASD participants in Enem. This report is divided into two parts to facilitate the development of the study. Firstly, we present a comparative analysis of accommodation procedures for specialized attendance in writing for students with ASD in Canada, the United States and the United Kingdom. The accommodation procedures were divided into three groups: infrastructure, assistive technology resources, and pedagogical accommodations for better comprehension and detailing. Secondly, we highlight the linguistic characteristics of the Brazilian ASD candidates based on the analyses of texts produced in the Enem 2017 edition.

[1] https://www.gov.br/inep/pt-br/assuntos/noticias/enem/mais-de-35-mil-terao-atendimento-especializado-no-enem-2022
[2] https://repositorio.ufjf.br/jspui/bitstream/ufjf/12102/1/anarcisadefreitasnascimento.pdf

Accommodations in writing for students with ASD in Canada, the United States and the United Kingdom

The study analysed educational documents from entities, universities and government education bureaus, in order to display leading practices and accommodations for ASD students both in basic and higher education systems. These analyses considered infrastructure, assistive technology resources and pedagogical accommodations received by ASD participants in basic education and also during test application. After the analyses, the main points that were innovative and practical in ASD assistance were outlined.

The main advances in Canada's ASD accommodations were the analysis of the participants' documents to define possible procedures to enable their attendance in writing exams and the use of assistive technology by these participants.

The most important aspects in the United Kingdom's ASD accommodations were the presence of an Oral Language Modifier (a professional responsible for interpreting and easing language aspects, such as metaphors, irony, jokes, inferences, implicit coherence, etc.), and the use of a prompter to guide participants in comprehension procedures.

The improvements identified in the USA were the use of assistive technology, the application of instructions in written test items, and loud voice reading of these instructions to ASD participants in tests.

Linguistic characteristics of Brazilian ASD candidates based on the analyses of texts produced in the Enem 2017 edition

This study involved 20 samples of ASD male and female participants in Enem 2017, distributed by the five Brazilian regions. The grades of the selected texts ranged from under the minimum score for approval (450 points) to next to the maximum grade (1,000 points). The analysis took in account linguistic characteristics that are common in ASD writing, such as pronominal inversion, phonological writing, echolalia, and textual text clipping, developed during my master's degree dissertation. These characteristics are reported in Nascimento (2020), based on Enem's text samples.

Data analysis and discussion

The two types of analyses – the experiences abroad and the textual analysis – resulted in an Educational Action Plan based on Nascimento (2020) for Inep's management teams so as to improve: the training plans for readers and transcribers of the exam for students with autism; the improvement of the logistical support of the exam, with the inclusion of the professional Oral Language Modifier and differentiated ambience resources, and the application of a written test on a computer.

As a result of the analyses we could suggest, as improvement, the inclusion of linguistic resources such as pronominal inversion, phonological writing, echolalia and textual text clipping, typical characteristics of students with ASD, empirically verified in the sample and essays, in the differential correction blueprint for ASD.

Final thoughts and suggestions for future studies

It is important to deepen the studies about the linguistic features of ASD participants, mainly those related to written activities. The growth in studies involving writing in large-scale exams for young and adult ASD test-takers might stimulate their participation in exams and give them support to express themselves in writing. In addition, it is highly relevant to research the impact of narratives or authorial texts in large-scale assessment for ASD participants, considering their school trajectory and the accommodations received during their primary education.

The use of assistive technology may be considered in the studies for the development of large-scale written tests because of its different and interactive resources for the design of written exams. Also, it is important to adjust the linguistic characteristics of ASD participants to the correction blueprint.

Last but not least, the presence of an oral language modifier, and of transcribers and readers, are fundamental to help ASD participants to overcome misunderstandings, irony, jokes, inferences, implicit information, metaphors and other linguistic barriers during test application.

Reference

Nascimento, A. de F. (2020). *Aprimoramento do atendimento especializado para pessoas com Transtorno do Espectro Autista na redação do Enem* [MA dissertation].

Bias is everywhere? An investigation into differential functioning on the item, rater and task level

Christine Troussart Van Bulck
KU Leuven, Belgium

Goedele Vandommele
KU Leuven, Belgium

Terry Willems
Radboud University, Netherlands

Anne van Asseldonk
Radboud University, Netherlands

Abstract

Certificate of Dutch as a Foreign Language (CNaVT) annually develops five examinations to evaluate Dutch language proficiency. Among these, the B2 *Educatief Startbekwaam* exam (STRT) is taken by candidates worldwide to gain entry into Higher Education Institutions (HEIs) in Belgium and the Netherlands, where Dutch serves as the primary medium of instruction.

This unique examination context necessitates a delicate balance: the need for authentic and contextually relevant tasks aligned with current practices in Belgian and Dutch HEIs, while simultaneously ensuring that cultural and other biases are minimized to ensure equitable testing for all candidates. Consequently, bias research becomes a critical facet of the standard STRT exam development and management process.

To guarantee fair testing for all candidates, CNaVT conducts annual differential item functioning (DIF) analyses at item level to investigate which items cause bias regarding age, sex, linguistic, cultural and educational background, and to disregard items that show bias. These analyses show DIF for approximately 4% of items. Cultural and linguistic bias is found on language accuracy scores, whereas content items are overwhelmingly free from DIF (Troussart Van Bulck et al., 2022).

However, as CNaVT tests language performances via integrated and task-based assessment, the absence of bias at item level might camouflage bias on other relevant aspects of a task-based assessment, i.e., the rater and task level. In our current research, we investigated whether bias is found at the rater and task level and if so, which contexts and/or linguistics tasks give rise to bias. For this analysis, we combined all datasets from 2017 to 2022 into one large dataset (N = 3,700; raters = 89; total tasks = 18). Differential rater functioning (DRF) analysis reveals non-systemic DRF, while differential task functioning (DTF) analysis shows that 6 out of 18 tasks reveal DTF in spite of the individual items not revealing DIF. Bias mainly concerns the candidate's native language, country of birth and country of assessment. Even though the results are encouraging, it remains important that task-based assessments monitor for DTF.

Introduction

Certificate of Dutch as a Foreign Language (CNaVT) offers five task-based exams ranging from the A2 to C1 level on the Common European Framework of Reference for Languages (CEFR, Council of Europe, 2001). The most administered exam is the B2 level *Educatief Startbekwaam* exam (STRT), which targets 17- to 20-year-old candidates who wish to gain access to Dutch Higher Education (HE) in Belgium and the Netherlands. The STRT exam is currently administered in 42 countries all over the world. This exam context leads to conflicting demands: the exam must consist of tasks that are authentic and relevant within the context of HE in Belgium and the Netherlands, while cultural (and other) bias must be minimized to ensure fair testing for all candidates.

Several steps are taken during the exam development process to avoid bias in the STRT exam tasks, such as the use of an international feedback group of language teaching and assessment professionals. In spite of these measures, a DIF analysis in Tiaplus revealed that 4% of all STRT exam items between 2017 and 2021 (N = 3,700, total items = 334) show DIF. Cultural and

linguistic bias is found on language accuracy scores, whereas content items are overwhelmingly free from DIF. We suspected that the STRT exam items are not inherently biased, but that bias occurs because of rater inconsistency (Troussart Van Bulck et al., 2022).

Research question

The current research question is twofold. We want to (1) investigate the suspected presence of differential rater function (DRF) in the STRT exam; and (2) investigate whether the STRT exam displays differential task functioning (DTF) even though content items appear free from bias, since the exam follows a task-based framework and the overarching task may display bias even though its individual items do not.

Methods

In order to investigate potential differential functioning at rater and task level, all datasets since 2017 were combined into one large dataset to detect previously unnoticed trends. The basic dataset used for score analysis consists of three facets, namely candidates (N = 3,700), raters (N = 89) and items (N = 246). This dataset was used to estimate measures in a Facets Rasch analysis.

We then added five bias specifiers to the dataset to analyze the role of the rater and the task, namely the candidate's age, native language, sex, country of birth and country of assessment. Data for age, native language, country of birth and country of assessment had to be recoded into broader categories so the minimal population size for each group could be achieved. Background information regarding the candidate's highest education level was also collected, but turned out to be unreliable due to the large differences among national education systems across the world. As a result, it was not included in the analysis.

A final facet was added for task ID (N = 18), bringing the total number of facets in the dataset to 9.

In order to test the first hypothesis that bias might be introduced at the rater level, a DRF analysis was conducted in Facets by investigating the bias interactions between each rater and the five bias specifiers. In order to test the second hypothesis that bias might be found at task level, a DTF analysis was conducted in Facets by investigating the bias interactions between each task and the five bias specifiers. Finally, tasks that showed DTF were reviewed among the CNaVT team members and presented to the audience at the 2023 International ALTE Conference.

Results

Rater bias

Regarding bias at rater level, there appears to be significant moderate to severe bias regarding the candidate's native language, age and country of birth and assessment for certain raters. Out of 89 raters, eight show significant bias for country of assessment, five for country of birth and four for native language. Zero raters show significant bias for age or sex.

Rater bias is not systemic, meaning that individual raters have different bias tendencies. With regard to bias for country of assessment and country of birth, this means that no specific region is affected more than others. Most major regions are favoured by some raters, while being disfavoured by other raters. With regard to bias for native language, it is mainly native Dutch speakers who are most affected. However, these speakers are favoured by some raters, while being disfavoured by other raters.

As a result, it is unclear whether rater bias has any impact on the candidates' total score. Because each candidate is rated by multiple raters, it is likely that various instances of rater bias cancel each other out or, at the very least, that the impact of the rater bias on the total score is minimized. In addition, rater bias only explains 0.10 to 0.24% of the variance among candidates.

Bias at task level

Out of 18 tasks, six show significant moderate to severe bias for country of assessment, four for country of birth, two for native language and one for age. Oral tasks are not more prone to bias than written tasks.

The results of the DTF analysis parallel the results of the DRF analysis. With regard to regional bias, no specific region seems to be affected more than others. With regard to bias for native language, native Dutch speakers are most affected, but native speakers overperform in one task, while underperforming in another task. Since all STRT exams consist of six different tasks, it is once again unclear to what extent bias at task level actually affects the participants' exam results. It is possible that various

instances of bias cancel each other out or, at the very least, that the impact of bias at task level on the candidate's total score is minimized.

Tasks that show DTF for multiple candidate groups were reviewed by CNaVT team members and presented to the ALTE audience in order to identify why bias emerged. The findings are mixed.

For some tasks, hypotheses about the potential source of bias could be formulated. For instance, according to the DTF analysis, candidates over 26 years old underperform on a writing task in which they have to write an application letter for the university's honours programme. Since honours programmes are a relatively new trend within Dutch-speaking HE, we suspect that older candidates might be less familiar with such programmes than younger candidates who are about to enroll in HE and are thus more familiar with the programmes currently offered.

For some other tasks, however, it is unclear why DTF occurs. For instance, Dutch native speakers and candidates in North America underperform on a speaking task about the candidate's preferred study methods, while candidates in Africa overperform on this task. Because candidates are free to choose their preferred study methods from a long list of common study methods, it seems unlikely that this task favours candidates in certain education systems over others. In such cases, it might be possible that DTF emerges because of statistical anomalies in the dataset. The fact that North America and Africa are both rather small candidate population groups supports this hypothesis.

Discussion

In this paper, we wanted to look into bias at rater and task level for task-based language exams giving access to Dutch-speaking higher education. The Facets DRF analysis of all STRT exams between 2017 and 2021 revealed significant moderate to large, but non-systemic DRF for 8 out of 89 raters. DRF concerns the candidate's native language, country of birth and country of assessment. These results are encouraging, as rater bias is highly personal and no group of candidates is affected more than others. We believe the lack of systemic rater bias might be due to the extensive rater training and the concise rating criteria. Furthermore, these results support our current practice of using multiple raters for each candidate so any potential rater bias is cancelled out or its impact on the candidate's total score is minimized.

Regarding bias at task level, 6 out of 18 tasks showed significant moderate to large DTF. DTF mainly occurs based on the candidate's native language, country of birth and country of assessment. We suspect that some instances of DTF might be the result of small candidate populations, since no likely source of bias could be identified upon review. Furthermore, we expected that oral tasks would be more prone to bias than written tasks, because candidates' background characteristics are more salient in an oral recording than in a written text. However, the DTF analysis showed no significant difference between oral and written tasks. These findings support our current practice of not anonymizing oral performances for rating.

The presence of DTF is noteworthy, since the DIF analysis of individual content items did not reveal significant bias. As a result, DTF is an important avenue of research for exams that are grounded in task-based assessment of proficiency.

Another interesting future avenue of research would be to investigate to what extent DTF actually affects the candidates' exam results and future opportunities. This could be done by adjusting the scores in the current dataset for DTF and identifying whether this adjustment leads to discrepancies in candidates' pass/fail results. Such an analysis would offer more insight in the severity of the DTF uncovered here.

Recommendations

We formulate the following recommendations based on the findings of our differential functioning analysis at rater and task level.

Because rater bias is highly individual, we believe it is important to ensure that each candidate is rated by multiple raters. If using multiple raters per candidate is not viable, we recommend conducting a DRF analysis and adjusting the affected candidates' scores if DRF is found.

In addition, we do not recommend anonymizing candidate recordings, as raters do not seem to be strongly affected by candidates' background characteristics that can be inferred from their oral performance.

Lastly, we strongly recommend that task-based exams conduct a differential functioning analysis at task level, since tasks as a whole can reveal bias that is not present at the level of the individual content items.

References

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge: Cambridge University Press.

Troussart Van Bulck, C., Wynants, J., van Asseldonk, A., Alferink, C., Vandommele, G., & Willems, M. (2022, 31 August). *Avoiding bias in CNaVT language exams through qualitative research procedures during the exam development process* [Conference presentation]. 9th International Conference on TBLT, Innsbruck, Austria.

Implementation of Frameworks

Aligning language education to the CEFR: Whys, whats and hows

Neus Figueras Casanovas
Universitat de Barcelona, Spain

Abstract

This paper addresses the content and recommendations in the Common European Framework of Reference for Languages (CEFR) and its Companion Volume (CEFR CV) (2001; 2020, respectively), what they mean, what they imply and how they have been implemented. It also argues why the uses of the CEFR should be revisited in order to 'go deeper into' a document which, although published 20 years ago and widely influential, is seemingly still not fully understood. Implementations of the CEFR by professionals working in different fields are discussed, and the Handbook (2022) for aligning education to the CEFR, which aims to contribute to a fuller understanding of the CEFR and to facilitate its use in language education is also examined. The processes to follow when aligning curricula, materials and assessments to the CEFR are outlined and readers are encouraged to put the Handbook to use in their own contexts and share their experiences.

The CEFR in 2023

ALTE 2023 Madrid included a good number of CEFR-related talks, which focused on its uses in different contexts and in different countries, mostly in the context of assessment and on the 'new' in the Common European Framework of Reference for Languages Companion Volume (CEFR CV; Council of Europe, 2020). This paper reaches beyond testing and assessment into language education and, because of this, takes some steps back in time to refer briefly to the work of the Council of Europe in the late 1970s and to the impact of the publication of the CEFR in 2001.

Despite the fact that CEFR level labels seem to have become common currency, other aspects in the document do not seem to have reached classrooms or language education policies (Berger, 2018; Little & Figueras, 2022; amongst others). The CEFR (2001) states how educational systems need to give importance to being comprehensive, transparent and coherent in their objectives and methods:

With regard to educational systems, coherence requires that there is a harmonious relation among their components:

- the identification of needs;
- the determination of objectives;
- the definition of content;
- the selection or creation of material;
- the establishment of teaching/learning programmes;
- the teaching and learning methods employed;
- evaluation, testing and assessment. (2001:7)

Although this breakdown of components is very clear and many would agree that it constitutes an admirable list of statements, the harmony it advocates cannot be seen in many language education policies and programmes. Little progress seems to have been made in 20 years, and we should ask ourselves not only why this has this been the case but also how the situation can be redressed.

On the one hand, the reason for the lack of a full implementation of the CEFR contents can be found in the rapid and rather foolhardy implementation of the level labels, with many users limiting their knowledge of the document to Table 1, Common Reference Levels in Chapter 3 (2001 p. 24). The CEFR is present in schools, but the impact has been weak; it may be present at the macro level, in curricula, but not so much in lesson plans; its approach and recommendations are referred to rather than applied in real fact; and when used the focus is on concrete elements, sections or descriptor scales.

On the other hand, the nature of the document itself, with the many criticisms received, deserved and undeserved, also needs to be considered. Practitioners found the document not easily accessible or user friendly, missed sublevels, could not use it with young learners or for higher education and could not cope with the vagueness in some descriptors. Researchers claimed that the document lacked a scientific basis and that the progression in the scales was not grounded in language acquisition theory.

The Council of Europe has been taking many steps to facilitate the dissemination of the contents of the CEFR and promote its use. As early as 2008, and following an Intergovernmental Forum on the impact of the CEFR in Europe in 2007, it released a Recommendation to the member states on how to foster good practices in the use of the document. The Council of Europe has also organized training events, funded projects and prepared guidebooks and documentation illustrating good practices in different contexts and for different purposes. The Platform of resources and references for plurilingual and intercultural education in the Council of Europe website and also the ECML (European Centre for Modern Languages) website show the rich array of events and documentation. More recently, and taking on board the criticisms and requests received and also the new needs in language education in Europe, the Council of Europe prepared a companion volume (CEFR CV) that was published in its final version in 2020. This publication complements the scales in the CEFR (Council of Europe, 2001), presents its key messages in a user-friendly form, and provides links and references to consult the chapters of the 2001 edition. It is important to highlight however, that in the CEFR CV, the main concern of the 2001 CEFR remains: 'It aims to facilitate transparency and coherence between the curriculum, teaching and assessment within an institution and transparency and coherence between institutions, educational sectors, regions and countries' (Council of Europe, 2020, p. 27).

This quote, which takes readers of the CEFR CV back to a key objective in the 2001 CEFR needs to be at the centre of all the developments and initiatives triggered by the new additions in the CEFR CV. Access to the CEFR CV should result in a renewed interest in the principles that underlie the CEFR's descriptive scheme, going beyond level labels and achieving coherence in language education systems. Rather than focusing on misunderstandings or on absences or gaps in the document, it is more productive to focus on what the CEFR can help improve (see North (2020) for an argued response to criticisms to the CEFR).

It seems that in 2023, more than two decades after the publication of the CEFR, the profession is where it was, somehow wiser but still faced with the need to tackle the rather complex task of really using the CEFR beyond language level labels and achieving coherence in language education systems, and addressing diversity in a transparent way. The document presented in the following sections aims at contributing positively to this endeavour.

What is alignment and why is it important?

Researchers (Biggs, 1996; Tyler, 1969; van Lier, 1996; amongst many others) have been arguing for the need of all the elements (purpose, teaching, assessment) involved in any education system to be aligned and, thanks to the publication of the CEFR CV, discussions on how to best improve language education seem to have come to the forefront. Amongst the many reactions to the publication of the CEFR CV, a symposium that took place in London in 2020 (see O'Dwyer, Hunke & Schmidt (2020) for a summary) stands out for the follow-up documents it has generated (Byram, Fleming, & Sheils, 2023; Little & Figueras 2022), the most salient one being the Handbook published in April 2022.

The London 2020 symposium concluded that alignment should apply not only to language tests but to policy, curriculum guidelines, curricula, syllabuses, textbooks and other teaching/learning resources. Unfortunately, and although these elements impact significantly on one another and on learning, curriculum developers, materials developers, teacher trainers and assessment specialists mostly work independently of one another. Alignment has obvious advantages, and takes us back to the Council of Europe's aim for language education (referred to in the preceding section). The answer to the question of why it is worth aligning language education to the CEFR is that it contributes to:

- achieving systemic coherence and transparency
- establishing a basis for principled comparison
- monitoring for purposes of quality assurance.

The convenors of the symposium took the responsibility as a steering group to put together a Handbook that would help interested parties to better understand what alignment entails and provide a structured set of procedures that helped complete an alignment process. They used O'Sullivan's (2020) concept of a comprehensive learning system – CLS – which integrates all the key elements involved in any education system as the basis. O'Sullivan (2020) argues that the success of any learning system depends on the close alignment of all the elements involved and points out that if one of these elements is in any way disconnected from the others, then the system is under threat.

The Handbook, published in April 2020 thanks to the collaboration of many professionals who contributed text and also provided feedback, organizes its contents so as to emphasize their relevance to different stakeholders working in different domains and contexts.

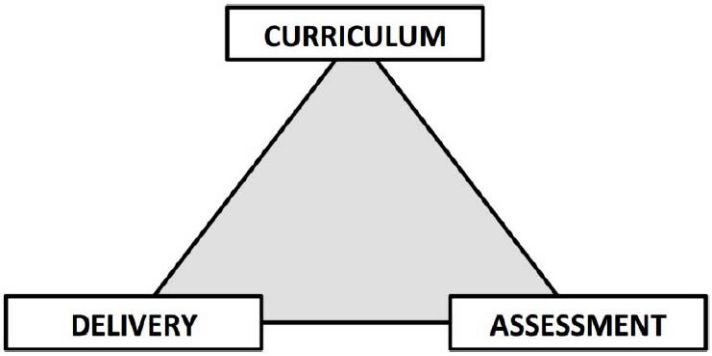


Figure 1 O'Sullivan's Comprehensive Learning System (CLS)

The three core elements of the CLS in the Handbook, with its illustrative triangle are described below.

Curriculum: Informal as well as formal

1. **Delivery**: Includes teacher selection; teacher training; accreditation; professional development and leadership; teaching and learning materials; the physical environment in which the delivery takes place
2. **Assessment**: Includes developmental assessment (diagnostic, aspects of progress, formative, etc.) and judgemental assessment (placement, aspects of progress, achievement, proficiency, etc.)

Before proceeding in more detail on the contents of the Handbook, there are three key points that need to be highlighted about the document:

- It focuses on the alignment of all elements involved in education.
- It is the result of collaboration between four leading organizations in the field (ALTE, British Council, EALTA, UKALTA).
- It is freely accessible online.

Alignment steps in the Handbook

The Handbook replicates to quite an extent the structure of the Council of Europe's 2009 Manual for Relating Examinations to the CEFR, as can be seen below. Yet, it differs from that document in that it goes beyond assessment and suggests different pathways depending on the needs and purposes of the user, their context and the resources at their disposal.

The different chapters in the Handbook describe five alignment steps: **familiarization**, an essential stage at the outset of any alignment exercise; followed by **specification**, which implies the analysis of the content(s) of any resource, existing or new, in terms of approach and coverage in relation to the categories presented in the CEFR; and by **standardisation**, the process of establishing that the main features of a given resource reflect a clear understanding of the relevant CEFR levels and descriptors. For those involved in all aspects of establishing an empirically-based link between a curriculum, a set of materials (e.g., textbook or online course), or an assessment or test, a fourth step is described, **standard setting**. The last, fifth step, **validation**, is presented as the continuous process of quality monitoring in order to gather the evidence to support claims of CEFR alignment.

Each chapter in the Handbook suggests different pathways, whether the user approaches the process individually, or whether the alignment process is planned as a group approach and involves one or more coordinator(s) and participants. Some practical advice and suggestions relating to these differing roles are included, including tasks to be completed with likely timings.

The last two sections in each chapter, the 'Guidelines for reporting' and the 'Notes for your own implementation', are meant to encourage reflection. The 'Guidelines' provide users with suggestions on how best to document their decisions and report the outcomes. The 'Notes' constitute a final reminder of the essential components of the activities presented and how to carry them out in a way that is relevant for their context and can guarantee success.

The appendix to the Handbook contains photocopiable summary forms to use and complete. This additional practical tool can assist users in their ongoing monitoring and validation throughout the alignment process, which can be useful when reporting on the process completed. The forms can be used as they are presented (photocopiable version) or adapted (editable version) to fit the needs of a particular alignment approach or resource.

Conclusion

The CEFR has been an indispensable reference point at all levels in language education since its publication in 2001. The CEFR and the CEFR CV, together with their many related documents, have become a *de facto* open-source apparatus that is not always used responsibly or in a valid way. Those involved in the many initiatives springing from the publication of the CEFR CV should take the opportunity to promote an accountable use of the CEFR apparatus which contributes to coherent language education systems.

The Handbook described in this chapter is just one initiative to improve language education, and focuses on one of the keys for success: alignment of all the elements involved. It is not possible to predict the impact of the document. However, given the huge impact of the 2001 CEFR and the interest raised by the CEFR CV, it is to be expected that this first edition will be put to use in alignment projects by many different stakeholders in the field of language education. We would like to invite those involved in such projects to share not only their outcomes but also their views on the usefulness of the Handbook. There are plans to host an event in October 2024 in Barcelona to present case studies and good practices in the use of the Handbook so that the resulting suggestions and proposals can be incorporated in a future edition.

References

Berger, A. (2018, 27 January). *What the new can-do descriptors can do for classroom assessment* [Conference presentation]. 6th EALTA CEFR SIG, Trinity College Dublin.

Biggs, J. (1996). Enhancing teaching through constructive alignment. *Higher Education*, *32*, 347–364.

British Council, UKALTA., EALTA., & ALTE. (2022). *Aligning Language Education with the CEFR: A Handbook*. Available online: http://www.ealta.eu.org/documents/resources/CEFR%20alignment%20handbook.pdf

Byram, M., Fleming, M., & Sheils, J. (Eds.). (2023). *Quality and equity in education. A practical guide to the Council of Europe vision for education for the plurilingual, intercultural and democratic citizenship*. Bristol: Multilingual Matters.

Council of Europe. (2001). *Common European Framework for Languages: Learning, Teaching, Assessment.* Cambridge: Cambridge University Press.

Council of Europe. (2008). *Recommendation CM/Rec(2008)7 to member states on the use of the Council of Europe's 'Common European Framework of Reference for Languages' (CEFR) and the promotion of plurilingualism.* Available online: https://search.coe.int/cm/Pages/result_details.aspx?ObjectId=09000016805d2fb1

Council of Europe. (2009). *Relating Language Examinations to the CEFR. A Manual.* Available online: https://rm.coe.int/1680667a2d

Council of Europe. (2020). *Common European Framework for Languages: Learning, Teaching, Assessment.* Strasbourg: Council of Europe Publishing.

Little, D., & Figueras, N. (Eds.). (2022). *Reflecting on the Common European Framework of Reference for Languages and its Companion Volume*. Bristol: Multilingual Matters.

North, B. (2020). Trolls, unicorns and the CEFR: Precision and professionalism in criticism of the CEFR. *CEFR Journal – Research and Practice*, *2*, 8–24.

O' Dwyer, F., Hunke, M., & Schmidt, G. (2020). The EALTA UKALTA 'Roadmap' conference— The CEFR: A road map for future research and development—meeting overview. *CEFR Journal – Research and Practice, 2*, 89–97.

O'Sullivan, B. (2020). *The Comprehensive Learning System.* London: British Council.

Tyler, R. (1969). *Basic Principles of Curriculum and Instruction*. Chicago: University of Chicago Press.

van Lier, L. (1996). *Interaction in the Language Curriculum: Awareness, Autonomy, Authenticity*. New York: Routledge.

Aligning a multimodal integrated speaking assessment task to the Common European Framework of Reference for Languages

Kim Anne Barchi
University of Padova, Italy

Mariana Jo Bisset
University of Padova, Italy

Abstract

The present study is a part of a wider pilot project aimed at aligning multimodal integrated speaking performances to the Common European Framework of Reference for Languages (CEFR) (Council of Europe, 2001; 2020) descriptors.

The study focuses on Task 2 of a three-task mandatory computer-based speaking test for students attending the University of Padova's Faculty of Education. The task involves a video and short article input to be summarized and discussed in the speaking performance.

85 recordings were transcribed and analysed. Discourse analysis measures were adopted, focusing on idea units (IUs), then compared to the CEFR B2 level descriptors in accordance with the test constructs.

Initial findings suggest that claims of test validity can be supported by evidence that the language elicited reflects the CEFR descriptors selected.

Introduction

The study is part of a wider research initiative to align an integrated task of a B2 speaking test (called 'TAL') at the University of Padova (Italy) to the Common European Framework of Reference for Languages (CEFR) (Council of Europe, 2001; 2020). In particular, this research involved mapping test-takers' performance by using content idea units (IU) to verify whether the language elicited could be linked to the descriptors of the CEFR selected in the test constructs.

The CEFR has long been accepted as one of the most influential models for comparing different language abilities (Figueras, 2012) although it was not designed as a ready-made rating tool (Fulcher, 2016), nor does it contain the descriptive adequacy required for test development (Alderson et al., 2004; Figueras, North, Takala, Verhelst, & Van Avermaet, 2005; Fulcher, 2004; Weir, 2005; among others). However, aligning tests to the CEFR has become fundamental in order to promote transparency for stakeholders and ensure reliability of tests and scores.

In higher educational settings, the necessity to design tests that replicate the demands students will encounter during their academic life (Brown, Iwashita, & McNamara, 2005) becomes even more evident when designing speaking tests, a skill which is considered complex to reliably assess (Luoma, 2004, pp. 27–28). A multimodal integrated task, i.e., where 'tasks are thematically linked, and input is provided as a basis for the response to be generated' (Lewkowicz, 1997, pp. 121), was thus created as part of the test, based on the argument that speaking skills can be more appropriately and authentically measured when in combination with other skills – such as listening, and/or reading (Hulstijn, 2011).

To investigate alignment with the CEFR, this study focuses on the analysis of elicited performances, assessed at B2 level by human raters, of one task of a three-task mandatory computer-based speaking test, taken by students attending the Faculty of Education.

Theoretical background

The CEFR and language testing

While there is no doubt that the CEFR (Council of Europe, 2001) has transformed and shaped the testing industry (Figueras, 2012; Figueras, Kaftandjieva, & Takala, 2013; Fulcher, 2008), its adequacy for test developers is still highly debated (for example, Figueras et al., 2005; Fulcher, 2004; Weir, 2005).

On one hand, the CEFR provides a 'descriptive framework, not a set of suggestions, recommendations, or guidelines' (Morrow, 2004, p. 7). It outlines learner proficiency and enhances communication among language practitioners (Council of Europe, 2001, p. 1), allowing for a common understanding of the crucial characteristics of learner levels (Council of Europe, 2009) and offering a 'practical, accessible tool that can be used to relate course, assessment, and examination content to the CEFR categories and levels' (North, 2004, p. 77).

Nevertheless, human judgment is required to interpret the CEFR descriptors and align tests to the framework, judgments that are characterized by uncertainty and easily subjected to 'bias and heuristics' (Eckes, 2012, p. 262). Furthermore, the CEFR does not describe test properties or item demands and is not based on a theory of item difficulty (Fulcher, 2004). The scales are intended to be illustrative only (North, 2004; Trim, 2012), and, although the Companion Volume (Council of Europe, 2020) has improved and expanded, it still 'leaves many contentious aspects [. . .] untouched' (Deygers, 2021, p. 190).

Aligning to the CEFR: Idea units

When it comes to discourse analysis, various measures are available; however 'researchers have been confronted with the fact that most grammars are based on a unit that is not defined for speech' (Crookes, 1990). The IU, developed by Kroll as 'a chunk of information which is viewed by the speaker/writer cohesively as it is given a surface form . . . related . . . to psychological reality for the encoder' (Kroll, 1977, p. 85), is an initial step towards the needs encountered when analysing speaking performances. Other studies have since adapted Kroll's IU, examining the speaking performance of integrated tasks in particular, allowing the measurement of key points included from the input text and the accuracy with which such information was reproduced and/or summarised (Frost, Elder & Wigglesworth 2012, Frost, Clothier, Huisman, & Wigglesworth, 2020); or used IUs to differentiate content-specific features, reflecting the concept of propositional complexity, i.e., the number of key information points a speaker encodes in a given language task to convey a given message content (Zaki & Ellis, 1999).

Aim

The present study aims at investigating whether performances elicited and assessed at B2 level presented discourse features, in terms of IUs, that align to the B2 descriptors included in the test construct.

Methods

The TAL B2 for the Faculty of Education

The TAL B2 speaking test for students attending the Faculty of Education consists of three tasks. It is computer-based and students record monologues in response to prompts. In this study, Task 2 – an integrated reading and video to speaking task – was analysed. The input topics relate to the test-takers' field of educational interest.

Four versions were analysed: Version A - Homeschooling, Version B – Standardised testing in Primary Schools, Version C – Parental involvement, Version D - Mixed vs Co-ed schooling. Both types of inputs relate to the same topic in each version.

Reading inputs count 300 words, while the video is a maximum of four minutes. Students are required to record a three-minute monologue, summarising the input sources and providing a personal opinion.

Participants

85 spoken performances were randomly selected from the larger corpus of recorded speaking tests of students attending the Faculty of Education (collected over a period of three years). Test-takers were 96% female, aged 22 to 32, with three outliers aged over 32.

Participants were assigned to four groups (A, B, C, D) depending on the task version they had completed. Test-takers shared a common L1, Italian, and had successfully passed the test.

The data set

The data collected consisted of:

- four integrated task inputs (four reading texts and four videos);
- 85 audio-recorded performances, elicited during official test setting, saved on the Moodle platform in m4a. format.

Data analysis

Test: input sources

The first step was to transcribe and map the reading and video inputs for IUs. IUs were adapted from Frost et al (2012; 2020). They included all clauses (also subordinate) or sub-clause variations (such as coordinated verb phrases; coordinated nouns or noun phrases connected to a common verb phrase; etc.) (see Frost et al., 2020, p. 6). Clarifications or explanations of the main IUs were counted within that IU.

The analysis was conducted by the two authors. All findings were compared and discrepancies discussed until mutual agreement was reached.

The IUs present in the different video sources varied: Video A had 16 main IUs, B had nine, C had five and D had 11. The reading inputs also varied: Texts A and B presented seven main IUs, C had 4 and D had 10. Moreover, three main ideas in Video A and D overlapped with Texts A and D respectively, while Videos and Texts B and C contained only one common main IU.

- Example of key idea from reading input:

Legally, every child in the UK needs to have an education.

- Example of key idea from video input:

In Walsall, the Walker family turned to home education when they felt mainstream school had failed them.

Performances

The 85 audio-recorded performances were transcribed using an online software (Otter.ai) and coded manually and separately by the two authors. Segments were compared and discussed until agreement was reached.

Test-taker IUs were divided into *accurate*, *distorted* or *other* (personal opinions, etc.) depending on how close they were to the IUs present in the input sources. IUs were considered accurate if they reproduced the same meaning, ignoring lexical or grammatical errors.

Test-taker IUs were further divided into groups depending on whether they simply replicated the input's IU (ARIU: accurate replication of IUs), combined two or more (ACIU: accurate IUs combined or AMP: macro-proposition of IUs) or repeated the idea verbatim (VIIU).

- Example of test-taker's ARIU:

And some parents choose homeschooling where public school fails.

- Example of test-taker's ACIU:

If having an education is a right and a duty at the same time, going to school, attending school to get a certification is no duty.

Variables were analysed by means of descriptive statistics.

Findings and discussion

The descriptors chosen in the test specifications for the integrated task were mapped. With regard to comprehension of the tasks' multimodal inputs, test-takers were expected to 'identify the main reasons for and against an argument or idea in a discussion conducted in clear standard speech' and to 'understand articles and reports concerned with contemporary problems in which the

writers adopt particular stances or viewpoints' (Council of Europe, 2020, p. 57). Regarding spoken production, test-takers were expected to 'explain a viewpoint on a topical issue giving the advantages and disadvantages of various options' (Council of Europe, 2020, p. 64).

Overall, test-takers were able to adequately produce an average of 12.08 IUs from the inputs (Standard Deviation: 5.44), with Versions A and B generating the most IUs (13.76 and 14.05 respectively), and Version D the fewest (8.29). Accurate replication of IUs amounted to 77.31%, whilst inaccurate video and reading replication of IUs totaled 10.47%. Original ideas were very limited (5.84 average IUs).

It should be noted that more IUs were recounted from the reading text (48.51% ARIU taken from reading input, 28.80% from listening), with only input B presenting an average equal division between reading and video input IUs.

Distorted IUs were minimal (less than TWO IUs on average) and tended to be connected to misunderstanding the video input. Verbatim recounting of input was limited (11.35% of total IUs), but mostly taken from the reading source (less than 1% from video input).

Overall, test-takers tended to recount IUs individually rather than summarising them as requested by the prompt. This is in line with other findings, such as Frost et al. (2012; 2020).

The initial data could indicate that the number of IUs elicited suggest an adequate use of the inputs and capacity to cognitively develop an argument, thus in alignment with the descriptors at B2 level.

Limitations

The limitations of the study are numerous. This study is only the initial step of a wider investigation. The differences in discourse density between the task inputs and the topic variety need further analysis. Furthermore, sample size and homogeneity of test-takers limit the generalizability of results.

Moreover, performances were selected from those evaluated at B2 level; however we did not investigate into actual rating procedures nor did we take into account the performance on the other two tasks of the test.

Conclusion and future directions

Overall, the average number of IUs provided by the test-takers could suggest an adequate use of the multimodal input sources, allowing for alignment with the descriptors chosen.

However, to support such a claim, performances at higher and lower levels should be analysed to explore potential correlations between the number of accurate IUs, how they are presented, i.e., recounted vs summarised, and proficiency level.

Future research could also benefit from the inclusion of additional measures, such as those of complexity, accuracy and fluency (CAF) and analysis of the influence of different sources (topic and text structure).

References

Alderson, J. C., Figueras, N., Kuijper, H., Nold, G., Takala, S., & Tardieu, C. (2004). *The development of specifications for item development and classification within the Common European Framework of Reference for Languages: Learning, teaching, assessment. Reading and listening.* Final report of the Dutch CEF construct project. Unpublished document.

Brown, A., Iwashita, N., and McNamara, T. (2005). *An examination of rater orientations and test-taker performance on English for Academic Purposes speaking tasks.* TOEFL Monograph Series MS-29. Princeton: Educational Testing Service.

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment.* Cambridge: Cambridge University Press.

Council of Europe. (2009). *Manual for relating language examinations to the Common European Framework of Reference for Languages (CEFR).* Strasbourg: Council of Europe.

Council of Europe. (2020). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment - Companion Volume*. Strasbourg: Council of Europe Publishing.

Crookes G. (1990). The Utterance, and Other Basic Units for Second Language Discourse Analysis. *Applied Linguistics*, *11*(2), 183–199.

Deygers, B. (2021). The CEFR Companion Volume: Between Research-Based Policy and Policy-Based Research. *Applied Linguistics*, *42*(1), 186–191.

Eckes, T. (2012). Examinee-centered standard setting for large-scale assessments: The prototype group method. *Psychological Test and Assessment Modeling*, *54*(3), 257–283.

Figueras, N. (2012). The impact of the CEFR. *ELT Journal*, *66*(4), 477–485.

Figueras, N., Kaftandjieva, F., & Takala, S. (2013). Relating a reading comprehension test to the CEFR levels: A case of standard setting in practice with focus on judges and items. *Canadian Modern Language Review*, *69*(4), 359–385.

Figueras, N., North, B., Takala, S., Verhelst, N., & Van Avermaet, P. (2005). Relating examinations to the Common European Framework: A manual. *Language Testing, 22*, 262–279.

Frost, K., Elder, C., and Wigglesworth, G. (2012). Investigating the validity of an integrated listening-speaking task: A discourse-based analysis of test takers' oral performances. *Language Testing*, 1–25.

Frost, K., Clothier, J., Huisman, A., & Wigglesworth, G. (2020). Responding to a TOEFL iBT integrated speaking task: Mapping task demands and test takers' use of stimulus content. *Language Testing*, *37* (1), 133–155.

Fulcher, G. (2004). Deluded by artifices? The common European framework and harmonization. *Language Assessment Quarterly*, *1* (4), 253–266.

Fulcher, G. (2008). Testing times ahead? *Liaison Magazine*, *1*, 20–24.

Fulcher, G. (2016). Standards and frameworks. In J. Banerjee and D. Tsagari (Eds.), *Handbook of Second Language Assessment* (pp. 29–44). Berlin: DeGruyter Mouton.

Hulstijn, J. (2011). Language proficiency in native and nonnative speakers: An agenda for research and suggestions for second-language assessment. *Language Assessment Quarterly*, *8*(3), 229–249.

Kroll, B. (1977). Combining ideas in written and spoken English: a look at subordination and coordination. In E. O. Keenan & T. L. Bennett (Eds.) *Discourse Across Time and Space*., S.C.O.P.I.L. No. 5. California: University of Southern California.

Lewkowicz, J. A . (1997). The integrated testing of a second language. In C. Clapham & D. Corson, D. (Eds) *Encyclopaedia of Language and Education. Vol. 7: Language Testing and Assessment* (pp. 121–130). Dortrecht: Kluwer.

Luoma, S. (2004). *Assessing Speaking*. Cambridge: Cambridge University Press.

Morrow, K. (Ed.) (2004). *Insights from the Common European Framework*. Oxford: Oxford University Press.

North, B. (2004). Relating assessments, examinations, and courses to the CEF. In K. Morrow (Ed.) (2004). *Insights from the Common European Framework* (pp. 77–90). Oxford: Oxford University Press.

Trim, J. L. M. (2012). The Common European Framework of Reference for Languages and its background: A case study of cultural politics and educational influences. In M. Byram & L. Parmenter (Eds.) *The Common European Framework of Reference: The Globalization of Language Education Policy* (pp. 14–34). Bristol: Multicultural Matters.

Weir, C. J. (2005). Limitations of the Common European Framework for developing comparable examinations and tests. *Language Testing*, *22*(3), 281–300.

Zaki, H., & Ellis, R. (1999). Learning vocabulary through interacting with a written text. In R. Ellis (Ed.), *Learning a Second Language Through Interaction* (pp. 153–169). Amsterdam: John Benjamins.

Mapping the SMEEA Gaokao tests to the CEFR

Jane Lloyd
Cambridge University Press & Assessment, United Kingdom

Graham Seed
Cambridge University Press & Assessment, United Kingdom

Abstract

This paper reports on a project to map six language tests produced by the Shanghai Municipal Educational Examinations Authority (SMEEA) to the CEFR. This project was undertaken to enable a comparison of the relative difficulty of each language. Additional project aims were to carry out training in the CEFR and CEFR mapping procedures, using a blend of online delivery and a cascaded training model. This project involved a series of linked training and mapping activities and workshops. In this paper the focus is on the practical aspects of a computer-mediated multilingual CEFR mapping project, and on the participant feedback. Some outcomes are included which are directly related to the CEFR, but not any outcomes which are confidential.

Introduction

In some situations, assessment bodies have a pre-existing test where the starting point may not have been the Common European Framework of Reference for Languages (CEFR, Council of Europe, 2001) but they wish to carry out the additional validation steps to determine alignment. Such is the case in this project, which was carried out by Cambridge University Press & Assessment (Cambridge) and SMEEA on six foreign language versions of the Shanghai Gaokao. The project used the scales and 'Can Do' descriptors found in the Companion Volume for the mapping activities and the updated 2020 version of the CEFR (Council of Europe, 2020). The forthcoming Chinese translation of the Companion Volume, coordinated by Xiangdong Gu and Zehan Chen of Chongqing University, China, with the support of the Council of Europe, was also supplied to SMEEA. It is believed that this project was the very first major use of the Chinese translation of the CEFR Companion Volume (CEFR CV). The 'Can Do' descriptors in English, French, German and Spanish were also supplied to SMEEA. Project participants were free to use whichever language version of the CEFR documentation they felt most comfortable accessing.

SMEAA exam overview

The six language versions of the Shanghai Gaokao are English, German, French, Spanish, Russian and Japanese. Each language version of the test is broadly similar in terms of structure, although minor differences occur in terms of item numbering, and in some subsections. Generally speaking, the written test paper contains several sections: Listening, Linguistic Knowledge, Reading, Summary Writing, Guided Writing and Translation. There is a Speaking test which contains a combined Speaking and Listening section.

Cambridge was provided with publicly available sample tests from the SMEEA battery of Gaokao language tests in order to gain an initial oversight into how the tests are organised. As a result, Cambridge staff were able to make an initial judgement of which of the CEFR's modes of communication each part of the test could fit into. To avoid cognitive overload, the different parts of the tests were loosely divided between the CEFR categorisations of reception, production, interaction, mediation, and communicative competence. This enabled the project participants to take each element in turn by examining how the CEFR approaches each element, and how those parts of the test can be mapped to the CEFR.

Project aims

The project had three main aims. The first was to familiarise participants with the CEFR by:

- introducing the CEFR's core conception of language learning
- introducing the CEFR's level framework of language proficiency

- inducting participants into the characteristics of input (listening and reading) and output (speaking and writing), as well as general communicative competence (e.g., vocabulary and grammar) for relevant CEFR levels.

Secondly, the project aimed to encourage participants to reflect on how the CEFR relates to specific parts of the tests to provide an overview of:

- how well each test aligns to the CEFR's theory of an 'action-oriented approach' to language assessment
- which CEFR scales each test maps to
- which CEFR proficiency levels each test maps to
- the similarities and differences of the different language versions of the tests.

The final aim was to train up key SMEEA staff and consultants in the use and application of the CEFR according to their context, so that they can carry out the mapping activity by cascading information to others; and also be able to plan how to use the mapping tools and activities in the future. It should be noted that a full-scale alignment of the tests to the CEFR, as set out in the Manual for relating language examinations to the CEFR (Council of Europe, 2009), was not the aim. Nevertheless, the mapping activity that was carried out was based on the principles within the Manual.

Practicalities of a multilingual blended project

A blended approach

Initial plans to run a series of workshops in person had to be adapted, due to travel restrictions brought about by the pandemic. Much of the interaction in the project was virtual, through online workshops, conferencing software and remote communication. A cascaded model was implemented to deliver the training to a small group of 'lead mappers'. This group consisted of two target language specialists for each of the six language versions. These 12 lead mappers received training from the Cambridge trainers online, and then cascaded the training and carried out further work with their specific language team in face-to-face workshops.

Stages followed

The training in the CEFR areas was delivered in this order: receptive skills; productive skills; interaction and mediation skills; and communicative competence. Training for each of these areas comprised the same main stages:

- pre-workshop activities for lead mappers
- a live online session led by Cambridge trainers to lead mappers, providing: training about the CEFR; discussion about which CEFR scales may be useful for mapping; discussion about what the items and tasks are targeting; how to map specific test sections to the CEFR
- a cascaded session led by the lead mappers to their language team
- additional sessions as appropriate for the language team to complete mapping
- an informal Q&A live online session for lead mappers with the Cambridge trainers
- opportunities to revisit mapping judgements at a later date.

Coupled with introductory and consolidation sessions, these activities formed 10 main sessions. All relevant documentation was available online. Presentations, with accompanying notes were provided in both English and Chinese.

The process of mapping to the CEFR

During live sessions, the Cambridge trainers explained the process of mapping. This process remained more or less the same for each test section. For Reading, Listening, Linguistic Knowledge and Translation sections, mapping was done at the item level. For Writing and Speaking sections, mapping was done at the level of the assessment criteria.

The mapping process was carried out in several steps within each language team: individual mapping was followed by pair discussion, then small group discussion, then in plenary. For the individual mapping, each member of the language team decided individually what each item[1] in the section is testing, which CEFR scale(s) are relevant to look at with regards to each item, and

[1] 'Item' here also refers to each grade in each assessment scale, for Writing and Speaking sections.

roughly which level each item was. Participants were advised to read the CEFR descriptors in the relevant scales, at, above and below the estimated level, for each item and then choose the CEFR level which is the best fit for each item. Following individual mapping, participants discussed their decisions in pairs, aiming to come to a compromise. Next, the process was replicated in groups of four or five people, and then at the whole-group level. The group could include up to three CEFR descriptors (from the same or from different scales) for each item judgement, as long as all descriptors were at the same CEFR level.

For Speaking and Writing mapping, another additional step was carried out. Candidate performances at different grades (especially the 'passing' grade) were reviewed. Each performance was marked again in two ways: using the usual assessment criteria for the SMEEA test; and using CEFR levels and descriptors only. The CEFR level given to the performance was compared with the CEFR level given to the grade, to see if they were the same. If they were not the same, participants had to explore the reason for this, and if necessary adjust either the grade given to the performance or the CEFR level given to the grade. Results were collated for all test sections and compared across languages. These results are confidential and cannot be shared here.

Evaluation

After having completed judgements for all test sections, a closing survey was administered to lead mappers. Responses showed that the general format of the project, with pre-workshop activities, training input sessions, and cascaded sessions to the language teams, worked well and took place in a demanding but achievable timeframe.

One of the key aims for the project was to increase familiarity and understanding of the CEFR. While only one of the lead mappers claimed to be very familiar with the CEFR before the start of the project, over the course of the project, familiarity with the CEFR clearly increased, so that by the end of the project, three quarters of the lead mappers stated that they were now very familiar with the CEFR, with the remaining quarter now responding that they were somewhat familiar.

Respondents also took the opportunity to note possible mismatches between the can-do focus of the CEFR and the selective focus of the SMEEA tests, which resulted in some difficulty in mapping, especially in Writing. One interesting perspective showed this difference:

'I feel that the CEFR focuses on affirming the content that students have mastered, while the test is mainly to find out students' deficiencies.' [Translated from Chinese]

Some respondents touched on considering changes in the Writing and Speaking assessment criteria to more fully reflect the CEFR. Overall, there was reflection on whether the fundamental basis for the test could be more in line with the CEFR's outlook: *'not based on errors but based on whether students can achieve a certain goal'* [translated from Chinese].

The actual process of coming to CEFR level judgements was seen as useful:

'Errors in judgement were inevitable, but we enjoyed the adequate and heated negotiation about these items.'

One respondent's comment sums up many of the others:

'After this benchmarking, I have a better understanding of the European standard framework, and I have a lot of reflections on my own teaching. I hope I can have more exchanges with experts face to face in the future.' [Translated from Chinese].

The multilingual nature of the project fitted in well with the CEFR's multilingual and plurilingual approaches. If all participants were able to communicate at a high level of proficiency in all seven languages used, it would enhance standardisation and a common understanding. As this is an unrealistic scenario, the language of the workshops with the Cambridge trainers was English and the common language of the cascaded workshops was Chinese, although language teams would also make use of the language of the test too. This limitation was mitigated with the translation of the materials from English into Chinese and the opportunity for the target languages to be used, rather than 'forcing' English, for example, on to all participants. We noted that for participants who are new to the CEFR, simply providing a version of the CEFR in a language they are more familiar with, such as Chinese, does not mean that their understanding of the terminology, particularly in the descriptors, can be taken for granted. SMEEA staff reported that the translated version sometimes complicated their understanding. A lack of examples in the descriptors made it hard for those who have no CEFR-related experience to understand or distinguish key terms.

It was the first time that SMEEA had carried out such a large-scale project, which involved more than 70 people and lasted over two months, blending online and offline delivery. Organisations have to decide whether to opt for highly intensive training, with little time for reflection or cascading in between sessions, or whether to space out sessions, and weigh up the risks of participants going off track. These considerations apply both to offline and online modes of delivery. Organisations need to weigh up which styles and modes of delivery will be efficient, and what will be effective, and accept that these may differ. A full alignment involving standard setting procedures would involve a much deeper investigation at each level for each language, and therefore would take a longer time and would likely be more costly. Nevertheless, it is believed that this mapping procedure has been useful in achieving its aims and has provided a satisfactory outcome of purpose for SMEEA.

References

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge: Cambridge University Press.

Council of Europe. (2009). *Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment. A Manual*. Strasbourg: Council of Europe Publishing.

Council of Europe. (2020). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Companion Volume*. Strasbourg: Council of Europe Publishing.

Validation of a high-stakes test: GA IESOL multiple-choice units[1]

Emanuela Botta
Department of Medicine and Psychology, Sapienza University of Rome, Italy

Snezana Mitrovic
Department of Medicine and Psychology, Sapienza University of Rome, Italy

Giusi Castellana
Department of Education, Roma Tre University, Italy

Abstract

The purpose of this study is the validation of the Listening and Reading units of the Gatehouse Awards (GA) Classic IESOL (International English for Speakers of Other Languages) examination at the CEFR Level B2, as criterion-referenced achievement tests, with the aim of reaching the multi-trait, multi-method approach of the examination. The data sets have been studied to examine the content validity, reliability of the scores and unidimensionality of the construct. Eight forms, four forms per unit, have been analysed using separate exploratory factor analyses, and they revealed excellent values of internal consistency for the listening unit. The reading test construct unidimensionality values are excellent for Versions E, F and G (henceforth VE, VF and VG), and acceptable for Version D (henceforth VD). In addition, a qualitative analysis of items has been performed and correlation analyses between the scores of the students who completed both tests and are positive and significant (0.621; <0.01). Further steps of the study will be the item response theory (IRT) equating procedure of the four versions.

Introduction and aims

The GA Classic IESOL is an internationally recognised examination of English as a foreign language testing the skills of listening, reading, writing and speaking separately at Levels A1 to C2 of the Common European Framework of Reference for Languages (CEFR, Council of Europe, 2001).

The aim of the study is the validation of the examination at the CEFR Level B2, as a criterion-referenced achievement test, using the multi-trait, multi-method approach (Bachman & Palmer, 1982; Campbell & Fiske, 1959). The sets of data that have been analysed to examine the content validity, reliability of the scores and the unidimensionality of the construct pertain to the multiple-choice units: listening and reading.

Listening and reading abilities as test constructs

Comprehension skills, listening and reading, are integral parts of most if not all internationally recognised examinations of English as a second language. These are most often administered as two separate tests, and 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).

As both listening and reading are internal processes that cannot be observed directly, what can be done is assess external behaviour (Alderson, 2000; Buck, 2001). Although the construct is never perfectible for the same reason, we can take a theory of listening and reading and then operationalise it in our tests: through the texts and the tasks we select and require the readers to perform. Consequently, generalisability can be achieved by appeal to theory, to the extent that it adequately reflects theory and the extent to which that theory is correct (Alderson, 2000). As we believe that the test performance is an indicator of an underlying competence (Buck, 2001), a description of these two abilities has been used as the basis for defining the two constructs.

[1] This paper is the result of the joint work of all three authors. The first two sections were written by Mitrovic, the third and fourth sections were written by Botta, and the final two sections were written by Castellana.

Research has also shown that although there is considerable overlap between listening and reading ability, they both have their unique aspects (Bae & Bachman, 1998; Buck, 1992) and for that reason are seen as two separate constructs.

In addition, there have been different attempts at identifying listening and reading sub-skills, starting from 1970s onward (Liu, Aryadoust, & Foo, 2022), proposing different solutions, that is, different sub-skills. For example, Wagner (2004) attempted to validate a two-factor model of listening: listening for explicitly stated information and listening for implicit information, which, however, proved to be less interpretable than a one-factor solution. Consequently, each of the comprehension skills is treated as a single construct.

Since it is the test-takers' abilities that we are most often interested in, the test construct can be defined as a description of the ability (Buck, 2001). English language awarding bodies define listening and reading abilities as test constructs in accordance with the CEFR definitions and levels. Specifically, the GA IESOL Classic B2 listening and reading units are mapped to the CEFR and its definitions of the listening and reading abilities at the B2 level.

Content domain specification, that is examination specification, being considered a necessary requisite for the test's construct and content validity (Bachman, 2002; Brown, 1996; Chalhoub-Deville, 2001), is provided in detail and includes functions and notions, grammar, discourse markers, topics, and key language items for the B2 level.

Method and participants

The B2 level listening and reading units each consists of three tasks, with a total of 22 items per unit. The data set used for this study originates from the online examinations held from October 2020 to November 2021. For each of the units, four forms were administered (VD, VE, VF and VG) and for each of the eight forms analyses have been performed. A total of 597 candidates completed a randomly assigned version of the listening test: 145 candidates completed form VD, 147 form VE, 162 form VF and 143 candidates completed form VG, while a total of 564 candidates completed a randomly assigned version of the reading test: 150 candidates completed form VD, 159 form VE, 144 form VF and 111 candidates completed form VG.

Results

In this paper, analyses of the reading tests are presented. Part of these analyses were performed for the listening test as well (Mitrovic, Botta, & Castellana, 2022). The analyses carried out are part of the validation of the whole examination, and for that reason, it is necessary to verify that the data fit the theoretical framework that the tests were based on and eventually provide information necessary for the design of future versions of the examination. To start with, exploratory factor analysis was performed for each of the versions to confirm that the items make up a unidimensional construct. The analyses were carried out using Mplus v.8.7, as it is a software that efficiently analyses categorial data (Muthén, 1983). The results were quite satisfactory – the fit index (RMSEA) is lower than .05 for all the versions, which is very good, and the ratio between the first and the second eigen value is quite high, as is the variance explained by the factor. The scree tests support the hypothesis of a single factor.

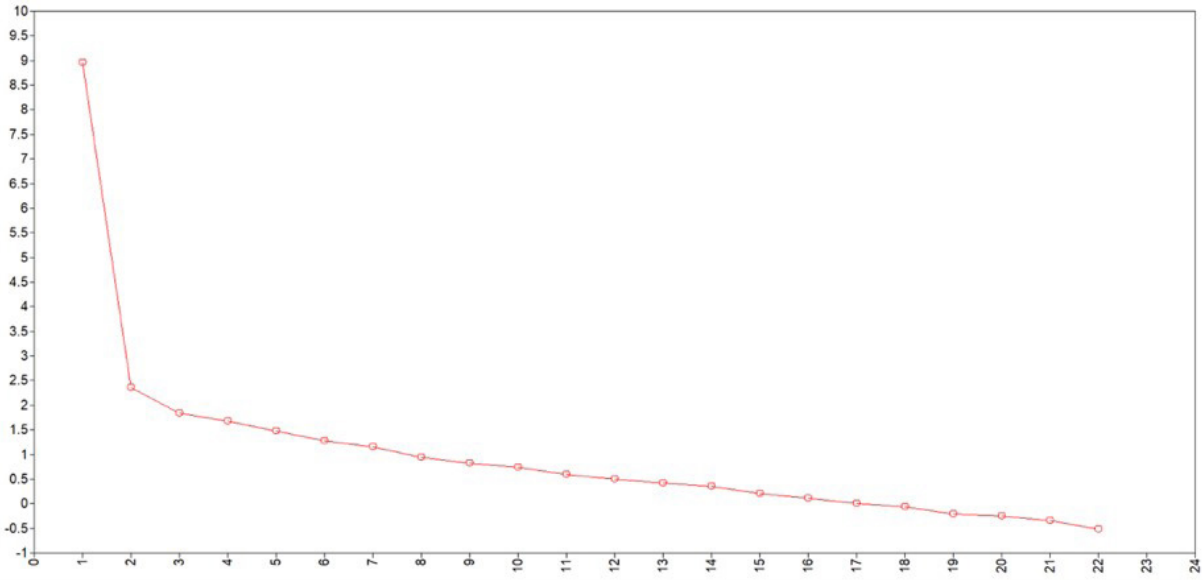


Figure 1 VE scree plot

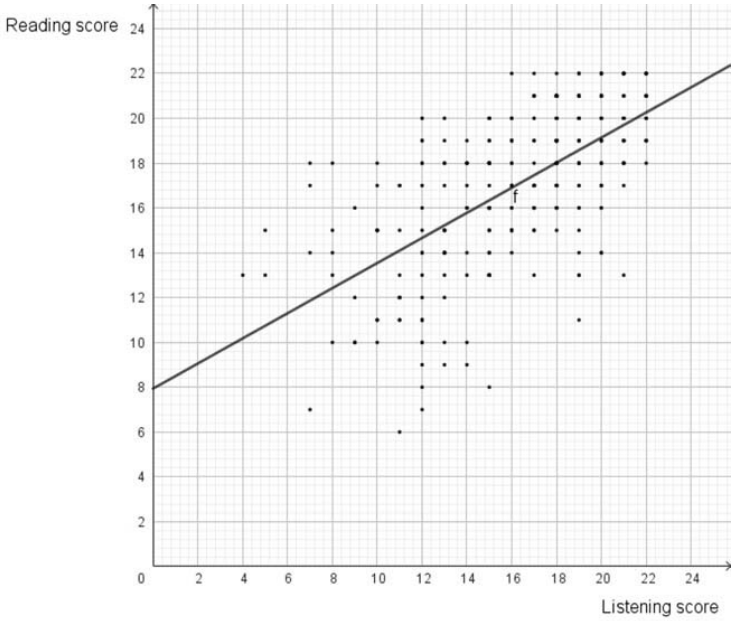


Figure 2 Correlation between listening and reading scores

In general, loadings are quite good or excellent, and there are only few items with loadings slightly below .300. Cronbach's alpha is between .700 and .800, that is adequate for this type of test.

Considering that the two units of the examination are part of the whole being validated, a correlation analysis between the listening unit and the reading unit, using a sub-sample of 47% of the total sample, was performed.

The correlation is positive and quite high (0.621). This confirms the previously mentioned research results, that there is considerable overlap between listening and reading ability and that the difference between the two is minimal. It also provides us with an external validation of the reading test.

Finally, all items were calibrated according to the Rasch model, setting the difficulty mean to 0. This made it possible to analyse each of the reading test forms and each of the items, in terms of model fit and distractor functioning. As regards the four versions, we can say that ANOVA between the proportions of correct responses confirms that the mean difficulty of the four forms is the same and that the differences are not statistically significant. Variability ranges of item difficulty values are slightly different but only few items of each form have values outside the common interval (-1.451; +1.825). Finally, in all four forms, it is evident that the candidates' skills (θ) are often higher than the difficulty (b) of the test, and this confirms that candidates prepare for the exam and only take it when they are actually ready (Figure 3).

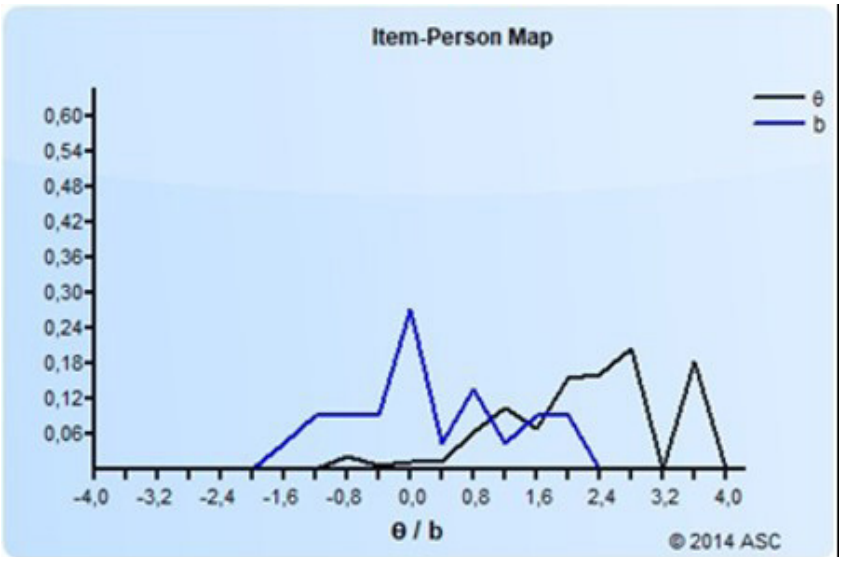


Figure 3 VE item/person map

These analyses provide valuable information on each item, in particular, goodness of fit, infit and outfit, item discrimination defined as point-biserial correlation coefficient of an item with all the rest of the same form (R >0.18) and the percentage of students who chose each of the possible question answers. For the fit indices, whose expected value is 1, the reference interval is 0.8–1.2. Low values indicate that possible question answers are not in line with the theoretical basis and that actual data are higher than the ones estimated by the model, while high values indicate a poorly fitting model, and undermine the validity of the measure, indicating that the actual data are lower than what the model estimates and cannot be predicted by the model.

Qualitative item analyses: Examples of good items and problematic items

Three examples, one of a well-constructed item and two with less satisfactory data results, are provided below. The first one (Figure 4) is the example of a well-constructed item.

Back in 1985, Coca-Cola decided to change its original recipe for the first time in 99 years. They wanted to make a sweeter version of Coca-Cola called "New Coke". The taste tests they did suggested it was a really good idea. 3) _____, they underestimated just how much people loved the original recipe and it became an absolute nightmare for the company.
a) However b) Additionally c) Equally d) Whereas

Figure 4 A well-constructed item

Candidates need to choose an option taking into consideration morphosyntax, as well as the register. The item has a quite good discrimination index (0.18) and excellent fit indices (infit = 1.097, outfit = 1.046). All the options are equally attractive.

The correct option, *however*, was chosen by 60% of the candidates. This is due to the right choice of distractors, which are in line with the ALTE *Materials for the Guidance of Test Item Writers* stating that options 'should have an approximately equivalent grammatical structure and level of complexity to one another' (1995, p. 116). The distractor that functions well here seems to be *whereas,* chosen by 23% of candidates as the correct answer. It is close to the correct answer *however* in terms of syntax, *whereas* being a conjunction, while *however* is a conjunctive adverb, and in terms of meaning they are both used to indicate contrast or differences. In addition, being a conjunctive adverb, *however* is followed by a comma, while *whereas* is used at the beginning of a subordinate clause and is not followed by a comma. The other two distractors are adverbs that have different meanings.

The following item is an example of an item that does not function very well: discrimination, infit and outfit indices are quite low. Consequently, 97% of the students chose the correct answer, while 'each wrong alternative should be attractive to at least some of the students' (Alderson, Clapham, & Wall, 1995). The three distractors do not seem to function (p = 1%), particularly *person*, not having been chosen by a single candidate. This is likely due to the adjective preceding the term Coca-Cola, 'iconic', which clearly announces the term and can hardly be linked to *service*, *person* and *belief*. It is for this particular reason that 'verbal clues which direct the candidate to the correct option ("word-spotting") should be avoided' (ALTE, 1995, p. 116). In addition, 'iconic' is not normally used with abstract nouns and *person* is obviously incorrect. The distractors also belong to different word groups.

[...] some like the change, and others say that changes shouldn't be made to things that aren't broken. It's not the first time a major food company has decided to change an iconic 2) _____. Here are some of our [....]

a) product b) service c) person d) belief

Figure 5 First problematic item

In the last item, the blank requires a verb, the correct answer being *argue*, with 37% of student answers – less than the first distractor, *ensure*, chosen by 47% of the students. The verb *argue* may have seemed too easy to the candidates as it is normally taught at the CEFR Level B1 meaning 'to give reasons', that is as a reporting verb, while at the CEFR Level B2 it is taught to mean 'disagree'. These two verbs seem to be too similar in meaning for the candidates to be able to tell the difference or seem to have a stronger meaning than the verb *argue*. As a result, the item discrimination index is negative, meaning that lower-ability students with lower-ability levels respond better than higher-ability ones, and infit and outfit indices are too high.

Nearly 40,000 Americans wrote to Coca-Cola to complain about the change and the company quicky backtracked three months later. However, to this day, Coca-Cola 4) _____ that the move was a great idea because sales of its original Coca-Cola increased a lot after its return.
a) ensure b) promise c) regret d) argue

Figure 6 Second problematic item

The above illustrate that preventive analyses of this kind are essential for every distractor before it is included in the test and a brief analysis of each response option or rationale for its inclusion needs to be provided beforehand.

Conclusion

Validation of an examination is a necessary process, as it is used to confirm that the test accurately measures the students' level of competence. As stated by Messick (1992, as cited in Weir, 2005) and Weir (2005), not many test makers provide validity evidence or perform validation studies of their tests.

The results of the analyses revealed good values of internal consistency for all four test versions of each of the units (listening and reading), confirming the hypothesis of unidimensionality of the two constructs and providing convergent validity through correlation between the two units. Since the sample was randomly selected and the versions randomly assigned, the identified differences can be attributed to chance.

The qualitative analyses of the item distractors that had lower than acceptable values demonstrate the importance of selecting plausible distractors while constructing multiple-choice questions in order to be equally appealing to the candidates so that the correct answer cannot be chosen by exclusion. For that reason, for each of the items, it is necessary to provide a Distractor Analysis, or 'a brief analysis of each response option or rationale for inclusion of specific response option with one item (one sentence at the most for each response option)' (Mullis, Martin, Cotter, & Centurino, 2013, pp. 12–13).

Similar analyses for the remaining units, writing and speaking, are planned, in order to arrive at the validation of the examination as a whole.

References

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

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language Test Construction and Evaluation.* Cambridge: Cambridge University Press.

ALTE. (1995). *Materials for the Guidance of Test Item Writers.* Available online: https://www.alte.org/resources/Documents/IWG%20July2005.pdf

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), 449–465.

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

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

Buck, G. (1992). Listening comprehension: construct validity and trait characteristics. *Language Learning, 42*(3), 313–357.

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

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

Chalhoub-Deville, M. (2001). Task-based assessment: Characteristics and validity evidence. In P. Skehan, M. Swain & M. Bygate (Eds.), *Applied Language Studies: Task-based Research* (210–228). New York: Longham.

Council of Europe. (2001). *Common European Framework of Reference for Languages*. Cambridge: Cambridge University Press.

Liu, T., Aryadoust, V., & Foo, S. (2022). Examining the factor structure and its replicability across multiple listening test forms: Validity evidence for the Michigan English Test. *Language Testing*, *39*(1), 142–171.

Mitrovic, S., Botta, E., & Castellana, G. (2022). *Validation of a High-Stakes Test: GA IESOL Listening Unit*. Available online: archive.headconf.org/head22/wp-content/uploads/pdfs/14514.pdf

Mullis, I. V., Martin, M. O., Cotter, K. E., & Centurino, V. A. (2013). *Item Writing Guidelines*. Boston: IEA, TIMSS & Pirls International Study Center, Lynch School of Education, Boston College.

Muthén, B. (1983). Latent variable structural equation modeling with categorical data. *Journal of Econometrics, 22*, 48–65.

Wagner, E. (2004). A construct validation study of the extended listening sections of the ECPE and MELAB. *Spaan Fellow Working Papers in Second or Foreign Language Assessment*, *2*, 1–23.

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

A flexible framework: Matching student assessments to the CEFR descriptors in a hybrid context

Steve Issitt
University of Birmingham, United Kingdom

Abstract

Intensive language programmes over the last three years have required changes in their modes of delivery while maintaining the efficacy and accuracy of their assessment methods. This presentation examines some the issues related to matching student assessments in these circumstances to CEFR descriptors and I outline how this was achieved at the University of Birmingham. A key focus of the discussion will be the production of transparent, user-friendly marking criteria which are congruent with the CEFR and adapted to the requirements of our intensive pre-departmental entry programme. I consider how a university may benefit from a programme of provision which aligns itself to an internationally recognised benchmark and discuss how this alignment is crucial to maintenance of the UK's position in the global international student arena.

Introduction: A variety of stakeholders

The various parties involved in English for Academic Purposes (EAP) provision in the UK higher education sector could be considered *stakeholders* (Freeman, 1984) and Reavill (1998) identifies 12 categories including the student, the employer, the family and dependants of the student, universities and their employees, the suppliers of goods and services to universities, the secondary education sector, other universities, commerce and industry, the nation, the government, taxpayers (national and local) and professional bodies. Such descriptions, if not proclaiming a hierarchy, do outline the contexts in which EAP provision takes place. A student perspective might examine their own experiences and consider impact on the home countries such as obtaining influential positions and helping to strengthen ties with the UK, creating greater understanding on linguistic, economic, cultural and political levels. From a university perspective, our teaching and assessment practices need to be unambiguous and linked to accepted international standards which the departments can understand and utilise. This procedure may contribute to the strength and resilience of the university and affords considerable benefit to the local economy. We might even suggest that a high-profile and quality EAP programme can produce a measurable boost to a town, city or region. When we take a national perspective, the benefits to stakeholders are staggering with a provision of 697,000 students in 2021/22, representing 15.7% of all UK undergraduates and 39.1% of all postgraduates (HESA, 2021). In economic terms alone this is a large figure and when add-ons such as local spending are considered, the total reaches over £40 billion (Universities UK, 2023). Regarding less directly measurable aspects there are also substantial benefits. For example, a study by the Higher Education Policy Institute (HEPI) (2022) who looked at *soft power* (Nye, 1990) found that 57 serving world leaders had experienced UK higher education, making the country's total second only to the USA. The impact is difficult to quantify but the UK's reputation and influence can surely be enhanced by providing educational qualifications and formative experiences of future world leaders. It is also important to maintain high-quality provision, uninfluenced by the financial importance of the student cohort who may be considered 'cash cows' (Lomer, & Mittelmeier, 2021) with their presence taken for granted. In addressing this problem, I suggest we should as practitioners, defend, promote and elaborate our EAP provision in various fora of academia, accreditation and quality control, recognizing the financial contributions that the students make to our higher educational system. A *Financial Times* (2022) report shows how international students effectively cross-subsidise areas of research and teaching, especially in Technology and Science where the cost of the courses are higher than the fees paid by domestic students. In addition, if EAP courses can ground their provision and assessment techniques in up to date, linguistically referenced contemporary data, this can only be of benefit to the research community, another valuable stakeholder.

EAP delivery and issues for writing assessment

UK university intensive EAP courses are often delivered in the form of *presessional programmes* of varying lengths and start dates, typically in spring and summer. These courses are generally delivered by specialist EAP teachers, often employed on a contract basis under the supervision of in-house staff. The courses themselves are open to accreditation by outside agencies such as the British Council who stipulate issues of quality control such as maximum class sizes and teacher qualifications, carrying out regular inspections. At the University of Birmingham, the last few years have witnessed a movement from entirely face-to-face delivery in 2019 with 1,800 students, to a wholly online provision in 2020. In 2021 the programme was 80% online and in 2022 a similar split but with more face-to-face teaching. The profile was very much the same in 2023 and for this period, around 1,000 students were in attendance. This has led to an increased focus on precision in provision and assessment as delivery methods have varied, and an increased need to match university departmental expectations to course and assessment content. It has also become increasingly important to offer academic departments clear and understandable assessments of students' competencies in practical skill areas and to ally end of course assessments to existing external frameworks. The most important of these, the Common European Framework for Languages (CEFR, Council of Europe, 2001; 2020), makes direct reference to EAP competencies such as: 'Can write clear, detailed text related to his/her field of interest, synthesising and evaluating information and argument from a range of sources' [Overall written production B2 level] (2001, p. 61) and: 'Can write an essay or report that develops an argument giving reasons in support for or against a particular point of view and explain the advantages and disadvantages of various options. Can synthesise information and argument from a number of sources' [Written reports and essays B2 level] (2001, p. 62).

These competencies have been directly incorporated into our teaching programmes and addressed in our assessment criteria. This is especially pertinent, as external examinations suites which often form the basis for admission decisions by the departments, may not test such abilities. In this sense, there is a need to *fill the gaps* and to give confidence to the various stakeholders in the international student domain that students are not being used purely for financial purposes and provided with inadequate programmes of EAP instruction. From an institutional position, it is also important that academic departments are presented with international students ready for study and given confidence that the EAP provision is of sufficient quality to equip the departmental entrants for the many challenges they face. One component of this provision is a relevant syllabus which I now outline.

The EAP reading and writing syllabus

Language competencies which recur over a student's academic career are prioritised and include ability to synthesise text and argument, text production and the creation of an authentic academic voice. Also important is text transformation, referencing and citation, and the ability to evaluate academic sources. Our 10-week reading and writing programme is presented in a 350-page materials book (student and teacher versions) and organised thematically (Technology in education, Communication in social media, Cross-cultural communication, and Truth: the reliability of evidence). The following syllabus overview outlines the programme.

You will develop the skills to:

Use general and academic vocabulary and collocations for reading and writing authentic academic texts.

Demonstrate controlled, accurate and flexible use of a wide variety of complex sentence structures.

Structure a paragraph logically . . . integrating and synthesising relevant supporting detail from a range of academic sources.

Critically evaluate a range of relevant complex academic texts to . . . effectively summarise, paraphrase and quote information without plagiarising.

Follow the Harvard conventions for citation and referencing.

Plan and logically structure a coherent piece of writing on a complex topic to develop an argument or express an opinion, evaluating different ideas from a range of academic sources.

Proactively identify own errors and respond effectively to feedback by proofreading and redrafting own work.

Flexibly apply a wide range of reading skills.

Identify and critically evaluate arguments and their supporting evidence in complex, extended academic texts.

The reading and response test, marking descriptors and marking examples

The above competences are assessed by use of a reading and response test which includes a requirement to answer a question under exam conditions such as that given below.

It is often said that, in order to combat fake news, students should be taught to think critically.

*Using the **sources provided**, define what critical thinking is and suggest how it can be taught. **Use the short texts provided** to support your answer. You should aim to write 500 words. You have two hours.*

The students' answers are assessed using the four strand marking descriptors described below. The example refers to a top band script equivalent to CEFR mid-C1 and IELTS 7.5 (see https://www.ielts.org).

Task achievement, argument and analysis. *Engages critically with generally sophisticated arguments and extremely effective integration of sources. Paraphrasing is very skilfully handled.*

Text organisation. *Very well structured, coherent text. Sequencing of ideas is very skilfully handled.*

Grammatical range and accuracy. *A wide range of complex structures to express precise meaning. Grammar and punctuation are well controlled and there are frequent error-free sentences.*

Vocabulary. *A very wide range of vocabulary to express a precise meaning. Errors in word choice and word formation are occasional and do not impact upon communication. Register is appropriate.*

Application of the descriptors

This example is taken from a marking standardisation session.

***Criterion*:** Task achievement, argument and analysis. ***Marks*:** 15/20 (IELTS 7.0, CEFR low C1)

***Comments*:** A convincing answer with arguments well-formed and good integration of sources, for example: *'. . . the former focus on clarifying and making unbiased judgements, while the latter emphasises that being a critical thinker needs to ask key questions, evaluate information and draw well-reasoned conclusions and solutions.'*

***Criterion*:** Text organisation. ***Marks*:** 17/20 (IELTS 7.5, CEFR mid-C1)

***Comments*:** *The text is very well structured, is easy to read and coherent. The ideas are well sequenced.*

***Criterion*:** Grammatical range and accuracy. ***Marks*:**17/20 (IELTS 7.5, CEFR mid-C1)

***Comments*:** There is a wide range of complex sentences, for example: *'In order to enhance students' critical thinking skills, it is necessary to define what they are'; '. . . it should be noticed that although students might realise the importance of critical thinking, they often have trouble using it.'*

***Criterion*:** Vocabulary. ***Marks*:** 15/20 (IELTS 7.0, CEFR low C1)

***Comments*:** There is a wide range of vocabulary and although there are some errors such as *'This essay mainly focus on . . .'* and *'A number of researcher engaged in'*.

Conclusion

The EAP presessional programme at the University of Birmingham has required adjustment over the recent, challenging years. It has offered sensitive provision, addressing the continuing practical needs of students, the requirements and expectations of the various academic departments, the parameters set by the common external frameworks and assessment targets established by external validation bodies. The numbers of international students have remained high, averaging around 1,000 for the last three years although the figures are down from a pre-pandemic peak of 1,800 in 2019. Extending this to a national level, the number of international students has risen significantly, generating over £40 billion (Universities UK, 2023) and contributing greatly to the UK's GDP. Finally, from a research perspective, there is a great opportunity to gather information, producing publications and conference papers using a large real-time database and to contribute to and benefit from contemporary studies. It is fair to say that the needs of several different stakeholders have been addressed and although hard evidence is difficult to ascertain, from a student perspective the results of our satisfaction survey are overwhelmingly positive and feature the following comments:

What were the important skills you've learned on your course?

Don't be shy to express opinions #18

Academic writing skills and express myself bravely #41

How far do you feel you've been stretched and challenged?

The pressure is good, the challenge is great #111

It was a big challenge to understand the teacher at first but gradually I got used to it #261

How much did your level of English improve during your course?

A lot but I still have a long way to go #158

References

Council of Europe. (2001). *Common European Framework for Languages: Learning, Teaching, Assessment.* Cambridge: Cambridge University Press.

Council of Europe. (2020). *Common European Framework for Languages: Learning, Teaching, Assessment.* Strasbourg: Council of Europe Publishing.

Financial Times. (2022, November 25). *University leaders defend benefits of overseas students*. Available online: www.ft.com/content/29c19c5c-db18-4929-a440-a247e7003faf

Freeman, R. (1984). *A Stakeholder Approach to Strategic Management*. New York: Pitman.

HEPI. (Higher Education Policy Institute). (2022). *Student academic experience survey*. Available online: https://www.hepi.ac.uk

HESA. (Higher Education Statistics Authority). (2021). *Where do HE students come from?*. Available online: https://www.hesa.ac.uk

Lomer, S., & Mittelmeier, J. (2021). *Cash cows or pedagogic partners? Mapping pedagogic practices for and with international students*. SHRE Research Report January 2021. Available online: https://srhe.ac.uk/wp-content/uploads/2021/02/Lomer-Mittelmeier-CarmichaelMurphy-FinalReport-SRHE.pdf

Nye, J. S. (1990). Soft Power. *Foreign Policy*, *80*, 153–171.

Reavill, L. R. P. (1998). Quality assessment, total quality management and the stakeholders in the UK higher education system. *Managing Service Quality: An Internal Journal, 8*(1), 55–63.

Universities UK. (2023, May 16). International students boost UK economy by 41.9 billion. *Universities UK.* Available online: https://www.universitiesuk.ac.uk/latest/news/international-students-boost-uk-economy

Overcoming challenges in aligning language assessments to standards

Fabiana MacMillan
WIDA at the University of Wisconsin-Madison, USA

David MacGregor
WIDA at the University of Wisconsin-Madison, USA

Jason A. Kemp
WIDA at the University of Wisconsin-Madison, USA

Abstract

Aligning language tests to external language proficiency standards is increasingly required by policy, often with high-stakes implications (e.g., English language requirements for settlement/naturalization or careers in aviation). In the US Kindergarten-Grade 12 (K-12) education context, federal law requires that states assess the English language proficiency of English learners with a test aligned to a state's English language proficiency standards. Regardless of context, since no one correct way to align a test to standards exists, test developers claiming alignment with standards must defend the consistent application of the selected alignment methodology (Fulcher, 2016). Webb (1997, 1999) provides a useful alignment framework, and the associated Web Alignment Tool (2005) can help conduct the study and analyze results. However, the methodology and tool were designed for alignment of content-area tests to standards. We discuss challenges faced in using Webb's alignment framework for a language proficiency test and how we addressed these concerns in our alignment study.

Introduction

Federal law in the US requires that all children in Kindergarten-Grade 12 (K-12) schools receive equal opportunities for academic advancement, including language support for students identified as English learners (ELs) as needed. The most current iteration of federal law, the Every Student Succeeds Act (ESSA, 2015), requires that each state 'provide for an annual assessment of English proficiency of all English learners' and that these assessments 'be aligned with the State's English language proficiency standards' (2015). By law, then, test developers must demonstrate such alignment, including updating the evidence when standards are updated or when assessments are revised. In this paper, we discuss the procedures used to align one such assessment, and the benefits and challenges we faced in applying a procedure developed for content-area tests to a language test.

Background

WIDA at the University of Wisconsin-Madison is a consortium of 41 state education agencies. Through a commitment to equity and social justice for culturally and linguistically diverse learners, WIDA provides members of the consortium with English language proficiency (ELP) assessments, English language development standards, and professional learning opportunities for K-12 educators. The WIDA English Language Development (ELD) Standards Framework (2007, 2012) link language development and academic content area. The five standards are: 1) Social & Instructional Language, 2) Language of Language Arts, 3) Language of Mathematics, 4) Language of Science, and 5) Language of Social Studies. ACCESS, our secure, large-scale summative ELP assessment, operationalizes the WIDA ELD Standards. ACCESS is not a test of content knowledge; it monitors students' progress in acquiring English language proficiency for academic contexts. ACCESS has four domain tests: Listening, Reading, Speaking and Writing. There are both paper and online versions of ACCESS. Both versions of the test are tiered (Tiers A, B, and C), and online ACCESS is semi-adaptive (Listening and Reading domain tests). Kindergarten ACCESS is a paper-and-pencil test. More than two million K-12 ELs take ACCESS annually.

Alignment method

Alignment has been viewed historically in terms of content validity, as a means to assure that the content in the assessment matches the content present in both the curriculum and classroom instruction (Madaus, 1983). More recently, alignment has been understood as a means of ascertaining that the items included in the assessment reflect the standards in three dimensions: match, breadth, and depth (Cook, 2006; Webb, 1997, 1999). In aligning ACCESS to the WIDA ELD Standards, we followed the Webb Alignment Procedure (Webb, 1997, 1999), which defines the three dimensions as follows:

Table 1: Webb's depths of knowledge

DOK level	*Description of level*
1	Recall & Reproduction
2	Skills & Concepts
3	Strategic Thinking & Reasoning
4	Extended Thinking

- Match refers to how well an assessment covers the standards. It is evaluated by calculating Categorical Concurrence (CAT), the average of the number of test items raters assign to specific standards. To meet this criterion, Webb suggests that an assessment must have at least six items measuring content from a standard.
- Depth refers to whether an assessment is as cognitively demanding as what students are expected to know and do per the standards. It is evaluated by calculating Depth-of-Knowledge Consistency (DOK), the percentage of items coded at the level of complexity of a standard. To meet this criterion, Webb suggests that at least 50% of the items corresponding to a standard must be at or above the level of knowledge of that standard. For this criterion, Webb defines four levels of cognitive depth, as shown in Table 1 (Hess, 2013).
- Breadth refers to how well an assessment covers the span of knowledge in the standards. It is evaluated by calculating two statistics, Range-of-Knowledge Correspondence and Balance of Representation. Range-of-Knowledge Correspondence is the percentage of objectives within the standard with at least one related assessment item. Balance of Representation is the extent to which items are evenly distributed across standards. To meet these criteria, Webb suggests a minimum of 50% for Range-of-Knowledge Correspondence and a value of at least .70 for Balance of Representation.

Procedures

Twenty experts external to WIDA participated in a three-day Zoom meeting in December 2021. We recruited panelists who had:

- deep knowledge of the WIDA English Language Development Standards (2007, 2012)
- great familiarity with the ACCESS assessment
- experience with alignment studies and/or similar standards and assessment activities.

After an initial training, panelists participated in a two-step process to assess the degree to which ACCESS is aligned to the Standards. In Part 1, panelists assigned linguistic difficulty levels (LDLs) to standards statements. The three LDLs within each skill domain are as follows:

- Level 1: Elementary Features
  - Limited to basic demands for processing formulaic English linguistic features
- Level 2: Intermediate Constructions
  - Basic to moderate demands for processing English linguistic features
- Level 3: Advanced Formulations
  - Moderate to sophisticated demands for processing English linguistic features

In Part 1, WIDA facilitators helped panelists in their grade-level group come to a consensus on the LDLs for each standards statement. First, panelists assigned LDLs to each statement independently. Then, panelists viewed each other's LDL ratings, discussed, and came to consensus on LDL assignments for each statement prior to moving on. In Part 2, panelists individually assigned LDLs to ACCESS test items and identified the standards statement(s) that best matched the test items. Facilitators discussed two or three test items with their group. This conversation allowed panelists to talk about any differences in perspective and helped promote a calibrated understanding of the process. Note, though, that for this step consensus was not required. Panelists entered data into the publicly available Web Alignment Tool (WAT). Upon completion of the study, data were extracted from the WAT and analyzed.

Affordances and constraints

A clear benefit of the Webb Alignment Procedure is that it provides a well-defined methodology for alignment with set procedures and objective measures of alignment. Additionally, the WAT allows panelists to enter their judgments individually, and provides a myriad of tables to analyze the results.

However, we also encountered several challenges in applying the procedure to a language proficiency test. First, the WAT allows for alignment across one dimension only, by aligning test items to the standards. However, due to the design of ACCESS, we were aligning it to the standards in two dimensions: first across all four domains, and second within each domain by proficiency level. To account for this two-dimensional approach, we had to download the data from the WAT and create Excel spreadsheets to do the analysis.

Second, since the DOK levels are based on cognitive abilities, they do not apply to language tests, as language proficiency is not an indicator of cognitive complexity. Therefore, rather than Webb's DOK levels, we used Cook's (2006) Linguistic Difficulty Levels (LDLs; Table 2).

Table 2: Cook's (2006) Linguistic Difficulty Levels

LDL	*Label*	*Description of level*
1	Elementary Features	Limited to basic ability to process and produce formulaic English linguistic features
2	Standard Constructions	Basic to moderate ability and facility to process and produce English linguistic features
3	Complex Formulations	Moderate to sophisticated ability and facility to process and produce English linguistic features

For match, because of the complex test design, we had to adjust the criteria, in some cases increasing the minimum needed to meet, and in some cases decreasing it. Table 3 shows the criteria we used for domain by proficiency level (PL). Finally, the breadth criterion is difficult to meet in an assessment such as ACCESS where the sampling of the standards is intentionally unbalanced.

Table 3: Categorical concurrence criteria for domain by PL

	Listening & Reading			*Speaking*			*Writing*		
PL	*No*	*Weak*	*Yes*	*No*	*Weak*	*Yes*	*No*	*Weak*	*Yes*
1	<2	≥2	≥3	<0.5	≥0.5	1	<0.5	≥0.5	1
2	<4	≥4	≥6						
3	<7	≥7	≥9	<0.5	≥0.5	1	<0.5	≥0.5	1
4	<4	≥4	≥6						
5	<2	≥2	≥3	<0.5	≥0.5	1			

Results

In analyzing the results within a domain by proficiency levels, we found that the match and depth criteria were met for almost all proficiency levels across the domains. Match was met for all PLs on 17 of the 20 test forms, and no test form had more than one PL with either no or weak match. Depth was met for all PLs on 10 of 20 test forms, and for all but one PL on six test forms.

On the other hand, the breadth criteria had mixed results. Range was met for all PLs on six of 10 Listening and Reading test forms, while no Listening or Reading test form had more than one PL that did not meet the criterion. On the other hand, none of the 10 Speaking and Writing test forms met the range criterion for all PLs. Meanwhile, balance was generally weak or not met for most PLs across all domains.

For the standards-across-domains alignment, match was strongly met for all standards across all clusters, and depth was strongly met for all standards across all clusters with one exception, where it was moderately met.

On the other hand, the breadth criteria had mixed results. Range was generally strong or moderate in all standards across all clusters, though in two cases it was found to be limited, while balance was limited or moderate in all cases.

Conclusion

As noted earlier, the methodology and tool Webb (1997, 1999) developed were designed for the alignment of content-area tests to standards. In order to use Webb's methodology for the alignment of standards to an English language proficiency test, we adjusted our criteria for determining alignment given test design and specifications. For example, we relied on Cook's (2006) guidance tailored to language proficiency assessments. Cizek, Kosh and Toutkoushian (2018) critiqued traditional alignment methods as they believe that these approaches to alignment are not grounded in the stated purpose of a testing program, yield results that are not consistently reliable, and include somewhat arbitrary criteria. We tried to address some of these critiques by incorporating elements of Cook's (2006) LDL instead of relying on Webb's (1999) DOK levels. However, we know there is more work to be done in this area.

Wolf, Bailey and Ballard (2023) called for greater nuance when examining the relationship that exists between language complexity and ELP assessments and standards. We will keep this perspective, among others, in mind as we prepare for our next alignment study. In December of 2020, WIDA released a new edition of our English Language Development Standards Framework. This new edition of the Standards is shaping updates to ACCESS. Per the US Department of Education's (2018) assessment peer review process, new standards or revisions to the test require evidence of alignment. WIDA will conduct another alignment study and submit the results to the Department of Education on behalf of the Consortium once the updated version of ACCESS is released.

References

Cizek, G. J., Kosh, A. E., & Toutkoushian, E. K. (2018). Gathering and evaluating validity evidence: The generalized assessment alignment tool. *Journal of Educational Measurement*, *55*(4), 477–512.

Cook, H. G. (2006). Aligning English language proficiency tests to English language learning standards. Report 5 in *Aligning assessment to guide the learning of all students: Six reports on the development, refinement, and dissemination of the Web alignment tool*. Washington, D.C.: Council of Chief State School Officers.

Every Student Succeeds Act. (ESSA). (2015). Available online: https://www.congress.gov/114/plaws/publ95/PLAW-114publ95.pdf

Fulcher, G. (2016). Standards and frameworks. In J. Banerjee & D. Tsagari (Eds.), *Handbook of Second Language Assessment* (pp. 29–44). Berlin: De Gruyter Mouton.

Hess, K. K. (2013). *A guide for Using Webb's Depth of Knowledge with Common Core State Standards*. Washington, D.C.: The Common Core Institute.

Madaus, G. F. (1983). *The Courts, Validity, and Minimum Competency Testing*. Boston: Kluwer-Nijhoff.

U.S. Department of Education. (2018). *A state's guide to the U.S. Department of Education's assessment peer review process*. Washington, D.C.: Office of Elementary and Secondary Education.

Webb, N. L. (1997). *Criteria for alignment of expectations and assessments in mathematics and science education*. Council of Chief State School Officers and National Institute for Science Education Research Monograph No. 6. Madison: University of Wisconsin, Wisconsin Center for Education Research.

Webb, N. L. (1999). *Alignment of science and mathematics standards and assessments in four states*. Research Monography No. 18. Madison: National Institute for Science Education at the University of Wisconsin-Madison.

WIDA. (2007). *English Language Proficiency Standards and Resource Guide, Prekindergarten Through Grade 12*. Madison: Board of Regents of the University of Wisconsin System.

WIDA. (2012). *2012 Amplification of the English Language Development Standards, Kindergarten–Grade 12*. Madison: Board of Regents of the University of Wisconsin System.

WIDA (2020). *WIDA English Language Development Standards Framework, 2020 Edition: Kindergarten-Grade 12*. Madison: Board of Regents of the University of Wisconsin System.

Wisconsin Center for Education Research (2005).

Web Alignment Tool. Available online: http://watv2.wceruw.org

Wolf, M. K., Bailey, A. L., & Ballard, L. (2023). Aligning English Language Proficiency Assessments to Standards: Conceptual and Technical Issues. *TESOL Quarterly*, *57*(2), 670–685.

Mediation: From theory to practice

Mary Luz Castro Pérez
Consejería de Educación, Formación Profesional, Actividad Física y Deportes de la Comunidad de Canarias

Joaquín Cruz Trapero
Centro de Estudios Avanzados en Lenguas Modernas, Universidad de Jaén

Olga Naranjo Travieso
Consejería de Educación, Formación Profesional, Actividad Física y Deportes de la Comunidad de Canarias

Adolfo Sánchez Cuadrado
Departamento de Lingüística General y Teoría de la Literatura, Universidad de Granada

Abstract

The publication of the Common European Framework of Reference for Languages Companion Volume (CEFR CV; Council of Europe, 2020) has sparked off an intense debate on mediation in different contexts. From the theoretical standpoint, the conceptualization of mediation is still challenging. From a practical perspective, integrating mediation in curricula and language assessment poses different questions and challenges. In this paper we propose a debate at three different levels.

First, we delve into the notion of mediation, a complex and multi-faceted mode of communication that complements the communicative competence construct for foreign/second/additional language learning. Second, we analyze the challenges that mediation poses in developing scales for its assessment. Finally, we discuss how mediation could impact one particular context, namely the Spanish Official Language Schools (OLS), which have included mediation in their curricula and in their language proficiency tests.

Introduction

Foreign language teaching and assessment in Spain is partially managed by a public network of 449 OLS (MEFP, 2021). These schools provide lessons in multiple languages and prepare high-stakes proficiency tests for levels A2 to C2. Every year, the OLS system caters for approximately 400,000 students (MEFP, 2019, p. 2), who find in it a quality alternative to private teaching and certification. Despite having a common curriculum for teaching (BOE, 2017) and common specifications for test design (BOE, 2019), the decentralization of Spanish education allows for a great deal of diversity and autonomy across OLS.

Following the publication of the Common European Framework of Reference for Languages Companion Volume (CEFR CV; Council of Europe, 2020), Spanish legislation made it compulsory for OLS to include mediation in their curricula (BOE, 2017) and proficiency tests (BOE, 2019). In the following sections, we describe how mediation has been operationalized in Spanish OLS, departing from theoretical and landing on practical grounds.

First challenge: Conceptualizing mediation for language assessment and testing

The vast array of mediation activities and strategies put forward in the CEFR CV sheds new light on language competence and allows the integration of many facets of language use that have been sometimes overlooked, e.g., cross-linguistic practices, interpersonal affective factors, co-construction of meaning or intercultural competence, just to name a few. Two major implications of this are: how mediation interacts with the other three modes of communication, and how mediation may foster a shift of focus when assessing communicative competence.

As for the first issue, while reception and production focus on the exchange of meaning in a unidirectional fashion and a marked tendency to asynchrony, and interaction includes the negotiation of meaning in a bidirectional fashion and a tendency to synchrony, mediation focuses on the construction of such meaning, which is a more elaborate way of accessing and reprocessing it. This operation can be both intrapersonal (as when we 'dialog' with ourselves to process meaning or when we help others

access meaning without any interaction), but it can also have a more social dimension, be it an interactive activity or in an unidirectional procedure, i.e., first reception and then production (Sánchez Cuadrado, 2022).

As for the second issue, among all the aspects that play a major role in language use, the ones that mediation clearly highlights are the conditions in which the communicative event takes place and the constraints that foster the use of language strategies. In a more traditional way, when assessing language competence in general language tests (i.e., not for specific purposes), the tendency has been to downplay task achievement criteria as compared to language-specific criteria. However, due to the complexity of the mediation construct, mediating strategies and abilities derived from the conditions and constraints of the task are of paramount importance, just as much as the linguistic competences which enable their effective deployment.

In conclusion, the jury is now out on whether it is time to turn to integrative assessment for us to be able to tackle these two dimensions of mediation, that is, its multifaceted nature and the subsequent need to re-balance task achievement and linguistic competences.

Second challenge: Developing scales for mediation

After the conceptualization of mediation, we discussed different challenges linked to its practical implementation. We discussed what happens when the need to assess mediation arises focusing on the context of the aforementioned Spanish OLS.

Assessing mediation is challenging in two different ways. Besides the normal problems that arise in the development and validation of assessment scales, we must also consider the challenges related to the multifaceted nature of mediation described in the previous section, and how such nature can be captured by a rating scale.

For the development of the scales we followed a protocol (Cruz, 2016) which contains six stages ranging from initial considerations about the shape of the scale (type of scale, number and characterization of dimensions, number of bands, etc.) to a two-sided validation process (qualitative and quantitative). We arrived at a scalable model that draws on the CEFR (Council of Europe, 2001, pp.181–182).

At this intermediate stage between theory (the conceptualization of mediation) and practice (effective assessment of mediation in the curriculum), finding consensus about the shape of the scales was particularly difficult. On the one hand, the opinion of all stakeholders should be heard so that the final product is not perceived as an imposition but as the result of a collaborative effort. On the other hand, creating working groups that are efficient while different opinions are taken into account may be difficult unless there is clear leadership with technical and practical authority to make decisions when multiple options are available.

When transferring mediation from the CEFR CV to operational scales, the main challenge was to capture the multifaceted nature of mediation and to determine whether it should be assessed independently or in an integrative way, as anticipated in the section above on conceptualization. We also noticed that while the conceptual framework of the CEFR and the CV is based on the notion of modes of communication, most language proficiency tests are based on a traditional four-skill approach. Finally, intralinguistic vs. crosslinguistic approaches to mediation were discussed. While the former seems to be more operational in the context of proficiency assessment, the latter is closer to the spirit with which mediation is described in the CV.

A case study: Official Language Schools in the Canary Islands

Finally, we illustrate the previous theoretical framework of mediation and its evaluation through the practical case of the OLS in the autonomous region of the Canary Islands, Spain. As mentioned in the introduction, these are public schools where adults can study a language, choosing from one of the eight offered: Arabic, Chinese, English, French, German, Italian, Portuguese, and Spanish as a second language. These schools not only teach languages but also develop certification proficiency exams from A2 to C2.

In the context of the publication of legal regulation in Spain concerning the curriculum of foreign languages for teaching, common specifications for test design and, consequently, regional regulation in 2018 (BOC, 2018) and 2022 (BOC, 2022), the OLS in the Canary Islands started implementing mediation both in teaching and testing.

The first step to undertake this challenging task was providing all the stakeholders with training to conceptualize mediation. This training started in October 2018 with a lecture to Canary OLS teachers by two experts from *Escuela Oficial de Idiomas de Elda* (Alicante, Spain) and continued, in February 2019, with the first Regional Conference, to which mediation experts from the Council of Europe and from the National and Kapodistrian University of Athens were invited. In July 2019, a team of five teachers received a week's training on mediation, which was the first step to an intensive cascade training plan taking place in all schools in the Canaries throughout 2019 to 2020.

While teachers were being instructed on mediation, the coordination team in the Canary Islands' Education Department started developing the specifications for the high-stakes certification proficiency construct. The decisions taken on the specifications, as well as the type of mediation (intralinguistic or crosslinguistic) and the mediation activities and strategies to be assessed, would be fundamental due to the backwash effect that these exams have on the daily teaching practice. Finally, as part of the certification of proficiency exams, mediation was to be assessed as the fifth part of these exams by a written task and a spoken task, with both tasks assessing intralinguistic mediation.

Once the decisions had been taken, it was time to develop the scales to assess these mediation tasks, which we have already referred to in a previous section. The first stage was deciding on the type of scales that would suit our tasks' specifications, finally opting to develop four analytical scales (from B1 to C2) for written mediation and four for oral mediation. The structure of these scales included a horizontal axis with five bands for the score and a vertical axis for the dimensions, each one containing two descriptors per band. The scales did not include descriptors for Bands 2 and 4 due to the fact that in our scalable model these correspond to plus levels, which are not always evenly developed in the CEFR or in the CEFR CV.

So far, there are two versions of these scales, the first one used in the exams held in 2019/2020 and 2020/2021 academic years and the second one in 2021/2022 and 2022/2023. The first version assessed two dimensions, Task Fulfillment and the Ability to Mediate, which was named Communication Management for the written mediation scales and Conveying Information for the oral mediation ones. The decision about not evaluating the linguistic competence was taken since those competences were already assessed in written and oral exams.

After implementing this first version of the mediation scales, we realized that they required some improvements mainly because the wording of some descriptors had proved to be confusing and because some of them were assessing the same aspects. In addition to that, we took the decision to merge the concept of Communication Management and Conveying Information into the more precise dimension of Mediation Ability, which is what we really assess. While redeveloping the wording of the descriptors, it proved to be truly useful to list keywords and ideas associated with each level in the CEFR. We underlined the keywords in the scales descriptors and wrote instructions for examiners on the other side of the paper; two simple things that worked remarkably well.

In spite of all the progress made on teaching and assessing mediation and evaluating mediation test design, there is still plenty of room for improvement, mainly on the area of mediation scales. Throughout the 2023/2024 academic year, the second version of these scales will be revised and we hope to be able to pilot and statistically validate them.

Conclusion

As we have seen, the conceptualization, teaching and assessment of mediation is a complex task. The main challenge lies in creating an assessment tool which is operational and, at the same time, captures the multifaceted nature of mediation. A pioneer attempt to include mediation in the curriculum and in assessment was developed in the OLS of the Canary Islands. The difficulties found along the way have led to a deep reflection on the very nature of mediation and to the implementation of innovative teaching and assessment methods which are slowly shifting their focus to the co-construction of meaning among users of language.

References

BOC (Boletín Oficial de Canarias, 16 October). (2018). Decreto 142/2018, de 8 de octubre, por el que se establece la ordenación y el currículo de las enseñanzas y la certificación de idiomas de régimen especial para la Comunidad Autónoma de Canarias. *Boletín Oficial de Canarias*, *200*, pp. 32590–33287.

BOC (Boletín Oficial de Canarias, 26 September). (2022). Orden de 15 de septiembre de 2022, por la que se regula la evaluación del alumnado de enseñanzas de idiomas y de las pruebas de certificación de idiomas de régimen especial en la Comunidad Autónoma de Canarias. *Boletín Oficial de Canarias*, *190*, pp. 33749–33771.

BOE (Boletín Oficial del Estado, 23 December). (2017). Real Decreto 1041/2017, de 22 de diciembre, por el que se fijan las exigencias mínimas del nivel básico a efectos de certificación, se establece el currículo básico de los niveles Intermedio B1, Intermedio B2, Avanzado C1, y Avanzado C2, de las Enseñanzas de idiomas de régimen especial. *Boletín Oficial del Estado*, *311*, pp. 127773–127838.

BOE (Boletín Oficial del Estado, 12 January). (2019). Real Decreto 1/2019, de 11 de enero, por el que se establecen los principios básicos comunes de evaluación aplicables a las pruebas de certificación oficial de los niveles Intermedio B1, Intermedio B2, Avanzado C1, y Avanzado C2 de las enseñanzas de idiomas de régimen especial. *Boletín Oficial del Estado*, *11*, pp. 2260–2268.

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge: Cambridge University Press.

Council of Europe. (2020). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment — Companion Volume*. Strasbourg: Council of Europe Publishing.

Cruz, J. (2016). *A Protocol to Design a CEFR-linked Proficiency Rating Scale for Oral Production and its App Implementation*. Jaén: Universidad de Jaén.

MEFP (Ministerio de Educación y Formación Profesional). (2019). *Estadística de las Enseñanzas no universitarias. Enseñanza de Lenguas Extranjeras. Curso 2017–2018*. Available online: https://bit.ly/3ydwDPT

MEFP (Ministerio de Educación y Formación Profesional). (2021). *Escuelas Oficiales de idiomas en España — Curso 2020–2021*. Available online: https://bit.ly/3Hpzx8x

Sánchez Cuadrado, A. (coord.) (2022). *Mediación en el aprendizaje de lenguas. Estrategias y recursos*. Madrid: Anaya.

From mediation to knowledge transformation: Expanding the construct of the reading-into-writing task

Alina Reid
Trinity College London, United Kingdom

Abstract

This paper examines mediation in the written mode enacted through the integrated reading-into-writing task type. Like the concept of mediation itself, the integrated writing task is rapidly gaining popularity by virtue of its similarity and pertinence to real-life writing tasks that students encounter in the academic domain. Language testing professionals are thus faced with new challenges in defining and operationalising the construct of the integrated writing task. The paper discusses how the CEFR descriptive categories of text mediation can be adapted, mixed and, indeed, *expanded* in order to create a reading-into-writing writing task that achieves a fuller representation of writing in the academic domain. Particular emphasis is placed on the cognitive process of *knowledge transformation*, known as the 'cognitive trademark' of academic writing (Flower, 1990). The paper defines knowledge transformation, highlighting its importance to the construct of academic writing, and suggests practical task design principles to elicit it.

Introduction

The addition of new mediation scales to the Common European Framework of Reference for Languages (CEFR) (Council of Europe, 2001) in its Companion Volume (CV) (Council of Europe, 2020) reflects the major shift towards skill integration, ever more prominent in communicative language teaching and assessment. Mediation holds particular relevance in educational contexts such as Content and Language Integrated Learning (CLIL), English as a Medium of Instruction (EMI) and English for Academic Purposes (EAP), where second language users interact with multiple information sources to co-construct meaning. When writing for educational purposes, students typically interpret, adapt, and synthesize information from external sources, combining it with existing knowledge to create new texts. Writing from sources requires a unique set of abilities which was sorely lacking from the CEFR descriptors (Chan, 2018). As such, the inclusion of the Mediating a Text scales is a welcome addition, and is likely to impact the definition and operationalization of integrated writing tasks.

However, language testing practitioners must avoid an overly broad or prescriptive adoption of the CEFR scales (Council of Europe, 2009). While the scales for Mediating a Text describe many of the essential competencies required to write from sources for academic purposes, there are also notable gaps and discrepancies. These stem from the CEFR's continued approach to describe proficiency as *context-free* yet *context-relevant* (Council of Europe, 2001). As the mediation scales are not specific to any particular context, practitioners must critically utilise the descriptors and reflect on competence within their specific contexts (Council of Europe, 2001; 2020).

This paper reports on the experience of drawing on the Mediating a Text scales in the process of developing an experimental reading-into-writing task at Trinity College London. Theoretical exploration delves into areas of congruence between the CEFR scales for mediation and real-life academic writing, highlighting potential opportunities to expand the construct of Text Mediation, with a focus on knowledge transformation — a key aspect of academic writing. From a practical task design perspective, the study discusses how knowledge transformation can be elicited and briefly presents empirical evidence of test-takers experiencing it during the task.

Mediating a text

Given the CEFR's major role in the field of language testing, it is becoming common practice to closely consult the scales from the very early stages of test and task design (Chan, 2018). When designing the experimental reading-into-writing task, the Trinity

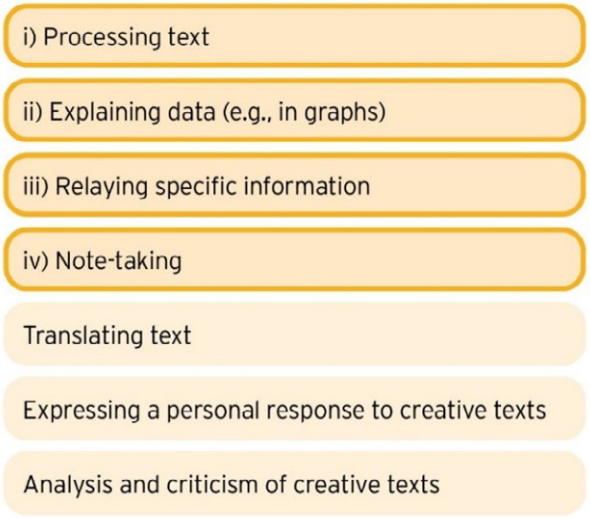


Figure 1 Areas of congruence between CEFR text mediation and academic writing

test development team turned to the mediation scales during the initial construct definition phase. This section briefly reports on the theoretical discussions held to define the task's underlying construct and the role played by the Mediating a Text scales.

The task targets academic writing, primarily in secondary and tertiary education settings. The scales were critically analysed to identify relevant areas of competence based on literature in academic writing. From the seven text mediation types, those related to translation or specific to literary and artistic studies were excluded, selecting only the four most applicable to general academic writing.

Processing text (i) introduces the ability to report on facts, arguments and viewpoints by paraphrasing, summarising, and synthesizing information from multiple texts. It is complemented by *explaining data* (ii), which describes the ability to verbalise information presented graphically. A further underlying ability is that of finding and relaying specific, relevant information from texts (iii). Note-taking (iv) introduces additional concepts related to accuracy and selectivity of source use. Together, the four scales introduce the essential competencies underlying academic writing. However, two points of disparity emerge.

The ability to discern and discard irrelevant content, ensuring relevance to the writing purpose, is of particular importance to the academic domain (Wette, 2021). The concept is mentioned in some scales and implied in others, but explicit and comprehensive description is lacking, and relevance is conceptualised in relation to the original text.

Furthermore, effective source writing necessitates transformative and adaptive use of sources to support the writer's own purpose (Wette, 2021). Here, the CEFR offers sporadic mentions of 'reformulating', 'reordering' and appropriacy to 'context' and 'communicative goal'. Despite such references, overall, the communicative goal of mediating text is conceptualized as creating another text which revolves almost entirely around the original. The writer is preoccupied not just with accurately transmitting the content of the original, but also promoting its purpose, its viewpoint, and even preserving its style and register. Adaptive and transformative use of sources is not conceptualised in the scales, as both the writer and the reader are assumed to be primarily interested in writing/reading about the original text. A visual representation of the overlap between the original text and the output, as conceptualized by the CEFR, might look something like Figure 2.

The output is a condensed version of the original, achieved through summary or omission of information. It represents source-driven writing where the mediator lacks writerly purpose, aiming solely to offer readers a more accessible rendition of the original text. However, academic writing differs significantly, as it involves transforming and building on source texts to support the writer's own purpose (Chan, 2018; Wette, 2021).

The CEFR scales for text mediation encompass many essential competencies for academic writing, thus serving as a valuable starting point in defining the task's construct. However, consulting broader literature on writing from sources for academic purposes revealed the need to extend the construct beyond source-driven writing.

Writing in the academic domain

Briefly put, writing for academic purposes means writing in order to learn and to display learning (Schleppegrell, 2004). Writing is text-responsible and intertextual (Leki, & Carson, 1994; Shaw, & Pecorari, 2013), as we slowly advance from writing independently

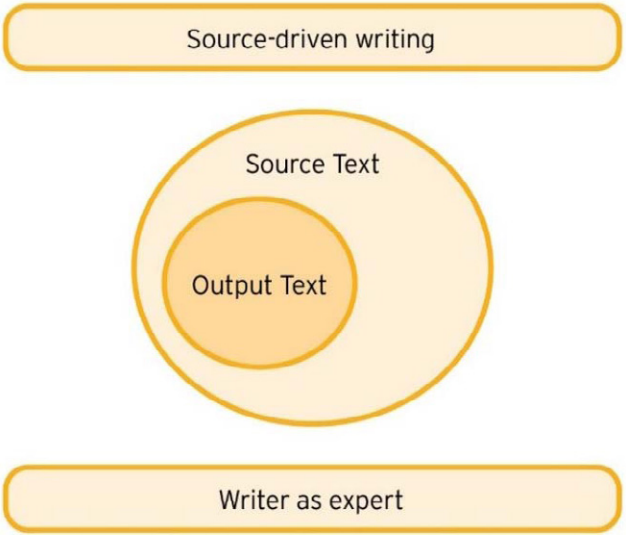


Figure 2 Input/output interaction in CEFR text mediation

and self-referentially about our pets and holidays, to integrating multiple sources and representing them responsibly. Academic writing is supported but not driven by sources. As we tackle the building blocks of academic discourse, we progress from source-driven writing (e.g., summary writing) towards finding our own voice and writing with a new communicative purpose (Wette, 2021). We are encouraged and assisted to progress from writing to tell of our existing knowledge towards writing to transform our knowledge (Scardamalia & Bereiter, 1987).

Knowledge transformation

Knowledge transformation (KT) is the 'cognitive trademark of academic discourse' (Flower, 1990, p. 249) and the purpose of academic pursuits. It represents critical literacy rather than basic literacy (Flower, 1990; Scardamalia & Bereiter, 1987). KT emerges while writing from sources if the writer approaches the task as a rhetorical problem-solving act. It manifests itself as **Dialogic Reading**, where the writer engages with source texts critically, entering a mutually transformative dialogue that enhances both their knowledge and the propositional meaning of the sources. During the writing phase, KT is seen in **Constructive Planning** *if* the writer activates meta-cognitive awareness of their own writing, turning off the autopilot and engaging in deliberate self-reflection. It is important to note that KT complements comprehension-response strategies rather than replacing them. Moreover, it is not guaranteed or essential, and not every reading-into-writing task will elicit it.

To stimulate KT, the task must present a rhetorical challenge that necessitates problem-solving for the writer (Flower, 1990). This, in turn, relies on task-writer characteristics, including source complexity, rhetorical function (e.g., summarising vs. synthesising), and the degree of similarity between sources and expected output. The degree of rhetorical innovation required determines the presence of KT. When the sources closely resemble the output, the writer's task is not one of innovation but rather replication of the original, similar to the CEFR conceptualisation of Mediating a Text.

To summarise, while KT is not guaranteed or always necessary, it remains the primary goal of real-world academic assignments. Although it is possible for students to get by without it, relying solely on basic comprehension-response writing often fails to meet academic expectations (Flower, 1990). Therefore, language tests measuring academic writing should consider KT in their construct definition and task design, particularly as we are entering the age of generative AI.

Practical task design considerations

The Trinity College London team aimed to create an experimental reading-into-writing task for the academic context. The objective was to elicit output-driven writing i.e., writing which serves a new communicative purpose supported by the source texts but not entirely driven by them. The purpose of writing is not to merely showcase understanding of the sources or relay the information faithfully. Instead, the purpose of writing is to accurately but selectively and adaptively use the sources to serve a new audience and writing purpose. Figure 3 presents a graphic representation of the overlap between the sources and the expected output.

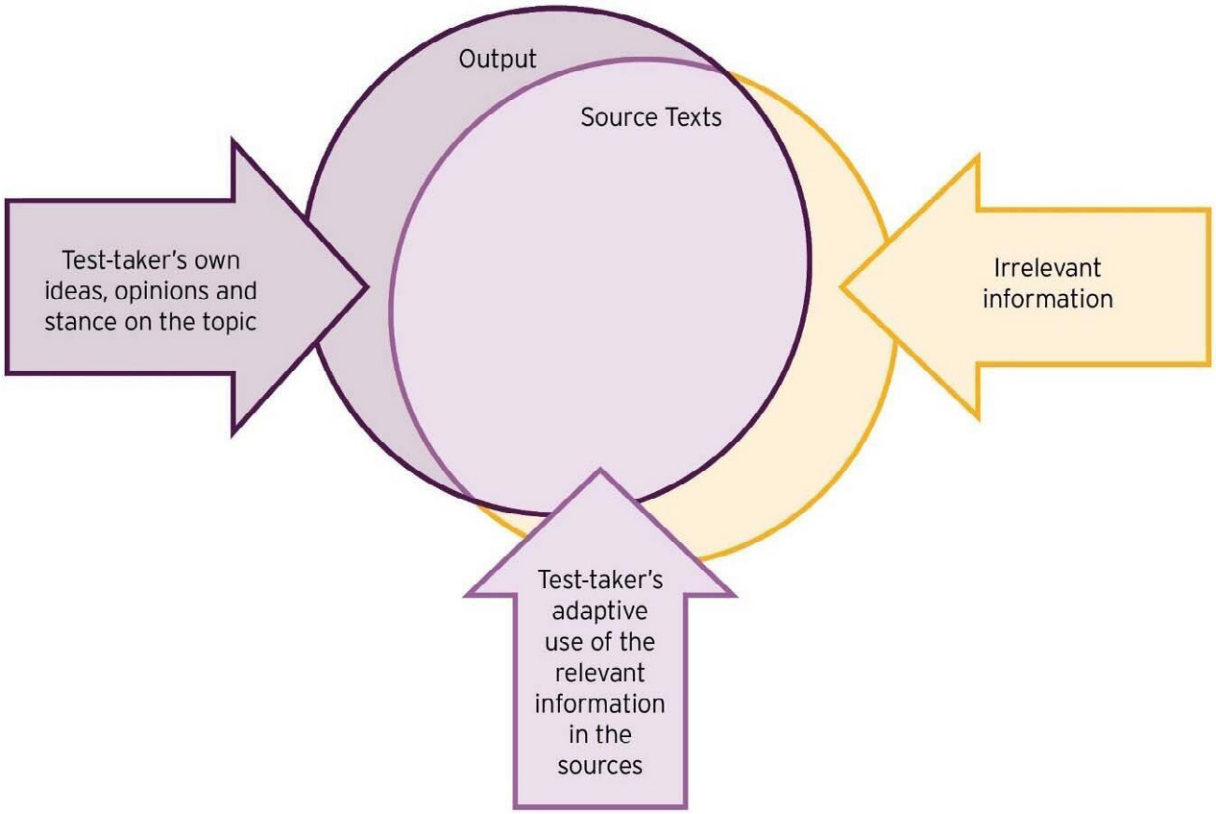


Figure 3 Input/output interaction in writing from sources

To achieve the expected output, the test-taker must read the sources, selecting relevant information, and then adapt it linguistically, conceptually, and rhetorically to suit the new communicative aim set by the prompt. It is worth stressing this adaptive and interpretative use of sources as it is crucial in creating the rhetorical challenge that elicits KT. Merely selecting and paraphrasing, or even summarising, are more likely to rely on knowledge-telling processes rather than transformation (Chan, 2011; 2018).

When designing the task and setting the specifications, special consideration was given to the overlap between the sources and the expected output. This overlap and similarity were carefully calibrated to elicit rhetorical innovation by varying textual characteristics like genre, purpose, audience, and topic (as illustrated in Table 1). This creates a rhetorical-problem space, requiring candidates to select and adapt relevant information, repurposing it to align with the new communicative aim and context.

Empirical validation

A study was carried out to examine the cognitive processes and strategies prompted by the task to assess its resemblance to real-life writing from sources. It involved 16 international students from various UK universities, who completed two task versions while reporting their cognitive activity through concurrent thinking-aloud and screen sharing. Although the study was considerably broader in scope, this paper focuses solely and briefly on the KT component.

Table 1: Input/output characteristics

	SOURCE TEXT	*EXPECTED OUTPUT*
Genre	Article	Essay
Purpose	Recommend / Inform	Argue
Audience	Hiring managers	Course tutor
Topic	How can you hire the best employees?	Should people start career preparations as young as possible?

Table 2: Examples of KT

Dialogic reading	'I don't think is one of the ways to solve the traffic. It's making more traffic instead of reducing it.'
Constructive planning	'Yeah, I think I stick to this method more. So, I think I should make it more clear that I believe in the first method more, but I do want to give them another option.'

KT was observed in the task performance, accounting for about 13% of the cognitive activity, while basic comprehension-response strategies occupied the remaining portion. This finding aligns with prior research indicating that KT complements, rather than replaces, basic literacy practices (Flower, 1990). **Dialogic Reading** occurred frequently (193 instances), with participants engaging in supportive/critical reading, involving inferences, evaluations, explications, examples, or counterexamples beyond mere paraphrasing. Table 2 illustrates an example where a participant adds their own inference and evaluation to the propositional meaning of the text. **Constructive Planning** was also evident (102 times), with participants reflecting on their writing purpose, self-evaluating, and considering alternative approaches. In the example below, upon reading the first draft of her text, this participant reflects on the clarity of her stance and comes up with a possible improvement, thus illustrating the type of metacognitive awareness characteristic of KT.

Concluding remarks

This paper briefly examined the usefulness and applicability of the new CEFR Mediating a Text scales to assessing academic writing. Areas of alignment as well as important gaps, particularly concerning KT, were identified. It was argued that, as an essential aspect academic discourse, KT should be accounted for in any task purporting to assess academic writing. Practical task design suggestions were provided, advocating for output-driven integrated writing tasks that involve rhetorical problem-solving to elicit KT, moving away from source-driven approaches. Empirical findings demonstrated test-takers experiencing KT in an experimental writing-from-sources task. It is hoped the paper will assist other practitioners in integrating and adapting the CEFR mediation scales to the context of academic writing.

Finally, a possible opportunity to expand the construct of Text Mediation, as currently conceptualized in the CEFR, emerges through the addition of a further competency that describes the ability to write from sources in order to learn and display learning.

References

Chan, S. H. C. (2011). *Research Note: Demonstrating cognitive validity and face validity of PTE Academic writing items Summarize Written Text and Write Essay*. London: Pearson

Chan, S. H. C. (2018). *Defining Integrated Reading-into-writing Constructs: Evidence at the B2-C1 Interface*. English Profile Studies Volume 8. Cambridge: UCLES/Cambridge University Press.

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge: Cambridge University Press.

Council of Europe. (2009). *Relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment (CEFR). A Manual*. Strasbourg: Council of Europe.

Council of Europe. (2020). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Companion Volume*. Strasbourg: Council of Europe Publishing.

Flower, L. (1990). Negotiating academic discourse. In L. Flower, V. Stein, J. Ackerman, M. J. Kantz, K. McCormick & W.C. Peck (Eds.), *Reading to Write: Exploring a Cognitive and Social Process* (pp. 221–252). New York: Oxford University Press.

Leki, I., & Carson, J. G. (1994). Students' perceptions of EAP writing instruction and writing needs across the disciplines. *TESOL Quarterly*, *28*(1), 81–101.

Scardamalia, M., & Bereiter, C. (1987). *The Psychology of Written Composition*. Mahwah: Lawrence Erlbaum.

Schleppegrell, M. J. (2004). *The Language of Schooling: A Functional Linguistics Perspective*. Mahwah: Lawrence Erlbaum.

Shaw, P., & Pecorari, D. (2013). Source use in academic writing: An introduction to the special issue. *Journal of English for Academic Purposes*, *12*(2), 83–154.

Wette, R. (2021). *Writing Using Sources for Academic Purposes: Theory, Research and Practice*. New York: Routledge.

Development of argumentative writing rating scale and its effectiveness in dynamic assessment

Hua Fu
Xi'an Jiaotong University, China

Yaru Meng
Xi'an Jiaotong University, China

Abstract

Argumentation ability is the key element for critical thinking and essential for argumentative essay writing. However, the fine-grained diagnosis and tailored mediation on argumentative writing have been underexplored. Thus, the present study aims to develop and validate a comprehensive rating scale based on Toulmin's Argument Model, to evaluate the quality of written argumentation at fine-grained level. Then, a rating scale-based mediation model (RS-MM) for argumentative writing was designed following the principle of gradualness in dynamic assessment (DA). 22 Chinese university students completed four argumentative essays with the first and last as pre-test and post-test. In the first three essays, they participated in one-on-one RS-MM informed DA sessions. The results revealed that the mediation model has a positive effect on most sub-measures of argument quality, including argument structure, argumentative elements and reasoning process. The results of a questionnaire showed significant improvement in the students' perceived argumentative writing abilities.

Introduction

Argumentative writing is an important type of genre in English writing, as it is associated with learners' critical thinking and academic performance (Wolfe, Britt, & Butler, 2009). It is also viewed as an important indicator of second language writing ability (Hirvela, 2017) and is usually used to assess English as Foreign Language (EFL) learners' writing proficiency (e.g., TOEFL, IELTS). However, most EFL learners encounter challenges in this regard, such as providing adequate justifications to their claims (Hsu, Van Dyke, Chen, & Smith, 2015), generating counterarguments to rebut the opposing views (Liu & Stapleton, 2020), and offering sufficient and persuasive evidence (Pessoa, Mitchell, & Miller, 2017). These challenges related to argument quality have a negative impact on the persuasiveness and effectiveness of argumentation.

Therefore, the instruction on the improvement of argument quality is urgently needed (Li & Zhang, 2022). However, there are two issues that influence the effectiveness of the argumentative writing instruction. First, the previous research on argumentative writing lacks fine-grained measurement of argument quality (Kathpalia & See, 2016; Stapleton & Wu, 2015). They assessed argumentation from limited aspects, such as relevance and acceptableness (Stapleton & Wu, 2015); relevance and correctness of evidence, presence of rebuttal (Kathpalia & See, 2016); and relevance of evidence to claims, number and soundness of evidence (Li & Zhang, 2022). Second, the existing research mainly implemented explicit instruction without fine-grained diagnosis and tailored mediation on argument quality, leading to the limited improvement in particular aspects of argumentative essays, such as the sufficiency of argumentation (Majidi, Janssen, & de Graaf, 2021), effectiveness of rebuttals (Liu & Stapleton, 2020), relevance of evidence to claims, and soundness of evidence (Li & Zhang, 2022).

In order to address the above issues, this study aims to develop a comprehensive and fine-grained rating scale for argumentative essays based on Toulmin's Argument Model (Toulmin, 1958, 2003), and construct a rating scale-based mediation model (RS-MM) for the follow-up dynamic assessment. The ultimate goal is to integrate the diagnosis and mediation of the undergraduate EFL learners' argumentative writing ability.

Dynamic assessment (DA) originates from Vygotsky's (1978) Sociocultural Theory of Mind, which highlighted that higher psychological functions are socially mediated. Lantolf and Poehner (2004) defined DA as the dialectical integration of assessment and instruction or mediation. Mediation refers to the assistance, support and feedback provided while difficulties arise in the collaborative interaction of more capable and less capable individuals. Following Vygotsky's (1978) idea of the zone of proximal

development (ZPD), the mediation should be graduated in explicitness to offer the appropriate assistance to facilitate learners' potential development (Aljaafreh & Lantolf,1994). According to this principle, a number of mediation models have been proposed (Ableeva, 2010; Aljaafreh & Lantolf, 1994; Meng & Fu, 2023; Zhang, 2023), where a series of mediation moves were arranged from most implicit (leading questions, hints, etc.) to most explicit (answers, explanations, etc.) to discover learners' ZPDs. As writing is a process of planning, drafting, revising and editing (Huang & Zhang, 2020), during which feedback and revision play important roles, DA can be applied as an approach to provide learners with tailored mediation or feedback to promote their argumentative abilities.

Method

22 Chinese university students (11 males and 11 females) aged from 18 to 22 participated in this study. They were at medium-high proficiency level of English based on their English scores from College English Test Band 4 (CET-4) in China. The first author was the mediator, who provided assistance or support in the interaction with the students.

The participants completed four writing tasks, with the first and fourth as the pre-test and post-test. For each of the tasks, they were required to write an argumentative essay with at least 250 words within 70 minutes. While writing, they are required to first, choose one side of the argument, and use specific reasons, evidence and examples to support their viewpoints. Then they need to provide reasons why the opposite side might disagree with them, and why the reasons of the opposite side were groundless.

The study consists of three stages: 1) developing the rating scale for argument quality, in reference to Toulmin's Argument Model, related research and expert judgement; 2) constructing a rating scale-based mediation model, in reference to the previous mediation typologies, the rating scale, the mediator and expert's suggestions; 3) implementing three DA sessions. Specifically, the 22 participants completed the first three argumentative essays, each followed with mediation, and a fourth argumentative essay as post-test without mediation. For each of the first three completed essays, the participants received two rounds of mediation: written mediation in the form of an added rating scale after their drafts via e-mails (i.e., most implicit prompts in highlighted descriptors needing revision in the rating scale) and face-to-face graduated prompts in student-mediator interaction through remote video meetings (i.e., arranged from implicit leading questions, hints, etc. to explicit explanations). They also filled in a questionnaire concerning their perceptions toward DA before and after the experiment. After gaining the scores of the first drafts of Essay 1 and the drafts of Essay 4, a series of paired samples t-tests were conducted to investigate whether there were significant improvements in the sub-measures of argument quality from the pre-test to the post-test. Then, a series of paired samples t-tests were also conducted based on the questionnaire data, to examine whether the participants perceived significant improvement in the argumentative writing after the DA procedures.

Results

The rating scale for argument quality of argumentative essays

In reference to Toulmin' s Argument Model (Toulmin, 1958, 2003), the previous rubrics for argumentative writing (Hadidi, 2023; Majidi et al., 2021; Stapleton & Wu, 2015), the evaluation framework of critical thinking in English argumentative writing (Geng, Yu, & Wang, 2021; Ma, 2021; Mu, 2016), this study developed an analytical rating scale for argument quality (Table 1). It is a 5-point Likert scale (Items 1–14) with a total score of 70. It was further revised through experts' and raters' suggestions. The pilot study showed that the reliability of the rating scale is high with Cronbach's alpha .89.

Rating scale-based mediation model for argument quality

Based on the mediation inventories of previous studies (Aljaafreh & Lantolf, 1994; Kushki, Nassaji, & Rahimi, 2022; Kushki, Rahimi, & Davin, 2022) and the six measures of the rating scale developed above, three mediation models were constructed (Table 2). Each model contains six to eight mediation moves arranged from most implicit to most explicit, aiming to guide learners to engage in the collaboration with mediators and trigger their critical thinking.

The effect of the RS-MM on argument quality of learners' essays

The results of paired samples t-tests revealed significant improvement on most sub-measures of argument quality from the pre-test to the post-test, except for relevance and clarity of central-claims, and relevance of sub-claims. This suggested that the mediation models were effective in improving most fine-grained features of argument quality, whereas the previous research found limited improvement in argument quality (Kathpalia & See, 2016; Li & Zhang, 2022). This can be attributed to the fact

Table 1: An analytical rating scale for argument quality

Measures	*Sub-measures*	*Descriptors*
1. Argument structure	Complexity	Using various argumentative elements as much as possible.
	Soundness	The organization of argumentative elements is sound\reasonable.
2. Central-claim	Relevance	The central-claim is relevant to the writing topic.
	Clarity	The central-claim is clear.
3. Sub-claim	Sufficiency	Providing multiple sub-claims (reasons) to support the central-claim.
	Relevance	The sub-claim is relevant to the central-claims.
	Clarity	The sub-claim is clear.
4. Evidence	Sufficiency	Providing multiple evidence to support the corresponding sub-claim.
	Reliability	The evidence comes from reliable sources, and they are correct.
	Relevance	The evidence is relevant to the corresponding sub-claims.
	Persuasiveness	The evidence is persuasive.
5. Rebuttal	Soundness	The reason for rebuttal is sound\acceptable.
	Effectiveness	The rebuttal is sufficient and then effective in refuting the opposite views.
6. Reasoning	Logic	The reasoning process from evidence to sub-claim is rigorous and reasonable.

that guided by the RS-MM, the mediator could provide specific feedback and guidance regarding the different sub-measures of argument quality during the mediator-student collaborative interactions. This process engaged the participants in enhancing their metacognitive awareness of argument quality gradually, contributing to the internalization of the concept of argument quality, and thereby substantial achievement in argument quality.

The results of the questionnaire also showed that the participants perceived their argumentative writing abilities improved significantly following the RS-MM informed DA sessions, especially with regard to the logic of reasoning, persuasiveness of evidence, and soundness of rebuttal.

Conclusion

According to Toulmin's Argument Model and the framework of DA, this study constructed a RS-MM for mediating learners' argument quality development in argumentative writing. The distinguishing feature of the model lies in that the mediating prompts are designed in relation to the particular quality requirements of argumentative writing, which makes the prompts specific and tailored to learners' individualized needs. Therefore, the mediation model has the advantage of improving the sub-measures of argument quality, which was corroborated by the three DA sessions, whereas the previous research revealed limited improvement in these aspects of argumentative essays (e.g., Liu & Stapleton, 2020; Majidi et al., 2021).

To sum up, the rating scale plays an essential role in this study. The employment of the rating scale for argument quality in constructing a mediation model is an innovative teaching practice in that it integrates language assessment standards with instructional intervention, linking the assessment of language abilities to the development of them. This broadens the application of the rating scale in language teaching and assessment, which is significant in facilitating the positive effect of it in advancing learner language development. Moreover, the rating scale was utilized in the evaluation of argumentative essays and the construction of the mediation model simultaneously, bridging the gap between summative assessment and formative assessment. In this way, the rating scale serves two functions, similar to the terms 'diachronic' and 'synchronic' (Poehner & Yu, 2022). In this study, the rating scale was used both as a summative assessment tool to assign scores to argumentative essays by raters, and a formative assessment tool to diagnose learners' areas of difficulties in argument quality and to provide fine-grained mediation. On the whole, the RS-MM is the product of the diagnostic assessment tool (i.e., rating scale) in the DA, representing the integration of diagnostic language assessment and DA, both of which target at the diagnosis of learner ability and promotion

Table 2: Rating scale-based mediation model for argument quality

Measure	*Sub-measure*	*Mediation model*	*Mediation move*
Argument structure	Complexity	Mediation model 1 Implicit ↓ Explicit	1. Ask students whether there is anything that needs improvement on the whole. 2. Narrow down to one of the six measures (i.e., argument structure\central-claim\rebuttal). 3. Ask students which sub-measure of the argument structure\central-claim\rebuttal needs improvement. 4. Point out the sub-measure that needs improvement. 5. Provide explanations using the key sentences in the essay. 6. Provide further explanations and suggestions\solutions.
	Soundness		
Central-claim	Relevance		
	Clarity		
Rebuttal	Soundness		
	Effectiveness		
Sub-claim	Sufficiency	Mediation model 2 Implicit ↓ Explicit	1. Ask students whether there is anything that needs improvement on the whole. 2. Narrow down to one of the six measures (i.e., sub-claim\evidence). 3. Ask students which sub-claim\evidence needs improvement. 4. Point out the sub-claim\evidence that needs improvement. 5. Ask students which sub-measure of the sub-claim\evidence need improvement. 6. Point out the sub-measure that needs improvement. 7. Provide explanations using the key sentences in the essay. 8. Provide further explanations and suggestions\solutions.
	Relevance		
	Clarity		
Evidence	Sufficiency		
	Credibility		
	Relevance		
	Persuasiveness		
Reasoning	Logic	Mediation model 3 Implicit ↓ Explicit	1. Ask students whether there is anything that needs improvement on the whole. 2. Narrow down to one of the six measures (i.e., reasoning). 3. Ask students which reasoning process from evidence to sub-claim needs improvement. 4. Point out the reasoning process that needs improvement. 5. Provide explanations using the key sentences in the essay. 6. Provide further explanations and suggestions\solutions.

of learner development. Future research can further refine the rating scale and validate it in other contexts. Research with larger sample sizes and control groups would provide more robust support for the effectiveness of the RS-MM.

References

Ableeva, R. (2010). *Dynamic assessment of listening comprehension in second language learning* [Unpublished doctoral dissertation]. The Pennsylvania State University.

Aljaafreh, A., & Lantolf, J. P. (1994). Negative feedback as regulation and second language learning in the zone of proximal development. *The Modern Language Journal*, *78*, 465–483.

Geng, F., Yu, S., & Wang, J. (2021). The influence of peer feedback on students' critical thinking abilities in argumentative writing. *Foreign Language World*, *204*(3), 37–45.

Hadidi, A. (2023). Comparing summative and dynamic assessments of L2 written argumentative discourse: Microgenetic validity evidence. *Assessing Writing*, *55*, 1–20.

Hirvela, A. (2017). Argumentation & second language writing: Are we missing the boat? *Journal of Second Language Writing*, *36*, 69–74.

Hsu, P. S., Van Dyke, M., Chen, Y., & Smith, T. J. (2015). The effect of a graph-oriented computer-assisted project-based learning environment on argumentation skills. *Journal of Computer Assisted Learning*, *31* (1), 32–58.

Huang, Y., & Zhang, L. J. (2020). Does a process-genre approach help improve students' argumentative writing in English as a foreign language? Findings from an intervention study. *Read and Writing Quarterly*, *36*, 339–364.

Kathpalia, S. S., & See, E. K. (2016). Improving argumentation through student blogs. *System*, *58*, 25–36.

Kushki, A., Nassaji, H., & Rahimi, M. (2022). Interventionist and interactionist dynamic assessment of argumentative writing in an EFL program. *System*, *107*, 1–13.

Kushki, A., Rahimi, M., & Davin, K. J. (2022). Dynamic assessment of argumentative writing: Mediating task response. *Assessing Writing*, *52*, 1–12.

Lantolf, J. P., & Poehner, M. E. (2004). Dynamic assessment of L2 development: Bringing the past into the future. *Journal of Applied Linguistics*, *1*, 49–72.

Li, H. H., & Zhang, L. J. (2022). Investigating effects of small-group student talk on the quality of argument in Chinese tertiary English as a foreign language learners' argumentative writing. *Frontiers in Psychology*, *13*, 1–14.

Liu, F., & Stapleton, P. (2020). Counterargumentation at the primary level: An intervention study investigating the argumentative writing of second language learners. *System*, *89*, 1–15.

Ma, L. (2021). Establishing evaluation framework for critical thinking in foreign language writing based on Delphi method. *Language Education*, *2*, 23–27.

Majidi, A., Janssen, D., & de Graaf, R. (2021).The effects of in-class debates on argumentation skills in second language education. *System*, *101*, 1–15.

Meng, Y. R. & Fu, H. (2023). Modeling mediation in the dynamic assessment of listening ability from the cognitive diagnostic perspective. *The Modern Language Journal*, *107*(S1), 137–160.

Mu, C. (2016). Investigating English major students' critical thinking ability in academic writing. *Modern Foreign Languages*, *5*, 693–703.

Pessoa, S., Mitchel, T. D., & Miller, R. T. (2017). Emergent arguments: A functional approach to analyzing student challenges with the argument genre. *Journal of Second Language Writing*, *38*, 42–55.

Poehner, M. E. & Yu, L. (2022). Dynamic assessment of L2 writing: Exploring the potential of rubrics as mediation in diagnosing learner emerging abilities. *TESOL Quarterly*, *56*(4), 1,191–1,217.

Stapleton, P. & Wu, Y. A. (2015). Assessing the quality of arguments in students' persuasive writing: A case study analyzing the relationship between surface structure and substance. *Journal of English for Academic Purposes*, *17*, 12–23.

Toulmin, S. (1958). *The Uses of Argument*. Cambridge: Cambridge University Press.

Toulmin, S. (2003). *The Uses of Argument* (Second edition). Cambridge: Cambridge University Press.

Vygotsky, L. S. (1978). *Mind in Society: The Development of Higher Psychological Process*. Cambridge: Harvard University Press.

Wolfe, C. R., Britt, M. A., & Butler, J. A. (2009). Argumentation schema and the myside bias in written argumentation. *Written Communication*, *26*, 183–209.

Zhang, Y. (2023). Promoting young EFL learners' listening potential: A model of mediation in the framework of dynamic assessment. *The Modern Language Journal*, *107*(S1), 113–136.

What is the future of plurilingual language assessment for 'monolingual' testing organisations?

Graham Seed
Cambridge University Press & Assessment, United Kingdom

Abstract

In the era of the 'multilingual turn' in language education (May, 2014), critics have noted the slow speed of recognition of plurilingual, code-switching and/or translanguaging practices within language assessment (e.g., Shohamy, 2011), and the difficulties in defining the construct of such an assessment (Saville & Seed, 2021). One particular source of vexation is how language tests, so historically entrenched in promoting proficiency along monolingual lines, can ever make a truly plurilingual turn. In Europe especially, practitioners have only recently identified how a 'plurilingual assessment' might work in practice (such as De Angelis, 2021; Seed & Holland, 2020).

This paper reports on how examples of code-switching are found even within what are supposed to be monolingual written test responses. It then looks at the nascent work already started as to how automated assessment may provide the best chances of creating a truly personalised plurilingual assessment (Nguyen, Yuan & Seed, 2022).

Introduction

The last decade has seen a shift in language education often known as the 'multilingual turn' (May, 2014), characterised by notions such as the move away from the idealised native speaker as a model and from siloed language lessons in school; the utilisation of plurilingual repertoires; embracing translanguaging in education; and increased research and practice on the subject (e.g., Piccardo, German-Rutherford & Lawrence (2021)). In Europe in particular, the publication of the Common European Framework of Reference for Languages Companion Volume (CEFR CV, Council of Europe, 2020), has increased the awareness of plurilingualism within L2 language education. Plurilingualism is defined by the CEFR CV as 'the dynamic and developing linguistic repertoire of an individual user or learner' (Council of Europe, 2020, p.30) in contrast to multilingualism, which is the 'coexistence of different languages at the social or individual level'. Plurilingualism shares similarities with the concepts of code-switching and translanguaging, and is often known as 'individual bilingualism' in North America. This paper will not go into detail regarding the nuances between the concepts, but takes the general concept of the importance of utilising all the languages known and partially known by a user when learning an additional language.

While these concepts were being discussed within the field of language education in the late 2010s, 'the contrast between the expanding use of multilingual practice in pedagogy, and the absence of multilingual approaches in assessment and evaluation measures is striking' (Schissel, De Korne, & López-Gopar, 2018, p. 341). Gorter and Cenoz (2017, p.243) agree: 'Tests should match actual language practices and multilinguals use resources from their whole linguistic repertoire. If teaching is going in the direction of a multilingual focus, assessment should also follow the same path.'

Academic critics have therefore been quick to point out the lack, or slow speed, of recognition or adoption of such practices within language assessment (e.g., Shohamy, 2011). Others have pointed out the difficulties in defining the construct of such an assessment, as well as the challenge of educational policies in promoting these (Saville & Seed, 2021). One particular source of vexation is how language tests, so historically entrenched in promoting proficiency in languages along monolingual lines, can ever truly make a plurilingual turn. This is all the more important given that ALTE's Principles of Good Practice (ALTE, 2020) places the 'plurilingual, pluricontextual language learner' at the centre of the reason for language assessment, and if real-life target language use demands utilisation of multilingual and plurilingual repertoires, then language assessment should reflect this in order to claim contextual validity.

A framework of plurilingual assessment

Seed (2020) proposed a simple framework of four categories for classifying assessments of plurilingual situations. One example is for assessments that require learners to use their plurilingual competence to demonstrate skills in two or more languages in one test. These tests require cross-linguistic mediation skills (North, 2021), such as the Greek KPG State Certificate of Language Proficiency Exams, which asks test-takers to read a text in Greek and summarise in English, for example.

Another category is assessments that require learners to use their plurilingual competence to demonstrate skills in other, non-linguistic subjects, such as science or maths. These tests often provide the input, and allow candidate output, in any language that the candidate feels comfortable to express themselves in. An example is a biology test in Belgium (De Backer, van Avermaet, & Slembrouck, 2017) which gives the question in both Dutch and Polish, to cater for Polish migrants. This type of test may be considered beneficial for fairness and social justice reasons: in this case, it is the knowledge of biology that is being tested, not the knowledge of language, and therefore it is unfair to Polish migrants who do not have a mastery of Dutch to only provide the questions in Dutch.

A further category is the assessment of the development of plurilingual competence, which is often more appropriately carried out by alternative forms of assessment, such as observations, portfolios and self-assessment. The new CEFR descriptors on plurilingual competence (Council of Europe, 2020) may assist, and Allgäuer-Hackl et al. (2018) provide some examples of this.

Seed & Holland (2020) go on to further exemplify real-world assessments of these categories. Separately, Melo-Pfeifer & Ollivier (forthcoming 2024) propose a continuum to categorise plurilingual assessment; De Angelis (2021) also provides guidance on constructing multilingual assessments, particularly in the classroom environment.

Plurilingual assessment in monolingual tests

A final category is the seemingly incongruent recognition of plurilingual competence to demonstrate skills of one language. This is of particular interest to historically 'monolingual' testing organisations who are increasingly facing calls to take note of their candidates' plurilingual abilities. It can be argued that an individual's plurilingual competence can, and in fact has to, exist alongside standard named languages, in order to communicate (Kunnan & Saville, 2021). In this respect, it is acceptable to measure the proficiency of a standard language for communicative purposes, in that it provides contextual validity. However, it is possible to test a standardised language while recognising an individual's plurilingual abilities.

Seed & Holland (2020) provide an example of a young learner's test of writing ability, where the assessment criteria do not mention accuracy of grammar and vocabulary, but rather the amount of effort required on the part of the reader to make sense of the response. One Chinese candidate transliterated a Chinese word where they did not know the English word they required. The result was therefore 'wrong' in that it was not English, but nevertheless the transliteration attempted to maintain a fluency in their writing and the assessment criteria did not penalise for this.

Seed (2019) challenged testing organisations firstly to recognise the value of test-takers' plurilingual competence by creating assessment criteria similar to the one described above; and secondly, not to promote their own tests in linguistic isolation but rather as part of creating a linguistic profile for individuals, demonstrated in a multilingual family of tests as showcased by ALTE and measured on the CEFR scale which explicitly promotes a profiled approach (Council of Europe, 2020, p. 38.).

One step further would be to explicitly allow instances of code-switching and translanguaging within assessments, i.e., in candidates' responses to prompts in speaking and writing tests. It could be argued that this is acceptable for assessments which aim to mirror target language use (TLU) situations where code-switching is the norm. In answer to the question of how this is possible when each TLU situation is unique, based on the different languages and languaging at play, a modern answer is the use of digital assessment which makes use of AI.

The use of AI and machine learning has the potential to create a more personalised, individualised learning and assessment experience for test-takers (Saville & Buttery, 2023). This could include the recognition of the individual linguistic repertoires of each learner, and how they affect the way in which learners approach an assessment task. However, to date, the machines have been almost exclusively trained on monolingual data, meaning that any examples of code-switching found in a test-taker's response marked by an automarker, are currently treated as errors. Language learning and assessment technologies built for code-switching are therefore in their infancy, and Nguyen et al. (2022) describe the nascent state of these endeavours. Rather than seeing cases of code-switching as errors, AI assessment tools should be able to first identify them as such, and then go on to either reward the linguistic use, and/or provide diagnostic feedback to aid the development of the learner's L2.

Code-switching examples in 'monolingual' tests

In order to train the automarkers, data of when, how and why test-takers actually code-switch needs to be gathered, particularly when they code-switch in responses in a supposedly monolingual test.

'Write & Improve' (Cambridge University Press & Assessment, n.d.) is a platform whereby users can submit a piece of (supposedly monolingual English) writing as a form of learning-oriented assessment, and receive automated, AI-generated, feedback in order to rewrite and improve their responses. Responses from this platform were analysed to ascertain if these test-takers used code-switching. Current AI language detection found 75 examples; however manual detection found 539 examples, excluding duplicates. This shows the current inability of AI to achieve the first step of recognising most examples of code-switching.

The languages of the code-switched words or phrases generally reflected the L1s of the test-takers in that there were 225 examples from Spanish, mirroring the large candidature from Latin America. The code-switch examples were also classified according to Poplack's (1980) framework of code-switching, which showed 148 inter-sentential examples, and 383 intra-sentential examples, of which 202 of the latter were at word level.

The examples were coded into the possible reasons why the test-taker may have code-switched. Of course, without additional qualitative analysis such as verbal protocols or questionnaires it is difficult to know exactly the reason, but it is possible to make an informed guess using the context of the rest of the response.

The reasons can be broadly categorised in two: firstly, 'genuine' code-switching, where the test-taker purposefully chose to use a non-English word or phrase; and secondly 'strategic' code-switching where the test-taker code-switched either to plan their response or to compensate for English words or phrases they did not know (similar to the example of the Chinese candidate above).

Table 1 shows the possible reasons for the code-switching occurrence, the number of occurrences of this type found, and an example to illustrate.

Conclusions and future directions

It is clear that the language testing industry is only at the very beginnings of AI detecting code-switching rather than treating examples as errors, as the first step towards the use of technologies to accept plurilingual repertoires within language assessment. Once the automarker is more successful at detecting examples, could it then detect 'genuine' instances to reward

Table 1: Code-switching occurrences in the dataset

Broad category	*Reason*	*Number of occurrences found*	*Example*
Genuine	Naming (food, festival, place, etc.)	170	*A day after that we have a big holiday 'Ден На Народните '*
	Direct quotes	17	*I would hear my mom saying to me as she gets dressed to work 'Rodrigo, estou de saida para o trabalho, beijo meu pretinho. Você vai para igreja hoje? Pode levar esta sacola com você?' I'd say yes.*
	Translating a word/ phrase	26	*Abu Dhabi is welcom to anyone and everyone ! As the Arabs say;* أهلا وسهلا بك يا عزيزي *- Welcome my fellows!*
	Other pragmatic or rhetoric reasons	73	*In conclusion, if I could choose a plece to be right now, it would doubtless be Veracruz. Vamos a Veracruz!*
Strategic	For planning purposes (e.g., text written in L1 to help formulation of ideas, then put into English)	170	*Creio que eu poderia ter um rendimento maior com um professor nativo em português, pois, dessa maneira o ensino e a compreensão da língua seria mais compreensíve* *I prefer learn English with a teacher whose native in my language, because I think would be better for talk with the teacher.*
	L1 intrusion or word/phrase unknown in English	78	*A lot of naukowców say that the best time to visit Sankt Petersburg is july, in order to see theirs 'white nights'.*

(or at the very least not penalise) the test-taker? Could it detect 'strategic' instances in order to provide diagnostic feedback to further enhance learning? It is also acknowledged that the examples shown here have come from written responses; a similar process for spoken responses would be more complex. Alongside all this work should come a greater discussion of how exactly AI-driven assessments can be more relevant to an individual's communication needs.

References

Allgäuer-Hackl, E., Brogan, K., Henning, U., Hufeisen, B., & Schlabach, J. (2018). *More languages? – PlurCur!: Research and practice regarding plurilingual whole school curricula.* Strasbourg: Council of Europe.

ALTE (2020). *Principles of Good Practice.* Available online: https://www.alte.org/

Cambridge University Press and Assessment (n.d.) *Write & Improve.* Available online: https://writeandimprove.com/

Council of Europe. (2020). *The Common European Framework of Reference for Languages: Learning, teaching, assessment. Companion Volume*. Strasbourg: Council of Europe.

De Angelis, G. (2021). *Multilingual Testing and Assessment.* Bristol: Multilingual Matters.

De Backer, F., van Avermaet, P., & Slembrouck, S. (2017). Schools as laboratories for exploring multilingual assessment policies and practices. *Language and Education, 30*(1), 1–14.

Gorter, D.M., & Cenoz, J. (2017). Language education policy and multilingual assessment. *Language and Education, 31*(3), 231–248.

Kunnan, A. J., & Saville, N. (2021). Setting Standards for Language Learning and Assessment in Educational Contexts. A Multilingual Perspective. In W. Ayres-Bennett & J. Bellamy (Eds.), *The Cambridge Handbook of Language Standardization* (pp. 496–516). Cambridge: Cambridge University Press.

May, S. (Ed.). (2014). *The Multilingual Turn Implications for SLA TESOL and Bilingual Education*. New York: Routledge.

Melo-Pfeifer, S. & Ollivier, C. (forthcoming 2024). Foreword. In. S. Melo-Pfeifer & C. Ollivier (Eds.), *Assessment of Plurilingual Competence and Plurilingual Learners in Educational Settings.* Abingdon: Routledge.

Nguyen, L., Yuan, Z. & Seed, G. (2022). Building Educational Technologies for Code-Switching: Current Practices, Difficulties and Future Directions. *Languages*, *7*(3), 220. Available online: https://doi.org/10.3390/languages7030220

North, B. (2021). Plurilingual mediation in the classroom. In E. Piccardo, A. Germain-Rutherford & G. Lawrence (Eds.), *The Routledge Handbook of Plurilingual Education* (pp. 319–336). New York: Routledge.

Piccardo, E., Germain-Rutherford, A. & Lawrence, G. (Eds.) (2021). *The Routledge Handbook of Plurilingual Education.* New York: Routledge.

Poplack, S. (1980). Sometimes I'll start a sentence in Spanish y termino en español: Toward a typology of codeswitching. *Linguistics*, *18*, 581–618.

Saville, N. & Buttery, P. (2023). Interdisciplinary collaborations for the future of learning-oriented assessment. In K. Sadeghi & D. Douglas (Eds). *Fundamental Considerations in Technology Mediated Language Assessment.* Abingdon: Routledge.

Saville, N., & Seed, G. (2021). Plurilingual Assessment. In E. Piccardo, A. Germain-Rutherford & G. Lawrence (Eds.), *The Routledge Handbook of Plurilingual Education* (pp. 360–376). New York: Routledge.

Seed. G. (2019, 8 November). *What is plurilingual assessment?* [Conference presentation]. ALTE 54th Conference Day, Ljubljana, Slovenia.

Seed, G. (2020). What is plurilingualism and what does it mean for language assessment?. *Research Notes*, *78*, 5–15.

Seed, G., & Holland, M. (2020). Taking account of plurilingualism in Cambridge Assessment English products and services. *Research Notes*, *78*, 16–25.

Schissel, J. L., De Korne, H., & López-Gopar, M. E. (2018). Grappling with translanguaging for teaching and assessment in culturally and linguistically diverse language learning contexts: Teacher perspectives from Oaxaca, Mexico. *International Journal of Bilingual Education and Bilingualism*, *24*(3), 340–356.

Shohamy, E. (2011). Assessing multilingual competencies: Adopting construct valid assessment policies. *The Modern Language Journal*, *95*, 418–429.

Towards multilingual language assessment: Adapting CEFR-J Can Do Tests

Yukio Tono
Tokyo University of Foreign Studies, Japan

Masashi Negishi
Tokyo University of Foreign Studies, Japan

Abstract

This paper outlines an ongoing project that aims to develop and adapt the CEFR-J Can Do tests for multiple languages, focusing on the testing process and preliminary outcomes. Based on the CEFR-J framework, the project created 'Can Do' tests for 26 languages, aligned with specific CEFR-J descriptors. 'Machine translations' and 'human evaluations and post-editing' were employed to adapt the English version of the tests. A pilot test for the Reading component provided insights into test effectiveness and identified translation challenges. Initial findings indicate variations in test performance across languages and reveal cultural adaptation requirements for specific items. This project promises to enhance language assessment and facilitate positive learning experiences across languages and proficiency levels.

Introduction

The CEFR-J represents a localization of the Common European Framework of Reference for Languages (CEFR, Council of Europe, 2001) tailored to English language education in Japan. This project has been consecutively awarded the Japan Society for Promotion of Science (JSPS) Grants-in-Aid for Scientific Research (KAKENHI) grants for five terms, with each term spanning four years (Negishi, Takada, & Tono, 2013; Tono, 2017). Over this period, our primary endeavor was to develop a unique set of 'Can Do' descriptors, grounded in the CEFR, to facilitate a more detailed self-assessment grid (cf. Table 2 in the CEFR (2001, pp. 26–29)). Notably, we introduced a Pre-A1 level preceding A1 and implemented subdivisions within each CEFR level up to B2, culminating in twelve distinct levels. In the updated CEFR Companion Volume (Council of Europe, 2020), the inclusion of the Pre-A1 level and the reintroduction of 'plus' levels, in part, acknowledge the impact of our work. Subsequent projects yielded the creation of Reference Level Description resources, such as the CEFR-J Wordlist and the CEFR-J Grammar/Text Profile. As a pivotal facet of this initiative, by 2020, we had crafted a series of 'Can Do' tests for English, and a sample test battery was made publicly accessible. This provision empowers local educational boards and commercial publishers to create their own assessments.

Concurrently, in 2014, the 'CEFR-J x 28' project was launched at Tokyo University of Foreign Studies (TUFS) under the umbrella of the Top Global University Project. Its primary objective was to repurpose resources initially designed for English to encompass the 27 additional languages offered as majors at TUFS. By harnessing technologies like multilingual machine translation and corpus resources – complemented by human evaluation and post-editing – we have produced the CEFR-J x 28 Wordlists and Phrase Lists (Tono, 2019). Additionally, we have developed multilingual versions of the CEFR-classified example sentence database, rooted in the *British Council/EAQUALS Core Inventory for General English* and the *Threshold Level* series. In 2020, the 'CEFR-J x 28' initiative transitioned into its subsequent phase, pivoting its focus to the adaptation of the CEFR-J 'Can Do' tests for English to the other 26 languages[1]. This paper serves as a provisional report of this ongoing venture, presenting both the test development process and the outcomes of pilot tests conducted with a cohort of approximately 170 students in seven foreign languages at the culmination of the 2022 academic year.

[1] Japanese was not included in the target language groups because most students at TUFS are native speakers of Japanese.

Multilingual test construction based on the CEFR-J

The basic test design of the 'Can Do' test

In our 'Can Do' tests, each test item is developed based on a specific 'Can Do' descriptor from the CEFR-J. In general, the descriptors in the CEFR encompass three pivotal elements: (a) *Performance*: This delineates what the learner is capable of achieving linguistically. (b) *Conditions*: These specify the circumstances under which the performance is rendered, such as whether assistance is provided by interlocutors or if reference tools are accessible. (c) *Criteria*: This aspect gauges the linguistic proficiency with which the learner executes the action. A comprehensive 'Can Do' test should encapsulate all these facets, supplemented by situational criteria that dictate the specific communicative context and relevant topics under which the performance is observed.

In the CEFR-J, each mode of communication (e.g., listening/spoken interaction/spoken production/reading/writing) at a given CEFR level is associated with two 'Can Do' descriptors. For instance, the Pre-A1 level for listening has two distinct descriptors. Consequently, there are a total of 100 descriptors (2 descriptors × 5 modes of communication × 10 levels) spanning from Pre-A1 to B2.2. We consciously chose to omit the C levels, as these tend to be domain-specific and present complexities that make test design challenging. In the development of our test tasks, we rigorously ensured that all three descriptor elements previously discussed were accurately represented.

Transforming the English version to multilingual versions

The subsequent phase involved constructing 'Can Do' tests for 26 languages other than English. Given that language teachers for each foreign language major were not versed in either language testing or the CEFR, it was imperative to equip them with knowledge of both. To facilitate this, we provided them with the original English version of the 'Can Do' tests as well as machine-translated versions in their respective languages. Each language team was tasked with verifying the accuracy of the translations. This process enabled each team to become acquainted with the target 'Can Do' descriptors, including the test content and format. Once they reviewed and understood the initial set derived from the English version, they proceeded to develop two more sets, adhering to the established format and design. As we will discuss in the results, some items rooted in the English version proved ineffective in certain target languages due to various reasons. Nevertheless, the multilingual versioning process was largely successful. By the end of the academic year 2022, we had accomplished the compilation of three test batteries for the majority of the languages. All test items have been catalogued in the item bank within the test management system designed specifically for this project.

Pilot test

In January and February of 2023, we conducted a pilot of the Reading component of the 'Can Do' test, involving 174 students. The languages tested included English (n = 25), French (n = 4), Polish (n = 31), Chinese (n = 49), Thai (n = 34), Filipino (n = 10), and Mongolian (n = 21). Students in their first year tackled tests ranging from Pre-A1 to A2.2, whereas those in their second year and beyond were assessed on levels from A2.1 to B2.2. Each test set consisted of 12 'Can Do' items targeting reading skills. The test was computer-based, framed in a multiple-choice format, and all participants completed it in a computer lab. Scoring was automated, with the test report detailing the specific 'Can Do' descriptors and indicating whether the participant passed or failed each item. Notably, we opted not to provide an aggregate score, instead highlighting the number of 'Can Do' tasks each student successfully completed.

The results of the test analyses

In this part, we report on the results of the test analyses for the Chinese and Polish versions for Pre-A1 to A2.2 levels. The Chinese version consists of 25 items and 48 first year students took the test. Some of the Chinese major students are returnees and others have Chinese parents. On the other hand, the Polish version consists of 24 items and 25 first year students took the test. These participants were all beginners.

Item analyses were based on classical test theory, including point-biserial correlation. Since the number of test-takers is quite small, we cannot employ item response theory. The average score of the Chinese version is higher than that of the Polish one. This indicates that the Chinese version needed higher-level items to see the achievement of the first year learning, whereas the Polish version seems to be quite appropriate for the students. This may be partly because some of the participants were not complete beginners, and also because Japanese learners can guess the meanings of the text based on their knowledge of the Chinese characters used in Japanese. All the Polish major students started learning the language from scratch. Both

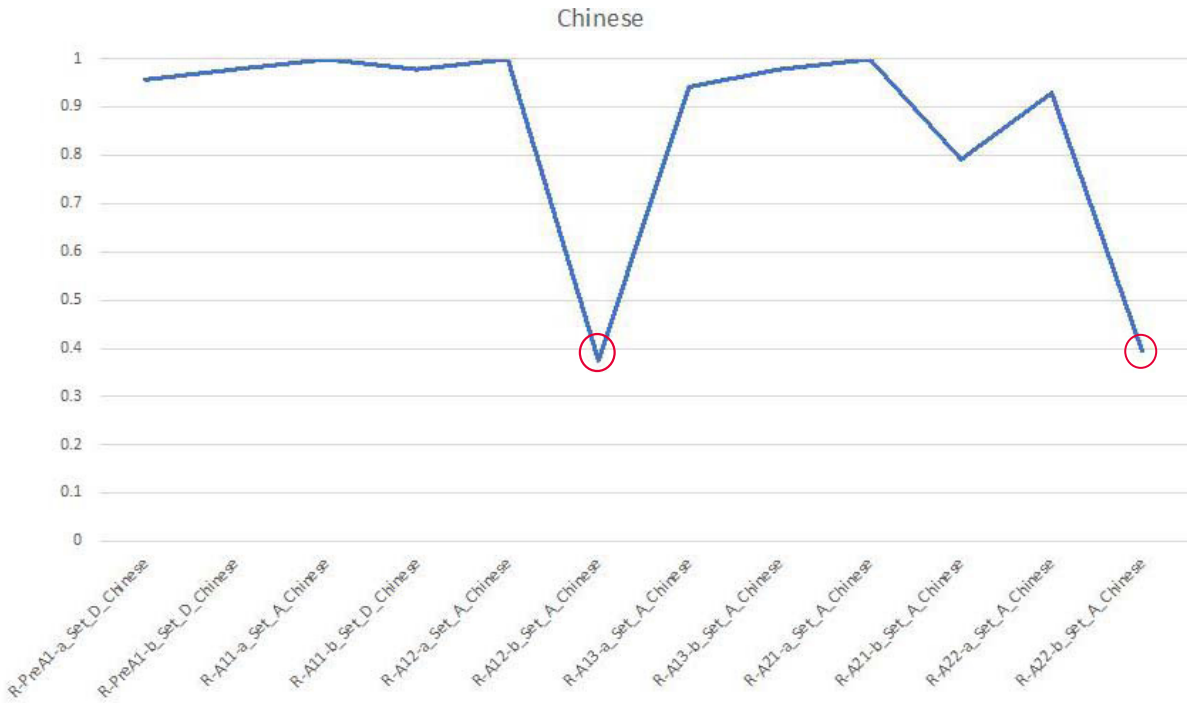


Figure 1 Item-Means of the Chinese version by CEFR-J 'Can Do' tests

of Cronbach's alpha coefficients are about 0.76. This means both of the versions are reliable for the limited number of items, indicating the potential usefulness of our test translation approach.

The Item-Means of the Chinese version by each section are shown in Figure 1. The vertical axis indicates the Item-Mean and the horizontal axis indicates the CEFR-J level. The Item-Means are the average scores of the items in each section, although some of them have only one item in the section. Red circles indicate the sections of items with lower means.

Figure 2 shows the Item-Means of the Polish version by each section. The patterns of the Chinese version and the Polish version are quite similar. Generally, the Item-Means of the sections of the higher CEFR-J levels are lower because they are supposed to be more difficult.

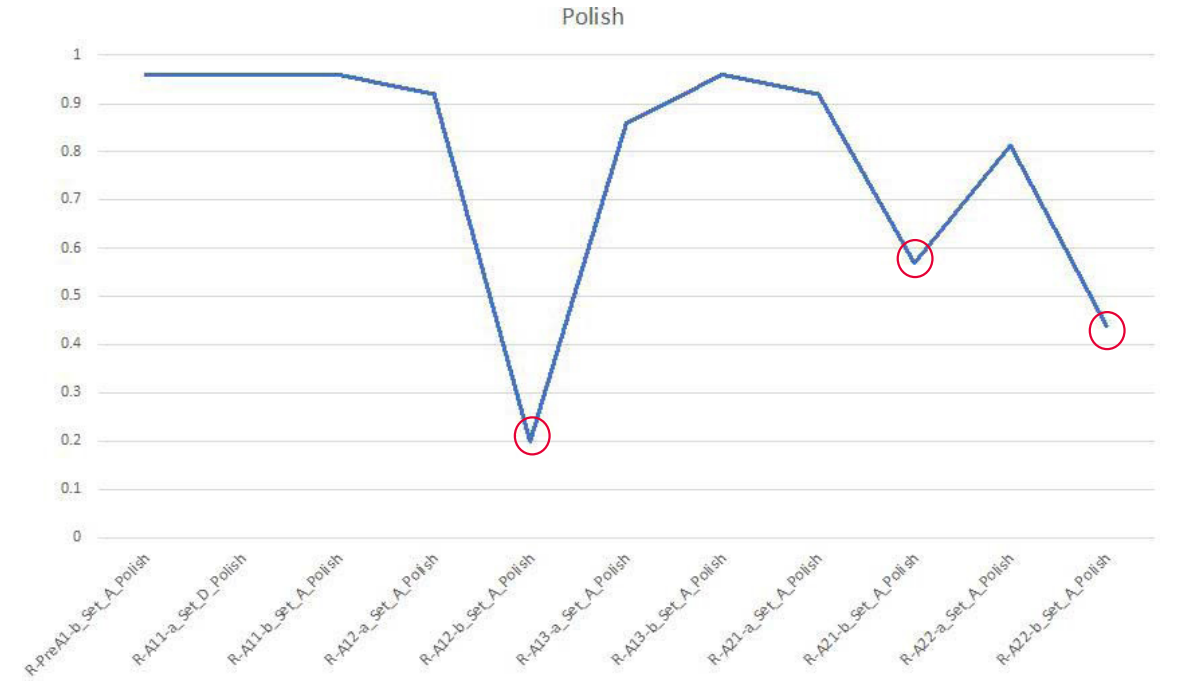


Figure 2 Item-Means of the Polish version by CEFR-J 'Can Do' tests

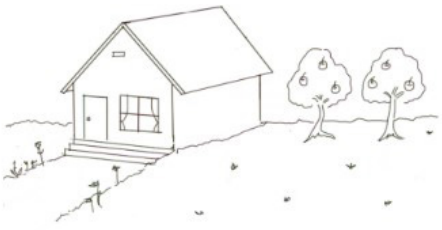
Let's take a look at the item for Reading-A1.2-b. The following is the original English version. The instruction goes like this, although the original instruction was given in Japanese:

> You keep in touch with friends you made while studying in the UK on social networking sites when you return home. You are now reading a message from Mary with a photo. Choose the most appropriate picture attached to the message from A to D.

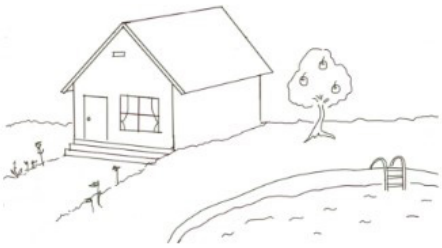
The reading passage is as follows:

Mary Williams
August 3 at 10:57 am
Hi, everyone!
We've just moved to our new house. Now we have a garden with an apple tree. The garden is large enough to have a swimming pool and we're planning to have one next summer.
Feel free to visit us!

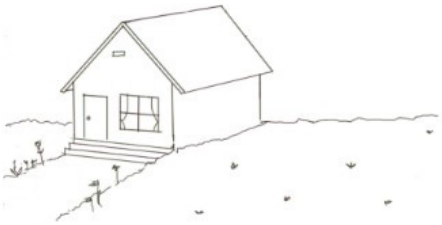
[A]



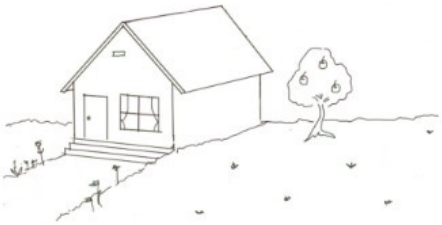
[B]



[C]



[D]



At first glance, this looks like an ordinary reading test item. However, it turned out to be quite a unique item when translated into other languages. For the Filipino version, the word 'apple tree' had to be changed to 'mango tree' because having an apple tree in a private garden in the Philippines is very rare. Such cultural adaptation was also necessary for some other languages.

The Item-Means of the Chinese and Polish versions of this item are much lower than the neighbouring items. The Item-Mean of this item in the Chinese version is 0.375. This is extremely low considering the CEFR-J level of the item. This is because, in the original Chinese translation, there was no reference to the number of trees. The corresponding Chinese word can be singular or plural. So both options A and D were correct. Many lower-level students chose option B simply because there is a reference to a swimming pool in the passage.

As for the Polish version, the Item-Mean is also low, but the point-biserial correlation is very high, indicating that the item discriminates very well. This is because only a handful of high-level learners got it right. A Polish professor stated that the linguistic structure of the relevant part might have been too difficult for most first year students.

Despite some issues of direct translation from English to the target language, overall, this translation approach worked quite well, and the study showed that item analysis can help us to identify problematic items.

Conclusion

This paper presents a provisional report on the CEFR-J 'Can Do' test development initiative. In 2024, we intend to administer the Reading component to the entire TUFS student body, encompassing approximately 3,500 students. By 2024, our goal is to finalize the Listening component and, if feasible, conduct a pilot test encompassing both Reading and Listening. As we move towards administering a comprehensive test covering all five modes of communication, we must navigate various complexities, including time allocation, the decision between total and sub-scores, the test's objectives (such as focusing on a single mode of communication at a specific level versus a broader proficiency assessment), and the assessment techniques for the production tests, namely speaking and writing. While the path ahead is challenging, the implementation of the 'Can Do' test promises to offer learners tangible benefits, fostering pragmatic communication goals and cultivating positive learning habits.

References

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge: Cambridge University Press.

Council of Europe. (2020). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Companion Volume*. Strasbourg: Council of Europe Publishing.

Negishi, M., Takada, T., & Tono, Y. (2013). A progress report on the development of the CEFR-J. In E. D. Galaczi & C. J. Weir (Eds.), *Exploring Language Frameworks. Proceedings of the ALTE Kraków Conference*, *July 2011* (pp. 135–163). Studies in Language Testing Volume 36. Cambridge: UCLES/Cambridge University Press.

Tono, Y. (2017). The CEFR-J and its Impact on English Language Teaching in Japan. *JACET International Convention Selected Papers*, *4*, 31–52.

Tono, Y. (2019). Coming Full Circle – From CEFR to CEFR-J and back. *CEFR Journal*, *1*, 15–17.

Common European Framework of Reference for Languages and Czech Sign Language Project APIV A 2019–2022

Denisa Lachmanová
Charles University in Prague, The Czech Republic

Vladimir Simon
Consultant

Radka Novakova
Charles University in Prague, The Czech Republic

Lucie Stadlerova
Independent linguist and teacher of the deaf

Romana Petranova
Czech Chamber of Sign Language Interpreters [Česká komora tlumočníků znakového jazyka]

Milena Cihakova
Pevnost – Czech Center of Sign Language [Pevnost – české centrum znakového jazyka]

Abstract

The project *Common European Framework of Reference linked to Czech Sign Language* started in 2019 in the Czech Republic. After four years of research, we are pleased to present three outcomes: a Framework of Reference for Sign Languages (FRSL, *Referenční rámec pro znakové jazyky*), Reference Level Descriptors for Czech Sign Language A1–B2 (RLDCSL, *Popisy referenčních úrovní A1-B2 pro český znakový jazyk*), and the website containing translation into Czech Sign Language (CzSL). In the FRSL we created new communicative situations related to communicative language activities (reception, production, interaction, mediation) that fit more precisely to real-world communication in sign languages. We also dealt with chapters including communicative language competencies (linguistics, pragmalinguistics, sociolinguistics). In the second document, we described CzSL and edited chapters, e.g., Overview of Grammar and Vocabulary of Czech Sign Language, Socio-Cultural Knowledge and Abilities, and Themes of Communication. In this paper, we want to share our experience and inspire other teams to dedicate their time to this topic.

Introduction

A team of Czech linguists, deaf and hearing, and users of Czech Sign Language (CzSL), present outcomes of a three-year project (2019–2022). The project was directed by the National Pedagogical Institute in the Czech Republic and the project was financed from the resources of the European Union. Three outcomes were completed: a Framework of Reference for Sign Languages (FRSL), Reference Level Descriptors of Czech Sign Language (RLDCSL) for Levels A1–B2, and a Glossary.

About the Common European Framework of Reference for Languages

The Common European Framework of Reference for Languages (CEFR, Council of Europe, 2001) and the CEFR Companion Volume (CEFR CV, Council of Europe, 2020) are at this point very well-established tools in the European context. They have also been adapted for several non-European languages. However, until now, most activities relating to the implementation of the CEFR and CEFR CV have concerned only spoken languages, i.e., those of the auditory-oral modality. When it comes to sign languages, which operate in the visual-gestural modality, it is possible to talk of two significant milestones relating to the adaptation of the CEFR. The first of these was the PRO-Sign project, which was completed by the European Centre for Modern Languages (ECML) in 2016. Another milestone was the increasing acceptance of the CEFR and the publication of the CEFR CV.

Indeed, the materials from the PRO-Sign project were made use of in the CEFR CV, where a separate chapter on sign languages and the illustrative descriptor scales were included, as well as texts explaining the theoretical background and key concepts relating to sign languages.

Project 2019–2022

The CEFR CV has thus been a major source of inspiration for the project called *Akční plán inkluzivního vzdělávání* (APIV A, 2019–2022), the outcomes of which we are summarising here.

In 2019, work aiming originally at the adaptation of the PRO-Sign outcomes to the context of CzSL began within the APIV A project, a project devoted to the inclusion of communities with first languages other than Czech. The author's team gathered three deaf and three hearing experts. Also, other teams were formed from deaf and hearing colleagues. In total, there were more than 50 people involved.

The outcomes

The realization was reached by the project team after several months of discussion and work, and it underpinned the decision to change the project's original objective. Accordingly, the planned outcomes of the project had to be conceptualized differently. The outcome concerning the problem of teaching and learning sign languages (Framework of Reference for Sign Languages, FRSL) (Nováková et al., 2022c) was separated from the outcome whose aim was to describe CzSL (Reference Level Descriptors of Czech Sign Language (RLDCSL) for Levels A1–B2 (Nováková et al., 2022b)). In the course of work on these documents, it became apparent that there was a need to compile a *Glossary* (Lachmanová, & Štádlerová, 2022a) containing the most important terms associated with them (Figure 1).

Figure 1 Three outcomes

The Framework of Reference for Sign Languages and Reference Level Descriptors of Czech Sign Language for Levels A1–B2

The FRSL and the RLDCSL were first produced in written Czech and then subsequently translated into CzSL. They were created primarily for users of CzSL: for example, teachers, and students of CzSL who are not deaf, as well as for experts and the informed public. They were also created for other users engaged in matters concerning CzSL, whether it be from a linguistic perspective, from a pedagogical perspective, or for those creating syllabi as well as possibly also for interpreters, etc.

The content of both documents, as well as the illustrative descriptor scales and their content, the focus, and the theoretical texts, reflect the most up-to-date linguistic understanding of CzSL (which the team of authors have to a large extent contributed to). They also reflect trends in sign language linguistics in general. At the same time, both documents have been prepared with both the Czech deaf community and the informed, hearing public in mind.

FRSL and RLDCSL differ in many respects but also share several key features. Both documents contain theoretical texts and illustrative descriptor scales, as well as texts introducing and explaining these scales. The documents complement each other, but the reasons and purposes they are used for will depend on the user and the user's objective. So, for example, the user might

explore the FRSL for a global perspective on sign languages, whilst a user starting from a specific standpoint regarding CzSL might begin by studying the RLDCSL.

The Framework of Reference for Sign Languages

The FRSL provides a theoretical grounding of sign language from a linguistic perspective. However, at the same time, it includes a practical view of life with and in sign language by means of a description of the sign language user's level of language competency. In other words, aside from the theoretical chapters concerning the modality of sign languages, the FRSL includes a description of what the user can do with and in the language in terms of communicative activities (reception, production, interaction, mediation) and which associated communicative strategies are employed. It also describes how these activities are carried out at each of the language proficiency levels (pre-A1 to C2). The FRSL describes knowledge and skills of the sign language user or learner without reference to a specific sign language. The authors of the document have tried to compensate for the more theoretical nature of the document by giving examples representing the general principles of sign languages in use (Figure 2).

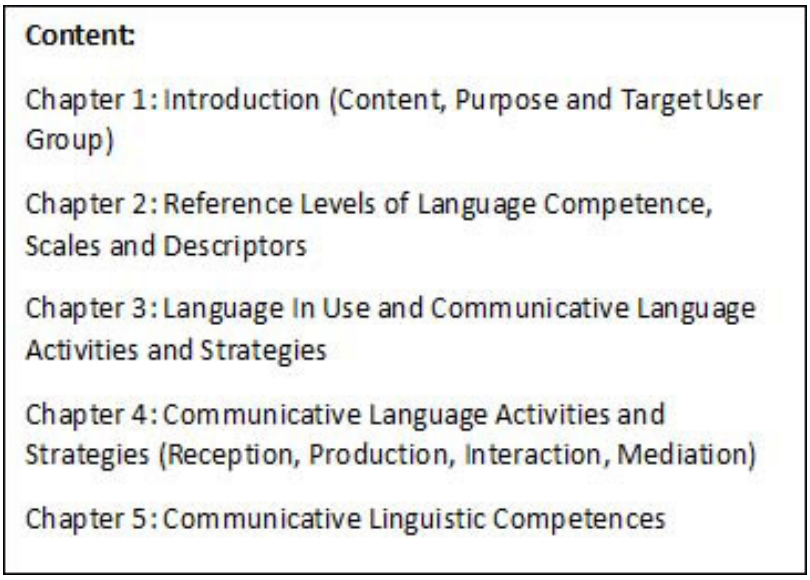


Figure 2 The FRSL content

Reference Level Descriptors of Czech Sign Language for Levels A1–B2

The RLDCSL is dedicated to CzSL and reflects the Czech culture and Deaf history and social environment (Figure 3).

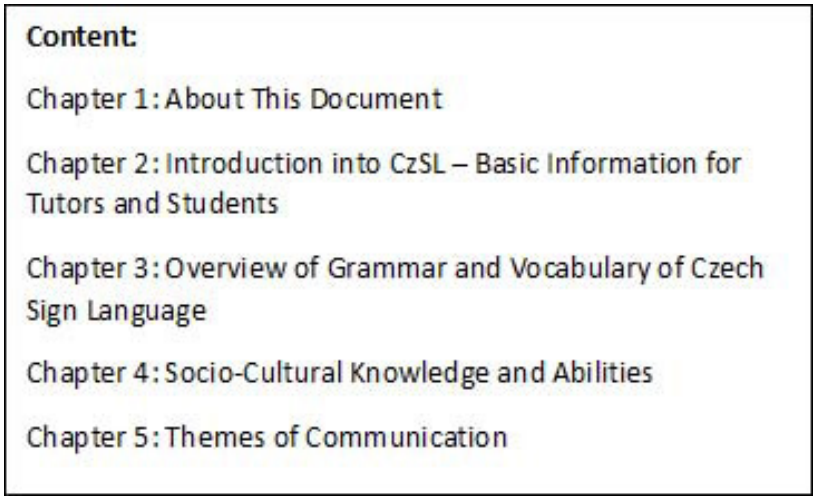


Figure 3 The RLDCSL content

Difficulties and tough moments

We faced several tough moments during the project. Some of them we called 'input problems'. For example, sign languages have no codification. There are no complex grammar summaries, dictionaries, books, or syllabi. There is a low number of studies, surveys, and investigations.

Other difficulties occurred during the process. We communicated in different languages in the team (Czech, CzSL, English) and all the texts were written in an academic way. We worked with specific CEFR terminology which all members of the team understand in the same way. CEFR terminology is unique to the CEFR context and the linguistic terminology of sign languages must also match. One of the most difficult moments was the translation into CzSL. This was the first time when this kind of academic text was translated into CzSL.

Pro-Sign 2016

	MONITORING AND REPAIR
C2	*Can backtrack and restructure around a difficulty so smoothly the interlocutor is hardly aware of it.*
C1	*Can backtrack when he/she encounters a difficulty and reformulate what he/she wants to say without fully interrupting the flow of signing.*
B2	*Can correct slips and errors if he/she becomes conscious of them or if they have led to misunderstandings.* *Can make a note of 'favourite mistakes' and consciously monitor output for it/them.*
B1	*Can correct mix-ups with the marking of time or expressions that lead to misunderstandings provided the interlocutor indicates there is a problem.*
	Can ask for confirmation that a form used is correct. *Can start again using a different tactic when communication breaks down.*
A2	*No descriptor available*
A1	*No descriptor available*

FRSL 2022

Monitoring and repair (draft translation)	
C2	*Same as C1*
C1	He can go back in the produced text and reformulate the statement without interrupting the flow of his speech. It can repair itself functionally, so there is no misunderstanding.
B2+	He can recognize minor errors in texts and mistakes that make comprehension difficult. He can often reverse these errors. He can retroactively correct his "typical mistakes".
B2	He can monitor his "typical mistakes" and become aware of errors leading to misunderstandings. He can usually correct an incorrect or inaccurately produced character and errors in language.
B1+	Can verbally ask for confirmation that the signs and utterances produced by him are comprehensible to the communication partner. In case of misunderstanding, he can reformulate the text or use other tactics (e. g., explanation with a more suitable example).
B1	Can verbally ask for confirmation that the signs and utterances produced by him are comprehensible to the communication partner. In case of misunderstanding, he can reformulate the text.
A2+	*The descriptor cannot be defined, because at this level the language user does not control such language resources and strategies that would enable him to implement the given skill.*
A2	*The descriptor cannot be defined, because at this level the language user does not control such language resources and strategies that would enable him to implement the given skill.*
A1	*The descriptor cannot be defined, because at this level the language user does not control such language resources and strategies that would enable him to implement the given skill.*
Pre-A1	*The descriptor cannot be defined, because at this level the language user does not control such language resources and strategies that would enable him to implement the given skill.*
	Note: Text – means signing production

Figure 4 Comparison of the scale Monitoring and Repair, Pro-Sign 2016 vs. FRSL 2022

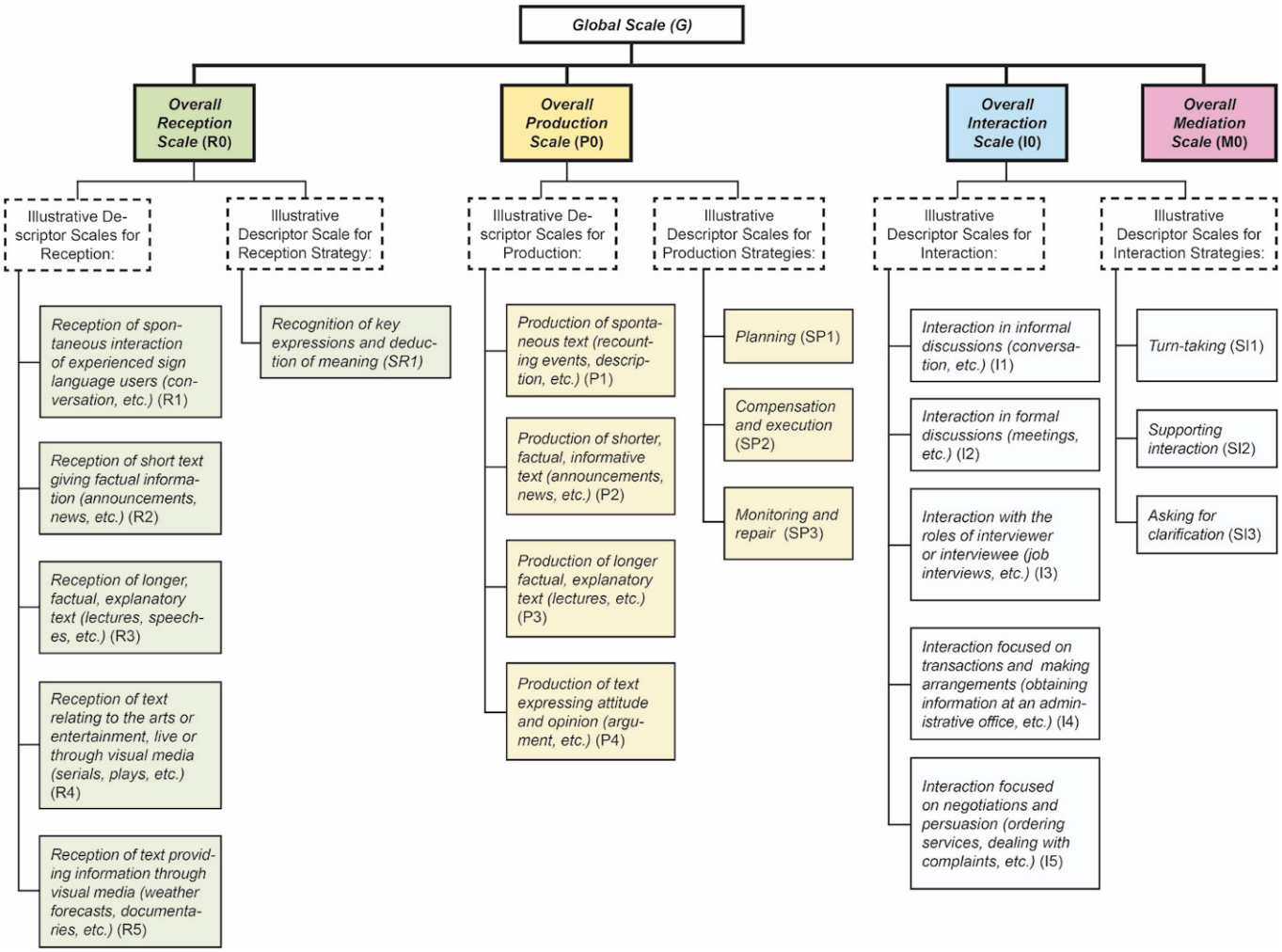


Figure 5 Overview of all scales included in the FRSL

1. **Deaf Community** (identity, Deaf users of SL, deaf places/spots)
2. **Relationships** (subcommunities, deaf organisations, name signs)
3. **Everyday Social Contacts and Rules of Communication with Deaf People**
4. **Everyday Life** (cultural events, tech. devices, interpretation, education)
5. **Life Situation** (no SL translation, labor market, social care in CZ, legislation)
6. **Shared Values** (Deaf values, history of Deaf community, Deaf art, religion, symbols)
7. **Social Conventions** (Deaf time, dining, 'clapping', 'whispering', sensitive topics)

Figure 6 Reference Level Descriptors of Czech Sign Language for Levels A1–B2 – Chapter Socio-Cultural Knowledge and Abilities

The original aim of the project was to translate existing outcomes of the PRO-Sign project and to provide accompanying texts and relevant examples for CzSL. After a few months, it became apparent a mere translation and addition of examples would not be sufficient to fulfil the project's objective. The principal reason for this was that PRO-Sign did not contain a complex enough picture of the particular modality of sign languages in contrast to spoken languages. Nor did it contain a complex enough picture of the socio-cultural, socio-linguistic, and pragmalinguistic particulars of users from the sign language community, nor a deeper linguistic picture of sign languages, and so forth.

After the first month of work and discussions, we agreed to create completely new scales and descriptors (Figure 4 and Figure 5), define new communication situations that are based on real communication in sign language, and adequately model gradually developing descriptors that can be more easily verified and tested.

We also focused on the hitherto little-explored areas of pragmalinguistic and sociolinguistic competencies and created new chapters, for example Socio-Cultural Knowledge and Abilities (Figure 6). Another important part of the project was the need to verify the created materials with the target group of users (validation groups and review phase). We had a deaf colleague who led the validation process and communicated with future deaf users of materials.

English summary on the web

We hope that summarized versions of the FRSL in English (Hulešová, 2022) and International Sign (based on the original Czech and CzSL versions) published on the website (https://cefr-czj.npi.cz/enis, https://cefr-czj.npi.cz/static/media/Framework_Reference_SL_Summary.3bd85ebf3019c2182d55.pdf) will be seen as an inspiration or guide for others wishing to create such expert documents in other national sign languages. In addition, it may be useful for organizations (including international ones) that are concerned with the teaching of sign language. Finally, it may help to provide common ground for organizations that deliver sign language instruction, or which are involved in sign language in some way.

References

Council of Europe. (2001). *The Common European Framework of Reference for Languages: Learning, Teaching, Assessment.* Cambridge: Cambridge University Press.

Council of Europe. (2020). *The Common European Framework of Reference for Languages: Learning, Teaching, Assessment: Companion Volume.* Strasbourg: Council of Europe Publishing.

Hulešová, M. (2022). *Framework of Reference for Sign Languages: Summary*. Prague: Národní pedagogický institut České republiky.

Lachmanová, D., & Štádlerová, L. (2022a). *Glosář*. Prague: Národní pedagogický institut České republiky.

Nováková, R., Petráňová, R., Štádlerová, L., Boccou Kestřánková, M., Čiháková, M., Hulešová, M., Lachmanová, D., Schormová, J., & Šimon, V. (2022b). *Popisy referenčních úrovní A1-B2 pro český znakový jazyk*. Prague: Národní pedagogický institut České republiky.

Nováková, R., Petráňová, R., Štádlerová, L., Boccou Kestřánková, M., Čiháková, M., Hulešová, M., Lachmanová, D., Schormová, J., & Šimon, V. (2022c). *Referenční rámec pro znakové jazyky*. Prague: Národní pedagogický institut České republiky.

Acknowledgments

The author of the material and all its parts, unless otherwise stated, is the collective of authors of Národní pedagogický institut České republiky (NPI CR) (successor organization of Národní ústav pro vzdělávání, NUV, Akční plán inkluzivního vzdělávání, IPs APIV A, project in the years 2017 to 2022). The work is licensed under CreativeCommons CC BY SA 4.0 – Attribution – Preserve – 4.0 International. The APIV A project was co-financed by the European Union.