Collated Papers for the

Madrid

ALTE 7th International Conference

Association of Language Testers in Europe

A L T E

ISSN 2789-2344

Collated Papers for the ALTE 7th International Conference, Madrid

Contents

Foreword

In 2020 ALTE celebrated 30 years as a multilingual association. The conference planned for Madrid in April 2020 on the theme of *Safeguarding the future of multilingual assessment* was intended as a celebration – looking back at the history, as well as looking forward to future prospects. The Covid-19 pandemic triggered a change of plan.

One year on, ALTE's 1st International Digital Symposium in April 2021 provides an opportunity to return to the theme in light of the Covid shock. On-line and 'hybrid' events will certainly be more common from now on, but ALTE will return to meeting in person as well. The next International Conference has already been confirmed and again will be co-hosted by Instituto Cervantes in Madrid in Spring 2023.

ALTE has a long-standing tradition of publishing the proceedings of its events and conferences and I am pleased to introduce this latest collection. Unusually, we have collated the papers that were due to be presented in person in Madrid even though the event itself was cancelled.

We are grateful to the authors who have contributed papers and hope that this collection will act as a positive showcase of presentations previously intended for the conference. They accepted the challenge of submitting and revising their papers in a short amount of time – so a big thank you for their efforts.

The collection is organised into four sub-themes based on the strands that the original conference was divided into.

Papers in the *Defining the Construct of Multilingualism in Language Assessment* section look at issues such as how multilingualism and plurilingualism can be encouraged in our field of language assessment, and how this construct can be defined more effectively. We also ask which policies and practices – such as the development of the CEFR Companion Volume – can be implemented to encourage this. The assessment of sign language also falls into this theme.

Fair and Valid Testing and Assessment is an ongoing issue, especially in today's ever-changing world. Papers in this section look at how we make testing fair and valid for all, despite the varied contexts where language assessment takes place, and how these aspects impact and benefit the test-taker, our primary stakeholder.

Recently there have been many *Innovations in Language Assessment*, and papers in this section look at developments seeking to make a meaningful impact on language assessment in the future. Many of the issues are to do with advances in technology, and how technology can be used wisely in our field. Outside of technology, other innovations discussed include the assessment of the *mediation mode* of communication.

The sub-strand of the conference entitled *Considerations of Ethics and Justice in the Multilingual Context of Assessment* was designed to address the question of how our profession can remain ethical and promote social justice in today's world. In this volume, we have one paper looking specifically at the language needs of refugees and migrants, a topic ALTE has paid particular attention to over the years.

As well as thanking the authors for their efforts, I'd also like to thank the reviewers from ALTE for their assistance, including: Emyr Davies, Thomas Eckes, Javier Fruns Gimenez, Marta Garcia, Tony Green, Cecilie Hamnes Carlsen, Gabriele Kecker, Cathy Taylor, Lynda Taylor, Koen Van Gorp and Dina Vilcu.

Thanks also go to the editorial team led by Graham Seed in the ALTE Secretariat.

I hope that readers will find the papers of relevance and that they contribute positively to ongoing discussions about the future of multilingual assessment.

Nick Saville,
Cambridge, April 2021

Defining the Construct of Multilingualism in Language Assessment

Testing and assessing endangered and less widely spoken languages

Marilena Karyolemou
University of Cyprus, Greece

Abstract

Taking as a case in point Cypriot Arabic (CyAr), a severely endangered language spoken in Cyprus, this paper discusses the results of the research project MapCyArS (Mapping Cypriot Arabic speakers: An investigation into linguistic demography and the sociolinguistic profile of Kormakiote Maronites), funded by the A G Leventis Foundation at the University of Cyprus (2017–2020), and highlights the need to change the way we consider language testing and assessment to fit the sociolinguistic situation of less widely spoken, oral and/or endangered languages and their communities. In the case of CyAr, we need to consider the fact that it is a language of oral tradition that has suffered severe reduction on all linguistic levels and it is not a first language for people under 40 anymore.

Introduction

Despite the rapid development and multiplication of research in the area of endangered and minority languages, testing and assessing such languages has rarely if ever been the focus of research (for notable exceptions see Borges, 2019; Borgia, 2009; Kahakalau, 2017). This lack of interest is probably due to the fact that the specificities of endangerment require the adaptation of assessment principles currently used for widely used languages. For instance, an obvious but not often discussed fact is that endangered languages with no written tradition can only be assessed orally, and oral assessment is already underrepresented in language assessment overall.

This paper briefly reports on the MapCyArS project (Mapping Cypriot Arabic speakers: An investigation into linguistic demography and the sociolinguistic profile of Kormakiote Maronites) financed by the A G Leventis Foundation at the University of Cyprus (2017–2020). The project aims at the design and development of an assessment test to evaluate proficiency in Cypriot Arabic or Sanna (hereafter CyAr), a severely endangered language spoken by members of the Maronite community in Cyprus. More specifically, it discusses the need: (a) to revise standard assessment tests based on frameworks such as the Common European Framework of Reference for Languages (CEFR) (Council of Europe, 2001) or others designed for dynamic and widely used languages, and devise approaches which are more suitable for such languages[1]; (b) to acknowledge the fundamentally bilingual nature of endangered communities and capitalize on the use of the dominant language and translanguaging as positive indicators of competence in the endangered language; and (c) to examine the possibility of using non-verbal tasks to assess competence.

Cypriot Arabic or Sanna

CyAr or Sanna is an oral variety of Arabic that has been spoken by the Maronite community of Cyprus for several centuries (Roth, 1986, 2000, 2004; Borg, 1985, 2004). All the speakers of this variety of Arabic originate from the village of Kormakitis (Kyrenia district) and almost all of them have been displaced after the 1974 Turkish invasion and were relocated to the south of the island within the Cypriot government-controlled areas (Karyolemou, 2019). Sanna belongs to the group of peripheral varieties of Arabic (Borg, 1985), and is considered to be one of the more divergent varieties of Arabic to date because of its centuries-long contact with (Cypriot) Greek ((Cy)Gr) (Kossmann, 2008; Grigore, 2019). Recognized by the Cypriot government as an indigenous minority language within the *Charter for regional or minority languages* of the Council of Europe in 2008, it has been undergoing revitalization since 2013, financed by the Ministry of Education, Culture, Sports and Youth (MECSY) (Karyolemou, 2019). MapCyArS is directly related to the Ministry's project, in the sense that it takes into account its results and complements them by working

[1] Although the CEFR is designed also for lesser used and taught languages, it has nevertheless been applied to very few, among them: Catalan, Galician, Basque, Friulan or Welsh, etc., which are officially recognized within their respective areas of use.

in an area that has not yet been tackled by the aforementioned project. Furthermore, it uses oral data from the Archive of Oral Tradition for CyAr under construction to assess various activities (Karyolemou, 2021).

MapCyArS

MapCyArS has a collaborative character and makes room for members of the community to work together with researchers both as part of an advisory focus group and as part of the research team itself (Karyolemou, 2021). It also takes into consideration both CyAr's oral nature and state of endangerment as well as the sociolinguistic situation of the community, and in particular the fact that it is not a native language nor actively used by people under 40. The project, therefore, excludes writing[2] and uses other means (oral speech, non-verbal tasks) to assess oral proficiency. It also capitalizes on the use of (Cy)Gr to assess oral comprehension in CyAr. Finally, taking into consideration the fact that most speakers under 40 have a partial and incomplete knowledge of the language, the research finds it more appropriate to measure proficiency in specific activities.

Evaluating speakers' proficiency

The test focuses on the following four areas: (a) knowledge of vocabulary, (b) oral comprehension in narratives, (c) oral comprehension in dialogue, and (d) oral production in dialogue. It uses: (a) audio and iconic material to assess knowledge of vocabulary, (b) non-verbal activities to assess oral comprehension in narratives, and finally, (c) performance in (Cy)Gr to assess CyAr conversational comprehension.

The test is organised into six internally rated activities with two to four levels of achievement in each. All the tasks evaluate oral comprehension and production. Four of them focus on oral comprehension, one on oral production and another one on the ability of speakers to accomplish specific grammatical/metalinguistic tasks.

Vocabulary I

For this activity, we used Microsoft Access with automatic rating of the answers. It consists of listening to a CyAr word and matching it with one of four images. Ten different sets of words (10 words in each set) of variable difficulty were created with the help of the focus group. To determine the degree of lexical difficulty we took into consideration the Swadesh list with additions from the Leipzig–Jakarta list for CyAr compiled by the MECSY research team in 2013–2014. The test is automatically graded. Words are organized as follows:

Category I: common words, everyday objects and kinship terms.

Category II: words for objects, places, buildings commonly found in a village.

Category III: words for traditional activities or customs, traditional gastronomy.

Category IV: words that concern a wide range of activities and objects.

Vocabulary II

The second task also evaluates mastery of vocabulary. It consists of listening to a CyAr word and identifying the corresponding item or activity on a thematic picture as follows:

Picture 1: Animals.

Picture 2: Places, objects and people of the village.

Picture 3: Everyday activities.

Picture 4: Traditional activities I.

Picture 5: Traditional activities II.

[2] CyAr is currently using a writing system based on the Roman alphabet with some additions from the Greek alphabet designed by Professor Alexander Borg in 2008. This alphabet was revised and standardized by the research team of the MECSY between 2013–2014 (Karyolemou & Armostis, forthcoming).

Oral comprehension of narratives

The third activity evaluates speakers' ability to understand short narratives in CyAr and perform specific non-verbal tasks. More precisely, the speakers are required to listen to a short narrative from the Archive of Oral Tradition for CyAr under construction. They are then given a set of pictures that illustrate the story and asked to put them in the correct order. All the stories were illustrated for the purpose of the test by a professional who produced a series of five to seven pictures for each narrative. The stories are of varying degrees of difficulty according to whether they report a series of events or include information about psychological or emotional shifts/developments, which are harder to identify. Each speaker is rated according to how many and which pictures they put in the right sequence.

Oral comprehension in conversational settings

The fourth activity evaluates oral comprehension in conversational settings and requires the presence of a native or fluent CyAr speaker whose role is to initiate a dialogue. The respondents are required to provide appropriate responses, using either (Cy)Gr, CyAr or some kind of translanguaging and are rated according to the relevance of their answers irrespective of how they do it. By accepting multiple means for this task, we acknowledge that speakers have similar verbal repertoires but variable degrees of competence.

Oral production in conversation

The fifth activity evaluates oral production in conversation and also requires the presence of a native or fluent CyAr speaker. The fluent CyAr speaker begins by asking simple questions such as name, parents' name, place of origin, place of residence etc., then proceeds with other questions in the form of a conversation. Respondents are required to answer using CyAr. They are rated according to their ability to use whatever resources they have in CyAr to provide a relevant answer rather than on their capacity to produce grammatically correct utterances.

Metalinguistic or grammatical competence

Finally, the sixth activity focuses on speakers' ability to accomplish orally simple grammatical or metalinguistic tasks, e.g., to conjugate verbs, to give the right form for a word when given its structural properties, to distinguish between the singular and the plural of nouns, verbs or adjectives, to explain a word etc.

All the tasks to be completed, types of input and output and internal rating are summarized in Table 1.

Table 1: Competence assessed, input and output, tasks and internal rating

	Focus	Input-Output	Task	Level	Competence
1	Isolated words	Verbal_{L1} vs Non-verbal	The task consists of listening to a word and choosing the right picture from a list of different pictures LEVEL 1 LEVEL 2 LEVEL 3	LEVEL 0	Does not understand any or very few words in Cat.I
				Understands most words in Cat.I and some in Cat.II	
				Understands most of Cat.I and II words and some in Cat.III	
				Understands most of the words in all categories	
2	Isolated words	Verbal_{L2} vs Non-verbal	The task consists of listening to a word and choosing the right item|activity out of a number of items|activities present on a picture LEVEL 1 LEVEL 2 LEVEL 3	LEVEL 0	Does not understand any words in CyAr
				Understands simple words widely used	
				Understands specialized vocabulary in relation to traditional activities	
				Understands a wide range of vocabulary	

	Focus	Input-Output	Task	Level	Competence
3	**Narrative**	**Verbal$_{L2}$ vs Non-verbal**	The task consists of listening to a story and choosing the right order of events in a set of pictures **LEVEL 1** **LEVEL 2**	**LEVEL 0**	Fails to sequence correctly
				Partially accomplishes the task by correctly sequencing some pictures	
				Successfully accomplishes the task by sequencing all pictures correctly	
4	**Conversation**	**Verbal$_{L2}$ vs Verbal$_{L1}$**	The task consists of using CyGr or translanguaging to provide short answers to questions asked in CyAr **LEVEL 1**	**LEVEL 0**	Fails to understand CyAr and correctly respond in SG[3]/(Cy)Gr
				Correctly understands CyAr and responds in SG/(Cy)Gr	
5	**Conversation**	**Verbal$_{L2}$ vs Verbal$_{L2}$**	The task consists of using CyAr to provide short answers to questions asked in CyAr **LEVEL 1** **LEVEL 2**	**LEVEL 0**	Fails to understand and correctly respond in CyAr
				Correctly understands common sentences and responds in CyAr	
				Correctly understands a wide range of conversation	
6	**Metalinguistic/ Grammatical**	**Verbal$_{L2}$ vs Verbal$_{L2}$**	The task consists of completing orally meta-linguistic or grammatical tasks such as declining a verb, giving the plural of a noun, translating etc. **LEVEL 1** **LEVEL 2**	**LEVEL 0**	Unable to accomplish any metalinguistic task
				Able to accomplish simple metalinguistic tasks	
				Able to accomplish complex metalinguistic tasks	

Conclusion

Designing an assessment test for an endangered oral language like CyAr is a complex objective that needs to take into consideration the oral character of CyAr, its endangered situation and the advanced state of its replacement by (Cy)Gr.

The oral character of CyAr suggests that we move away from reading and writing as means to evaluate oral skills. Its endangered situation suggests that traditional approaches in testing and assessment by level of competence (A1, A2, etc.) should be abandoned in favour of a more flexible type of evaluation. The advanced state of CyAr replacement by (Cy)Gr leads towards a more flexible use of language resources to evaluate speakers' proficiency, and invites us to capitalize on the use of (Cy)Gr or translanguaging to do so. The project is, therefore, also an opportunity to re-examine the baselines of language assessment used for modern languages in order to adapt them to conditions of orality and severe endangerment.

References

Borg, A. (1985). *Cypriot Arabic: A historical and comparative investigation into the phonology and morphology of the Arabic vernacular spoken by the Maronites of Kormakiti village in the Kyrenia district of north-western Cyprus.* Stuttgart: Komissionsverlag Steiner Wiesbaden.

Borg, A. (2004). *A Comparative Glossary of Cypriot Maronite Arabic (Arabic-English) With an Introductory Essay*. Handbook of Oriental Studies Section 1: Near and Middle East volume 70. Leiden: Brill.

Borges, R. (2019). Rapid automatized picture naming as a proficiency assessment for endangered language contexts: Results from Wilamowice. *Journal of Communication and Cultural Trends*, *1*(1), 1–25.

[3] SG = Standard Greek

Borgia, M. (2009). Modifying assessment tools for Ganöhsesge:kha: Hë:nödeyë:stha, a Seneca culture-language school. In J. Reyhner & L. Lockard (Eds.), *Indigenous Language Revitalization: Encouragement, Guidance and Lessons Learned* (pp. 191–210). Arizona: Northern Arizona University.

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

Grigore, G. (2019). Peripheral varieties. In E. Al-Wer & U. Horesh (Eds.), *The Routledge Handbook of Arabic Sociolinguistics* (pp. 117–133). Abingdon: Routledge.

Karyolemou, M. (2018). Language revitalization, land and identity in an enclaved Arab community in Cyprus. In. S. Drude, N. Ostler & M. Moser (Eds.), *Language Documentation and Description. Proceedings of the 22nd Annual Conference of the Foundation for Endangered Languages* (pp. 14–18) Foundation for Endangered Languages XXII. London: EL Publishing. Retrieved from: www.elpublishing.org/book/endangered-languages-and-the-land

Karyolemou, M. (2019). A story at the periphery: Documenting, standardizing and reviving Cypriot Arabic. *International Journal of the Sociology of Language*, *260*, 1–14.

Karyolemou, M. (2021). *MapCyArS 2017-2020. Mapping Cypriot Arabic speakers: An investigation into linguistic demography and the sociolinguistic profile of Kormakiote Maronites*. Nicosia: University of Cyprus Editions.

Karyolemou, M. (forthcoming). Teaching endangered languages of oral tradition: How and what to assess?. *Applied Linguistics.*

Karyolemou, M. & Armostis, Sp. (forthcoming) Langue ancienne, écriture nouvelle: codification et standardisation de l'arabe de Chypre. *Cahiers du Centre d'Etudes Chypriotes*, *50*.

Kahakalau, K. (2017). Developing an indigenous proficiency scale. *Cogent Education*, *4*. Retrieved from: www.tandfonline.com/doi/full/10.1080/2331186X.2017.1377508

Kossmann, M. (2008). Parallel morphological systems due to the borrowing of paradigms. *Diachronica, 27*(3), 459–488.

Roth, A. (1986). Langue dominée, langue dominante: à propos de deux scénarios d'extinction ou d'expansion de l'arabe. *Hérodote*, *42*, 65–74.

Roth, A. (2000). Un usage linguistique en voie d'éviction. Observations sur la "réduction" syntaxique et stylistique dans le parler arabe de Kormakiti. In *Chypre et la Méditerranée orientale. Formations identitaires: perspectives historiques et enjeux contemporains* (pp. 127–137). Actes du colloque tenu à Lyon, 1997, Université Lumière-Lyon 2, Université de Chypre. Lyon: Maison de l'Orient et de la Méditerranée Jean Pouilloux.

Roth, A. (2004). Le parler arabe maronite de Chypre: observations à propos d'un contact linguistique pluriséculaire. *International Journal of the Sociology of Language*, *168*, 55–76.

Assessing writing across multilingual contexts in Southeast Asia

Louise Courtney
Australian Council for Educational Research (ACER) 2021

Abstract

This paper introduces the regional assessment called SEA-PLM and explains the challenges faced by test developers in terms of construct definition, task development and scoring. Prior to the field trial, ACER test developers hypothesised that some features of writing are comparable across languages, while others are language specific. However, after removing poorly functioning items post field trial, the main study showed that all the features of writing assessed were comparable across languages. This provides the foundation for an argument that cross-language comparisons of writing literacy are possible by using generic criteria for scaling.

Introduction

This paper introduces the regional assessment called SEA-PLM and explains the challenges faced by test developers in terms of construct definition, task development and scoring, as well as the processes used for translation. Some recent results of the main study for student writing and language issues are presented.

What is SEA-PLM?

The Southeast Asia Primary Learning Metrics (SEA-PLM) is a new comparative learning assessment programme, designed by and for countries in Southeast Asia. It provides robust evidence about how children in Southeast Asia perform against regional measurements in reading, writing and mathematics in Grade Five towards the end of primary school. After field trials in 2017, the SEA-PLM main survey was implemented at the end of the 2018–2019 school year. Six countries from the region were involved: Cambodia, Lao PDR, Malaysia, Myanmar, The Philippines and Vietnam. The results of the first round of the SEA-PLM main study were released in December 2020 in *SEA-PLM 2019 Main Regional Report, Children's learning in 6 Southeast Asian countries*. This document is also available in digital format at www.seaplm.org

The programme generated reliable data and evidence for monitoring student learning outcomes across and within countries. It also provided an understanding of what factors help or hinder effective learning along children's school journeys in participating countries. To keep the assessment relevant, the SEA-PLM programme is designed so that children's achievement can be measured over time through subsequent cycles of assessment, with the next one proposed for 2023.

Contextual surveys

Surveys were also conducted to gather contextual information about the students being assessed. The results of these surveys enable research into other educational questions through regional comparison of learning environments, children's experiences, school practices and specific issues such as whether students' mother tongues are the same as the language of instruction at their school. Participating countries are now better able to monitor the profile of disadvantaged sub-groups of children and schools at the end of primary years, and they also can correlate learning achievements with this contextual information.

Construct definition, task development and scoring

Orientation

The SEA-PLM Assessment Framework adopts a 'literacy orientation', focussing on literacy and numeracy skills that students will need for life and for study beyond primary school. This means that writing assessment tasks reflect real-life tasks that students need to perform to communicate effectively in writing in their own languages.

Languages

The SEA-PLM assessment of Writing Literacy is particularly novel. There has never been a cross-language writing instrument administered across the Southeast Asian region. Indeed, there have been very few other international assessments that have attempted to directly compare writing literacy across multiple languages. The source versions of all SEA-PLM 2019's assessment materials were prepared in English, then translated, trialled and quality assured by regional experts. Tests and questionnaires were administered in the official language(s) of instruction in Grade Five in each country, as determined by each Ministry of Education in collaboration with the SEA-PLM Secretariat. Table 1 presents the final list of test languages by country.

Table 1: Test languages by country

Countries	*Language(s) tested*
Cambodia	Khmer
Lao PDR	Lao language
Malaysia	Chinese, Malay, Tamil
Myanmar	Myanmar language (Burmese)
Philippines	English
Vietnam	Vietnamese

Assessment design for writing tasks

In defining the construct, to make clear what we were aiming to assess, test developers at ACER had to create a **framework**, or test design, that would allow us to compare student performance across languages. There were many things to be considered such as linguistic and cultural differences in terms of producing written texts, as well as teaching and learning styles and community values. The main task was to gather evidence about what students could do, in terms of writing, rather than collecting evidence of their technical knowledge of areas such as formal grammar. Before designing the assessment, ACER test developers asked these five key questions:

1. What is writing?

2. What are suitable **tasks** that would apply to writers of all languages?

3. What are the main features of writing, **regardless** of language?

4. What aspects of writing might **not** be able to be assessed in a variety of languages?

5. In what **contexts** should the writing tasks be set?

Responding to the key questions

What is writing?

The definition of **writing** developed was: *Writing literacy is constructing meaning by generating a range of written texts to express oneself and communicate with others, in order to meet personal, societal, economic and civic needs.*

It is important to note that this practical, literacy focus means that the emphasis is on creating *meaning* for *communication*. As such, efforts such as copying slabs of text or even single words without imparting meaning are not valued as much as efforts where students create individual meaning by thinking and combining words together to produce original text, however brief.

What are suitable tasks that would apply to writers of all languages?

In answer to this question, test developers concluded that the following **tasks** would apply to students writing in any of the languages.

- Narrative. For example, writing a story in response to a picture.
- Descriptive. For example, describing a well-known event or experience in writing.

- Persuasive. For example, giving an opinion on something in writing.
- Instructional. For example, writing sentences about how to do a familiar task.
- Transactional. For example, a real-life task such as writing a note to communicate.

These tasks are not unusual for student assessments of writing. However, an extra one was added to provide access to the test for lower-performing students. This was:

- Labelling. For example, writing a word to correspond to a picture of a basic and well-known item.

What are the main features of writing, regardless of language?

Given the literacy focus of the writing tasks, our research suggested that the main abilities that could be compared across languages were to assess if the student producing the writing could:

- Generate ideas.

  Writing tasks typically require the creation, selection and crafting of ideas. The quantity and quality of the ideas, and their suitability for the task, are parts of this skill.

- Control text structure and organisation.

  Different text types have different structures. Effective writers select a suitable organisational form for the particular writing task.

- Manage coherence.

  Good writers are able to structure texts so that the links between ideas are clear. They produce a logical progression of ideas that express meaning, as well as through writing features such as reference, and lexical features such as discourse markers and connectives.

- Use vocabulary.

  Writing involves knowledge of words and an understanding of how they can be used in specific contexts. Good writers utilise a wide vocabulary to present ideas precisely and concisely.

- Control [of] syntax and grammar.

  Good writers produce grammatically correct, meaningful sentences and use a range of syntactic structures.

(*SEA-PLM 2019 Global Citizenship Assessment Framework*, UNICEF & SEAMEO, 2017, pp. 35–39)

What aspects of writing might not be able to be assessed in a variety of languages?

Prior to SEA-PLM, researchers did not yet know whether certain language-specific features, such as spelling, character formation/handwriting, punctuation and register could be compared across languages. We decided to call these aspects of writing: 'Other language-specific features.' However, after the main study was conducted, the data showed that cross-language comparisons of writing literacy capacity in terms of language-specific features *are* possible, using generic criteria for scaling. The information gained from the language-specific features items was therefore able to be used when developing the band descriptors. This 'language-specific' aspect of writing could be a fruitful area for future detailed research when comparing writing between various languages.

In what contexts should the writing tasks be set?

The SEA-PLM programme uses three contexts: personal contexts, local contexts, and wider-world.

Scoring the writing tasks

In the SEA-PLM writing assessment, students must write something for every task. This differs from many other tests where multiple-choice questions are used to assess students' knowledge of grammar, punctuation and spelling, for example. The writing section presents tasks requiring students to produce a written response. Each task may assess a range of criteria with differing numbers of scale score points available to be awarded, depending on the quality of the writing produced. For example, a labelling task may only assess the ability of the student to *use vocabulary*. Can the student write the word in their language of instruction for a given picture? In contrast, an everyday communication task, such as writing a note, might be assessed according to three criteria such as *controlling text structure and organisation* and two of the other language-specific features, such as *spelling* and *handwriting/character formation.* Each specific criterion may contain up to four score points. This technique of assessing writing provides SEA-PLM countries with a large, focused and detailed amount of information about what their students can actually do in terms of producing writing in their language of instruction.

Measuring student proficiency in writing

A SEA-PLM proficiency scale of eight bands was developed using the psychometric data. The description of each band describes what student writers can do. For instance, in the lowest band (Band 1 and below) students have only limited ability to present ideas in writing. Students who are in the higher bands have demonstrated varying proficiencies in writing literacy skills, with those in Band 8 and above able to write cohesive texts with detailed ideas and a good range of appropriate vocabulary. Students in the higher bands are therefore working towards meeting the SEA-PLM definition of writing literacy, 'constructing meaning by generating a range of written texts to express oneself and communicate with others, in order to meet personal, societal, economic and civic needs' (UNICEF & SEAMEO, 2019, p. 30).

How did students perform?

Student performance varied greatly across each of the six SEA-PLM countries. Each country displays their own unique level of student ability in writing, which depends on a multitude of factors, including relative wealth and the condition of many aspects of the education system. Approximately 9% of students who sat SEA-PLM 2019 performed at Band 7 or Band 8 or above, the highest two bands. The middle four bands have similar proportions of students in them; approximately 62% of all students fall into one of the middle bands. Below this, 30% of students averaged out across the SEA-PLM participating countries are in the **lowest** band for writing literacy.

From our data, ACER test developers can see that even after five years of schooling, *more than half* of participating students in the region are in the lowest bands, 1 to 3. These students could only produce limited writing, with simple ideas and basic vocabulary, in the SEA-PLM assessment. These students are falling far behind, unlikely to catch up, despite being near the end of their primary education. As communication is an important 21st century skill, the need for improvement in writing skills across the region is vital.

What does the data show about language policy?

Many surveys were attached to the SEA-PLM 2019 assessment, including one data about the language of testing and the main language spoken by students at home. This information was then correlated with test scores for all subjects across the countries (*SEA-PLM 2019 Main Regional Report*, UNICEF & SEAMEO, 2020, p. 72). Across five of the six participating countries, on average, children who reported that the language of instruction (and of the test) was the same as the language spoken at home outperformed children who spoke another language, on average, in all domains. The effect was most pronounced for writing performance in generally lower-performing countries, with increased scores by 10–20 points when the language spoken at home was the same as the language of instruction.

Learning outcomes are affected by a whole array of factors that go beyond that of the school environment. The socio-cultural context of the different countries involved in SEA-PLM means that many countries in the region are made up of populations with hundreds of national languages. The complex discussion between mother tongue and/or the adoption of a common language of instruction continues to be fiercely debated. However, it is no surprise that when students are assessed in a language different to their mother tongue, they perform under disadvantaged circumstances.

Conclusion

SEA-PLM is the first large scale assessment of its kind, designed to reflect Southeast Asian students' contexts, and assess their skills in writing, reading, mathematics and global citizenship after five years of primary schooling. The writing assessment in particular is of note to educational researchers as there have been very few international assessments that have attempted to directly compare writing literacy across multiple languages. It appears that researchers were successful in placing students' writing in different languages on the same scale, by using generic criteria for scaling.

One important aim of the assessment is to promote cross-border exchange on learning and education policies and to help countries to identify, prioritise and address educational challenges in important policy areas, such as curriculum development, resource allocation, teacher training, classroom practices, and planning at both national and sub-national levels. Further research will reveal if this comes to fruition now that the data is widely available.

References

UNICEF., & SEAMEO. (2017). *SEA-PLM 2019 Global Citizenship Assessment Framework* (1st ed.). Bangkok: UNICEF/ SEAMEO – SEA-PLM Secretariat.

UNICEF., & SEAMEO (2020). *SEA-PLM 2019 Main Regional Report, Children's learning in 6 Southeast Asian countries.* Bangkok: UNICEF/SEAMEO – SEA-PLM Secretariat.

Promoting a multilingual policy in secondary schools in three languages: The case of Israel

Tziona Levi
Ministry of Education, Israel

Dvora Harpaz
Levinsky College, Israel

Abstract

Israel is a multicultural, multilingual society. A question arises as to whether a multilingual education policy should be introduced in schools that differ from the current situation in which each language is taught independently regardless of the teaching of other languages, with the advantage that a multilingual policy in schools can lead to openness and tolerance among learners.

We present an example of a trilingual policy in 26 secondary schools in the northern Israeli periphery where three languages, English, Arabic and Hebrew, are taught reciprocally. The initiative is informed by ecological thinking as an approach to educational change leading to the integration of collaboration, interaction and the development of language disciplinary teaching teams. 57 English, Arabic and Hebrew language department heads were trained to lead a change in the schools' policy for language teaching.

The initiative was accompanied by a mixed methods study to examine whether and how language instruction had changed following the training of school language department heads. Study results showed changes in the planning and implementation of language teaching in the schools in question on the path to developing a school multilingual policy. They also showed a significant shift in the language department heads' perception of their role as local policy and change leaders.

Theoretical background

Israel is a culturally and linguistically diverse society. This is a result of (1) massive waves of Jewish immigration from many countries; (2) the local Arabic-speaking population; (3) the proliferation of Hebrew as the national language; and (4) the arrival of migrant workers and refugees mainly from Africa and East Asia (Or and Shohamy, 2016). It is expected that such a society would uphold the values of tolerance, linguistic justice and rights (Shohamy, 2014), and multicultural education would seem to be the way to ensure openness to and respect for diversity in order to increase social equality and strengthen democratic discourse (Einav, 2020).

Assuming that a multicultural/multilingual society wishes to train its graduates to master additional languages beyond their mother tongue (Shohamy & Spolsky, 2003), a change is suggested for Israel, in which each language is currently taught in isolation (Haskel-Shaham & Shaniak, 2015). Salsa-Murcia and Olshtain (2014) mention that such a beneficial change is possible to enable learner use of another language efficiently and competently.

Nevertheless, it seems that insufficient attention is being paid to developing a multilingual approach to language teaching and learning in Israel's education system. Haskel-Shaham and Shaniak (2015, p. 177) claimed that 'there is currently no common denominator for teaching the various languages . . . there is a tendency to see each "player" in isolation when it comes to policy for language learning in high school.' This needs to be rectified by strengthening knowledge about language acquisition and awareness of the development of teaching/learning skills among active language teachers and not just among language teaching experts in the academic setting (Donitsa-Schmidt & Inbar-Lourie, 2014).

This notion is reinforced in the study on multilingual policy in Israeli secondary schools conducted by Shohamy and Tennenbaum (2019). They found that teachers and students perceive multilingualism as a value with important benefits for the individual and for society as a whole. Their feelings were that the three main languages (Hebrew, Arabic and English) should be taught throughout the years of schooling, with options for additional languages. However, this study also mentioned obstacles to building a multilingual policy in Israel, such as the fact that EFL teachers are encouraged in their training to avoid using the first language in class (although clearly they often do), and have no training to work with immigrants with other first languages (e.g., Russian, French, Amharic, see Haim, 2014). Moreover, as of yet there are no documented guidelines for effective use of first languages to enhance second/additional language acquisition. Thus, it is important to develop programs that provide present and future teachers with a theoretical and practical foundation upon which to build and implement a multilingual policy for their schools.

It is with this data in mind that one of the school chains in Israel decided to devise a plan to strengthen its schools' multilingual policy and validate it through the evaluation research presented here. The research focused on seeking what was common to the teaching of the key languages – Hebrew, Arabic and English (Haim, 2014) – and on the need to promote a multilingual context which would consider social mobility and linguistic choices (Olshtain & Nissim-Amitai, 2008).

This study was grounded in ecological thinking (Keaney, 2006), an approach to educational change that assimilates principles such as reciprocity, collaboration, teacher interaction, and developing learning communities of teacher teams. The essence of the desired change involved a change in both language teaching pedagogy and the perception of the language teacher's role in imparting literacy skills. According to Sarason (2011) this would also involve changes in routines of the behavioral and organizational culture in school, and not just in students' achievements.

The aim was to generate significant and sustainable long-term change in how the three language department heads for Hebrew, Arabic and English perceived their roles in collaboration with the school management. Hence the process posed a challenge that was both pedagogical and organizational. In terms of content, the study sought to integrate principles of education policy to support the development and implementation of a joint agenda for teaching first and second/other languages for all participants in the program.

Keaney (2006), presenting a systemic educational ecological process based on reciprocity among all entities at school, saw this as different from the more familiar 'bottom up' and 'top-down' models of change. She considers school relationships as a spiral process of mutual dependence among learners with collaboration and interaction between teachers, and where community members (including researchers) take part. The emphasis is on the dual role of all participants as interactive stakeholders on the one hand, and as reflectors who are aware of their functioning and their knowledge systems and values on the other. Furthermore, she claimed that systemic change begins with one subsystem that then drives other subsystems (for example, the community of English language department heads interacting with another group of language department heads) to establish new mechanisms (joint in-service training). This may enable ongoing contact among teachers from different disciplines regarding common issues such as shared knowledge, visions, values and norms, as they collaborate to promote change and enhance their learning. This, of course, requires professional development (e.g., Shulman, 1986), as there are accumulative influences of teacher knowledge acquired in social-collaborative settings (such as the community of language department heads) that will positively affect students' future achievements.

Research context

The schools in question had no pedagogical or organizational infrastructure to help promote awareness of a multilingual policy, thus this research accompanied an initiative based on two assumptions: 1) different languages have several common components (Haim, 2014), and 2) familiarity with school cultures is possible thanks to knowledge of the target language, for example Arabic (Amara, 2014). This would seem to be an imposed change as defined by Hargreaves (2011) since it was initiated by the school network management staff in the hope that it would be accepted by the language department heads who understood the need. There were two reasons for the choice to work with the department heads for English, Hebrew and Arabic: (a) they could develop study processes and class management practices (Lalas, 2007); (b) as department heads, they are part of the school's middle management responsible for the work of their staff (Barak-Medina, Avni, & Ben Arie, 2011). This choice was further influenced by Adey (2000), who claimed that department heads are usually undertrained, but with the proper support, their potential for improvement of their management skills would grow.

Thus, the language department heads were being trained so that together they could build a school language education policy. The training included raising awareness of the need for reciprocity in language teaching at school and for a multi-year plan for the development of a multilingual education policy. The program evaluated in the current research included a year-long in-service course addressing issues such as common language disciplinary topics, tools for implementing change at school, establishing and leading learning communities that focus on language issues, and means for developing a dialogue with students to impart common language skills.

In order to obtain more details of the implementation of the changes for an initiative that combines a multilingual policy of three languages and the use of a school pedagogical language to enable the work of the three language disciplines, the current research applied both qualitative and quantitative methods. It sought to ascertain whether and how the training promoted individual and group learning and development in order to establish a language education policy in the participating schools. The study also sought to examine the degree of self-efficacy that developed among the department heads during the year. Thus, it can also serve as an example of school activity that promotes multilingual education.

The research questions that guided this study were as follows: Has there been a change in language instruction in the school following the training of language department heads and, if so, what is the evidence of this change to display a multilingual policy? Can the department heads lead such a change in school policy?

Quantitative and qualitative data were obtained from participants' performance tasks, open-ended feedback forms and a mixed method questionnaire with open and closed questions administered at different times during the training period.

Research population

Participants were 57 department heads from 26 secondary schools with 31 Jewish, 22 Arab and four Bedouin populations who participated in an in-service training course.

Findings

The quantitative data indicate an enhanced sense of efficacy in creating change: 79% noted they felt they now had the tools to lead the change 'to a very great extent', and another 21% 'to a great extent'.

Analysis of the qualitative data yielded four main categories and key themes:

(1) the language department heads' role: evidence of a sense of self-efficacy; (2) understanding the double role as discipline team leader and as a school change leader: taking responsibility, finding solutions, needing outside knowledge to understand their role; (3) sense of change to enable the lead from a single language disciplinary team to a team collaborating on work plans and divisions of labor with another language in school; (4) reciprocity in promoting language teaching, engaging with the other language teams in school, collaboration, building routines and sharing knowledge.

Data revealed several attempts in developing work plans and initiatives promoting a multilingual school approach such as: writing a trilingual dictionary, class programs to expand vocabulary learning in three languages, conducting a bilingual debate, and teaching common literacy topics such as reading strategies in three languages.

Participants' responses in the qualitative data indicated other changes, such as:

'I do not have to look only at English as a second language, but at the whole linguistic curriculum in my school.'

'Cooperation among teachers from different languages can lead students to sharing insights and skills from one language to another.'

'It is very important to turn the team into a professional community in which peers work together to promote a certain goal rather than on their own . . . peer teaching helps improve language learning because teachers see what other teachers are doing and share their own knowledge, so the whole group advances.'

The sample outcomes displayed by the department heads' products were early indications of school initiatives such as writing argumentation, teaching logical sequences, understanding rhetorical devices and so forth in all three languages, which signaled the implementation of a multilingual policy in the schools. The findings also reinforced the creation of a multidirectional process thanks to the ecological thinking, the components of which were:

1. Developing a multilingual policy.

2. The need to implement change in the school setting.

3. Training language department heads.

4. Awareness of the need for reciprocity in language teaching.

5. Developing a language policy in the school, and so forth.

The research findings showed that language department heads can generate a school change in the teaching of the three languages as one whole if they attend appropriate training informed by the concept of a multilingual policy for a school context. The promotion of such a policy might also advance collaboration in the teaching of the language disciplines taught at school in order to further establish a multidisciplinary-multilingual approach.

References

Adey, K. (2000). Professional development priorities: The views of middle managers in secondary schools. *Educational Management & Administration*, *28*(4), 419–431.

Amara, M. (2014). The 'Yad B'yad' model of bilingual education: Vision and challenges. In S. Donitsa-Schmidt & O. Inbar-Lurie (Eds.). *Issues in Language Teaching in Israel Part 2* (pp. 56–77). Tel Aviv: MOFET/CLIL.

Barak-Medina E., Avni, R., & Ben Arie, Y. (2001). *Developing Mid-level Leadership at School*. Retrieved from: avneyrosha.org.il/resourcecenter/resoursesdocs/

Donitsa-Schmidt, S., & Inbar-Lurie O. (Eds.). (2014). *Issues in Language Teaching in Israel, Part 2*. Tel Aviv: MOFET/CLIL.

Einav, Y. (2020). Education towards multiculturalism in Israel. In R. Sagie, L. Biberman-Shalev, & Y. Gilat (Eds.), *Education in Multicultural Spaces* (pp. 69–86). Tel Aviv: Resling.

Haim, O. (2014). Transferring language components of academic competence in three languages for Russian-speaking immigrants. In S. Donitsa-Schmidt & O. Inbar-Lurie (Eds.), *Issues in Language Teaching in Israel Part 1* (pp. 125–150). Tel Aviv: MOFET/CLIL.

Haskel-Shaham, I., & Shaniak, M. (2015). *Teaching Language in Israel: Between Vision and Reality – The Evolution of Language Policy*. Tel Aviv: Tel Aviv University/Bar Ilan University.

Hargreaves, A. (2011). Inclusion and exclusion in educational change: Teachers' emotional responses and their implications for leadership. In G. Fisher & N. Michaeli (Eds.), *Change and Improvement in Education Systems* (pp. 162–184). Jerusalem: Branco-Weiss Institute/Avney Rosha Institute/Urim.

Keaney, S. (2006). *Ecological Thinking: A New Approach to Educational Change*. Tel Aviv: MOFET/CLIL.

Lalas, J. (2007). Teaching for social justice in multicultural urban schools: Conceptualization and classroom implication. *Multicultural Education, 14*(3), 17–21.

Olshtain, E., & Nissim-Amitai, F. (2008). Language acquisition in the multicultural and multilingual context. *Hed Hahinuch Hehadash, 94*, 3–17.

Or, I. G., & Shohamy, E. (2016). Asymmetries and inequalities in the teaching of Arabic and Hebrew in the Israeli educational system. *Journal of Language and Politics*, *15*(1), 25–44.

Sarason, B.S. (2011). Regularities and behaviors. In G. Fisher & N. Michaeli (Eds.), *Change and Improvement in Education Systems* (pp. 70–87). Jerusalem: Branco-Weiss Institute/Avney Rosha Institute/Urim.

Salsa-Murcia, M., & Olshtain, E. (2014). Discourse-based approaches: A new framework for second language teaching and learning. In S. Donitsa-Schmidt & O. Inbar-Lurie (Eds.), *Issues in Language Teaching in Israel Part 1* (pp. 98–124). Tel Aviv: MOFET/CLIL.

Shohamy, I. (2014). Language policy and language and social justice in Israel. In S. Donitsa-Schmidt & O. Inbar-Lurie (Eds.), *Issues in Language Teaching in Israel, Part 1* (pp. 68–97). Tel Aviv: MOFET.

Shohamy, E., & Spolsky, D. (2003). A new language education policy in Israel: From mono to multilingualism?. In Y. Dror, D. Nevo & R. Shapira (Eds.). *Changes in Education: Guidelines for Education Policy in Israel for the Second Millennium* (pp. 193–208). Tel Aviv: Ramot, Tel Aviv University.

Shohamy, I., & Tennenbaum, M. (2019). *Research-led Policy: Empirical Underpinning for an Educational Multilingual Policy*. Retrieved from: meyda.education.gov.il/files/LishcatMadaan/stoneReport2Final.pdf

Shulman, L. S. (1986). Paradigms and research programs for the study of teaching. In M.C. Wittrock (Ed.), *Handbook of Research on Teaching* (3rd ed., pp. 3–36). New York: Macmillan.

La incorporación del anglicismo en la evaluación y certificación del español

Marta García
Cursos Internacionales de la Universidad de Salamanca, Spain

Abstract

Planteamos en estas páginas una reflexión sobre la importancia del extranjerismo, en general, y del anglicismo, en particular, en el mundo de la evaluación y la certificación lingüística. En el caso del español, encontramos cada día en los medios de comunicación y en las conversaciones entre particulares términos extraídos del inglés. Algunos son admitidos por la Real Academia de la Lengua, otros no, pero su uso está muy extendido. ¿Qué debe hacer el evaluador y el profesor cuando se encuentra uno de estos términos en el texto o en la expresión de un candidato? ¿Debe corregirlos? ¿Debe enseñarlos? ¿Es consciente el candidato de que su uso es posible? A estas y a otras preguntas intentaremos responder en las páginas siguientes.

Introducción

Que el inglés es la lengua global no se discute. Que es la lengua que ha penetrado en nuestra realidad hasta encontrar acomodo, tampoco. En este avance sin freno, como no podía ser de otra manera, también ha llegado a la enseñanza del español como LE/L2 y, del mismo modo, a la vida del profesor, que cada día se encuentra con más términos del inglés en los textos orales y escritos, y a la del evaluador, que debe calificar pruebas orales y escritas de candidatos que viven esta misma realidad inusitada (Echazú y Rodríguez, 2018). Por lo tanto, cabría preguntarse qué debemos hacer los evaluadores si aparecen ciertos términos en inglés en las producciones de nuestros candidatos, qué criterio debemos seguir ante la invasión de estos términos que, en muchos casos, han venido para quedarse (Calvi, 2013), ¿Debemos aceptarlos cuando evaluamos una lengua que no es el inglés? ¿Debe aparecer alguna mención en los diferentes descriptores de las certificaciones lingüísticas? ¿En los repositorios que usamos para crear tareas? ¿Podemos hablar de *translanguaging* en algunos casos?, ¿multilinguismo?, ¿plurilinguismo?

Metodología

Con el objeto de aclarar un poco la cuestión del anglicismo en la evaluación y certificación del español, y para intentar llegar a alguna conclusión que pueda ser útil, hemos elaborado el siguiente itinerario de trabajo. En primer lugar, hemos elaborado una lista con 19 palabras inglesas admitidas por la Real Academia de la Lengua Española (a partir de ahora, RAE) y otra lista con otras 19 palabras inglesas que no están admitidas por la RAE. Los términos que recogen las listas aparecen en la prensa económica con bastante frecuencia. Hemos acotado el campo del léxico al mundo de los negocios por ser uno de los que más anglicismos recoge. En investigaciones anteriores (García, 2021), intentamos establecer una relación de frecuencia de uso entre el anglicismo y su correlato en español para ver cuál de los dos se usaban más con el fin de intentar llegar a conclusiones que nos permitieran tener una referencia de rentabilidad de uso en cada caso. Creemos que, si un anglicismo es más usado que su correlato, no debería penalizarse por usarse en su lugar. Si es más usado, es porque se encuentra más vivo en la lengua de nuestros estudiantes, mientras que el uso del correlato en español podría llegar a provocarle, en algún caso, una interferencia comunicativa o alguna estridencia en su producción oral o escrita, ya que es posible que haya más conocimiento social del anglicismo que del correlato.

Hemos entregado estas dos listas de palabras a 53 estudiantes de negocios de nivel avanzado de diferentes lenguas maternas: japonesa (10), inglesa (20), china (8), portuguesa (3), alemana (2) y neerlandesa (1). Los estudiantes debían indicar, en primer lugar, si conocían el término en inglés y, en segundo lugar, debían escribir el correlato en español. Nos interesaba esclarecer si, en sus producciones escritas y orales, incluirían uno u otro. La idea es que, si un candidato a una certificación lingüística necesita escribir, por ejemplo, sobre la comunicación en el siglo XXI, lo normal es que emplee exclusivamente palabras de la lengua que va a certificar. Pero lo cierto es que el estudiante puede no conocer la palabra en español, pero sí en inglés, y, curiosamente, puede que ese término esté aceptado por la RAE. La inercia del estudiante es emplear la palabra en español, pero no sabe que podría hacerlo en inglés con palabras aceptadas por la RAE. Así, por ejemplo, el término *wifi* está aceptado por la RAE, y el candidato intentará evitarlo porque cree que la palabra no es española y desconoce que esté admitida. Creemos que no solo deberíamos hablar de palabras admitidas por la RAE, sino de palabras admitidas socialmente, aunque la RAE todavía no se haya pronunciado.

Por tanto, si se emplean palabras admitidas por la RAE y palabras no admitidas por la RAE, pero sí aceptadas socialmente, sería labor del docente explicarle a su aprendiz que puede utilizar ciertas palabras en inglés en contextos profesionales y que puede también, por tanto, utilizarlas en sus exámenes de español. Al evaluador le correspondería no penalizar la aparición de esos términos en las producciones orales y escritas.

En segundo lugar, hemos contado con la ayuda del otro protagonista, el profesor y evaluador experto de español LE/L2. En concreto, hemos contado con 26 profesores/ evaluadores para responder un pequeño cuestionario con las siguientes preguntas:

> Cuando calificas o corriges una producción escrita de un estudiante de tu clase o de un candidato de un examen de certificación lingüística y encuentras en el texto un extranjerismo, ¿qué haces? ¿Corriges y escribes el equivalente en español, si es un texto de clase? ¿Lo penalizas si aparece en un examen?, o, dependiendo del extranjerismo que sea, ¿lo dejas o lo corriges/penalizas?

La segunda parte de la encuesta mostraba diez anglicismos en la que los profesores/ evaluadores debían señalar cuáles pensaban que estaban admitidos por la RAE y cuáles no, con el fin de averiguar si, realmente, el hecho de estar aceptado o no por la RAE puede suponer una penalización al uso de anglicismos por parte de los candidatos si los docentes y evaluadores conocen los términos aceptados. Los anglicismos aceptados por la RAE eran los siguientes: *feedback*, *manager*, *stock*, *freelance*, *wifi*, *manager,* y los no aceptados, *online*, *newsletter*, *banner*, *community manager* y *partner*.

Estudiantes y candidatos frente al anglicismo

Si bien el número de muestras recogidas no es muy amplio, nos sirve para iniciar un tema de estudio para próximas investigaciones que nos permita ampliar las conclusiones iniciales a las que podemos llegar en esta primera incursión en el tema. El cuestionario estaba dividido en dos partes. La primera parte contenía anglicismos frecuentes en la clase de español con fines específicos (de ahora en adelante, EFE) admitidos por la RAE: *blister, blog, feedback, freelance, hacker, handicap, holding, leasing, lobby, manager, marketing, ranking, slogan, spam, sponsor, spot, stock* y *wifi*. La segunda parte contenía los siguientes anglicismos no admitidos por la RAE, pero también de frecuente aparición en la clase de negocios: *banner, benchmarking, cash, CEO, Community manager, coworking, crowdfounding, email, Know how, networking, newsletter, online, outsourcing, packing, partner, report, retail, start up y training.*

Una vez pasados los cuestionarios, hemos comprobado si los estudiantes conocían los términos en inglés. Fuera del estudio de datos recogidos, hemos dejado, obviamente, a los estudiantes anglófonos, porque su conocimiento de estos términos era prácticamente total, aunque, hemos de señalar que términos como *crowdfunding* resultaban desconocido para 11 de los 19 encuestados, y el término *outsourcing* para 6 de todo ese grupo de lengua materna inglesa.

En cuanto a los otros términos, entre los 34 estudiantes no nativos de lengua inglesa, tendríamos que destacar que, de los términos admitidos por la RAE, tan solo *blister* es desconocido para muchos de ellos, en concreto, para 19 de los 34. De los términos no aceptados por la RAE, entre los no nativos de inglés, 12 de los 34 desconocen los términos *benchmarking*, 13 *community manager*, 14 *coworking*,14 *outsourcing* y 6 *start up*. Como primera conclusión, podríamos decir que este resultado puede deberse a que son términos muy nuevos en el mundo de los negocios. Su conocimiento no está tan asentado entre los aprendices de ELE/L2, aun estando aprendiendo español para fines específicos. Respecto a los datos que extraemos sobre su conocimiento del correlato, son, por decirlo de alguna manera, un tanto decepcionantes: tan solo encontramos tres términos con cierto conocimiento por parte de los estudiantes: *cash* (efectivo) (25), *email* (correo electrónico) (33), *slogan* (eslogan) (19) y *online* (en línea) (17). No deja de sorprendernos que conozcan los correlatos de algunos de los anglicismos más asentados en nuestra lengua, cuando los propios nativos emplean con más frecuencia la forma inglesa que la forma española, como puede ser el caso de *online* (García, 2021).

Docentes y evaluadores frente al anglicismo

Como ya dijimos anteriormente, 26 expertos en docencia y evaluación de español LE/L2 respondieron a un breve cuestionario que tenía como finalidad, por un lado, conocer su actitud como docente y evaluador ante los términos en inglés, ya que esto nos permitiría poder extraer como conclusión si muestran cierta severidad o benevolencia en el aula o a la hora de calificar una producción.

De las respuestas recogidas, podemos señalar que 11 respondieron que, cuando aparece en clase un extranjerismo, lo corrigen y escriben el equivalente en español, lo cual puede mostrarnos cierta "intolerancia" al término en inglés. Sin embargo, solo 6 lo penalizarían si lo encontraran en un examen. La mayoría señaló que dependía del extranjerismo, es decir, podía penalizarlo o corregirlo dependiendo del término, lo cual nos hace pensar que los profesores y evaluadores son más permisivos con términos que ocupan un espacio consolidado en el uso de nuestra lengua, por ejemplo, si están aceptados por la RAE y, en base a esto, toman una decisión. Algunos hicieron alguna anotación al margen para justificar su calificación o corrección dependiendo del nivel del estudiante o candidato.

Por otro lado, les pedimos que señalaran cuáles de los siguientes términos: *feedback*, *manager*, *stock*, *freelance*, *wifi*, *online*, *newsletter*, *banner* y *community manager*, están aceptados por la RAE para poder comprobar si su conocimiento o desconocimiento de la aceptación del término por parte de la RAE puede condicionar la permisividad ante el término. Los resultados mostraron que, del grupo de docentes evaluadores, 19 señalaron *manager*, 24 *wifi*, 17 *stock* y 12 *freelance*. Sin embargo, de las palabras no aceptadas, 17 señalaron *online* como aceptada. Recordemos que, para este término, su correlato era de los más conocidos por los estudiantes, pero entendemos que, por lo tanto, el término en inglés, si se encontrara en una producción, no sería penalizado y, quizá, tampoco se le enseñaría su correlato. El hecho de que no haya unanimidad por parte de los profesores ya es un dato a tener en cuenta.

Conclusiones

Nuestra realidad ha cambiado notablemente en los últimos 20 años. Los movimientos migratorios, los programas de movilidad, los diferentes cambios sociales y las consecuentes nuevas realidades han propiciado la proliferación de nuevos trabajos y la revisión de los ya existentes en el mundo del aprendizaje, enseñanza y evaluación las de lenguas. Estos cambios han afectado, muy especialmente, a la metodología de la enseñanza y a su evaluación, y les han dado protagonismo a conceptos como el plurilingüismo o la mediación. Así, en el *Companium volume with new descriptors*, (Consejo de Europa, 2020, p. 162), estas dos cuestiones, el plurilingüismo y la mediación lingüística, se presentan como las herramientas que permiten a los aprendices desenvolverse en las situaciones comunicativas que se les presentan en esos nuevos contextos sociales. A la vista de los resultados que hemos obtenido en nuestras encuestas, a pesar de ser tan solo una pequeña cala de estudio, podemos concluir que los alumnos conocen pocos correlatos, ya estén aceptados por la RAE o no. Por lo tanto, debería permitirse a los aprendices que tienen, en muchos casos, un perfil plurilingüe, usar sus repertorios lingüísticos. Deben emplear su conocimiento de otras lenguas para buscar diferencias y similitudes y sacar provecho, de este modo, a ese andamiaje creado a partir de su repertorio plurilingüe, sin pensar si el término está aceptado o no en español. Por eso, creemos que el candidato podrá hacer uso del anglicismo como estrategia para comunicarse eficazmente porque, en la mayoría de los casos, será un saber compartido que le permitirá llegar a donde quiere llegar.

Por otro lado, hemos comprobado que el docente y evaluador no conoce bien qué anglicismos acepta la RAE y cuáles no. Sin atender a esa distinción, tiende a enseñar en clase los correlatos en español, pero acepta los anglicismos en los exámenes porque es permeable al uso y sabe que la competencia plurilingüe no es estática. Como consecuencia de esta permisividad, primará la estrategia del candidato sobre la norma, porque se encontrará en los textos de los aprendices muestras reales de lengua, de ahí que podamos concluir que los anglicismos comunes se aceptan en la evaluación, ya estén reconocidos o no por la RAE, y a pesar de cierta resistencia de los profesionales de la enseñanza de español a su empleo generalizado.

Por último, al igual que sucede con el español y el inglés, habría que tener en cuenta en la certificación lingüística de otras lenguas posibles extranjerismos que, provenientes de otras lenguas, se han hecho un hueco en el uso diario y se encuentran implantados en la lengua materna del candidato, aunque no estén aceptados por las academias de las lenguas correspondientes.

Bibliografía

Calvi, M. V. (2013). El léxico de la enseñanza de ELE con fines específicos. En J. Gómez de Enterría (Coord.), *V Jornada-Coloquio de la Asociación Española de Terminología (AETER)*. Recuperado de: cvc.cervantes.es/lengua/aeter/conferencias/calvi.htm

Consejo de Europa. (2020). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Companion Volume with new Descriptors.* Strasbourg: Council of Europe.

Echazú, E., y Rodríguez, R. (2018). *Primer glosario de comunicación estratégica en español.* Recuperado de: fundeu.es/wp-content/uploads/2018/02/Glosario-de-Comunicaci%C3%B3n-Estrat%C3%A9gica-Fund%C3%A9u.pdf

García, M. (2021). La incorporación del anglicismo en la enseñanza del español de los negocios. En M. Saracho-Arnáiz y H. Otero-Doval (Eds.), *Internacionalización y enseñanza del español como LE/L2: plurilingüismo y comunicación intercultural* (pp. 999-1,012). Oporto: ASELE.

Enhancing bilingual education and CLIL in Italy: The impact of Cambridge International upper-secondary school programmes on the Italian curriculum[1]

Letizia Cinganotto
INDIRE, Italy

Juliet Wilson
Cambridge International, United Kingdom

Alessandra Varriale
Cambridge International, United Kingdom

Abstract

The paper will present the general outline and main aims of a research project promoted by Cambridge International in cooperation with INDIRE (Italy) and endorsed by the Italian Ministry of Education, aimed at investigating a wide range of aspects related to Cambridge International upper-secondary (ages 14 to 19) programmes.

The aim of the research project is to collect information about the impact of Cambridge International upper-secondary programmes on a wide range of dimensions, such as teachers' professional development and needs; students' learning outcomes and needs; school leaders' perceptions and reactions; organisational issues; recognition and appreciation of the programme by the Ministry, universities and the world of work; and long-term plans and perspectives. The research focuses on the impact of upper-secondary school programmes on the Italian school curriculum, as a tool for supporting bilingual education and Content and Language Integrated Learning (CLIL) methodology, made compulsory in upper-secondary school by Decree 88/89 dated 2010. The research uses multi-methods including an online survey to principals, teachers and students and focus groups with the various actors involved.

The following research questions will guide the project:

1. How do teachers in Italy decide which International General Certificate of Secondary Education (IGCSE), Advanced Subsidiary (AS) Level and Advanced (A) Level subjects to implement and why?

2. How do Cambridge International upper-secondary programmes (IGCSE, AS and A Level) help Italian students to perform in their national curriculum?

3. How does studying for Cambridge International qualifications help develop the English language proficiency of Italian students?

4. How can Cambridge International help school leaders in Italy understand the added value of Cambridge International programmes of study and assessment?

Due to the COVID-19 pandemic, the research had to be rescheduled, that is why in this contribution only the general outline, methods and expected outputs of the project will be highlighted.

Introduction

In recent years more and more Italian schools are adopting international programmes, with the aim of fostering the international dimension, which has become so important for the development of the 21st century skills of the students and for their global

[1] The authors wish to thank the other members of the research group: Fausto Benedetti, Patrizia Garista (INDIRE, Italy), and Stuart Shaw and Carla Pastorino (Cambridge International).

mobility and employability. Among the most popular international programmes adopted by the schools, Cambridge International Advanced Subsidiary Levels (AS Levels), Cambridge International Advanced Levels (A Levels) and IGCSE programmes are attracting an increasing number of schools.

Cambridge International Advanced Subsidiary Levels (AS Levels) and Cambridge International Advanced Levels (A Levels) are subject-based qualifications usually taken in the final two years of high school. Cambridge International AS Level is usually a one-year programme of study, while Cambridge International A Level typically takes two years. Assessment takes place at the end of each programme. Most subjects can be started as a Cambridge International AS Level and extended to a Cambridge International A Level. The syllabuses are international in outlook but retain local relevance. They have been created specifically for an international student body with content to match the needs of a wide variety of schools and to avoid cultural bias and intercultural misunderstanding.

The Cambridge International IGCSE is the world's most popular international qualification for students aged 14 to 16 years. It inspires students to love learning, helping them discover new abilities and a wider world. It is offered in a wide range of subjects and the syllabuses are flexible and easily adaptable with other curricula.

IGCSE for CLIL in Italy

In Italy, the Cambridge International IGCSE is the choice of over 350 state schools which have an international section. Teachers are trained by Cambridge International to teach the IGCSE programmes in a variety of subjects alongside the national curriculum with different methodological approaches. In 2019, over 25,000 Italian students aged between 16 and 17 years took at least one Cambridge International IGCSE qualification including an IGCSE in English as first or second language. Since 2010, the Cambridge International IGCSE programmes and qualifications have successfully helped Italian teachers to prepare students for the mandatory ministerial Content and Language Integrated Learning (CLIL) exam at the end of upper-secondary school. At present, over 20 of the best Italian universities in Italy recognise the Cambridge International IGCSE in English and English as Second Language as proof of English language proficiency.

One of the reasons why the Cambridge International IGCSE programmes are so popular among Italian secondary schools is that they provide the teaching of a subject in English as a Foreign Language, choosing among a wide range of different subjects. Although the Cambridge International IGCSE subjects are typically developed, taught and assessed according to the British curriculum and standards, it can represent a very effective way to implement CLIL methodology (Coyle, Hood, & Marsh, 2010; Marsh, 2009).

CLIL is often used as an umbrella term, referring to a wide range of teaching strategies which all place the learner as the real protagonist of the learning pathway, such as teachers' professional development and needs; students' learning outcomes and needs; school leaders' perceptions and reactions; organizational issues; recognition and appreciation of the programme by the Ministry, universities and the world of work; and long-term plans and perspectives.

CLIL provision is currently expanding in Europe, as highlighted by the latest Eurydice report 'Keydata on teaching languages at school in Europe' (2017), considering the benefits of CLIL both for learners and teachers. Students can easily improve their linguistic competences, enhancing the delivery of subject content at the same time. From the teacher's perspective, the introduction of CLIL into the curriculum can bring an enormous improvement in teaching practices and can open up a whole new world of material and resources.

CLIL has been compulsory in Italy since 2010 in all upper-secondary schools (Cinganotto, 2016; 2018) and the Council Recommendation on a comprehensive approach to the teaching and learning of languages (2019) mentioned CLIL provision in Italy, highlighting the democratic and inclusive approach adopted by the Italian Ministry of Education, as 'CLIL is for all', which means that no admission test is required in the classes where CLIL is mandatory by law.

Considering the added value of CLIL into the school curriculum and the quality of IGCSE programmes, a lot of schools have chosen Cambridge International programmes with the aim to implement CLIL methodology, also taking advantage of the benefits of well-known international programmes accredited by a large number of universities, companies and other stakeholders.

The research carried out by the National Institute for Documentation, Innovation, Educational Research in Italy (INDIRE) in cooperation with Cambridge International, with the endorsement of the Italian Ministry of Education, has been designed within this framework, with the aim to find answers to the following research questions:

1. How do teachers in Italy decide which Cambridge International IGCSE, AS and A Level subjects to implement and why?

2. How do Cambridge International upper-secondary programmes (IGCSE, AS and A Level) help Italian students to perform in their national curriculum?

3. How does studying for Cambridge International qualifications help develop the English language proficiency of Italian students?

4. How can Cambridge International help school leaders in Italy understand the added value of Cambridge International programmes of study and assessment?

Outline of the research

Due to the COVID-19 pandemic, the timeline originally planned for the development of the project had to be rescheduled and postponed. Therefore, in the following paragraphs, only the general outline of the research will be highlighted, with reference to the methods and expected outputs.

Methods

The research is based on a multi-method approach, including interviews and focus groups with principals, teachers, students on a wide range of dimensions, such as teachers' professional development and needs; students' learning outcomes and needs; school leaders' perceptions and reactions; organizational issues; recognition and appreciation of the programme by the Ministry, universities and the world of work; and long-term plans and perspectives.

Data collection in this field requires the adoption of methods and techniques capable of restoring the subjectivity implicit in the concept of satisfaction. With this aim, a group of 'qualitative' tools that allow the researcher to correctly interpret the opinions expressed by users will be used for qualitative analysis. By using an accurate methodological approach, the researcher avoids the risk of imposing their own point of view on users.

Out of all the different qualitative methods, it has been decided to adopt the focus group method, which uses group dynamics to obtain data on a specific subject. Focus groups are generally considered to be a particularly effective technique in the field of educational and professional updating.

Preliminary interviews with a sample of school leaders, teachers and students will be arranged in order to gather information about critical and relevant issues to be investigated in depth through focus groups. This characterization of the thematic areas will turn out to be crucial for conducting the focus groups. The general outline of the interviews will be shared among the researchers, in order to agree on the most important areas to be discussed with the different informants. Interviews will be video-recorded and specific consent forms will be collected for this aim.

By using a 'theoretical' sampling approach, first the reference variables will be identified, and then a sample of school leaders, teachers and students will be selected, according to their availability, to be involved in focus groups.

The discussion will be planned as follows: opening phase (introductions, explanation of research aims, description of the process, ice-breaking and preliminary questions), central phase (focused on assessments of the main aspects of the programme), and final phase (conclusions of the group work, suggestions and recommendation for improvement).

The interview will be co-created by INDIRE researchers and Cambridge International researchers, on the basis of a particular method aimed at getting as close as possible to the participants' cognitive and representative world. This will be supported by the active role of the researcher-interviewer who will be able to choose to intervene in the discussion by using specific communication techniques for collecting information and supporting the flow of communication (open questions, focusing, classification, summaries, reformulations, echoes), thus contributing to collecting research material.

Data analysis

The data collected will be analyzed according to the grounded hermeneutic approach (Glaser, 2007; Strauss, & Corbin, 1990), which attempts to throw light on all the aspects that constitute and make daily events understandable, grounding itself on data from different types of communication (verbal, paraverbal, non-verbal).

Indeed, this approach underlines the hermeneutic dimension of the construction of reality by emphasizing some of its main aspects: individuals give meanings to their actions and these meanings are important in order to understand human behaviour. The term 'meaning' does not refer solely to what is verbalized but also to actions and practices (e.g. presence or active participation in a discussion).

Of course, the meaning obtained is never neutral but is the result of the subject's background: their environment, social structures, personal history, and experiences. A final element to bear in mind during data analysis and interpretation is that the meanings of actions are seldom fixed, clear and intelligible or limited to fixed categories. Instead they are negotiable and change over time, according to context and individual. For this reason, an interpretation of qualitative data is necessary.

The analysis of recordings will allow the identification of recurrent topics, by distinguishing individual opinions from what received the group's approval, starting from the interview outline and from the elements which emerged during the focus group.

Outputs and conclusions

A research report will collect and illustrate all the data gathered during the different sessions of the project. A section of the report will be devoted to the general recommendations and guidelines coming from the different stakeholders: school leaders, teachers, students. The report will be useful for highlighting the strengths and weaknesses of Cambridge International programmes in Italian schools, considering their possible added value to the school system. It will also be interesting to understand how the Cambridge International IGCSE programmes can support the international dimension and the implementation of the CLIL methodology in the Italian schools.

References

Cinganotto, L. (2016). CLIL in Italy: A general overview. *Latin American Journal of Content and Language Integrated Learning, 9* (2), 374-400.

Cinganotto L. (2018). *Apprendimento CLIL e interazione in classe*. Rome: Aracne editrice.

Coyle, D., Hood P., & Marsh D. (2010). *CLIL: Content and Language Integrated Learning*. Cambridge: Cambridge University Press.

Eurydice. (2017). *Key data on teaching languages at school in Europe*. Luxembourg: Publications Office of the European Union.

Glaser, B. G. (2007). *Doing Formal Grounded Theory: A Proposal*. Mill Valley: Sociology Press

Marsh, D. (2009). Foreword. In Y. R. de Zarobe & R. M. Catalán (Eds.), *Content and Language Integrated Learning – Evidence from Research in Europe* (pp. vii–viii). Clevedon: Multilingual Matters.

Strauss, A. L., & Corbin, J. (1990). *The Basics of Qualitative Analysis: Grounded Theory Procedures and Techniques*. Newbury Park: Sage.

Incorporating English and Spanish mediation tasks into language proficiency exams

Caroline Shackleton
Centro de Lenguas Modernas de la Universidad de Granada, Spain

Adolfo Sánchez Cuadrado
Centro de Lenguas Modernas de la Universidad de Granada, Spain

Nathan Paul Turner
Centro de Lenguas Modernas de la Universidad de Granada, Spain

Abstract

The recently published CEFR Companion Volume places greater emphasis on the language user as a social agent operating within a multilingual globalized world, where languages like English and Spanish are used in an increasingly *lingua franca* environment. Given the relevance of mediation to *lingua franca* scenarios, its inclusion in language proficiency tests should be seen as an opportunity to expand the test construct, with a view to creating positive washback through the inclusion of these real-world competencies. Nevertheless, given the complexity of constructs representing social uses of language, the inclusion of mediation tasks in language proficiency tests poses serious challenges. The present report outlines how the evidence-centered design approach to test development proposed by Mislevy, Steinberg and Almond (2002) can be drawn upon in order to provide a measurable theoretical construct for mediation.

Background

After two decades of a markedly assessment-oriented use of the 2001 Common European Framework of Reference for Languages (CEFR, Council of Europe, 2001) the expanded Companion Volume (CV, Council of Europe, 2020) aims at promoting its implementation within the fields of teaching and learning, the two other goals of the framework. One of the key areas focused on in the 2020 update is mediation, which remained rather unspecified in the previous document. In fact, the previously somewhat blurry conceptualization of this mode of communication has given way to a multi-faceted and complex construct, which sees an equation of language learning and language use ('user/learner'), together with a focus on the KSAs — knowledge, skills and abilities — needed to carry out real-life tasks ('acts'). The new document further places a particular emphasis on context, task fulfilment, the socially-oriented use of language and the self-regulation of a user as a social agent.

It is clear that such a high degree of complexity could be seen as an impediment to the implementation of this mode of communication. Prior attempts, such as the KPG exams in Greece (Dendrinos, 2006), mainly dealt with cross-linguistic textual mediation. Nevertheless, there is still a patent and growing interest in exploring the full potential of mediation (see, for instance, the EALTA CEFR SIG 2018 report). Furthermore, mediation has been introduced in some settings as part of governmental language-planning decisions, e.g. the Spanish government's inclusion of mediation as part of official accreditation exams in its state-funded language schools.

The inclusion of mediation in the language curriculum can help foster language learning in four major ways: (a) it establishes a purposeful and context-driven use of language; (b) it helps redefine the concept of language skill and brings it to a wider and overarching competence model (Barrett, 2016, p. 23) and to a macro-functional perspective, in conjunction with reception, production, and interaction; (c) it taps into an integrative view of the language learner, who not only needs to master their language competence but also has to deal with otherness, make use of appropriate intercultural and affective strategies, and uphold and foster democratic values; and (d) it gives grounding to the implementation of teaching methodologies such as content and language integrated learning (CLIL) and scenario-based learning, plurilingual education (the cross-linguistic dimension), and English as a *lingua franca* — widely used as a linguistic mediating tool in plurilingual contexts.

Nevertheless, mediation, with its inherent focus on the social dimension of language, its multi-modal and holistic approach to language use, and its focus on the co-construction of meaning with others, poses a real challenge to the development of reliable assessment instruments. While it certainly enriches the construct of language competence, its multi-faceted nature requires not only a stable and reliable measurement instrument but also a thorough definition of the traits to be measured.

Theoretical framework and task design

Evidence-centered design (ECD) offers a logical and systematic framework for designing performance assessments, and aims to include validity claims during the initial design and development stages, thereby making it particularly useful for measuring new constructs (Zieky, 2014). ECD is a methodology incorporating best practice procedures; it implements Messick's (1989) approach to measuring constructs but with a focus on test construction. In the first layers of ECD, *domain analysis* and *domain modelling* are undertaken in order to inform valid task creation for mediation. In the domain analysis layer, important knowledge, skills and abilities for the construct of mediation are identified and prioritized. The domain modelling layer then deals with the production of a list of the claims we wish to make about test-takers and to describe the necessary evidence or observable data on which the claims will be based. In this way, a chain of reasoning linking evidence to claims about target constructs is directly incorporated at the beginning of the assessment development process (Riconscente, Mislevy, & Corrigan, 2015). For our project, a number of small-scale pilot studies were carried out during this iterative process, which has been called the 'thinking stage' of ECD (McNamara and Roever, 2006, p. 23). Prototype tasks were developed and piloted, and insights from verbal reports and focus group discussions with both students and teachers were fed back into the task development cycle in order to make improvements on those features of the tasks which would better elicit the required evidence.

Our aim is to devise a series of possible task shells for inclusion in proficiency exams which will cover the construct in relevant authentic settings. Cleary, useful assessments should be context specific. The present study took place in the context of CEFR-related English and Spanish CertAcles proficiency exams provided by the University of Granada's modern language centre, where the majority of the test taker population are 18–25-year-olds taking the test for degree accreditation or access to mobility programmes. The task shell outlined in the present paper is a B2 written response task.

In terms of domain analysis, a rich description of the construct is necessary if we are to be clear about the required knowledge, skills and abilities (KSAs). Here, we looked directly to the CV in order to better define which KSAs were specifically relevant to mediation.

The CV (2020, p. 176) includes three different types of mediation among its big picture claims:

- cognitive mediation (facilitating access to knowledge, awareness and skills)
- interpersonal mediation/mediating communication (establishing and maintaining relationships; defining roles and conventions in order to enhance receptivity, avoid/resolve conflict and negotiate compromise)
- textual mediation (transmitting information and argument: clarifying, summarising, translating etc.).

Clearly, mediating communication is an extremely important and necessary skill in a *lingua franca* environment. Therefore, in order to create a context-specific prototype task shell, one specific area of mediating communication was chosen — that of facilitating communication in delicate situations and disagreement — and an initial list of relevant KSAs was drawn up and prioritised in order to provide a list of sub-claims important to this particular domain. These were, the:

- ability to take into account the intended audience
- ability to (help to) identify common ground
- ability to highlight possible areas of agreement
- ability to highlight possible obstacles to agreement
- ability to attenuate and use diplomatic language
- ability to give additional explanations if necessary
- ability to ask for repetition and clarification
- ability to summarise/paraphrase
- ability to discuss similarities and differences
- ability to pose neutral questions
- ability to highlight key information
- ability to get clarification on what is meant.

This list of abilities would then form the basis for the observable evidence considered acceptable for showing that a student has the desired KSAs. The task model outlines tasks which the student performs to elicit the desired evidence. Task features need to be decided in order to identify potential work products, which should be authentic and relevant to the student population. Again, we looked to the CV (2020, Appendix 6, p. 220), which provides potential situations for personal, public and educational domains (in our context the occupational domain is not relevant).

For the first task shell, we chose the personal domain, and looked to create a context relevant to our target population's experience and needs. To this end, we opted specifically for the mediation of disputes with landlords and tenants in the context of an Erasmus-style university flat share, regarding disagreements or misunderstanding over payments (e.g., how and when rent is due, how and when bills are paid, what services are available, etc.), damages to fittings and fixtures or property (whose fault, who should pay, how much deposit should be returned, etc.), or house rules, among others.

The first version of the task developed required the candidate to write to both landlord and tenant simultaneously in order to help resolve the following situation:

> The candidate receives a series of SMS messages from a flatmate about a broken water heater that the landlord refuses to pay for, as well as an email from the landlord threatening eviction.

However, on trialling the task with our first focus group, it was found that the task did not in fact elicit the authentic situation for mediation desired. The participants did not see writing to both parties at the same time as realistic, and there was an obvious bias in support of the tenant. Following Mislevy and Haertel (2006, p. 7), who state that, 'cycles of iteration and refinement both within and across layers are expected and appropriate', changes were made to the task in an attempt to make it more realistic and balanced. Subsequently, the second focus group provided much more positive feedback about the authenticity of the task, the final version of which can be seen at this link: bit.ly/30eFNw0

Results

The task was then trialled (N=7). 'Think alouds' were conducted at the macro-planning stage of the writing task, and transcribed and examined for salient themes.

Four major themes emerged from the data:

1. Role of the mediator

   Example: 'The first thing I thought about was to take into account everything that my flatmate had said on the one hand, and relate to what the landlord says . . . Try to contrast the two positions, erm, I want to try and create, I don't know – links, bridges between what each person is saying, give an explanation or try to soften the message at least.'

2. Positioning

   Example: 'At the beginning I was completely on my friend's side because if the water heater has broken and the landlord won't mend it – of course the landlord isn't right. But then I started thinking about . . . he hasn't paid his rent, so the landlord has a right to be angry . . . and after reading the landlord's email . . . we pay late and if this wasn't the case, maybe I wouldn't understand the landlord, and I would want to support my friend more.'

3. Drawing on prior knowledge

   Example: 'Yes, this is a real situation; I don't know what the law says here, but appliances in the house are the landlord's responsibility, by law . . . in England everything that is in the house . . . If a tenant breaks a chair, for example, then no because it's his fault . . . but appliances, the fridge, the cooker and the water heater are the responsibility of the landlord.'

4. Additional explanations/justification

   Example: 'There are a lot of things to do. I'm thinking about some kind of justification I can give and I haven't got the information about the contract so I'm going to invent it . . . Maybe say that we have a guarantee that covers general expenses.'

The task was then piloted with a group of volunteers (N =24) with small groups of L2 English (10 students) and Spanish learners (14 students) from high B1 to low C1 (mostly B2), and the scripts were analysed using the checklist of KSAs in order to determine whether the required observable evidence was indeed elicited by the task. The results can be seen in Table 1, together with a few examples taken from both the responses. Here, the KSAs shown in uppercase were added after examining the scripts, as it was felt that they were pertinent to the successful completion of the task. It can be seen that, with the exception of the KSA 'discuss similarities and differences', all other KSAs were elicited to a greater or lesser degree by the task.

A few examples representing the observable evidence taken from the written responses are shown below:

> Facilitate continued communication:
>
> 'Maybe we can meet and talk about the situation. We will explain our reasons and problems and you do the same.'
>
> Attenuate and use diplomatic language:
>
> 'I have to say that I am really sorry for this situation. I understand that you might be a bit irritated because of my friend's way of answering.'
>
> Give additional explanations if necessary:
>
> 'He is under a big pressure because of his recent divorce and many times he is overacting.'
>
> Highlight possible areas of agreement/solutions (suggest a compromise):
>
> '. . . suggest signing a new contract which would be satisfactory for both parts or it might be convenient to pay a secure company between us, detailing in the new contract.'
>
> Summarise/paraphrase:
>
> 'I would agree that the language he has used is unacceptable and I also understand he has been late with his part of the rent a few times and that is also disrespectful towards our rent agreement.'
>
> Explain cultural references:

'It might be a cross cultural communication problem. She might seem rude, but she is only direct (it's her culture).'

Table 1: Frequency of KSAs used in written responses

KSA	*Total/24*
Take into account the intended audience	23
FACILITATE (CONTINUED) COMMUNICATION	21
Attenuate and use diplomatic language	21
Give additional explanations if necessary	21
Highlight possible areas of agreement/ SOLUTIONS (SUGGEST A COMPROMISE)	17
Highlight key information	17
Summarise/paraphrase	10
Pose neutral questions/MAKE REQUESTS	10
Highlight possible obstacles to agreement	8
Get clarification on what is meant	7
Identify common ground	3
Explain cultural references	2
Discuss similarities and differences	0

Conclusion

It has been claimed that current proficiency tests are rapidly becoming obsolete in our current, globalized world (McNamara & Shahomy, 2016), because they fail to represent how languages are actually used in contemporary society (Seidlhofer, 2011). This study represents a sincere attempt to address such deficits; in addressing the challenge of expanding our test construct, we believe it to be very much in line with the spirit of the CV, and its goal of broadening the scope of language education. Indeed, a keen awareness of the powerful washback effects of proficiency tests led us to adopt a systematic approach, with the aim of producing a series of task shells which would cover an appropriate range of relevant *lingua franca* mediation scenarios that could later be sampled for each exam administration. The inclusion of this type of task, together with an adequate sampling of the construct, obliges those teachers who would teach to the test to teach to the construct itself, and as such should encourage the acquisition of real-world abilities, thereby leading to positive washback on teaching and learning (Shackleton, 2021). As the overarching premise of this project was to encourage positive washback, it was decided that observable KSAs should form the basis of the assessment rubric. As such, we were particularly concerned not to merely pay lip service to mediation by simply including it in the assessment scales as task achievement. As part of the next stage of this process, we are currently engaged in the validation of an analytical rating scale which places mediation centre stage. It is our hope that by following such a systematic approach, it will be possible to produce mediation tasks with an underlying measurable construct, thereby allowing for the development of an evidence-based assessment rubric.

References

Barrett, M. (2016). *Competences for Democratic Culture. Living Together as Equals in Culturally Diverse Democratic Societies*. Strasbourg: Council of Europe.

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

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

Dendrinos, B. (2006). Mediation in communication, language teaching and testing. *Journal of Applied Linguistics*, *22*, 9–35.

EALTA (2018). *The CEFR Companion Volume with New Descriptors: Uses and Implications for Language Testing and Assessment. Report on the VIth EALTA CEFR SIG*. Bristol: Multilingual Matters.

McNamara, T. & Roever, C. (2006). *Language Testing: The Social Dimension*. Malden: Blackwell Publishing.

McNamara, T. & Shohamy, E. (2016). Language testing and ELF: Making the connection. *English as a Lingua Franca: Perspectives and Prospects* (pp. 227–234). Berlin: De Gruyter.

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

Mislevy, R., L., & Haertel, G. D. (2006). Implications of evidence-centered design for educational testing. *Educational Measurement: Issues and Practice*, *25*, 6–20.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2002). Design and analysis in task-based language assessment. *Language Testing*, *19*(4), 477–496.

Riconscente, M., Mislevy, R., L., & Corrigan, S. (2015). Evidence-centered design. In S. Lane, M. R. Raymond, & T. M. Haladyna (Eds.), *Handbook of Test Development* (pp. 40–63). Abingdon: Routledge.

Shackleton, C. (2021). Planning for positive washback. In S. Hidri (Ed.), *Perspectives on Language Assessment Literacy: Challenges for Improved Student Learning* (pp. 220– 240). Abingdon: Routledge.

Seidlhofer, B. (2011). *Understanding English as a Lingua Franca.* Oxford: Oxford University Press.

Zieky, M., J. (2014). An introduction to the use of evidence-centered design in test development. *Psicología Educativa*, *20*(2), 79–87.

Lost in translation: Translatability of the CEFR-J based English tests

Masashi Negishi
Tokyo University of Foreign Studies, Japan

Abstract

In 2018, Tokyo University of Foreign Studies (TUFS) started a project called the CEFR-J x 28, in which we attempt to align our language programmes with the CEFR-J, a modified version of the original CEFR to support the learning, teaching and assessment of English in Japan. This study aims to investigate the translatability of the CEFR-J based English tests to other languages taught at TUFS. In order to streamline the test production process, reading and writing tests of English were initially translated into other languages with Google Translate, and then the results were checked and revised by native speakers of those languages. Most of the tests needed modification after the initial machine translation. Features specific to the language are ingrained in the test, and many of these are not automatically translatable. While the approach was quite useful for translation of texts in the tests, some of the test items, especially the writing test items, required modification regarding the settings of the tasks.

Introduction

The CEFR-J was originally developed to apply the Common European Framework of Reference for Languages (CEFR, Council of Europe, 2001) to English language education in Japan (Negishi, Takada, & Tono 2013; Negishi & Tono, 2016). The CEFR describes language proficiency using Can Do descriptors and divides language proficiency into six levels: A1, A2, B1, B2, C1, and C2, whereas the CEFR-J describes 12 levels: Pre-A1, A1.1, A1.2, A1.3, A2.1, A2.2, B1.1, B1.2, B2.1, B2.2, C1, and C2, branching the lower levels of the CEFR. The first version of the CEFR-J Can Do descriptors was published in 2012, and thereafter related research has generated such tools as the 'CEFR-J Wordlist', 'CEFR-J Grammar Profile', 'CEFR-J Text Profile', 'CEFR-J based tests'[1], etc.

Tokyo University of Foreign Studies (TUFS) has been teaching 28 languages as major languages. However, the teaching and assessment of these languages has traditionally been conducted independently, and the achievement level of each language could not be compared. In order to sort out this problem, standardization of teaching and assessment using the CEFR-J has been underway since 2018. This project is called the CEFR-J x 28. So far, three types of e-learning resources were created, i.e. (1) a flash-card app for learning vocabulary, (2) an online syntax writing tool for the study of grammar and vocabulary, and (3) an online spoken and written production corpus collection tool (Tono, 2019). Tono (2019, p. 16) states: 'The evaluation of our multilingual resource development based on the CEFR-J is yet to be seen, but the approach taken by the CEFR-J x 28 project is moving in a promising direction in that resource-rich languages such as English could give support to under-resourced languages in terms of language teaching and learning content and methods'.

Research

Assessment tools used

In this study, the reading and writing tests aligned with CEFR-J Can Do descriptors were used. The reading test consists of 12 sets of test items that correspond to the Can Do descriptors of CEFR-J Pre-A1, A1.1, A1.2, A1.3, A2.1, and A2.2 with two sets of items for each level. The writing test consists of 16 sets of test items that correspond to the Can Do descriptors of CEFR-J Pre-A1, A1.1, A1.2, A1.3, A2.1, A2.2, B1.1, and B1.2 also with two sets of test items for each level. These levels were selected because the CEFR-J project had already developed tests that corresponded to these levels, and also because most of the non-English language learners at TUFS learn their major language from scratch.

[1] The CEFR-J based tests are available at www.cefr-j.org/download.html#cefrj_testasks

Research methods

In this study, we proposed translating existing CEFR-J based English tests into other languages and tested the feasibility of this approach. Specifically, CEFR-J-based English reading tests and writing tests were translated into Japanese, Chinese, Thai, Filipino, Indonesian, Malay, Myanmar, Mongolian, Russian, French, Spanish, Portuguese and Brazilian Portuguese, using Google Translate, and the translation results were subsequently checked and corrected by native speakers. As regards the translation of short texts, there is not much difference between employing Google Translate and employing native speakers from the beginning, but the longer the texts, the more benefits can be gained by using Google Translate, such as saving time, money, and energy. There are a number of reasons why Google Translate was chosen for this study: (1) since Google Translate is free of charge, anyone can use it; (2) the result of the translation is generally of quite good quality; and (3) Google Translate had already played a pivotal role in the production of some of the resources in the CEFR-J x 28 project (Tono, 2019). The translation by Google Translate was conducted between June 2019 and January 2021. The native speakers of those languages in this project were TUFS graduate students in linguistics and applied linguistics.

Examination of translation results

Google Translate results

Google Translate helped us at least get a rough idea of the test. However, the results of the translation, some of which were judged as inappropriate by native speakers, could not be used as they were, so they needed some modifications. One such modification is to do with a T-V distinction as in French, Spanish, and Portuguese. These languages have second-person pronouns that distinguish varying degrees of politeness, social distance, and familiarity toward the addressee. For example, when you talk to someone in French, you use *tu* if you have an intimate, amicable, and/or equal relationship with the person, and you use *vous* if you have a respectful and/or distant relationship with the person. Since Modern English does not have a T–V distinction, with the exception of a few dialects, translators need to make a choice with T and V in translating texts from English to languages with a T-V distinction. When Google Translate translates English texts to those languages, it seems that the translation tool is likely to select V forms in the target languages at the time of writing. However, in some cases, the selections were not appropriate in the given contexts.

In a similar vein, when an English newspaper was translated into Japanese, Google Translate translated it with polite verb-endings (*-desu, -deshita,* and *masu*) but it would be common to use neutral ones (*-da, -datta,* and *-dearu*) in Japanese newspaper articles. Therefore, the verb endings also needed to be changed.

Reading tests

Problems that turned up in the translation of the reading texts and the reading tasks are as follows:

1. Proper nouns

   Proper nouns, such as people's names and place names in English texts, naturally became inappropriate for the texts of the target language. People's names in English may be replaced with typical names in each language. However, the original names may have been chosen with special intentions or meanings, and therefore names should be carefully chosen.

2. Picture books

   A nursery rhyme from 'Mother Goose' was chosen to represent 'a picture book that is already familiar through oral activities' in the Can Do descriptor of Pre-A1. A typical 'picture book' in the target language must be selected, but it may be difficult to decide on one typical book.

3. Uppercase or lowercase letters

   The task of looking for letters in the map, which is based on a Pre-A1 Can Do descriptor 'I can recognise upper- and lowercase letters printed in block type.', did not apply to languages without uppercase/lowercase distinction e.g. Japanese and Chinese, and did not function for such languages.

4. Language level and language selection for a specific text type

   The degrees of translatability of the task of reading a sign at the entrance of a café (A1.1) were different from language to language. Although the English sign was translated to many of the languages without any problems, there were reports that the language difficulty level of such signs is considerably higher in Chinese. It was pointed out that in the Philippines, English would be used instead of Filipino for such a sign in a café, and also that texts such as online texts for summer camps in A2.1 would be written not in Filipino but in English.

5. Objects in the reading texts

   There is a reference to apple trees in the garden in an A1.2 reading test, which was not a problem for Japanese, Chinese, Russian, French, etc. but it was reported that it is not common for apple trees to be found in the garden in Thailand, Malaysia, Indonesia and the Philippines, so mango trees were proposed. Therefore, this kind of fact checking is always essential for test translation.

Writing tests

Since writing tests do not necessarily include texts in the target language, many of the key problems were in the settings of writing tasks. Problems that turned up in the translation of the writing tests are as follows:

1. Purposes and settings of the communication

   The CEFR-J based tests, adopting the action-oriented approach, specify the purposes and settings of the communication and therefore they needed to be modified in the translation of most of the writing test items. One such example is 'You will study at an English language school for a short period of time' (A1.1) would have to be modified as a Japanese test to say: 'You will study at a Japanese language school for a short period of time.'

2. Writing systems

   Test takers of a Pre-A1 writing item are required to write down words as they listen to the spellings of the words. This item was successfully translated into languages with a phonemic writing system e.g. Filipino and French. However, it did not function with languages using logograms such as Chinese characters.

3. Style

   Greetings of letters and e-mails are different from language to language. A typical English greeting is 'Dear . . .', for which corresponding translations are easily found in such languages as French, Spanish, Portuguese and Russian. However, the direct translation of this English greeting is not common in some languages, e.g. Japanese and Indonesian.

4. Language selection

   In Filipino, it was pointed out that, as in reading, texts above a certain level are not written in Filipino. Accordingly, the higher the CEFR-J level, the more unnatural it is to perform the writing tasks in Filipino. However, it was pointed out that Filipino would be chosen for online interaction on a social networking site (Council of Europe, 2020).

Conclusion

The results of Google Translate are not perfect at the time of writing but are considered to be sufficient to streamline the process of test translation. However, the results must always be checked by experts and/or native speakers of the language and the necessary modifications must be made. In reading tests, this approach was quite useful for translation of texts. On the other hand, in the case of writing tests it was necessary to change the settings of the tasks. Furthermore, in this study, we attempted to translate the tests assuming Japanese learners were taking the tests of non-English languages. However, if you translate the tests for speakers of the other languages, further modifications may be necessary; at least test instructions need to be translated into the language the target test-takers are capable of understanding, e.g. their L1.

The CEFR-J framework and various tools based on it have been developed with the aim of adapting the CEFR to English language education in Japan. Many of the challenges identified in the CEFR-J-based test translation may arise from differences between the languages and between the societies in which the languages are used.

There are several advantages to developing multilingual tests in this way. One is to help develop tests for languages that do not have a standardised proficiency test. Another is to relate these test results to CEFR-J (CEFR) levels.

This project has just begun, and there are several challenges. At this moment, the scope of this project is narrow in a number of respects. Although translated tests are available for Japanese, Chinese, Thai, Filipino, Indonesian, Malay, Myanmar, Mongolian, Russian, French, Spanish, Portuguese and Brazilian Portuguese, it is necessary to expand the range of the languages. Also, levels are limited to the lower levels of the CEFR-J. If translation is done to test items of upper levels, other issues may arise. For example, the questions, keys, and options of the current multiple-choice B1.1–B2.2 reading test items are all written in English, and therefore the relationships of those elements and the relevant parts in the reading texts need to be carefully examined. Also, it may be difficult to find native speakers of the target language who are capable of understanding higher-level English texts. In addition, we have translated the reading and writing tests, but we need to see what will happen to the translation of the listening and speaking tests. In addition to this expansion of coverage, it may be necessary to check the difficulty of the test items that were created. This is because the difficulty of a test item in one language may not be maintained in another. Finally, we asked just one native speaker to review the results of Google Translate, but experience tells us there is a great deal of variability of the judgements of translation, so we need to ask multiple translators who, if possible, are familiar with language teaching and assessment, and to give them opportunities to discuss the validity of the translation. Despite some of these challenges, however, this attempt might provide a new methodology for the development of multilingual tests.

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 with New Descriptors*. Strasbourg: Council of Europe.

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 2001* (pp. 135–163). Studies in Language Testing volume 36. Cambridge: UCLES/Cambridge University Press.

Negishi, M., & Tono, Y. (2016). An update on the CEFR-J project and its impact on English language education in Japan. In C. Docherty & F. Barker (Eds.), *Language Assessment for Multilingualism, Proceedings of the ALTE Paris Conference, April 2014* (pp. 113–133). Studies in Language Testing volume 44. Cambridge: UCLES/Cambridge University Press.

Tono, Y. (2019). Coming full circle – from CEFR to CEFR-J and back. *CEFR Journal – Research and Practice*, 2–17.

Assessing the visual communication and Flemish Sign Language development in deaf two-year-olds

Beatrijs Wille
Ghent University, Belgium

Abstract

This paper highlights the international literature and the modality-specific reasoning behind the adaptation procedure of the recently developed Visual Communication and Flemish Sign Language checklist. The original Visual Communication and Sign Language (VCSL) Checklist originated in the USA has been subjected to a detailed adaptation process to fit the Flemish context (Belgium). The linguistic considerations and adjustments are mainly based on the internal complexity of certain linguistic phenomena, such as negation and interrogation, and missing communicative and linguistic milestones. This adaptation fills a distinct gap in the current field of sign language assessments and will be part of a new multimodal assessment to measure young deaf children's communication and language progress in Flanders, independent of chosen language(s).

Visual communication

The key element of early language development is the accessibility of parents' communication and language output or in other words the communication and language perception of children. Parents' output can be actively and passively perceived through hearing and sight by their hearing children. However, a focus on parental output and children's input is even more relevant to the situation of deaf children as they need to actively perceive meaningful communication – initially – mainly through sight.

Overall, research has shown that the acquisition of early visual communication and sign language skills – even prior to cochlear implantation – is beneficial for further language development and it facilitates the development of a second/spoken language (Mitchiner, Nussbaum, & Scott, 2012). Thus, the visual communication learning process takes time and can sometimes - if not acquired smoothly - hinder the acquisition of any language; spoken or signed. The strong visual character of deaf children's communication and joint attention plays a very important role in early interactions between parent and deaf child, such as gaze direction and sustained attention. About 95% of deaf children have hearing parents who have no or hardly any experience with the visual communication needs of their deaf infants (Mitchell & Karchmer, 2004; van den Bogaerde, 2000). Hence, the minority of deaf children has deaf parents. To capture the attention of their deaf children, deaf parents have the cultural and visual knowledge to implement child-directed communication in their daily interactions. The highly visual nature of early interactions with deaf children can be illustrated by a recent American study showing that deaf children of deaf parents score significantly higher in following their parents' gaze than their hearing peers (Brooks, Singleton & Meltzoff, 2019). Furthermore, child-directed language plays an important role in these early interactions and visual training (Lieberman & Mayberry, 2015; Lieberman, Hatrak, & Mayberry, 2014; Masataka, 2000). This less complex use of language facilitates language acquisition and may play a positive affective role in children's early social and emotional development (Masataka, 2000). An increase in language accessibility — regardless of the language offered — can be achieved through communication strategies; also called attention strategies (Lieberman & Mayberry, 2015; Wille, Van Lierde, & Van Herreweghe, 2019).

Joint attention — mentioned above as one of the essential conditions to optimize and enrich early language input for deaf children — refers to the shared focus between the child, the parent (i.e., the linguistic input) and an object or event (i.e., the non-linguistic input). 'Joint attention serves as a foundation for developing communicative competence and is one basis for the development of early social and cognitive skills' (Lieberman, 2012, p. 2). In addition to social and cognitive influence, joint attention is also seen as an important predictor of later language skills (Butterworth, 1991; Harris & Chasin, 2005). The development of joint attention can be divided into several phases, from attention labeled as being disengaged and passive to more coordinated and full triadic joint attention interactions whereby the three elements – i.e. child, parent, and object or event – actively alternate (Bakeman & Adamson, 1984). '[. . .] In order to acquire signs, they [sign-exposed children] need to see both a sign and a contingent nonverbal context that will serve to elucidate its meaning' (Harris & Chasin, 2005, p.1,116). For deaf children, the linguistic and non-linguistic information is (initially) perceived mainly visually. These children must actively learn to perceive and link two visual stimuli. This sequential learning is a fundamental difference from the usual information processing in hearing children. They integrate auditory and visual sensory perceptions, that is a cross-modal integration where no sensory collision occurs when perceiving linguistic and non-linguistic information (Piaget, 1952). To promote sequential learning and joint attention, parents can implement communication strategies in their interactions with the deaf child. Recent research in Flanders has shown that deaf parents can demand and sustain attention more quickly and for a longer period of time in interaction with their deaf child, meaning that deaf parents have relevant modality and culture-related knowledge that can support hearing parents in developing child-centered interactions (Wille et al., 2019). Furthermore, deaf fathers displayed more implicit strategies and preferred a wait-and-see attitude compared to a more active controlling attitude found in the participating mothers, which is consistent with interactional research within hearing families (Barton & Tomasello, 1994; Wille et al., 2019).

It can be stated that child-directed communication, joint attention, and parental communication strategies contribute to an optimal communication and language input. This can then change the quality of the child's language development and may result in a richer communication; both for spoken and sign language acquisition.

Thus far, however, we have been unable to systematically assess deaf children's early communicative development in Flanders (Belgium). Thus, recent efforts have led to the development of a combined checklist which allows assessment of both the early visual communication (VC) and the Flemish Sign Language (VGT[1]) skills of young deaf sign-exposed children during their first two years.

Worldwide, only a few reports have been made on assessments of sign language acquisition in deaf children this young (Haug & Mann, 2008). The current checklist builds on an intra-modality adaptation that offers many modality-specific advantages, such as the presence of items measuring children's attention to visual and tactile stimuli, pointing, manual babbling, communication strategies etc. The standardized Visual Communication and Sign Language (VCSL) checklist for American Sign Language served as a template for the development of this diagnostic instrument (Simms, Baker, & Clark, 2013). The intra-modality adaptation process consisted of multiple steps before formal testing was possible, namely determining the suitability of the test, translation, checking for linguistic and cultural differences, pretesting, and standardization. No cultural differences were found. The current adaptation focused on the first two years of deaf sign-exposed children growing up in Flanders (Belgium).

Sign language development

Along with deaf children's early communicative skills, this checklist also measures children's language progress. For example, the early lexical development of signing children is characterized by phonological processes due to their developing motor skills and control (Meier, Mauk, Cheek, & Moreland, 2008). In isolation the phonological parameters - i.e. hand shape or handshape, movement, location, palm orientation and a fifth non-manual parameter (Stokoe, 1960; Vermeerbergen, 1997) – are meaningless elements. However, specific combinations of these five parameters give meaning, resulting in the formation of a true sign (Sandler, 2012). The formation process of a phonological repertoire is expressed in manual babbling, i.e. hand and finger babbling (Petitto, 2000). The phenomenon of babbling, first observed around the age of 10 months, is not exclusive to speech, but a universal phenomenon in which vocal and manual babbling coexist, have the same function, and follow a similar course over time (Petitto & Marentette, 1991; Petitto, 1997; Masataka, 2000; Wille, Mouvet, Vermeerbergen, & Van Herreweghe, 2018). In addition, research has shown a clear developmental pattern within the manual parameters where more complex parameters are replaced by less complex ones (Morgan, 2014; Wille et al., 2018). The most accurate parameter is found to be location, followed closely by movement, whereas handshape is considered least error resistant. This is especially visible when a marked handshape is replaced by an unmarked handshape (C, A, S, 1 or 5) (Boyes Braem, 1990; Mann, Marshall, Mason, & Morgan, 2010; Marentette & Mayberry, 2002; Wille et al., 2018). Unmarked handshapes tend to be more frequent and motorically and visually simpler than the marked handshapes.

Furthermore, the non-manual parameter adds specific grammatical meaning where mainly the facial expression, and to a lesser extent the whole body, carries the meaning of a sentence (Vermeerbergen, 1997). Reilly's research into the development of non-manual marking highlights the presence of communicative non-manual items acquired as one package with the manual parameters (Reilly, 2006). Only later, when children have a more complex grammatical development, do they acquire the specific meaning and function of the isolated non-manual items; i.e. 'hands before face' (Reilly, 2006, p. 286). The gesture comes two months before the YES nodding. Children then acquire the sign in isolation (hands), in order to make the link with the non-manual expression (face). Previous research by Boyes Braem (1994), McIntire (1994) and Emmorey (2002) indicates that children acquire these non-manual expressions - coupled with the corresponding grammatical structures - during the age range of 1.03 to 6.00. The development thus depends on the time when specific lexical signs and their corresponding structure

[1] The Flemish Sign Language has been unanimously recognized by the government in 2006 as an official, fully-fledged and natural language, which is unrelated to the surrounding spoken language Dutch (Vermeerbergen & Van Herreweghe, 2008).

are acquired. Some structures are acquired earlier (e.g. WHERE?), while it may take children years to correctly acquire the specific isolated function of other grammatical structures; e.g. the internal structure of negations: NO versus NONE (Anderson & Reilly, 2002).

Visual communication and Flemish Sign Language

In order to ameliorate the effectiveness of the checklist within the Flemish context and to represent the detailed complexity in early language development, some well-informed adaptations were made to the milestone sequence. The linguistic analysis resulted in the division of some items into two separate standalone items, mainly based on their level of complexity, and two fundamental items concerning early interaction and language were added.

First, in the initial checklist the development of children's pointing behavior was not correctly represented. On the one hand, early pointing behavior appears to be mainly towards objects and not towards 'the self'. The difference in acquisition between 'pointing to objects' and 'pointing to self' became clear from previous research (Chen Pichler, Lillo-Martin, & Gölgöz, 2018; Lillo-Martin & Chen Pichler, 2018; Wille et al., 2018), and is now clarified in the checklist as two separate items. On the other hand, the checklist also appeared to be missing a fundamental item, namely 'sign + pointing'. During previous interactional research among a deaf family in Flanders, this milestone was clearly identified as an intermediate stage between one-sign and two-sign utterances (Wille et al., 2018). Second, the checklist did not clearly take into account the hands before face phenomenon (Reilly, 2006, p. 286). The fact that this is a multi-step acquisition process is now reflected in the checklist. Third, the pilot study indicated that there might be a complexity difference in children's responses to single commands. Children respond to an assignment in different ways and elicit an isolated head nod (or head shake), an action or a head nod along with the corresponding action. In order to evaluate this complexity transparently, children's difference in answers can now be checked in the checklist. Fourth, another element of complexity was initially not taken into account in the original ASL-checklist. Based on internal complexity the item 'Answers questions WHERE? and WHAT?' was split into two separate items based on the data analyses that responding correctly to where-questions was found earlier and more often — e.g. by a pointing gesture — than to the what-questions. Finally, children's share in parent-child interactions, especially the initiation of interactions has been shown to be an extremely important aspect of children's early communication skills (Mouvet, Matthijs, Loots, Taverniers, & Van Herreweghe, 2013). Mouvet and colleagues described this communicative and interactional function in young deaf children growing up in deaf families (before the age of 24 months), while initiative taking was a rare find in deaf children in hearing families. Therefore, an item assessing children's initiations was added to the checklist.

Conclusion

With this adapted guide, the processing of visual information, the development of communication skills (e.g. communication strategies, joint attention, etc.) and the acquisition of early Flemish Sign Language milestones can be monitored and mapped for all young deaf children (0–2 years). The VGT VCSL checklist is considered part of the first multimodal assessment for Flemish Sign Language and deaf children.

More information on the checklist, data and pilot study following the adaptation process:

- Wille, B., Allen, T., Van Lierde, K., & Van Herreweghe, M. (2020). Using the adapted Flemish Sign Language VCSL checklist. *Journal of Deaf Studies and Deaf Education, 25*(2), 188–198.
- Wille, B., Allen, T., Van Lierde, K., & Van Herreweghe, M. (2020). The first Flemish Sign Language diagnostic instrument for children up to 2 years: Adaptation from American Sign Language. *American Annals of the Deaf*, *165*(3), 311–334.

References

Anderson, D., & Reilly, D. (2002). The MacArthur Communicative Development Inventory: Normative data for American Sign Language. *Journal of Deaf Studies and Deaf Education*, *7*(2), 83–106.

Bakeman, R., & Adamson, L. (1984). Coordinating attention to people and objects in mother-infant and peer-infant interaction. *Child Development*, *55*, 1,278–1,289.

Barton, M., & Tomasello, M. (1994). The rest of the family: The role of fathers and siblings in early language development. In C. Gallaway & Richards B.J. (Eds.), *Input and Interaction in Language Acquisition* (pp. 109–134). Cambridge: Cambridge University Press.

Boyes Braem, P. (1990). Acquisition of the handshape in American Sign Language: a preliminary analysis. In V. Volterra & C. Erting (Eds.), *From Gesture to Language in Hearing and Deaf Children* (pp. 107–127). Washington, DC: Gallaudet University Press.

Boyes Braem, P. (1994). *Einsührung in die Gebärensprache und ihre Erforschung*. Internationalen Arbeiten zur Gebärensprache und Kommunikation Gehörlosen. Band II. Hamburg: SIGNUM.

Brooks, R., Singleton, J. L., & Meltzoff, A. N. (2020). Enhanced gaze-following behavior in deaf infants of deaf parents. *Developmental Science*, *23*(2), e12900.

Butterworth, G. (1991). The ontogeny and phylogeny of joint visual attention. In A. Whitten (Ed.), *Natural Theories of Mind: Evolution, Development and Simulation of Everyday Mindreading* (pp. 223–232). Oxford: Blackwell Publishing.

Chen Pichler, D., Lillo-Martin, D., & Gökgöz, K. (2018, June). *Points to Self by Deaf, Hearing and Coda Children* [Conference presentation]. Third International Conference on Sign Language Acquisition, Istanbul.

Emmorey, K. (2002). *Language, Cognition, and the Brain: Insights from Sign Language Research*. Mahwah: Lawrence Erlbaum Associates.

Harris, M., & Chasin, J. (2005). Visual attention in deaf and hearing infants: The role of auditory cues. *Journal of Child Psychology and Psychiatry*, *46*(10), 1,116–1,123.

Haug, T. & Mann, W. (2008). Adapting tests of sign language assessment for other sign languages – a review of linguistic, cultural, and psychometric problems. *Journal of Deaf Studies and Deaf Education*, *13*(1), 138–147.

Lieberman, A. (2012). *Eye-gaze and Joint Attention*. Research Brief No. 5. Washington, DC: NSF supported Science of Learning Center on Visual Language and Visual Learning, Gallaudet University.

Lieberman, A., & Mayberry, R. (2015). Studying sign language acquisition. In E. Orfanidou, B.Woll, & G. Morgan (Eds.), *Research Methods in Sign Language Studies: A Practical Guide* (pp. 281–299). London: Wiley-Blackwell.

Lieberman, A., Hatrak, M., & Mayberry, R. (2014). Learning to look for language: Development of joint attention in young deaf children. *Language Learning and Development*, *10*(1), 19–35.

Lillo-Martin, D., & Chen Pichler, D. (2018, June). *It's Not All ME, ME, ME: Revisiting the Acquisition of ASL Pronouns* [Conference presentation]. Formal and Experimental Advances in Sign Language Theory (FEAST) colloquium, Venice.

Mann, W., Marshall, C. R., Mason, K. & Morgan, G. (2010). The acquisition of sign language: The impact of phonetic complexity on phonology. *Language Learning and Development*, *6* (1), 60–86.

Marentette, P., & Mayberry, R. (2000). Principles for an emerging phonological system: A case study of acquisition of American Sign Language. In C. Chamberlain, J. Morford & R. I. Mayberry (Eds.), *Language Acquisition by Eye* (pp. 71–90). Mahwah: Lawrence Erlbaum Associates.

Masataka, N. (2000). The role of modality and input in the earliest stage of language acquisition: Studies of Japanese Sign Language. In C. Chamberlain, J.P. Morford, & R.I. Mayberry (Eds.), *Language Acquisition by Eye* (pp. 3–24), Mahwah: Lawrence Erlbaum Associates.

McIntire, M. (1994). *The Acquisition of American Sign Language by Deaf Children*. SLS Monographs. Burtonsville: Linstok Press.

Meier, R.P., Mauk, C.E., Cheek, A. & Moreland, C.J. (2008). The form of children's early signs: Iconic or motoric determinants?. *Language Learning and Development*, *4*(1), 63–98.

Mitchell, R., & Karchmer, M. (2004). Chasing the mythical ten percent: Parental hearing status of deaf and hard of hearing students in the United States. *Sign Language Studies*, *4*, 138–163.

Mitchiner, J., Nussbaum, D.B., & Scott, S. (2012). *The Implications of Bimodal Bilingual Approaches for Children With Cochlear Implants*. Washington DC: Visual Language and Visual Science of Learning Center, Gallaudet University.

Morgan, G. (2014). On language acquisition in speech and sign: Development of combinatorial structure in both modalities. *Frontiers in Psychology*, *5*, 1–8.

Mouvet, K., Matthijs, L., Loots, G., Taverniers, M. & Van Herreweghe, M. (2013). The language development of a deaf child with a cochlear implant. *Language Sciences*, *35*, 59–79.

Petitto, L.A. (1997). In the beginning: On the genetic and environmental factors that make early language acquisition possible. In M. Gopnik (Ed.), *The Inheritance and Innateness of Grammars* (pp. 45–69). Oxford: Oxford University Press.

Petitto, L.A. (2000). On the biological foundation of human language. In K. Emmorey & H. Lane (Eds.), *The Signs of Language Revisited: An Anthology in Honor of Ursula Bellugi And Edward Klima* (pp. 447–471). Mahwah: Lawrence Erlbaum Associates.

Petitto, L.A. & Marentette, P. (1991). Babbling in the manual mode: Evidence for the ontogeny of language. *Science*, *25*, 1,483–1,496.

Piaget, J. (1952). *The Origins of Intelligence in Children* (M. Cook, Trans.). New York: W.W. Norton & Co.

Reilly, J. (2006). How faces come to serve grammar: the development of nonmanual morphology in American Sign Language. In B. Schick, M. Marschark & P.E. Spencer (Eds.), *Advances in Sign Language Development by Deaf Children* (pp. 262–290). New York: Oxford University Press.

Sandler, W. (2012). The phonological organization of sign languages. *Language and Linguistics Compass*, *6*(3), 162–182.

Simms, L., Baker, S., & Clark, M.D. (2013). The standardized Visual Communication and Sign Language checklist for signing children. *Sign Language Studies*, *14*(1), 101–124.

Stokoe, W. (1960). *Sign Language Structure: An Outline of the Visual Communication Systems of the American Deaf*. Studies in Linguistics: Occasional Papers 8. Silver Spring: Linstok Press.

van den Bogaerde, B. (2000). *Input and interaction in deaf families*. Doctoral dissertation. University of Amsterdam.

Vermeerbergen, M. (1997). *Grammaticale Aspecten van de Vlaams-Belgische Gebarentaal*. Gentbrugge: Cultuur voor Doven.

Vermeerbergen, M., & Van Herreweghe, M. (2008). *Wat Gewenst/Geweest is: Organisaties van en voor Doven in Vlaanderen Bevraagd Over 10 Thema's*. Ghent: Academia Press/Fevlado-Diversus.

Wille, B., Mouvet, K., Vermeerbergen, M., & Van Herreweghe, M. (2018). Flemish Sign Language development: a case study on deaf mother-deaf child interactions. *Functions of Language*, *35*(2), 269–302.

Wille, B., Van Lierde, K., & Van Herreweghe, M. (2019). Parental strategies used in communication with their deaf infants. *Child Language Teaching and Therapy, 35*(2), 165–183.

Assessing strength of vocabulary knowledge in deaf and hearing children using Finnish Sign Language

Laura Kanto
University of Jyväskylä, Finland

Wolfgang Mann
University of Roehampton, United Kingdom

Abstract

This study investigated strength of vocabulary knowledge in 31 deaf and hearing children aged 4 to 15, who had acquired Finnish Sign Language (FinSL) from birth. Children's understanding of different form-meaning mappings and lexical-semantic organisation was assessed by using an adapted version of the web-based British Sign Language Vocabulary Test (BSL-VT; Mann, 2009), the FinSL-VT (FinSL) (Kanto, Syrjälä, & Mann, 2021). The BSL-VT was developed to measure strength of vocabulary in the form of deaf children's understanding of different mappings between form and meaning for single signs. It consists of four tasks, each of which taps a different degree of strength of vocabulary knowledge, namely: *meaning recognition*, *form recognition*, *form recall*, and *meaning recall.* Findings from the adapted version for FinSL showed a hierarchy of the degrees of difficulty for these tasks which is comparable to results previously reported for other signed languages, including BSL and American Sign Language (ASL). This result strengthens previous claims that children's signed vocabulary knowledge develops gradually (rather than instantaneously) and that the form-meaning mapping paradigm is appropriate for use with signed languages. Finally, the study also provides new insights into the adaptation process of tests from one signed language to another.

Introduction

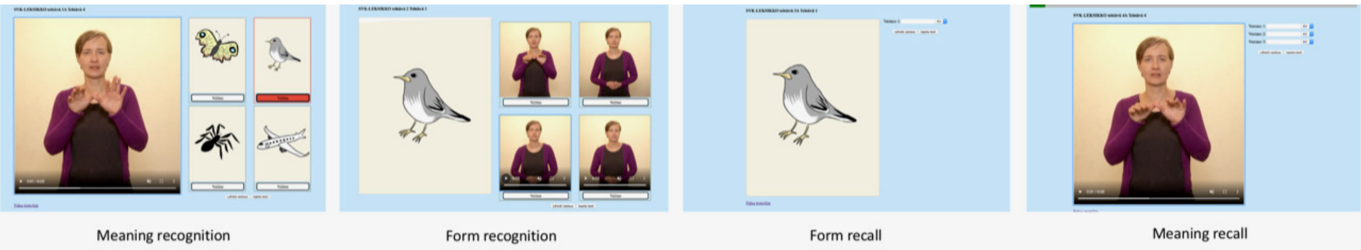
Vocabulary knowledge fulfils a key role in children's successful language acquisition and often serves as a predictor of later language development such as reading comprehension. One way of studying vocabulary development is to examine both *quantitative* and *qualitative* features of vocabulary knowledge (see the review by Yanagisawa & Webb, 2019). The quantitative features of vocabulary knowledge refer to the number of different lexical items in the vocabulary of the child (vocabulary breadth). The qualitative features of vocabulary knowledge, on the other hand, refer to knowledge related to meaning, strength of different semantic connections, and grammatical and usage properties of different lexical items (Laufer, 2013; Lin & Morrison, 2010; Qian & Lin, 2019; Schmitt, Cobb, Horst, & Schmitt, 2015).

While children acquire new lexical items (spoken or signed) with increasing speed, the newly acquired items need to be organized in a structured manner in their mental lexicon to maintain a steadily growing vocabulary. To facilitate this, children follow a form-meaning mapping approach by which the phonological form of a lexical item gives the child access to its meaning and, at the same time, meaning provides access to the phonological form of an item (Clark, 2009). Previous research has established that children start forming different semantic connections by combining newly acquired lexical items in a structured manner in their mental lexicon. This semantic network consists of strong links that are formed between lexical items that have closely related semantic meaning and weak(er) links between items that share fewer semantic relations (Clark, 2009; Madole & Oakes, 1999). As children's vocabulary grows, so does the number of relations that are generated as they (children) start structuring their mental lexicons.

Studies on language acquisition show that children's early semantic networks tend to contain mainly context-dependent semantic relations between acquired lexical items that are strongly bounded through context (Lin & Morrison, 2010). These types of relations have been referred to as *syntagmatic*. Syntagmatic relations include words that form a syntactic sequence (e.g., cold-outside) or words that share a thematic relationship with the stimulus (e.g., cold-sweater, cold-winter, apple-eat). A second type of semantic relations draws on categorical links between words in children's lexical networks (e.g. synonym, coordinate, subordinate); these relations are referred to as *paradigmatic* (Mann, Sheng, & Morgan, 2016). In most children's lexical-semantic networks we generally find both syntagmatic and paradigmatic semantic relations. Syntagmatic relations tend to be more context-dependent and triggered by actual perceptual and conceptual experiences whereas paradigmatic relations have been regarded as more abstract relations that are, in part, informed by linguistical knowledge. Consequently, children tend to start using paradigmatic relations at a later age.

Given the importance of vocabulary, the assessment of vocabulary knowledge forms a critical part of assessing children's language skills and development. Vocabulary assessment provides important information for parents, practitioners and clinicians working with the child. When assessing children's vocabulary knowledge, it is important to evaluate both quantitative and qualitative features. However, there is a general shortage of vocabulary tests that do this, which is particularly apparent in the case of signed languages.

The aim of the present study is to investigate vocabulary knowledge in a language and population both of which have been rarely studied: Finnish children (deaf and hearing) that use Finnish Sign Language (FinSL). We specifically focus on children's understanding of different form-meaning mappings by using a multi-layered assessment format originally developed for another signed language, British Sign Language (BSL). The web-based British Sign Language Vocabulary Test (BSL-VT; Mann, 2009) was developed to measure the degree of strength of the mappings between form and meaning for items in the lexicon of signing children aged 4–15, drawing on research on spoken languages by Read (2010) and Laufer and colleagues (Laufer & Goldstein, 2004). The BSL-VT consists of four tasks each of which taps a different degree of strength of vocabulary knowledge, namely: *meaning recognition* (matching a sign to four pictures), *form recognition* (matching a picture to four signs), *form recall* (picture naming), and *meaning recall* (repeated sign association). The BSL-VT has been adapted for American Sign Language (ASL) (Mann, Roy & Morgan, 2016) and, more recently, for FinSL (Kanto et al., 2021) (see Figure 1).



Meaning recognition

Form recognition

Form recall

Meaning recall

Figure 1 Screenshots of FinSL-VT tasks, adapted from the four BSL-VT tasks described in the paragraph above

Method

The web-based BSL-VT by Mann (2009) was adapted for FinSL following the steps outlined by Mann, Roy, et al. (2016) and Enns and Herman (2011) and piloted with altogether 37 children (22 deaf and 15 hearing) between the ages of 4–15 (average age: 9). All children had at least one deaf parent and were acquiring FinSL from birth. They represent a unique group of children that are exposed from birth to a signed language. In comparison, the majority of deaf children (90–95%, Mitchell & Karchmer, 2004) grows up in hearing families where parents may or may not use signed language. The collected included children's performance scores from the four web-based FinSL vocabulary tasks (two receptive + two expressive). Each of the tasks contained 60 items. These items were the same across tasks (see Figure 1). The test was presented on a laptop by a team of deaf and hearing administrators all of whom were native or near native signers. In addition, demographic data was collected from parents via a questionnaire.

The responses to both receptive vocabulary tasks were automatically saved onto an Excel datasheet on the web server. In the meaning recognition task, children see a sign and need to match it to one of four images. This format is similar to existing, standardised vocabulary tests for spoken languages, e.g., The Peabody Picture Vocabulary test (PPVT, Dunn & Dunn, 1997) or the Receptive One Word Picture Vocabulary Test (ROWPT, Dunn, Dunn, Whetton, & Burley, 1997). In the form recognition task, children are asked to match a stimulus image to one of four signs. For the production tasks, responses were typed as Finnish glosses into a text box on the laptop screen by the test administrator and scored as either '0', '0.5', or '1'. All signed responses were video-recorded, as well. In the form recall task, children see an image on the computer screen and are asked to produce the corresponding sign. This format is similar to the Expressive One Word Picture Vocabulary Test (EOPVT, Brownell, 2000), a standardised test for spoken English. The response was coded as correct and scored as '1' if the child produced the expected FinSL sign to name the target item. Responses were coded as partially corrected and scored as 0.5 if the child produced a sign that was outside the immediate range of expected answers but still suggested that the child knew the meaning of the target (e.g., when the target sign was FRIEND and the child signed MAN). Finally, the meaning recall task is a repeated meaning association task where participants see the target sign in FinSL presented on their screen and have to supply three different FinSL signs with an associated meaning. Children's responses were scored as '1' if the child produced a paradigmatic response, e.g., a synonym (happy–excited), antonym (strong–weak), coordinate (cherry–strawberry), subordinate (bird–swan), or superordinate (mouse–animal).

Children's responses were scored as '0.5' if the child produced a syntagmatic response (e.g., hospital– doctor, mouse–small).

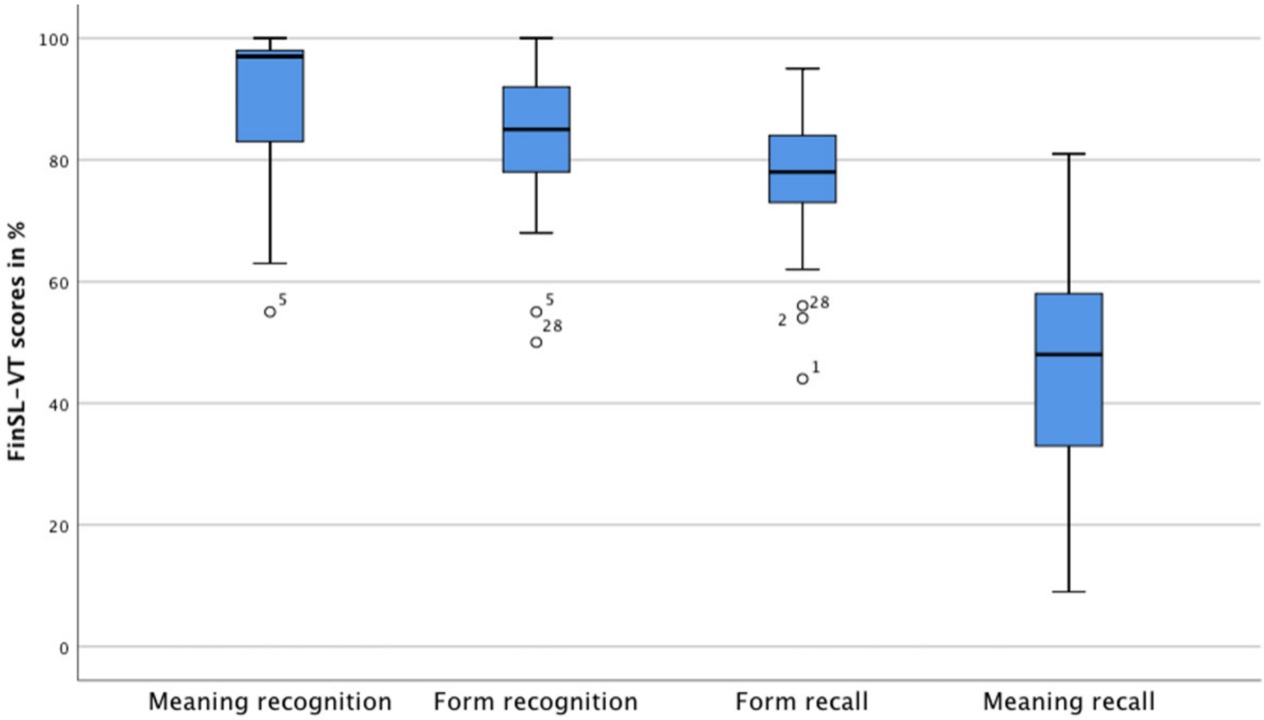


Figure 2 Bar graph displaying FinSL-VT mean scores in percentages by task

Results

Findings showed a hierarchy of difficulty between the four tasks, which is concordant with results reported previously for BSL and American Sign Language (ASL) (Mann & Marshall, 2012; Mann, Roy, et al., 2016). According to this hierarchy, children performed best on the meaning recognition task and the lowest on the meaning recall task. The scores of the meaning recall tasks were considerably lower compared with other tasks. In addition, children's within-group performance on this task varied more notably compared to the other three tasks.

A Pearson correlation between age and performance on the four tasks showed strong, positive relationships (Cohen, 1988) all of which were statistically significant (see Table 1). When carrying out partial correlations between the different tasks, controlling for age, the following correlations remained significant: meaning recognition and form recognition, R(8) = .836, *p*←.01, meaning recognition and form recall, R(8) = .470, *p*←.01, form recognition and form recall, R(8) = .520, *p*←.01, form recall and meaning recall R(8) = .459, *p*←.01.

Table 1: Correlation between age and FinSL-VT raw scores

Variable	*Meaning recognition*	*Form recognition*	*Form recall*	*Meaning recall*
Age	0.696** (.456–.865)	0.703** (.504–.840)	0.711** (.525–.832)	0.670** (.492–.800)

**p←.01, BCa bootstrap 95%; confidence intervals reported in parentheses.

Additionally, a second Pearson correlation was carried out between age and performance on the meaning recall task to examine the development of semantical networks in children's vocabulary. Results showed strong, positive relationships that were statistically significant for both types of semantic relations (paradigmatic, syntagmatic) (see Table 2). This relationship is slightly weaker for paradigmatic relations which is perhaps not too surprising as children usually start forming these relations at a later age. At the same time the number of mistakes and missing answers (both of which were scored as '0') decreased with age, showing a strong negative correlation that was statistically significant.

Discussion

The reported findings of this study indicate that children's performance on the FinSL tasks correlated even after partialling out age. These results suggest that all tasks tap the vocabulary knowledge of children. Furthermore, findings showed a hierarchy of the degrees of difficulty in four different form-meaning mapping tasks in FinSL-VT that is comparable to results previously reported for other signed languages, that is BSL and ASL (Mann & Marshall, 2012; Mann, Roy, et al., 2016; Mann, Sheng, et al., 2016). The findings strengthen previous claims by Mann and Marshall (2012) that children's signed vocabulary knowledge develops gradually (rather than instantaneously) and that this developmental progress can be assessed, using a form-meaning mapping paradigm which has shown to be appropriate for signed languages. Furthermore, they indicate that the strength of vocabulary is a multidimensional feature that contains both quantitative and qualitative aspects. For this reason, (signed) vocabulary knowledge should be seen as a continuum where the strength of vocabulary knowledge changes from one end to another and that this continuum can be expressed through a hierarchy of children's understanding of form-meaning mappings. The presented findings from FinSL are in line with previous reports from two other signed languages, BSL and ASL, which reported an increase in strength of children's (understanding of) mappings between form and meaning over time. Our new results from FinSL are important because they validate the form-meaning mapping model for use with signed languages.

Table 2: Correlation between age and different type of answers for the meaning recall task

Variable	*Number of paradigmatic answers*	*Number of syntagmatic answers*	*Number of mistakes and missing answers*
Age	0.415* (.114–.653)	0.476**} (.225–.687)	-0.721** (-.826–-.580)

**p←.01, *p←.05, BCa bootstrap 95% confidence intervals reported in parenthesis.

Finally, the presented study also provides new insights into the adaptation process of tests from one signed language to another and shows this process to be a reliable and valid way to develop assessment tools for lesser researched signed languages such as FinSL (for a more detailed report of the psychometric properties of the FinSL, see Kanto et al., 2021).

References

Brownell, R. (2000). *Expressive One-word Picture Vocabulary Test.* Novato: Academic Therapy Publications.

Clark, E. (2009). *First Language Acquisition* (2nd ed.). Cambridge: Cambridge University Press.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioural Sciences*. Hillsdale: Laurence Erlbaum Associates.

Dunn, L. M., & Dunn, L. M. (1997). *PPVT-III: Peabody Picture Vocabulary Test*. Circle Pine: American Guidance Service.

Dunn, L.M., Dunn, L.M., Whetton, C., & Burley, J. (1997). *The British Picture Vocabulary Scale* (2nd ed.). Slough: NFER-Nelson.

Enns, C., & Herman, R. (2011). Adapting the assessing British Sign Language Development: Receptive Skills Test into American Sign Language. *Journal of Deaf Studies and Deaf Education, 16,* 362–374.

Kanto, L., Syrjälä H., & Mann, M. (2019). *FinSL-VT.* Retrieved from: viittomatesti.cc.jyu.fi/

Kanto, L., Syrjälä, H., & Mann, W (2021). Assessing vocabulary in deaf and hearing children using Finnish Sign Language. *The Journal of Deaf Studies and Deaf Education*, *26*, 147–158.

Laufer, B. (2013). Vocabulary and writing. In C. A. Chapelle (Ed.), *The Encyclopedia of Applied Linguistics*. New Jersey: Wiley-Blackwell.

Laufer, B., & Goldstein, Z. (2004). Testing vocabulary knowledge: Size, strength, and computer adaptiveness. *Language Learning*, *54*(3), 399–436.

Lin, L. H. F., & Morrison, B. (2010). The impact of the medium of instruction in Hong Kong secondary schools on tertiary students' vocabulary. *Journal of English for Academic Purposes, 9*(4), 255–266.

Madole, K. L., & Oakes, L. M. (1999). Making sense of infant categorization: stable process and changing representations. *Developmental Review*, *19*, 263–296.

Mann, W. (2009). *British Sign Language Vocabulary Test (BSL-VT)*. Unpublished test. City University London.

Mann, W., & Marshall, C. (2012). Investigating deaf children's vocabulary knowledge in British Sign Language. *Language Learning, 62*, 1,024–1,051.

Mann, W., Roy, P., & Morgan G. (2016). Adaptation of a vocabulary test from British Sign Language to American Sign Language. *Language Testing, 33*, 3–22.

Mann, W., Sheng, L. & Morgan, G. (2016). Lexical-semantic organization in bilingual developing deaf children with ASL-dominant language exposure: Evidence from a repeated meaning association task. *Language Learning*, *66*, 827–899.

Mitchell, R. E., & Karchmer, M. (2004). Chasing the mythical ten percent: Parental hearing status of deaf and hard of hearing students in the United States. *Sign Language Studies*, *4*(2), 138–163.

Qian, D.D., & Lin, L. H. F. (2019). The relationship between vocabulary knowledge and language proficiency. In S. Webb (Ed.) *Routledge Handbooks in Vocabulary Studies* (pp. 66–80). Abingdon: Routledge.

Read, J. (2010). *Assessing Vocabulary*. Cambridge: Cambridge University Press.

Schmitt, N., Cobb, T., Horst, M., & Schmitt, D. (2015). How much vocabulary is needed to use English? Replication of van Zeeland & Schmitt (2012), Nation (2006) and Cobb (2007). *Language Teaching, 50*(2), 212–226.

Yanagisawa, A., & Webb, S. (2019). Measuring depth of vocabulary knowledge. In S. Webb (Ed.) *Routledge Handbooks in Vocabulary Studies* (pp. 371–386). Abingdon: Routledge.

Fair and Valid Testing and Assessment

Post-entry language assessment in higher education: ensuring fairness by including the voice of the test-taker

Jordi Heeren
KU Leuven, Belgium

Dirk Speelman
KU Leuven, Belgium

Lieve De Wachter
KU Leuven, Belgium

Abstract

Test-taker feedback has always been a part of test validation. Although on its own it can never be sufficient as evidence for validity, adding test-taker perceptions to other sources of validity evidence can reveal potential sources of construct underrepresentation or construct-irrelevant variance. In this article, we report on a study into the perception of a low-stakes, post-entry Dutch academic vocabulary and reading screening test used to screen all starting students in a university in Flanders. Students' perception of the test administration, the representativeness of its language, difficulty and tasks for the academic domain as well as their perception of the usefulness of the instrument and the feedback will be examined. Gender differences, and differences according to language background and higher education experience will be taken into account. To conclude, we will address how test-taker feedback can help to support the claims connected to score interpretation and use.

Introduction

Validity in language testing studies has largely been investigated using statistical analysis and psychometric studies. Test-taker feedback used to be labelled as 'face validity' and has largely been rejected by language testers (Bachman & Palmer, 1996). Iwashita and Elder (1997) mention the possibility of conflicting evidence between test-taker feedback and other sources of evidence and the fact that test-taker feedback can be subjective and influenced by test-takers' background variables and level of language proficiency. Nevertheless, test-taker perception and the acceptance of the test by test-takers can play an important role in establishing the representativeness of a measurement. Test-takers are the stakeholders that are directly affected by the use of a test (Bachman & Palmer, 1996) and if they do not perceive an assessment as credible, they may reject it (Iwashita & Elder, 1997). Furthermore, including test-taker feedback into the validation of tests can even be considered a democratic practice (Shohamy, 2001). In recent argument based validity frameworks, where validity evidence is accumulated through a chain of inferences, test-taker feedback forms one of the possible sources of evidence to back the warrants underlying different claims (Bachman & Palmer, 2010; Fox & Cheng, 2015; Knoch & Elder, 2013).

In argumentative validation frameworks, information on test administration and the clearness of instructions functions as part of the observation or evaluation inference, i.e. that the test score is an adequate reflection of the observed test behaviour (Kane, 1992; Knoch & Elder, 2013; Knoch, Elder, & O'Hagan, 2016). Only test-takers can provide information on how they perceive the administrative procedures, the clearness of instructions, procedures and the test environment (Bachman & Palmer, 2010; Cheng & DeLuca, 2011; Gardiner & Howlett, 2016; Yan, Thirakunkovit, Kauper, & Ginther, 2016). It can also help to guarantee that results are not influenced by technical issues or students' lack of computer skills. An investigation into student perception of administration procedures can also be used to uncover or explain differences according to student backgrounds (Doe, Cheng, Fox, Klinger, & Zheng, 2011; Iwashita & Elder, 1997; Zheng, Klinger, Cheng, Fox, & Doe, 2011). In addition, test-takers can also provide information on their perception of the test as an adequate reflection of the intended construct (Bradshaw, 1990; Cheng & DeLuca, 2011; Fox & Cheng, 2007; Fox & Cheng, 2015). As Iwashita & Elder (1997) show, this type of evidence should be interpreted with care and should be triangulated with other sources. Perceptions can be influenced by test scores and lead to self-serving bias, i.e. poor-performing students blaming their lower score on the test construct or tasks. Lastly, test-takers can also provide insight into what Bachman and Palmer (1996) call 'impact'. Here, the focus is on the educational utility of the instrument and its scores

(Cheng & DeLuca, 2011; Sinclair, Larson, & Rajendram, 2019). This is especially relevant in the context of post-entry language assessment (PELA) as the focus of this type of tests is not admission, but to identify possible at-risk students. In this study, test-taker feedback on a post-entry Dutch academic reading and vocabulary (ARV) screening test for all starting university students is examined.

Test-taker feedback on a post-entry academic reading and vocabulary screening test

Instruments

To measure incoming students' academic language proficiency, several university faculties use the ARV screening test at the start of the academic year. This is a practical screening instrument to be administered on a large scale, with the potential to screen all incoming students using limited resources (De Wachter & Heeren, 2013). The test used in this study consists of 25 mostly selected-response academic reading and vocabulary items with a time limit of 30 minutes. While the test is normally administered under supervised conditions on campus, in the year in which the study was conducted, students filled in the test at home due to the Covid-19 pandemic. On completion, students immediately receive their result and a feedback message provided by their faculty.

The post-test questionnaire on test-taker perception, consisting of three sections, was piloted in 2019–2020 and received ethics approval from the institutional review board. Students are invited to complete the questionnaire immediately after they receive their result. The first section comprises three questions and investigates the perceived usefulness of the test, the clearness of test procedures and test anxiety. Each item uses a 4-point Likert-scale (strongly agree, rather agree, rather disagree, strongly disagree). The second section comprises 15 questions on the difficulty of the test and item types, the impact of the time constraints and contextual factors such as background noise and technical issues. Two 4-point Likert-scales were used (strongly agree, rather agree, rather disagree, strongly disagree; and very difficult, rather difficult, rather easy, very easy). The third section contains eight questions about the usefulness of the feedback and the interpretability of the score. All questions in the third section use the same 4-point Likert-scale from the first section. The questionnaire ended with one open-ended question: 'do you have any further comments about the test?'.

Analyses

In total, 1,663 students filled in the questionnaire and gave their consent. Descriptive analyses with frequencies were used, as well as statistical analyses (Mann-Whitney-U tests with the rank biserial correlation (*r*) as effect size) to establish possible differences according to gender, home language and higher education experience (HEE). Spearman rank correlations between item responses and test scores were calculated to investigate the influence of students' test scores, thereby also uncovering potential instances of self-serving bias. Responses to the open-ended question were coded based on a coding system developed during the pilot. Three major categories were identified: test-taker characteristics, test characteristics, and test environment characteristics. One coder coded all 161 comments, a second rater randomly coded 44 (27%) of the comments. Inter-coder reliability as exact agreement was 87.5%. The total number of given codes was 195.

Findings

Table 1 shows that during test administration students experienced few technical difficulties and almost all students found that they had sufficient computer skills. Task instructions were clear for most students. When the perception of the representativeness of the screening test for the target language use (TLU) domain was investigated, students mostly agreed that the language used in the test tasks is representative for the language in the academic domain. They agreed less on the representativeness of the tasks themselves. This is not surprising since the test is a practical instrument, limited mostly to selected-response items. The time constraints in the test system also seem to have had an effect on test-takers' experience. Although many test-takers found the time constraints reflective of those of the academic year, many students also indicated that the time limit had a considerable impact on their test behaviour. The open-ended questions revealed that some students felt that the time limit substantially increased their stress level. In addition, 63.32% of the students also perceived the test as very difficult or rather difficult. With regards to the usefulness of the information yielded by the test, the majority of the test-takers found the screening test useful and their score clear. Most students also found the feedback clear and useful; however, some of the answers to the open-ended questions indicate that the feedback should perhaps be more elaborate.

The correlations with screening scores reveal a high positive correlation between students' screening test score and students' perception of the test score as a correct reflection of their academic reading ability (ρ=.52). This might be a reflection of what is

Table 1: Frequencies of student responses to a selection of questionnaire items

	n	Strongly disagree	Rather disagree	Rather agree	Strongly agree
There were no technical issues in the testing system	1,663	6.19%	13.05%	17.56%	63.20%
My computer knowledge was sufficient for this test	1,663	1.38%	2.53%	14.67%	81.42%
Task instructions were clear	1,663	0.24%	4.39%	32.17%	63.20%
The language in the test was representative of the language in my discipline	1,663	1.02%	14.49%	58.75%	25.74%
The test tasks were representative of the tasks I will have to do during my courses	1,663	3.55%	31.87%	51.05%	13.53%
The time constraints were representative of the time pressure during the academic year	1,662	5.78%	25.15%	50.60%	18.47%
I changed the way I filled in the test because of the time limit	1,663	11.55%	20.57%	39.33%	28.56%
The usefulness of the test was clear before we started	1,663	1.74%	11.73%	42.33%	44.20%
My test score was easy to interpret	1,663	1.92%	11.06%	42.51%	44.50%
The feedback on my score was clear	1,662	4.03%	18.59%	49.94%	27.44%
The feedback on my score was useful	1,662	4.33%	25.57%	50.66%	19.43%

called self-serving bias where poor performing participants see their performance as a result of external factors, in this case the appropriateness of the test construct. Moderately positive correlations were also found: students with a lower screening test score reported finding their score less easy to interpret (ρ=.35) and reported a lower test effort (ρ=.44). Items on the difficulty of the test and the test items show a large negative correlation, with the item on test difficulty being the largest (ρ=-.60).

When background differences for test perception are investigated, several trends become apparent. Firstly, differences according to gender indicated that female students found the administration procedures (*U*=309716, *p*←.001, *r*=-.10) and task instructions (*U*=299380, *p*←.001, *r*=-.13) more clear than male students. They also found the test more useful (*U*=318282.5, *p*=.002, *r*=-.08) and thought it gave them more insight into their abilities (*U*=325468, *p*=.026, *r*=-.06). While they found the time limit more representative (*U*=311268.5; *p*←.001, *r*=-.10), they reported their test behaviour being more affected by it (*U*=306098.5, *p*←.001, *r*=-.11). They were also significantly more nervous (*U=261289.5*, *p*←.001, *r*=-.24), which is in line with earlier studies that show that test anxiety appears to be higher in women (Cassady & Johnson, 2002; Chapell, et al., 2005; Zeidner, 1990). Secondly, differences also appeared according to students' language background. Multilingual students reported their test behaviour being more affected by the time limit (*U*=204003, *p*=.003, *r*=-.10) and were more nervous than monolingual students (*U*=188256, *p*←.001, *r*=-.17). They perceived the screening test as more difficult (*U*=208394.5, *p*=.009, *r*=-.08) as well as most of the academic vocabulary item types. They agreed less on the fact that they had sufficient computer skills (*U*=245123, *p*←.001, *r*=.08) and were bothered more by background noise (*U*=248151, *p*=.004, *r*=.10). They also indicated less often that their test score reflects their ability (*U*=256880; *p*←.001 *r*=.14), found the language use less authentic (*U*=242632.5, *p*=.020, *r*=.07), and their score less easy to interpret (*U*=255065.5, *p*←.001 *r*=.13). Students with HEE found the usefulness of the test more clear (*U*=174407, *p*=.030, *r*=.08), the test result more reflective of their abilities (*U*=177736.5, *p*=.005, *r*=.11), and the language used in the test more representative than students without HEE (*U*=172894, *p*=.047, *r*=.07). They were also less nervous (*U*=133200, *p*←.001, *r*=-.17) and looked at the clock less often (*U*=137375, *p*←.001, *r*=-.15). They found the test easier (*U*=118155, *p*←.001, *r*=-.27) as well as the different academic vocabulary item types. This is to be expected as they already have experience with the vocabulary encountered in higher education. In all, the effects found are very small, although the effect size is slightly larger for test anxiety and difficulty.

Conclusion

Although the questionnaire yields valuable test-taker insights, the need to triangulate the information from the student questionnaires is apparent. For example, despite multilingual students and women being more nervous about the screening test, quantitative analyses should be used to prove that this does not lead to test bias. Additional evidence on representativeness and impact can consist of corpus studies, verbal protocol analyses or the use of qualitative research methods and other stakeholder groups. The results of this study point out that most students perceive the test as useful, the test procedures as clear and the test language, despite a narrow range of tasks, as reflective of the language use in the TLU-domain. These lines of evidence can thus provide additional backing for the warrants that the test administration procedures are adequate; the test content, in this case mostly its language, is representative for that of the TLU-domain; and that the screening test provides students with useful information at the start of their university education. Test-taker feedback, and even broader, stakeholder feedback, can thus provide crucial insights for test validation and fairness. As Bradshaw (1990, p. 27) concludes on the topic of stakeholder input: 'These groups appear to have opinions to offer, and are willing to offer them'.

References

Bachman, L.F., & Palmer, A. (1996). *Language Testing in Practice.* Oxford: Oxford University Press.

Bachman, L.F., & Palmer, A. (2010). *Language Assessment in Practice.* Oxford: Oxford University Press.

Bradshaw, J. (1990). Test-takers' reactions to a placement test. *Language Testing*, *7*(1), 13–30.

Cassady, J. C., & Johnson, R. E. (2002). Cognitive test anxiety and academic performance. *Contemporary Educational Psychology*, *27*(2), 270–295.

Cheng, L., & DeLuca, C. (2011). Voices from test-takers: Further evidence for language assessment validation and use. *Educational Assessment*, *16*(2), 104–122.

Chapell, M. S., Blanding, Z. B., Silverstein, M. E., Takahashi, M., Newman, B., Gubi, A., & McCann, N. (2005). Test anxiety and academic performance in undergraduate and graduate students. *Journal of Educational Psychology*, *97*(2), 268–274.

De Wachter, L., & Heeren, J. (2013). Een taaltest als signaal. De ontwikkeling en implementatie van een strategische taalvaardigheidstoets. [A language test as a signal: the development and implementation of a strategic language test.] *Levende Talen Tijdschrift*, *14*(1), 19–27.

Doe, C., Cheng, L., Fox, J., Klinger, D., & Zheng, Y. (2011). What has experience got to do with it? An exploration of L1 and L2 test-takers' perceptions of test performance and alignment to classroom literacy activities. *Canadian Journal of Education*, *34*(3), 68–85.

Fox, J., & Cheng, L. (2007). Did we take the same test? Differing accounts of the Ontario Secondary School Literacy Test by first and second language test-takers. *Assessment in Education*, *14*(1), 9–26.

Fox, J., & Cheng, L. (2015). Walk a mile in my shoes: Stakeholder accounts of testing experience with a computer-administered test. *TESL Canada Journal*, *32*(9), 65–86.

Gardiner, J., & Howlett, S. (2016). Student perceptions of four university gateway tests. *University of Sydney Papers in TESOL*, *11*, 67–96.

Iwashita, N., & Elder, C. (1997). Expert feedback? Assessing the role of test-taker reactions to a proficiency test for teachers of Japanese. *Melbourne Papers in Language Testing*, *6*(1), 53–67.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin*, *112*(3), 527–535.

Knoch, U., & Elder, C. (2013). A framework for validating post-entry language assessments (PELAs). *Papers in Language Testing and Assessment*, *2*(2), 48–66.

Knoch, U., Elder, C., & O'Hagan, S. (2016). Examining the validity of a post-entry screening tool embedded in a specific policy context. In J. Read (Ed.), *Post-admission Language Assessment of University Students* (pp. 23–42). New York: Springer.

Shohamy, E. (2001). Democratic assessment as an alternative. *Language Testing*, *18*(4), 373–391.

Sinclair, J., Larson, E. J., & Rajendram, S. (2019). 'Be a Machine': International graduate students' narratives around high-stakes English tests. *Language Assessment Quarterly*, *16*(2), 236–252.

Yan, X., Thirakunkovit, S. P., Kauper, N.L., & Ginther, A. (2016). What do test-takers say? Test-taker feedback as input for quality management of a local oral English proficiency test. In J. Read (Ed.), *Post-admission Language Assessment of University Students* (pp. 113–136). New York: Springer.

Zeidner, M. (1990). Does test anxiety bias scholastic aptitude test performance by gender and sociocultural group?. *Journal of Personality Assessment*, *55*(1/2), 145–160.

Zheng, Y., Klinger, D. A., Cheng, L., Fox, J., & Doe, C. (2011). Test-takers' background, literacy activities, and views of the Ontario Secondary School Literacy Test. *Alberta Journal of Educational Research*, *57*(2), 115–136.

Examining IELTS score consistency across testing sites: The question of test fairness

Linyu Liao
University of Macau, China

Abstract

The current study examined IELTS score consistency across testing sites as a means of exploring the fairness of large-scale international language tests. Specifically, 77 Chinese test-takers who sat IELTS twice in two Asian testing sites within 30 days voluntarily provided their test-retest scores for analysis. These test-retest scores were analysed using descriptive analyses, paired-samples t-tests, paired-samples correlation analyses, and generalisability analyses in SPSS and mGENOVA. These analyses together showed a low cross-site score consistency, especially on the Writing and Speaking sections. Such inconsistency is likely to pose a threat to test fairness and validity. Test developers, therefore, need to further investigate the sources of such inconsistency and take measures accordingly to improve the reliability of test scores and safeguard test fairness and validity.

Introduction

Reliability is a crucial aspect of test fairness, especially for large-scale high-stakes tests administered around the world. A large volume of research, therefore, has investigated the factors that may impact the reliability of test scores, including raters (e.g., Blackhurst, 2004), tasks (e.g., O'Loughlin & Wigglesworth, 2003), and scoring procedures (e.g. Falvey & Shaw, 2006). However, the potential impact of different testing sites, as an overall factor that includes a variety of variables such as raters, tasks, and physical conditions, on score consistency has long been neglected, despite its obvious importance in ensuring test fairness for tests that are large-scale and administered worldwide.

In large-scale high-stakes assessments, it is common for test-takers to repeat test taking in order to meet a particular cut-score for university admission or other purposes (Barkaoui, 2017). A few studies have investigated these test repeaters' score change between test occasions and examined the sources of such change (e.g., Barkaoui, 2019; Cho & Blood, 2020; Green, 2005; Lin & Chen, 2020; Wilson, 1987; Zhang, 2008). In most of these studies, the time interval between test occasions was quite long, lasting for months and even years. Only Zhang (2008) investigated the Internet-based Test of English as a Foreign Language (TOEFL iBT) repeaters' score change across test occasions within 30 days. However, it is important to conduct research that limits test-retest time intervals to be within a short time, because only by doing so is it possible to find out whether the score change is caused by test takers' actual change of language proficiency or other construct-irrelevant factors. Considering these research gaps, the current study aims to investigate the consistency of IELTS scores gained from two testing sites within 30 days. The specific research question is as follows: To what extent are IELTS scores consistent across testing sites?

Research methods

The data used in this study were collected from 77 Chinese test-takers who sat IELTS (Academic) twice in two Asian testing sites within 30 days. One of the testing sites was a city in mainland China, while the other was a major city in a Southeast Asian country or Hong Kong. After sitting the test twice in different testing sites, these test takers voluntarily provided their test-retest scores (including total scores, i.e., overall band scores, and band scores on the Listening, Reading, Writing, and Speaking sections) for analysis. The test and retest time window was limited to be within 30 days, so as to avoid actual and substantial proficiency improvement during this period. The test-retest scores from these 77 test-takers were then used to conduct descriptive analyses, paired-samples t-tests and paired-samples correlation analyses with SPSS and generalisability analyses with mGENOVA.

Results

The results of data analyses are presented in the following tables. Table 1 shows the statistics of test scores gained from different testing sites. The mean of section scores and total scores ranged from 5.41 to 6.28. It was clear that Reading and Listening scores (6.01–6.28) were higher than Writing and Speaking scores (5.41–5.80), which reflects what usually happens with Chinese test-takers. As indicated by skewness and kurtosis, the test scores had no gross violation of normality.

Table 2 shows the results of the paired-samples t-tests. According to this table, the test-retest scores in the Reading and Listening sections were not significantly different (p → 0.05), but the total scores, Writing scores, and Speaking scores were all significantly different across the two testing sites (p ← 0.01).

Table 1: Statistics of test scores (n=77)

	Min	*Max*	*Mean*	*SD*	*Skewness*	*Kurtosis*
S1Reading	4.5	9.0	6.23	1.15	.41	-.58
S1Listening	4.5	9.0	6.01	1.00	.97	.93
S1Speaking	4.0	7.0	5.41	.611	.05	.03
S1Writing	4.0	7.5	5.46	.59	.59	1.96
S1Total	4.5	8.0	5.84	.76	.60	.54
S2Reading	5.0	9.0	6.28	1.02	.72	-.05
S2Listening	4.5	9.0	6.12	1.04	.96	.36
S2Speaking	4.5	7.5	5.80	.71	.44	.04
S2Writing	4.0	7.5	5.75	.58	-.37	1.32
S2Total	5.0	8.5	6.14	.79	.68	.35

Note: S1: Site 1, a city in mainland China; S2: Site 2, a city in a Southeast Asian country or Hong Kong

Table 2: Differences between test-retest scores from different testing sites (n=77)

		Paired differences							
			Std. error mean	*95% Confidence Interval of the difference*					
Mean		*SD*		*Lower*	*Upper*		*t*	*df*	*Sig. (2-tailed)*
Pair 1	S1Reading & S2Reading	-.05	.75	.09	-.22	.12	-.60	76	.55
Pair 2	S1Listening & S2Listening	-.10	.64	.07	-.25	.04	-1.42	76	.12
Pair 3	S1Speaking & S2Speaking	-.39	.57	.07	-.52	-.26	-5.98	76	.00
Pair 4	S1Writing & S2Writing	-.29	.65	.07	-.44	-.14	-3.94	76	.00
Pair 5	S1Total & S2Total	-.31	.36	.04	-.39	-.22	-7.35	76	.00

Table 3 shows the results of paired-samples correlation analyses. According to this table, the correlation between test-retest total scores and scores on the Reading and Listening sections were relatively high (r: 0.77–0.80), but the correlation between test-retest scores on the Writing and Speaking sections were relatively weaker (r = 0.39 and 0.63 respectively).

Table 4 and Table 5 present the results of generalisability analyses. According to Table 4, test-takers themselves, i.e., their language ability, explained a great majority of the total variance in Listening, Reading, and total scores of IELTS (76%–80%) but only explained 36% and 54% of the total variance in Writing and Speaking scores respectively. The remaining 64% and 46% of the

Table 3: Correlation between test-retest scores from different testing sites (n=77)

		Correlation	*Sig.*
Pair 1	S1Reading & S2Reading	.77	.00
Pair 2	S1Listening & S2Listening	.80	.00
Pair 3	S1Speaking & S2Speaking	.63	.00
Pair 4	S1Writing & S2Writing	.39	.00
Pair 5	S1Total & S2Total	.89	.00

variance in Writing and Speaking scores were caused by construct-irrelevant factors (i.e., site, person-site interaction effect, and error).

Table 4: Source of test score variance (n=77)

Source of variance	*Listening*		*Reading*		*Writing*		*Speaking*		*Total score*	
	Variance component	*% of total variance*	*Variance component*	*% of total variance*	*Variance component*	*% of total variance*	*Variance component*	*% of total variance*	*Variance component*	*% of total variance*
p	0.90	76%	0.84	80%	0.14	36%	0.27	54%	0.53	82%
i	0.00	0%	0.00	0%	0.04	10%	0.07	14%	0.05	8%
pi, e	0.28	24%	0.21	20%	0.21	54%	0.16	32%	0.07	11%
Total	1.18	100%	1.05	100%	0.39	100%	0.50	100%	0.65	101%*

Note: p: person, referring to test-takers; i: site; pi: person-site interaction; e: error

*Percentages of components variance do not always add up to 100% due to rounding.

Table 5 shows that the cross-site Speaking scores, and especially the Writing scores, had relatively low generalisability and dependability ($E\rho^2$ = 0.77 and 0.56 respectively, Φ = 0.70 and 0.52 respectively). By contrast, the cross-site Listening, Reading, and total scores had high generalisability and dependability ($E\rho^2$: 0.86–0.94, Φ: 0.86–0.90).

These results together suggested a high cross-site consistency on the Listening and Reading scores but a low cross-site score consistency on the Writing and Speaking scores. As for the cross-site total scores, although the correlation, generalisability, and dependability were high, the difference between them was significant.

Table 5: Generalizability ($E\rho^2$) and dependability coefficient (Φ) for test scores across testing sites (n=77)

Skills	$E\rho^2$	*Φ*
Listening	0.86	0.86
Reading	0.89	0.89
Writing	0.56	0.52
Speaking	0.77	0.70
Total	0.94	0.90

Discussion

Similar to Wilson (1987) who found that testing sites (inside or outside America) were significantly associated with TOEFL scores, this study also found that test-takers' achieved different total scores, and especially different Writing and Speaking scores, in different testing sites (inside or outside mainland China) within 30 days. In Wilson's study, since the test-takers repeated TOEFL within 1 to 60 months, it was not known whether the score difference between the two tests was caused by test-takers' actual change of language proficiency or other construct-irrelevant factors. However, in the current study, as all test-takers repeated IELTS within 30 days only, it could be speculated that such score inconsistency was probably caused by construct-irrelevant factors related to the difference between testing sites.

As the score difference between the two tests lay primarily in the Writing and Speaking sections, it was reasonable to suppose that the score inconsistency across testing sites was mainly related to writing- and speaking-specific factors such as rating reliability, which concerned leniency or severity of raters and the standardisation of the rating process. Lin and Chen (2020) found that test-takers' writing scores tended to be stable within a 6-month period. This finding further suggests that the inconsistency of IELTS Writing and Speaking scores across testing sites within 30 days was unlikely to be caused by the change of test-takers' actual writing and speaking ability.

It is also possible that the inconsistency of IELTS Writing and Speaking scores was caused by measurement error inherent in tests. Although any test scores contain measurement error (Cho & Blood, 2020), it seems that writing and speaking tests that involve qualitative evaluation of language production tend to contain a greater degree of measurement error than listening and reading tests that have only a single correct answer. Maybe due to this reason, the test-takers' Listening and Reading scores showed high consistency across testing sites.

Some people may argue that the difference in the IELTS total, Writing, and Speaking scores across testing sites was small (0.29–0.39) in the nine-band scoring system of IELTS. However, considering that the test-takers' English proficiency was already at the intermediate level, as shown by the mean scores, it was unlikely for the test-takers to significantly improve their language ability or scores within a short time. Moreover, due to the rounding up policy adopted by IELTS, a score change by 0.25 may lead to a half-band improvement in the final score, which may determine whether the test-takers can achieve the cut-score for university admission or other purposes. Therefore, the IELTS score inconsistency across testing sites found in this study may cause test unfairness and have a huge impact on test-takers.

Conclusion

The current study examined IELTS score consistency across testing sites as a means of exploring the fairness of large-scale international language tests. Multiple types of analyses of the 77 Chinese test-takers' IELTS scores gained from different testing sites within 30 days together showed a low cross-site score consistency, especially on the Writing and Speaking sections. Such inconsistency is likely to pose a threat to test fairness and validity. Test developers, therefore, need to further investigate the sources of such inconsistency by, for example, exploring test-takers' actual experiences of taking and retaking IELTS at different testing sites, and take measures accordingly to improve the reliability of test scores and safeguard test fairness and validity.

References

Barkaoui, K. (2017). Examining repeaters' performance on second language proficiency tests: A review and a call for research. *Language Assessment Quarterly*, *3*(4), 420–431.

Barkaoui, K. (2019). Examining sources of variability in repeaters' L2 writing scores: The case of the PTE Academic writing section. *Language Testing*, *36*(1), 3–25.

Blackhurst, A. (2004). IELTS test performance data 2003. *Research Notes*, *18*, 18–20.

Cho, Y., & Blood, I. A. (2020). An analysis of TOEFL primary repeaters: how much score change occurs?. *Language Testing*, *37*(4), 503–522.

Falvey, P., & S. D. Shaw. (2006). IELTS Writing: revising assessment criteria and scales (phase 5). *Research Notes*, *23*, 7–12.

Green, A. (2005). EAP study recommendations and score gains on the IELTS Academic Writing test. *Assessing Writing*, *10*, 44–60.

Lin, Y. M. , & Chen, M. Y. (2020). Understanding writing quality change: a longitudinal study of repeaters of a high-stakes standardized English proficiency test. *Language Testing,37*(4), 523–549.

O'Loughlin, K., & Wigglesworth, G. (2003). Task. design in IELTS Academic Writing Task 1: The effect of quantity and manner of presentation of information on candidate writing. In R. Tulloh (Ed.), *IELTS Research Reports 2003 Volume 4* (pp. 88–130). Retrieved from: www.ielts.org/-/media/research-reports/ielts_rr_volume04_report3.ashx

Wilson, K.M. (1987). *Patterns of Test Taking and Score Change for Examinees who Repeat the Test of English as a Foreign Language.* Research Report No. RR-87-03. Princeton: Educational Testing Service.

Zhang, Y. (2008). *Repeater Analyses for TOEFL iBT.* Research Memorandum 08-05. Princeton: Educational Testing Service.

'Broken Finnish': Speaker L1 and its recognition affecting rating in National Certificates of Language Proficiency test in Finnish

Sari Ahola
University of Jyväskylä, Finland

Mia Halonen
University of Jyväskylä, Finland

Abstract

As many European countries have language proficiency requirements for obtaining citizenship, language testing is a possible source of social inequality. The 'Broken Finnish' project has been set up to ensure test fairness by addressing the proficiency rating in the National Certificates of Language Proficiency (NCLP) test for Finnish in Finland, with a special focus on perceptions of pronunciation and 'accent' in relation to the examinee's L1 and how the raters recognise them. We explore if and how these perceptions affect the proficiency ratings. We are also interested in studying where the perceptions might arise from. In this paper, we present results from one L1 group: Thai speakers.

Introduction: Accent perceptions in societal gatekeeping

Like many European countries, Finland uses language proficiency requirements as one of the gatekeepers for citizenship. Consequently, language testing is a possible source of social inequality. To ensure the fairness of a test system, in the project *'Broken Finnish': Accent perceptions in societal gatekeeping* (Academy of Finland; 2018–2022, www.jyu.fi/hytk/fi/laitokset/solki/broken-finnish/in-english), we study the rating process in the National Certificates of Language Proficiency (NCLP) test in Finland, with a focus on perceptions of pronunciation in relation to the examinee's L1 and how the raters recognize it. We study whether there is any bias towards test-takers, and whether the correct recognition of the speakers' L1 affects the ratings of oral proficiency. We will focus on the ratings of Thai L1 speakers and show 1) how the correct recognition of the Thai speakers' L1 influenced their rating, and 2) how the raters described the performances of the speakers.

Context of the study: Finnish National Certificates of Language Proficiency (NCLP) test

The NCLP is a test system for adult second and foreign language learners overseen by the Finnish government's official system for language proficiency testing in a total of nine languages. Since 1994, there have been approximately 130,000 test-takers, and there are approximately 8,000 L2 Finnish test-takers per year.

The organizations responsible for the test system are Finnish National Agency for Education (EDUFI) and University of Jyväskylä (Centre for Applied Language Studies). It is independent of any syllabus or curriculum, but applies the Common European Framework of Reference for Languages (CEFR) (Council of Europe, 2001). NCLP has its own rating scale varying from 1–6 in four skills. Levels 3 and 4 form the so-called intermediate level which equals CEFR Levels B1and B2, which are the citizenship threshold in Finland; since 2003, CEFR Level B1 in Finnish or Swedish has been the language requirement for Finnish citizenship.

The focus of our project is the speaking test of Finnish at the intermediate level. The performances are rated along seven criteria for speaking (aligning with the CEFR): 1) general criteria (a holistic criterion covering six other, more analytical criteria), 2) fluency, 3) flexibility, 4) coherence and cohesion, 5) vocabulary (range, accuracy, idiomaticity), 6) pronunciation and phonological control, and 7) grammatical accuracy.

Data of the project

In order to explore whether the test-takers face any biased rating, we designed a data set focusing on test-takers of five different L1s: Arabic, Estonian, Russian, Thai, and Finland Swedish. Informed by previous research, we knew that the speakers of these languages face negative stereotyping in Finland (e.g. Jaakkola, 2009; McRae, Bennett, & Miljan, 1988; Reuter & Kyntäjä, 2005; Sirkkilä, 2005.) Our hypothesis was that, because of these general stereotypes, these groups might also face biased rating.

The data were collected through an online platform in 2015 and 2016. Altogether 50 (10 for each L1) examinees' speaking performances were rated by 44 certified raters. The performances were rated both holistically (the general criterion) and separately for the six analytical criteria presented above. The raters were also asked to write down the speaker's L1, to indicate how certain they were about the assumption, and to describe the speaker, the performance and the bases for their assumptions of the L1.

The three-fold interaction between L1, the various criteria and the L1 assumption (whether it was accurate, that is, 'recognized', or not) was studied using multi-faceted Rasch analyses. In addition to these statistical analyses done by the project's statisticians, we analyzed how the performances were described in the open answers.

Findings: L1 recognition affects on the rating of Thai L1 speakers

The results confirm the hypothesis in that the correct recognition of the L1 affected the rating in the case of Thai speakers. If the L1 was recognized, there was a significant decrease in the rating of pronunciation as well as a significant increase in the ratings of fluency and the general criterion. Furthermore, the other criteria seem to be affected along with recognition, but these changes were only tendencies, not reaching statistical significance. In this section, we will present some of the raters' reasoning for the recognition as well as descriptions of the affected skills, pronunciation, fluency and general holistic impression, drawing on Ahola's previous research (in press).

On one hand, the L1 Thai female speakers (n=8) were very well recognised by the raters. The recognition seemed to be based on the raters' experience in teaching and testing L1 female Thai speakers. On the other, L1 Thai male speakers (n=2) were not recognised by the raters but the suggested L1s included Russian, Somali and Arab speakers. The most probable reason for this bias in gender-related (non)recognition is simply that the raters do not have experience in teaching or rating L1 Thai male speakers because there are not many in Finland.

Negatively rated skill: Pronunciation

The criterion that was negatively rated, and thus rated lower when L1 was recognized, was pronunciation. The pronunciation was heavily criticized and the learning of comprehensible pronunciation was described as difficult or even impossible:

> For the speakers of these languages it (pronunciation) is so difficult to learn that they will never manage to raise their level any higher. One feels sorry for them.
>
> Source language can be strongly heard in the pronunciation. It requires attentiveness and patience from the listener and certainly repetition from the speaker to be understood in everyday situations.

The challenges in comprehension were described, for example, on the sound level. The raters described problems in producing consonants and consonant clusters (e.g. *r* = *l*, *-ts-* , *-st-*) which are typical problems of Thai language speakers and, more generally, speakers of tonal languages in Finnish (e.g. Aho, Toivola, Karlsson, & Lennes, 2016). In addition to sounds, reasons for comprehension challenges were put down to prosodic features, like high pitch or speech rhythm. Deviant prosody is known to easily reveal learners' L1 (Anderson-Hsieh, Johnson, & Koehler, 1992; Giles & Rakić, 2014).

> The Thai appear to find it difficult to produce speech. Stress is often on each word or individual words and there are pauses between words.
>
> Speech rhythm, intonation and phonology sound Asian throughout.
>
> Pronunciation also sounds naive, like little children's speech.
>
> A girly way to speak and intonation remind me of Thai speakers.

As can be seen in the last two extracts, the prosodic features were connected to the perceptions and stereotypes of Thai women. In these descriptions the raters describe more qualities of an imagined speaker than the performance itself.

Positively rated skills: Fluency and holistic impression (general criterion)

Fluency is connected to the amount of speech and the speakers' ability to fill the given time with speech as well as to the speech rate (e.g. Fillmore, 1979; Kormos & Dénes, 2004). This 'productivity' leads naturally to more output to be assessed. The Thai L1 speakers produced speech rather actively and there was more speech production compared to speakers of other L1s. This means that they were able to show their production fluency even though they had shortcomings in other language skills (Ang-Aw & Goh, 2011; May, 2006; Pollitt & Murray, 1996). The raters perceived and described this productivity in terms of fluency, which they also rated higher when they recognized the speaker as a Thai L1 speaker.

> Produces a lot of content. Difficult to assess – appearance of fluency but a lot of vagueness.
>
> Speech is fluent and there is a relative amount of speech production. Searches for words to an extent and the speech is partly listing items.
>
> Speaks nonstop without pausing to make things into logical wholes.
>
> The verb is often missing and the expressions are constructed by putting words together.
>
> A lot of speech together, nonstop. The speech is a little choppy structurally.

As we can see, perception of fluency does not depend on, for example, the perception of the level of proficiency of pronunciation, proven by the fact that fluency was rated higher while pronunciation was rated lower. Fluency is often compared to general proficiency or they are even used as synonyms for each other. This is seen in the behavior of the raters: they granted the Thai L1 speakers higher ratings in the general criterion, in their holistic impression of proficiency, than the ratings of the analytical criteria would have predicted.

Summary and implications

Our data showed that the recognition of the L1 of the speakers and the raters' perceptions of the speakers as a group had an effect on rating in the case of Thai L1 speakers. This is not an unexpected or surprising result as there is an extensive pool of research on rater bias based on speakers' background (see, e.g. Brennan & Brennan, 1981; Carey, Mannel, & Dunn, 2011; Cargile & Giles, 1997; Kang & Rubin, 2009; Lev-Ari & Keysar, 2010; Lindemann, 2005; Munro, 2003; Reid, Trofimovich, & O'Brien, 2019; Toivola, 2011).

Previous studies have often addressed so-called 'primed' ratings where the samples are preceded with, for example, real and fake photos of the (disguised) speakers and the raters are nonprofessionals in relation to language proficiency assessment. Our research differs from most of those in that it is done in a high-stakes test context and with trained raters. However, despite the education and experience in rating, the raters of our research are vulnerable to (possibly unconscious) stereotyping and the differences in their degree of linguistic awareness (Niedzielski & Preston, 2000; Preston, 1996).

Comprehensibility as a phenomenon is both a listener-specific and speaker-specific feature, and, unlike intelligibility, sensitive to various aspects independent of language or proficiency, like stereotyping and prejudices (see, e.g., Isaacs & Trofimovich, 2012; Munro & Derwing, 1995; Riney, Tagaki, & Inutsuka, 2005). In general, our results show that in addition to the formal criteria, there are some hidden, implicit or unconscious criteria: 'the raters' own' criteria, such as 'grammatical accuracy equals high proficiency', 'you cannot expect more of them', or a 'pity factor'. The rating seems to be dependent on the expectations of proficiency level, which seem to be set much lower than for many other L1s, e.g. for Estonians (see Ahola, 2020). These lower expectations, then, apparently connect to the image of the Thai L1 speakers living in rural peripheral areas as stay-home mothers and wives given less opportunities to learn Finnish. This is naturally a stereotyped view on the speaker group and does not apply to everyone, but according to statistics and previous research, this seems to be the case with this particular speaker group (e.g. Lumio, 2014; Shinyella, 2012; SVT, 2018)

No test is perfectly reliable, but it is obvious that in this kind of high-stakes test, there should not be any kind of bias involved in the rating process. However, it is also obvious that, where there are humans involved, there will also be emotions and attitudes. Now that we have more understanding of the bias, we are able to develop rater training further, also in relation to these more unconscious and sensitive areas, such as stereotyping. This will be implemented by giving the raters even more feedback on rater bias and raising awareness of rating behavior and hidden criteria, as well as training with more samples of different L1 speakers.

References

Aho, E., Toivola, M., Karlsson, F., & Lennes, M. (2016). Aikuisten maahanmuuttajien suomen ääntämisestä. *Puhe ja kieli, 32*(2), 77–96.

Ahola, S. (2020). Sujuvaa mutta viron kielen vaikutusta. *Virittäjä*, *124*(2), 217–242.

Ahola, S. (in press). Yleisten kielitutkintojen arvioijien käsityksiä thainkieliseksi tunnistettujen suomenoppijoiden suullisesta kielitaidosta, *Puhe ja Kieli*, *40*(4), 203–224.

Anderson-Hsieh, J., Johnson, R., & Koehler, K. (1992). The relationship between native speaker judgments of nonnative pronunciation and deviance in segmentals, prosody, and syllable structure. *Language Learning*, *42*(4), 529–555.

Ang-Aw, H. T., & Goh, C. C. M. (2011). Understanding discrepancies in rater judgement on national-level oral examination tasks. *RELC Journal, 42,* 31–51.

Brennan, R. L., & Brennan, D. J. (1981). Measurements of accent and attitude towards Mexican-American speech. *Journal of Psycholinguistic Research 10,* 487–501.

Carey, M. D., Mannel, R. H., & Dunn, P. K. (2011). Does a rater's familiarity with a candidate's pronunciation affect the rating in oral proficiency interviews? *Language Testing 28*(2), 201– 219.

Cargile, A. C., & Giles, H. (1997). Understanding language attitudes: Exploring listener affect and identity. *Language & Communication, 17*(3), 195–217.

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

Fillmore, C. J. (1979). On fluency. In C. J. Fillmore, D. Kempler & W.S-Y. Wang (Eds.), *Individual Differences in Language Ability and Language Behavior* (pp. 85–102). London: Academic Press.

Giles, H., & Rakić, T. (2014). Language attitudes: Social determinants and consequences of language variation. In T. M. Holtgraves (Ed.), *The Oxford Handbook of Language and Social Psychology* (pp. 11–26). Oxford: Oxford University Press.

Isaacs, T., & Trofimovich, P. (2012). Deconstructing comprehensibility: Identifying the linguistic influences on listeners' L2 comprehensibility ratings. *Studies in Second Language Acquisition, 34*(3), 475–505.

Jaakkola, M. (2009). *Maahanmuuttajat suomalaisten näkökulmasta. Asennemuutokset 1987–2007*. Helsinki: The City of Helsinki.

Kang, O., & Rubin, D. L. (2009). Reverse linguistic stereotyping: Measuring the effect of listener expectations on speech evaluation. *Journal of Language and Social Psychology*, *28*(4), 441–456.

Kormos, J., & Dénes, M. (2004). Exploring measures and perceptions of fluency in the speech of second language learners. *System*, *32*, 145–164.

Lev-Ari, S., & Keysar, B. (2010). Why don't we believe non-native speakers? The influence of accent on credibility. *Journal of Experimental Social Psychology*, *46*, 1,093–1,096.

Lindemann, S. (2005). Who speaks 'broken English'? US undergraduates' perceptions of non-native English. *International Journal of Applied Linguistics*, *15*, 187–212.

Lumio, M. (2014). Hymyn takana – thainmaalaiset maahanmuuttajat ja suomalais-thainmaalaiset avioliitot. In E. Heikkilä, P. Oksi-Walter & M. Säävälä (Eds.) *Monikulttuuriset avioliitot sillanrakentajina* (s. 36–51). Turku: Siirtolaisinstituutti.

May, L. A. (2006). An examination of rater orientation on a paired candidate discussion task through stimulated verbal recall. *Melbourne Papers in Language Testing, 11*(1), 29–51.

McRae, K. D., Bennett, S. E., & Miljan, T. (1988). *Intergroup Sympathies and Language Patterns in Finland: Results from a Survey*. Helsinki: Suomen Gallupin julkaisusarja.

Munro, M. J. (2003). A primer on accent discrimination in the Canadian context. *TESL Canada Journal, 20*(2), 38–51.

Munro, M. J., & Derwing, T. (1995). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning, 45*(1), 73–97.

Niedzielski, N. A., & Preston, D. R. (2000). *Folk Linguistics*. Berlin: De Gruyter Mouton.

Pollitt, A., & Murray, N. (1996). What raters really pay attention to. In M. Milanovic & N. Saville (Eds.), *Performance Testing, Cognition and Assessment. Selected Papers from the 15th Language Research Testing Colloquium* (pp. 74–91). Studies in Language Testing volume 3. Cambridge: UCLES/Cambridge University Press.

Preston, D. R. (1996). Whaddayaknow – modes of folk linguistic awareness. *Language Awareness*, *5*, 40–74.

Reid, K. T., Trofimovich, P., & O'Brien, M. G. (2019). Social attitudes and speech ratings: Effects of positive and negative bias on multi-age listeners' judgments of second language speech. *Studies in Second Language Acquisition 41*(2), 419–442.

Reuter, A., & Kyntäjä, E. (2005). Kansainvälinen avioliitto ja stigma. In T. Martikainen (Ed.), *Etnisyys Suomessa 2000-luvulla* (pp. 104–125). Helsinki: Finnish Literature Society.

Riney T. J., Tagaki, N., & Inutsuka, K. (2005). Phonetic parameters and perceptual judgements of accent in English by American and Japanese listeners. *TESOL Quarterly, 39*(3), 441–466.

Shinyella, T. (2012). *Kaksikulttuurista arkea suomalais-thainmaalaisissa lapsiperheissä. Selvitys suomalais-thainmaalaisten lapsiperheiden tilanteesta sekä erityispiirteistä, -tarpeista ja -haasteista.* Helsinki: Monikulttuuriyhdistys Familia Club ry.

Sirkkilä, H. (2005). *Elättäjyyttä vai erotiikkaa. Miten suomalaiset miehet legitimoivat parisuhteensa thainmaalaisen naisen kanssa?*. Jyväskylä Studies in Education, Psychology and Social Research 268. Jyväskylä: University of Jyväskylä.

SVT (2018). *Suomen virallinen tilasto: Väestörakenne*. Helsinki: Tilastokeskus.

Toivola, M. (2011). *Vieraan aksentin arviointi ja mittaaminen suomessa*. Helsinki: The University of Helsinki.

Adolescent test-taker characteristics: A qualitative validation study of the Singapore-Cambridge General Certificate of Education Ordinary-Level Chinese Language Examination reading paper

Yun-Yee Cheong
University College London Institute of Education, United Kingdom

Abstract

This qualitative validation study investigates the characteristics of adolescent test-takers sitting the reading paper of the Singapore-Cambridge General Certificate of Education Ordinary-Level Chinese Language Examination (GCE 1160). Specifically, evidence gathered through semi-structured interview and document analysis was organized around the following three issues: adolescence and adolescent literacy, reading motivation, interests and authenticity, and new forms of reading literacy. Although evidence suggested that the GCE 1160 reading examination is designed with adequate knowledge of adolescence and adolescent literacy, several threats to validity were identified, including the relatively low appeal of the passages to test-takers, and their relevance, authenticity, and sensitivity to new forms of reading literacy. Drawing on a socio-cognitive validation framework, the study analyses these threats to validity and outlines practical directions whereby policy makers and test designers might ensure validity in the reading examination.

Introduction

This qualitative validation study investigates the characteristics and needs of adolescent test-takers sitting the reading paper of the Singapore-Cambridge General Certificate of Education Ordinary-Level Chinese Language Examination (GCE 1160[1]). Specifically, evidence gathered through semi-structured interview and document analysis was organized around the following three issues: adolescence and adolescent literacy, reading motivation, interests and authenticity, and new forms of reading literacy. Whilst there are many test-taker characteristics that can be examined, the study highlights the defining characteristics of GCE 1160 test-takers, namely, adolescents in a highly modernized society with a rich tapestry of language and culture.

The following section begins with a brief overview of the background to this validation study. The research questions are also set out. This is accompanied by an introduction to Weir's (2005) socio-cognitive framework which was used to guide this study. Next is a section on research methods. The ensuing section presents and analyses the findings, leading into discussions on the directions in which reading examinations could develop in the future.

Background to the problem

Ethnic Chinese constitute the majority of the population in Singapore at 76.0% (Prime Minister's Office, Singapore Department of Statistics, Ministry of Home Affairs, Immigration and Checkpoints Authority, & Ministry of Manpower, 2019). Under the government policy of bilingual education, first adopted in 1966, English is taught as a first language in all schools and is also the medium of instruction for most subjects. At the same time, it is mandatory for most Chinese in Singapore to study Chinese as

[1] The subject code for the Singapore-Cambridge General Certificate of Education Ordinary-Level Chinese Language Examination was GCE 1162 prior to May 2016. After this date, the subject code was changed to GCE 1160. For ease of reference, GCE 1162 and GCE 1160 are referred to collectively as GCE 1160 in this study.

a second language (CL2) at primary and secondary school levels. Ethnic Singaporean Chinese, however, exhibit a full spectrum of language proficiency, from speaking Chinese as a first language to speaking Chinese as a foreign language (Chin, 2011; Tan, 2016). The Singapore Ministry of Education (MOE) has established that the percentage of Chinese students from English-speaking families has steadily increased since 1980, reaching 71% in 2019 (Chan, 2020). Studies imply that home language background and language proficiency are closely related, and the shift of home language from Chinese and Chinese dialects to English could lead to a decline in Chinese language standards (Guo, 2011; Tan, 2011).

The complex and fast-evolving Chinese language environment in Singapore has brought about major education reforms (Cheong, 2019). Whilst MOE has expended considerable effort to revamp the CL2 curriculum and associated pedagogies, national examinations have been less centre stage. Yet, to demonstrate that national CL2 examinations cohere with the characteristics and needs of Singaporean students, validation studies need to be carried out regularly.

In response to the need for ongoing validation studies, this research investigates the test-taker characteristics component of the reading paper of the GCE 1160 examination. Using guidelines from Cambridge Assessment English, the GCE 1160 examination (which is administered and scored twice a year in Singapore over May/June and October/November) has the largest number of test-takers among all Chinese language national examinations at secondary school level. In the examination, test-takers are assessed on all four language skills: reading, writing, listening and speaking. This study focuses on the reading paper of the examination, from this point referred to as the GCE 1160 reading examination, which carries the highest weighting of 35% of the entire GCE 1160 examination.

Three research questions (RQs) which define the scope of the study were identified:

(1) Is the examination supported by knowledge of adolescence and adolescent literacy?

(2) Does the examination appeal to the reading interests of Singaporean adolescents; is it relevant and authentic, paralleling the real-life needs of Singaporean adolescents?

(3) Does the examination take into account new forms of reading literacy?

Validity and Weir's socio-cognitive framework

Validity is the hallmark of measurement quality in testing (Koretz, 2008; Newton & Shaw, 2014). Among the many validation frameworks, Weir's socio-cognitive framework (Khalifa & Weir, 2009) has been extensively used across a range of validation studies, including studies conducted by Cambridge Assessment English for its suite of English as a second language examinations. The framework comprises six validation components, namely test-taker characteristics, cognitive parameters, contextual parameters, scoring, criterion-related, and washback and impact components, which are arranged in temporal sequence according to the stages in an examination cycle. As addressing all six components in the framework is impossible in a single study, this study focuses solely on test-taker characteristics, whilst noting its relation with other components in the framework.

Research methods

The study draws on semi-structured interviews and document analysis. Across the pilot and main studies, a total of 22 stakeholder interviewees were selected using a purposive maximum variation sampling technique (Patton, 2015). The interviewees included an elite policy group with privileged access to the detailed test specifications and procedures. Interviews were also carried out with secondary school CL2 teachers and students, whose perspectives are seldom considered in validation processes. In addition, opinions were sought from experts in the field of CL2 reading and assessment.

The resultant dataset provided a rich and informative collection of stakeholder views and opinions about the test-taker characteristics of the GCE 1160 reading examination, 11 hours of recording, and nearly 5,800 lines of transcription. The transcribed data were read and tagged using NVivo10, guided by a list of codes. Subsequently, findings from the semi-structured interviews were compared with those gleaned from the analysis of documents. Documents examined are held or published mainly by the MOE, the Singapore Examinations and Assessment Board (SEAB) and Cambridge Assessment English. The selection of documents included recent Chinese language and mother tongue languages reviews, reports, press releases, speeches, and SEAB presentations.

This study abided by the *British Educational Research Association Ethical Guidelines for Educational Research* (British Educational Research Association, 2018). All necessary measures were taken to ensure that the study was conducted in an ethically defensible manner.

Findings and discussion

The validity evidence collected through semi-structured interview and document analysis centres on adolescence and adolescent literacy, reading motivation, interests and authenticity, and new forms of reading literacy, as guided by the three RQs.

Adolescence and adolescent literacy

According to education theorist Chall (1996), individuals often advance through a series of stages in reading development. Normally by the end of secondary education (15–17 years old), progressing adolescents should be able to purposefully extract and interpret information from a variety of fiction and non-fiction texts to learn new ideas, to gain new knowledge and to experience new feelings. They should also begin to recognize that texts embrace multiple viewpoints and be able to discern differences in perspective.

Evidence from the study revealed, however, that a significant percentage of students are unable to exhibit these key skills when sitting the GCE 1160 reading examination. Interviewees observed that there are three main types of struggling test-takers. There is a group of test-takers who can read the examination passages with reasonable speed and accuracy. They, however, lack the vocabulary and higher-order thinking skills needed to infer beyond the literal meaning of the passages and to answer the more challenging items. Next, there are test-takers who can decode some of the characters (字) and words (词) but lack the required fluency to complete the GCE 1160 reading examination within the stipulated one and a half hours. As a result, most of their attention and time is spent on character and word identification at the expense of comprehension. Last, there are the weakest test-takers who have never successfully passed through the decoding stage. Reading is slow and halting, characterized by frequent stops at unfamiliar characters. Their reading level is several years below their grade placement and completing the items is extremely difficult for them.

The problem, then, is twofold. On one hand, there must be items eliciting higher-order thinking skills that are critical to adolescents in the 21st century; on the other hand, there have to be enough items for the less able students to attempt. Put differently, a delicate balance must be maintained among test items of varying difficulty. This need should be clearly exhibited in the test specifications.

Reading motivation, interests and authenticity

The second issue concerns motivation and authentic assessment. Evidence from the study indicated that most Singaporean adolescents are not energized or activated toward reading in the Chinese language, devoting very little time and effort to it, as they do not perceive reading in the Chinese language as a vital aspect of their daily lives or their future. Furthermore, the evidence gathered revealed that the main barriers to reading are that Singaporean adolescents nowadays tend to be occupied with homework, co-curricular activities and most of all, screens – Internet, mobile phone applications, games, and Instagram, to name a few.

Reading requires an effort that initially many adolescents are reluctant to make. It is only when the literacy needs and reading interests of adolescents are understood that reading examinations can be designed to be worth teaching to. Most interviewees agreed that the GCE 1160 reading examination could benefit from a wider range of item types, such as information transfer, matching headings, summarizing and comparing multiple texts, forming a closer alignment with the real-world reading needs of adolescents. Interviewees also spoke of the possibility of 'repackaging', i.e. giving the GCE 1160 reading examination a modern revamp. Passages that are more relevant and relatable, and therefore potentially more appealing to adolescents, could be introduced. Suggestions put forth by the interviewees include extracts from canonical texts, lifestyle articles, expository texts on science, geography and history intended for general readers, and texts on contemporary culture including film, art and literature.

New forms of reading literacy

The third issue is the relationship between adolescents and new literacies. Global economies, new technologies and exponential growth in information are rapidly transforming the world. Central to this shifting landscape is the appearance and spread of the Internet. In Singapore, Internet usage is so prevalent that 88.4% of Singaporeans now use the Internet (Internet World Stats, 2020).

The meaning of literacy has evolved with the widespread use of the Internet (Coiro, Knobel, Lankshear, & Leu, 2008). To have been literate yesterday, in a world defined primarily by relatively static book and print technologies, does not guarantee full literacy today in an online age of information and communication. Policymakers, however, have not yet fully considered the implications that Internet technologies have for testing and assessment.

Evidence collected in this validation study suggested that online reading was thought to place greater demands on critical thinking and analysis than traditional offline reading as adolescents need to evaluate the level of accuracy, reliability and information bias. In addition, the act of reading on the Internet is perceived as a more active process than traditional offline reading. Students often read on the Internet to solve problems and answer questions. Initiated by a specific purpose, they sift through disparate sources to locate the information that meets their needs. Last, many interviewees pointed out that reading online is often a more collaborative and integrated process. When adolescents engage in online reading and research, they usually work collaboratively or solicit help from others online. Unfortunately, these distinctions between online reading and traditional offline reading are rarely captured by traditional reading assessments, including the GCE 1160 examination.

Conclusion

In conclusion, a deeper understanding of potential test-taker characteristics and needs should be established in the conceptualization phase of an examination, failing which, policy makers and test designers risk facing threats to validity. Several threats to the validity were identified in the study, including the relatively low appeal of the passages to test-takers, and their relevance, authenticity, and sensitivity to new forms of reading literacy. Questionnaires and information sheets could be developed and used alongside the GCE 1160 reading examination to gather valuable information about test-takers, such as their reading exposure, habits and strategies. Feedback from test-takers on the examination could also be elicited through post-examination surveys, focus groups and protocol analysis, and used in modifying passages and items where necessary.

It is important to recognize the following two caveats that affect the validation study. The first caveat acknowledges that different evaluators might well reach different judgements, even on the basis of the same corpus of evidence and analysis (Newton, 2017). The second caveat is that the impact of a high-stakes national examination is widespread, therefore, reaching a definitive verdict of sufficient validity is necessarily a collective public responsibility. While this in-depth qualitative research has examined the views and opinions of an important group of stakeholders, the study should be complemented by validation studies of a larger scale.

Acknowledgements

The author gratefully acknowledges the helpful comments and suggestions of Professors David Scott, Paul Newton, Catherine Snow and Ms Katharine Carruthers OBE in preparing this paper.

References

British Educational Research Association. (2018). *Ethical Guidelines for Educational Research*. Retrieved from: www.bera.ac.uk/publication/ethical-guidelines-for-educational-research-2018-online

Chall, J. S. (1996). *Stages of Reading Development*. Texas: Harcourt Brace.

Chan, M. (2020, January 3). English, mother tongue and the Singapore identity. *The Straits Times*. Retrieved from: www.straitstimes.com/opinion/english-mother-tongue-and-the-spore-identity

Cheong, Y. Y. (2019). Singapore's Chinese language education and assessment policy. *Journal of Chinese Language Education* [华文学刊], *17*(2), 1–33.

Chin, C. K. (2011). *Chinese Language Curriculum and Pedagogies of Singapore*. Nanjing: Nanjing University Press.

Coiro, J., Knobel, M., Lankshear, C. & Leu, D. J. (Eds.). (2008). *Handbook of Research on New Literacies*. Mahwah: Lawrence Erlbaum.

Guo, X. (2011). Chinese language teaching in Singapore: Variations and objectives [华文教学在新加坡]. *Journal of Chinese Language Education* [华文学刊], *9*(1), 1–16.

Internet World Stats. (2020). *Usage and population statistics*. Retrieved from: www.internetworldstats.com/stats.htm

Khalifa, H., & Weir, C. J. (2009). *Examining Reading: Research and Practice in Assessing Second Language Reading*. Studies in Language Testing volume 29. Cambridge: UCLES/Cambridge University Press.

Koretz, D. (2008). *Measuring Up: What Educational Testing Really Tells Us*. Cambridge: Harvard University Press.

Newton, P. E. (2017). *An approach to understanding validation arguments*. OFQUAL report. Retrieved from: assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/653070/An_approach_to_understanding_validation_arguments.pdf

Newton, P. E., & Shaw, S. D. (2014). *Validity in Educational and Psychological Assessment*. London: Sage Publications.

Patton, M. Q. (2015). *Qualitative Research & Evaluation Methods: Integrating Theory and Practice*. California: Sage Publications.

Prime Minister's Office, Singapore Department of Statistics, Ministry of Home Affairs, Immigration and Checkpoints Authority, & Ministry of Manpower. (2019). *Population in Brief 2019*. Retrieved from: www.strategygroup.gov.sg/files/media-centre/publications/population-in-brief-2019.pdf

Tan, C. L. (Ed.). (2011). *From Practice to Practical: Teaching and Learning of Chinese as a Second Language*. Nanjing: Nanjing University Press.

Tan, C. L. (2016). The present: An overview of teaching Chinese language in Singapore. In K. C. Soh (Ed.), *Teaching Chinese Language in Singapore: Retrospect and Challenges* (pp. 11–26). Singapore: Springer.

Weir, C. J. (2005). *Language Testing and Validation: An Evidence-based Approach*. Basingstoke: Palgrave Macmillan.

A ‘netnographic’ study of test impact from the test-takers’ perspective: The case of a translation test

Chengyuan Yu
University of Macau

Cecilia Guanfang Zhao
University of Macau

Abstract

This study investigates the perceived fairness of a translation test in China from the test-takers’ perspective, focusing specifically on test-takers’ perceptions and interpretation of the (intended or unintended) use of the test, as well as the difficulties associated with test taking and preparation. Following a netnographic approach (Kozinets, 2010), naturalistic and unobtrusive data from online postings published by prospective and past test-takers are inductively analyzed. Results revealed that many test-takers reported taking and using the test for purposes beyond translator certification. Challenges included the lack of access to the scoring rubrics, specific guidance on the use of official preparation materials, and sufficient test-taking time. This study can inform the development of the translation test in question, and points to the importance of transparency and information accessibility in enhancing perceived test fairness and the need for further validation of translation tests to bring about a positive impact on language learning.

Introduction

The testing of translation has been a long-standing practice that can be dated back to 1913, according to Weir (2005), when translation tasks were part of the Cambridge Assessment English Certificate of Proficiency in English (CPE), the first contemporary English as a second/foreign language (ESL/EFL) test. Although translation tasks are still in use on some ESL/EFL proficiency tests and locally-designed translation tests abound, research on the testing of translation ability is scarce. While various models of translation assessment have been proposed by translation scholars (e.g. Fawcett, 2000; House, 2015; Leuven-Zwart, 1989), the validity of such models is often challenged, even within the profession (Karoubi, 2016). Some translation scholars, for example, question the reliability, validity, and usefulness of existing translation assessments, and call for an urgent collaboration between translation scholars and language testers to ensure quality assessment based on empirical evidence (e.g. Angelelli & Jacobson, 2009; Waddington, 2004; Williams, 2001).

As a response, this study empirically investigated the (mis)use and impact of a translation test, China Accreditation Test for Translators and Interpreters: Translation category (CATTI-T). Administered by the Ministry of Human Resources and Social Security of China, CATTI-T certifies prospective and in-service translators and is taken by over 200,000 people annually. Of the four levels and seven language pairs in the CATTI-T suite, the present study focused on the English-Chinese translation test at the two lower levels, i.e., Level II and III, as they attract and affect more than 90% of the CATTI-T test-taker population. Specifically, the study examined test impact and fairness (Kunnan, 2018) from the test-takers’ perspectives, focusing on test-takers’ reported (mis)uses of the test and their test taking and preparation experiences and practices.

Two specific research questions (RQs) are proposed:

(1) What are test-takers’ reported or observed localized uses of CATTI-T in China?

(2) What are the reported challenges associated with and practices adopted for test taking and preparation?

Methods

Netnography and the dataset

This study employed a 'netnographic' (a coined term combining net and ethnographic) approach that is known as a special ethnography for online communities (Kozinets, 2010). Different from traditional qualitative research methods that collect data in the forms of interviews, field notes, or journals, all the data utilized in a 'netnography' are naturalistic, unobtrusive, and captured from online forums. The data of the present study were collected from postings published by prospective and past test-takers in a one-year period on two large online forums in China, i.e. Baidu Tieba and WeChat Official Account, where CATTI-T test-takers can share experiences, ask questions, exchange information, and express opinions about the test. For each post shared on these platforms, viewers could also write comments and responses. Such data were thus interactive, featuring communication among prospective test-takers, test-takers in the process of 'last-minute' test preparation, and previous test-takers who had already been certified. The final dataset, including all relevant online posts and viewer comments, contained 106,373 Chinese characters (with some occasional use of English terms) and was imported into NVivo 11 for analysis.

Data analysis

Data were coded by the first author following an iterative inductive coding paradigm that features both descriptive coding and pattern coding (Saldaña, 2013). During the first cycle, data were read and reread carefully to arrive at relevant descriptive codes that correspond to the research foci presented in the two RQs. Basically, the descriptive coding condensed the large body of qualitative data into descriptive labels in the form of single words or short phrases. For example, the codes corresponding to RQ1 included phrases that summarized test-takers' reasons for taking CATTI-T and the different uses of the CATTI-T certificate. The first-cycle descriptive coding resulted in an inventory of descriptive codes. For the second cycle of coding, all the descriptive codes and the text that the codes summarized were read again and compared with one another to be categorized into a smaller number of analytic units. The product of the second-cycle coding was two short lists of categories that can be used to answer the RQs. Following Mackey and Gass's (2016) suggestion, data were re-coded two months later to improve the rigor of data analysis, leading to the combination, relabeling, and adjustment of the first-round coding results. To avoid any possible bias in the coding process, the coding results were also shared and discussed with the second author, which leads to some further re-labeling both authors eventually agreed upon.

Results

Reported test uses

Four categories of reported test uses were identified through the two cycles of coding. The first category is *to find good jobs*. As an accreditation test, most test-takers claimed that their purpose of taking CATTI was to find a well-compensated job and secure a career in the translation industry. They believed that the certificate is a valuable 'door knocking' tool in the job market, as can be seen in the following posts.

> A CATTI Level II certificate can help to catch HR's eyes in some translation companies. (Baidu 34)

> I have graduated for three years without a well-paid full-time job. I hope this certificate could save me. (Baidu 34)

Not only was the CATTI-T certificate used for job hunting in the translation industry, but it is also interesting to note how prospective test-takers believed it can be used for job seeking in general. As reflected in the quote below, some commented on how, as a government-acknowledged certificate, CATTI-T may give people an edge when they apply for jobs in state-owned or government entities, regardless of whether the job is translation-related.

> If you plan to get employed in state-owned companies or government departments, not necessarily to do translation work, this certificate could help, according to some official documents. But I don't know if it is also the case in your city. (Baidu 2)

The second major category for reported test use is *to fulfill personal interest*. Some test-takers do not intend to enter the translation industry at all, have jobs in other fields, and do not intend to use the certificate to find them a good job in any field. They reported that they took CATTI-T merely for personal interest. For example, one clinician who passed the test said that '[he] simply enjoyed the process of learning about translation and didn't know what [he] could do with the certificate or if [he] could use this certificate to make money' (WeChat 12). Another certificate holder mentioned that 'doing translation is my hobby and he passed without any preparation' (WeChat 1).

The third category is *to demonstrate general English proficiency*. Many test-takers believed that translation is an advanced language skill. In their opinion, the CATTI-T certificate has more value than certificates of other English language tests in China, and it is also more affordable than major international English proficiency tests:

> This test is cheaper to take than such internationally-recognized tests as IELTS and TOEFL. If you want to demonstrate your language proficiency, this can be a good choice. (Baidu 2)

> Having a certificate of translator sounds great! It could prove that you are good at English and it speaks louder than College English Test (CET)-6 that most [Chinese students] have. (Baidu 2)

The fourth main reported category of test use is to *promote learning through the test*. Many English-major test-takers believed that preparing for CATTI-T can push them to practise their translation ability, because translation tasks are used on various testing occasions, for example, the postgraduate entrance examination.

> I registered for this test to push myself to practise translation because I am preparing for the postgraduate entrance examination for an MTI program. (Baidu 21)

In summary, many test-takers reported taking and using the test for purposes beyond translator certification, including using it as proof of their general English proficiency, taking it out of personal interest, for job seeking in general, or to promote learning through the test.

Challenges for test preparation and test taking

Analysis of data revealed four main types of challenges associated with test preparation and test-taking, as reported and shared by prospective and past test-takers on the net. The most commented-on among test-takers is the lack of opportunities to learn about and prepare for the test due to various reasons. One such reason test-takers noted was the lack of access to explicit evaluative criteria:

> This is the fourth time that I took the test. My score for the practice task has gradually increased for the last three times. I scored 56 last time, but I only got 30 this time. I don't know what happened. (Baidu 43)

> I think it is important to have separate scores for English-Chinese and Chinese-English translation respectively. Otherwise, I won't know how my translation is scored . . . Maybe the raters are not the same kind of creatures as us. I think I did a good job this time. But I failed again. (Baidu 47)

The insufficient information about scoring criteria made it difficult for test-takers to properly prepare for the test. The difficulty and format of the practice tasks in the official textbooks were perceived by many as different from the actual test tasks. Furthermore, the quality of the practice tasks is regarded by many as a problem. The large volume of complaints regarding the textbooks suggest that they sell well, despite the perceived problems.

> The textbooks are too difficult. I can assure you what you will encounter in the real test will be much easier in terms of vocabulary and grammar . . . I would like to remind you that there are some mistakes in the official textbooks. (WeChat 3)

> I regret to buy the official textbooks. The task format in it is totally different from what's in the real test. (WeChat 8)

Some test-takers even speculated that this test might be norm-referenced, rather than criterion-referenced as specified in the official test syllabus. As one complained on Baidu, 'maybe there are too many test-takers this time. Even if I think it's easier this time, they still allow only a certain number of test-takers to pass' (Baidu 43).

In addition to the lack of *opportunities to learn* (Kunnan, 2018), many test-takers also complained about the *insufficient test-taking time* as a particular challenge, as evidenced by the various strategies used by test-takers in reaction to the insufficient time.

> If you are better at Chinese to English translation, you should work on this part first. The time allotted is very limited. I think you shouldn't let yourself lose points on what you are good at simply because of insufficient time. (WeChat 11)

> To make sure that you can finish the test, it is recommended that you take particular note of the time when you practice. (WeChat 17)

> When you decide to take CATTI Translation, the most important thing you need to keep in mind is to keep practicing. You need to allocate at least three hours every day to mock testing practice to automatize your test-taking strategy. (WeChat 5)

Participating in online courses and guest lectures was perceived as useful by some test-takers, which suggests potential unfairness due to the unequal distribution of such resources and the potential extra economic burden. The test-takers reported:

> You can take some online courses and try to remember some techniques. (Baidu 6)

> I found the online courses very good. It taught me how to treat translation flexibly. (Baidu 74)

Many test-takers mentioned *rote memorization of chunks* that are rarely used in real communication because the actual test has a pre-test task that focuses on measuring general English proficiency in the form of multiple-choice questions.

> It is not a waste of time to memorize so many synonyms because you will encounter them in the pre-translation section. For those who don't have a large vocabulary size, such a task can be quite difficult. (Baidu 81)

Discussion and conclusion

Focusing on a less researched type of language test and employing a unique 'netnographic' approach, this study examined the local uses and impact of a translation test from the test-takers' perspective based on unobtrusive digital data. Test-takers reported misuses of the test, criticisms of the lack of access to various resources and information, and inappropriate test preparation practices that have important implications for language assessment developers, users, and researchers.

First, findings from the study highlight the importance of transparency and access (e.g., intended test use, scoring rubrics, and official materials) in test administration and use. Regardless of the socio-economic status of test-takers or their institutions, adequate and equal opportunities for learning and appropriate access and administration need to be ensured to enhance test validity, promote positive impact, and ultimately establish test fairness and justice (Kunnan, 2018). Second, the test-taker perspective adopted in this study also reminds us that involving test-takers in test design is an important step in forming a cyclic communication loop among test developers, users, and other stakeholders (e.g. Hamp-Lyons, 2000). Test-takers' involvement and input could not only provide validity evidence and improve the technical quality of a test, but it could also uncover otherwise overlooked misuses-in-context and malpractices that may jeopardize the intended purposes and benefits of having a test in the first place.

References

Angelelli, C. V., & Jacobson, H. E. (2009). Introduction. In C. V. Angelelli & H. E. Jacobson (Eds.), *Testing and Assessment in Translation and Interpreting Studies: A Call for Dialogue between Research and Practice* (pp. 1–9). Amsterdam: John Benjamins.

Fawcett, P. (2000). Translation in the broadsheets. *The Translator, 6*(2), 295–307.

Hamp-Lyons, L. (2000). Fairness in language testing. In A. J. Kunnan (Eds.), *Fairness and Validation in Language Assessment Selected Papers from the 19th Language Testing Research Colloquium, Orlando, Florida* (pp. 30–34). Studies in Language Testing volume 9. Cambridge: UCLES/Cambridge University Press.

House, J. (2015). *Translation Quality Assessment: Past and Present*. New York: Springer.

Karoubi, B. (2016). Translation quality assessment demystified. *Babel: International Journal of Translation, 62*(2), 253–277.

Kozinets, R. V. (2010). *Netnography: Doing Ethnographic Research Online*. California: Sage Publications Ltd.

Kunnan, A. J. (2018). *Evaluating Language Assessments*. Abingdon: Routledge.

Leuven-Zwart, K. (1989). Translation and original: Similarities and dissimilarities, I. *Target: International Journal of Translation Studies, 1*(2), 151–181.

Mackey, A., & Gass, S. M. (2016). *Second Language Research: Methodology and Design* (2nd ed.). Abingdon: Routledge.

Saldaña, J. (2013). *The Coding Manual for Qualitative Researchers* (2nd ed.). California: Sage Publications Ltd.

Waddington, C. (2004). Should student translations be assessed holistically or through error analysis?. *Lebende Sprachen*, *49*(1), 28–35.

Weir, C. J. (2005). *Language Testing and Validation: An Evidence-based Approach*. Basingstoke: Palgrave Macmillan.

Williams, M. (2001). The application of argumentation theory to translation quality assessment. *Meta*, *46*(2), 327–344.

Predicting item difficulty in the assessment of Welsh

Emyr Davies
CBAC-WJEC, Wales

Abstract

Ensuring that items in a test are at an intended level of difficulty is challenging, a challenge which is greater when testing small numbers of candidates. Item writers must include different aspects of a construct and aim for an appropriate balance of more/less difficult items. When testing languages with fewer test-takers, predicting how items will perform is important, as opportunities for pretesting are limited. A small-scale research project was undertaken to determine how well a group of examiners and item writers could predict the difficulty of items in receptive tests of Welsh at B1 and B2 levels, using item-level data from live past papers as a baseline. The panel was asked to estimate the difficulty of 320 items in total. They were also invited to a workshop to discuss why certain items had not performed as predicted and to infer which factors made them difficult to predict. The paper will present the results of the analysis.

Introduction

This paper summarises a research project initiated in 2019. The purpose was to explore how accurately examiners and item writers can predict the difficulty of items in language tests, i.e. how accessible the items are. Ensuring that test items are of an intended difficulty is a challenge for any test provider, and our ability to predict the accessibility of items directly impacts the reliability and quality of our tests. If we can better predict how candidates will cope, and how individual items will perform, our assessments will be more consistent across time. Another reason for exploring this topic was to identify the factors which affect item difficulty, including linguistic and cognitive factors, and to understand why certain items were less predictable. This in turn would help prepare candidates, improve resources and train future item writers.

In any test, a balance between easy, moderate and difficult items is needed – at task level, paper level (assuming that skills are tested separately) and across papers (if these are combined to give one final score or grade). This is achieved in different ways in the Welsh tests, which are the focus of this study. Items are scrutinised by a team of experienced examiners whose judgements are captured in a standard-setting exercise. Items are then pre-tested, i.e. candidates who are at the appropriate level take tests under the same conditions and the results are analysed and compared with their performance in a live test. Items may then be accepted, revised or discarded. In the case of the Welsh tests, pretesting of items occurs three years before their use in live papers. If items then behave differently in a live test to how they were predicted to behave, then they can be deleted or grade boundaries changed to account for unexpected levels of facility. Post-exam analysis of item-level data often shows that items behave in unexpected ways, despite scrutiny and pre-testing. From a test developer's point of view, the less post-exam adaptation required, the better. It is also better in terms of the candidate experience and perception, i.e. that they feel that the tests themselves are no more difficult or easy than those taken in the past. In the case of Welsh, the number of candidates is small and the number of candidates willing to undertake pre-tests is consequently smaller. This necessarily limits the value of statistics obtained through pretesting. Therefore, we must rely to a greater degree on expert judgement, and the ability of experts to predict difficulty takes on greater importance.

The main research question was: how well can item writers and examiners predict the level of difficulty in passive tests, i.e. how accessible items in Reading, Listening and Knowledge of Language tests would be. Related questions were also explored, e.g. Which factors affected item difficulty? Were different question formats, different skills or different level tests more difficult to predict than others? Were some people better predictors of difficulty than others? Would training improve predictions? How could any findings be used to improve exam development, resources or teaching?

Literature review

There are not many research studies focusing specifically on this topic. Sydorenko (2011) examined how well item writers judged the difficulty of items against actual difficulty. This was based on Viewing comprehension tests of Russian learners in a university context in America. The number of candidates and items was small but here analysis showed that 'the correlation between actual item difficulty and predicted difficulty level was low' (Sydorenko, 2011, p. 49). The item writers in this study (in this study at least) were not good predictors of how their items would perform. Loukina, Yoon, Sakano, Wei, and Sheedan (2016) examined how analysing the features of a text in an English language Listening test could be used to predict how items would perform, rather than expert judgement. They found that software tools which analysed text complexity were more successful as predictors of item difficulty than test developers' 'intuition'. Beinborn, Zesch and Gurevych (2014) also looked at applying automatic prediction of difficulty in c-tests (a type of cloze test) and found that the model used performed similarly to human judges. Such text analysis tools are not yet available for Welsh, but if developed, could be used to underpin expert judgement in future.

Method

A suite of tests at different levels is available for adult learners of Welsh, from A1 to B2.

The method used for this study focused specifically on two levels (B1 and B2) of Reading, Listening and Knowledge of Language tests. The reasons for excluding the A1 and A2 tests were that candidates taking the A1 tests tend to be at a higher level than that required of the test and therefore there was not a sufficient range of ability for the purposes of this study. The A2 tests were undergoing revision at the time and were therefore excluded. The item-level data from the B1 and B2 live tests sat in 2015, 2016, 2017 and 2018 were collated, i.e. actual data about how items had performed in the live tests (data which were not in the public domain). Then, a panel was invited to participate in the study, including item writers, examiners, markers of the relevant tests and other 'non-experts' who had some experience in teaching – 13 in total.

There were three stages to the study:

- asking the panel to estimate the difficulty of 160 items (this was done remotely)
- attending a workshop in person to get generalised feedback from the first stage and a detailed discussion of the factors affecting items whose difficulty was least well predicted
- asking the panel to estimate the difficulty of a further 160 different items (also done remotely).

One of the research questions was to establish whether the workshop training would improve predictions. Also, panellists were asked to estimate their own confidence levels in their judgements for each item. There were 320 different items used in total and participants were asked to estimate difficulty on a five-point scale for each item:

1. Very difficult – 0–20% of candidates would get the item correct.
2. Fairly difficult – 20–40% of candidates would get the item correct.
3. Medium level of difficulty – 40–60% of candidates would get the item correct.
4. Fairly easy – 60–80% of candidates would get the item correct.
5. Very easy – 80–100% of candidates would get the item correct.

Items were mostly dichotomous, and the small number of partially correct answers (in the Reading tests) were not accounted for.

In these tests, the item-level data following use as 'live' exams is routinely analysed using a number of measures, including the 'facility factor', which is the percentage of candidates giving a correct response. Therefore, WJEC researchers could correlate the panellists' predictions with the observed facility of each of the 320 items. Participants were asked to confirm that they did not remember the items in question nor any analysis associated with them, so that we could be confident that they were not influenced by previous knowledge or experience. As an indication of the number of candidates involved in the live tests, these varied between 200 and 250 for B1, and between 60 and 100 for B2.

Summary of findings

The participants' ability to predict difficulty varied. Generally, they were very good at predicting items at both ends of the range, i.e. items which were very difficult and very easy. Experienced examiners and markers were better predictors of item difficulty, as might be expected. There was a general tendency to overestimate the level of difficulty, i.e. items were estimated to be more difficult than they actually were. Predictions improved after the workshop, but predictions made by some participants did not improve or even became less accurate. Analysis of the confidence levels of participants did not show any significant results: those who lacked confidence at Stage 1 also lacked confidence at Stage 3, and the most confident were not necessarily the best predictors of difficulty. Generally, multiple-choice items were more difficult to predict; Listening test items were also more difficult to predict than items on the Reading and Knowledge of Language tests.

There was an abundance of data for analysis. 14 different papers were used, a total of 320 individual items and 10 panellists (excluding the three 'non-experts'). Using Spearman's rank correlations we had 140 different correlations between expert predictions and candidate performance. These had a mean value of 0.26, a standard deviation of 0.25 and varied from -0.51 to 0.82. When we pooled the expert judgements and took the correlation between the mean expert response and candidate performance, we had 14 values (one for each paper) with a mean of 0.45 and a standard deviation of 0.26 with values varying from -0.02 to 0.91.

Factors affecting difficulty

At the workshop, participants were asked to consider which factors may have affected the behaviour of individual items which had not been predicted accurately. These items were selected across different skills, levels and item formats. Participants were asked to discount affective and other factors for the purposes of this study, i.e. those beyond the control of item writers, including:

- literacy and ability of candidates (including speed of reading and writing)
- motivation and attitude to the test
- physical factors, e.g. tiredness or illness
- amount of preparation by the candidate, including exam technique, timing, etc.
- factors related to particular requirements, e.g. hearing or vision.

The Knowledge of Language tests were more easily controlled and were found to be a more predictable measure of candidates' knowledge of vocabulary, collocations, grammar, etc. Many more factors were considered by participants to have affected candidates' ability to answer in the Reading and Listening tests, and an initial shortlist identified nearly 50 factors, including linguistic, textual, 'content' and cognitive factors, which could be found to affect item accessibility. These categorisations are not definitive, and it could be argued that some belong to more than one category.

Linguistic factors included: vocabulary; syntax; register; dialect, colloquial or literary forms; use of English loan words and code switching; and key words hidden by initial consonant mutation.

Textual factors included: sentence or text length; clarity of structure; 'signposting' of answers in text; and repetition or paraphrasing of key information.

'Content' factors included: density of information; topic familiarity; abstract or academic nature of content; redundant information; and need for background or cultural knowledge.

Cognitive factors included: needing to infer answers; ambiguity in questions; distracter strength; 'cognitive load' (having to remember or reason the answer); and predictability of the answer.

Further factors affecting Listening tests included: speed of delivery; scripted or unscripted text; turn-taking or interrupted discourse; background noise; time to process the answer; tone, pitch and clarity of speakers; amount of filler or redundant language; and whether keywords could be isolated above the 'noise'.

This list is by no means exhaustive and there is considerable overlap. It was clear that for each item, multiple factors were at work and it was often unclear which factors were making certain items difficult to predict. For example, a keyword in the Welsh text could be less identifiable because of initial consonant mutation, and may not be repeated or paraphrased, or not emphasised clearly by a speaker; an item may have strong distracters and may also be on an unfamiliar topic, etc. As noted by Alderson (2000, p. 70): 'identifying text variables which consistently cause difficulty is a complex task. Clearly at some level the syntax and lexis of texts will contribute to [. . .] difficulty, but the interaction among syntactic, lexical, discourse and topic variables is such that no one variable can be shown to be paramount.'

Conclusion

This project was initiated to explore issues which are important for language testing practitioners. The participants were given individualised feedback (after the final stage), and the workshop certainly raised awareness of the multiple factors to be considered in designing test items. For item writers, the workshop highlighted the need for questions to be worded clearly and positively – badly worded questions should not be one of the factors affecting difficulty. It also highlighted that key words or phrases (i.e. those which need to be identified in order to answer an item) should not be obfuscated, and should be limited to the core vocabulary list for the relevant level. For us as a UK award organisation, the key finding was to highlight the value of pre-testing and the need to obtain as much data as possible, given the limited number of test-takers. The expert judgement of item writers is valuable but needs to be supported by other evidence, especially when evaluating difficulty. No amount of scrutiny nor pre-testing can ensure that all items perform exactly as predicted, but pre-testing is a useful indicator. Further analysis of the data collected in this study and dissemination of findings are ongoing, as well as training item writers and examiners involved in the Welsh tests. The issues explored are of relevance to all language test practitioners, but particularly so where candidate numbers are relatively small.

Acknowledgements

Thanks to Lillie Halton, Joanna Maziarz and Alun Evans from WJEC Research Department for their work on this project. Thanks also to the participants in the project panel.

References

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

Beinborn, L., Zesch, T., & Gurevych, I. (2014). Predicting the difficulty of language proficiency tests. *Transactions of the Association for Computational Linguistics*, *2*, 517–529.

Loukina, A., Yoon, S. Y., Sakano, J., Wei, Y., & Sheedan, K. M. (2016, December). *Textual complexity as a predictor of difficulty of listening items in language proficiency tests* [Conference presentation]. Paper presented at COLING 2016, the 26th International Conference on Computational Linguistics, Osaka, Japan.

Sydorenko, T. (2011). Item writer judgments of item difficulty versus actual item difficulty: A case study. *Language Assessment Quarterly*, *8*, 34–52.

Item performance in context: Differential item functioning between pilot and formal administration of the Norwegian language test

Ga Young Yoon
University of Oslo, Skills Norway

Abstract

In this study, differential item functioning (DIF) was examined between the pilot and formal reading comprehension tests in the Norwegian language test (Norskprøven), using a log-likelihood ratio test method. Purification method was conducted to clean the invariant items and to improve the precision of DIF-item detections. The results revealed 10 DIF items with a large effect size out of 39 dichotomous reading proficiency items that were examined. A different amount of DIF was found in different levels of ability, i.e., non-uniform DIF. DIF items also showed a tendency to vary more in item discrimination than in item difficulty. Regarding the descriptive analysis with item features and context, i.e., item format and count of words, there was no clear evidence of those factors being related to DIF. However, more items among the anchor items were piloted in two different levels, in contrast to the DIF items. Therefore, the shifts in individual item performance seem to be more related to test administration and calibration design rather than to the item features.

Introduction

Differential item functioning (DIF) occurs when an item has different probabilities that it will be correctly answered conditional on the same ability level for different groups (van der Linden, 2017). The DIF items can lead to biased results between subgroups (Kim, 2001). Although DIF items do not necessarily indicate a whole study bias, they may contain important information about the subgroups being examined or the potential sources of bias (Kim, 2001; van der Linden, 2017).

The Norwegian language test (Norskprøven) for adult immigrants is developed and assessed by Skills Norway (www.skillsnorway.no) at the request of the Norwegian Ministry of Education and Research. Norskprøven is a popular large-scale assessment with around 25,000 test-takers each year, and it measures the language proficiency of Norwegian as a second language (Birkeland, Midtbø, & Ulven, 2019). Four different aspects of language proficiency are assessed in Norskprøven: reading, listening, writing, and oral communication. In this study, the investigation is limited only to the reading proficiency test.

The backgrounds of the test candidates for Norskprøven are highly diverse with respect to age, immigration status in Norway, language, and education background (Birkeland et al., 2019; Moe & Verhelst, 2017). Based on the various backgrounds of the test candidates, it is important that Norskprøven be fair and precise in measuring Norwegian language skills in order to provide the appropriate opportunities and help for those who need it. The formal test of Norwegian reading proficiency in Norskprøven is designed as multistage adaptive testing (MST) (Yan, von Davier, & Lewis, 2014). One of the intentions behind using a multistage test design in the administration of the test is that it would be better suited for a heterogeneous group of test-takers (Moe & Verhelst, 2017). Moreover, MST provides reduced test length, while maintaining the necessary reliability (Sadeghi & Abolfazli Khonbi, 2017).

Reading comprehension tests of Norskprøven are criterion-referenced tests, in which the cut-scores of different levels of the test are carried out with standard-setting procedures (Moe & Verhelst, 2017). The levels of difficulty used for the reading test items of Norskprøven are B2, B1, A2, A1, and under A1, which are built upon the Common European Framework of Reference for Languages (CEFR; Council of Europe, 2001). To transfer the correct levels of the test from the standard-setting procedure, and to distribute the new items in the correct levels of standards, *pilot tests* are used for pre-equating. Pre-equating means that the item parameters are estimated beforehand for use in the operational setting, which is *the formal test* in this study (Davey & Lee, 2011). Pre-equating is more practical than post-equating because the items are already calibrated for the operational use (Davey & Lee, 2011). However, since the items are already located in each position based on the previous calibration, *context effect* is a potential issue (Davey & Lee, 2011). Context effect means that item performance and characteristics are sensitive to the specific way in which an item is presented in the test, for instance wording, format, item position in the test, sequencing of the items, and specific features (Davey & Lee, 2011; Leary & Dorans, 1985; Yen, 1980). A question that follows is whether the parameter estimates are tied to the specific context in which an item was pretested or generalized to remain valid across the contexts in which an item may appear in operational tests (Davey & Lee, 2011).

Importantly, the pilot test of Norskprøven is administered in a low-stakes situation for the examinees. Pilot tests are primarily used as practice by test-takers and are provided at no extra charge with the language courses they are attending. The formal test administration of Norskprøven is a high-stakes assessment, which means that the test can be used to make crucial decisions about individuals or aggregated to make decisions about groups (Kingston & Kramer, 2013). A certain level of successful certification in Norskprøven is one of the requirements in applying for Norwegian citizenship and admission to higher education in Norway (Skills Norway, 2017, see also www.udi.no/en/). Different engagement in test-taking in high- and low-stakes situations is one of the potential sources of bias in estimation (Pokropek, 2016; Ulitzsch, von Davier, & Pohl, 2019; Wang & Xu, 2015). In low-stakes situations, examinees experience few or no consequences from their test performance and therefore may not be fully engaged when responding to the test items (Ulitzsch et al., 2019). In such situations, examinees may exhibit the disengaged behavior of omitting responses and randomly guessing. Previous studies suggest that the disengaged behavior of examinees should be regarded as a different construct than proficiency in low-stakes assessments (Pokropek, 2016; Ulitzsch et al., 2019; Wang & Xu, 2015). When this is neglected in the measurement procedure, person and item parameter estimates can be biased (Ulitzsch et al., 2019). Moreover, engagement probabilities tend to vary across items, i.e., items evoked disengaged behavior to a different degree. For instance, the probabilities for correct guesses were differently shown on different item formats, such as multiple-choice items and open response items (Ulitzsch et al., 2019).

In the light of previous research findings, our research questions (RQs) are:

1. How do the items shift from the pilot test to the formal test?
    a. Are there influential DIF items with a large effect size?
    b. What is the nature of DIF items, in terms of direction and behavior of parameter estimates?
2. What are the potential factors related to the item shifting from the pilot test to the formal test?
    a. Are the factors such as item format, count of words and item position showing differences between DIF and anchor items?

Method

Sample and data

Response test data of Norwegian reading proficiency from formal and pilot tests, test structures, item contents and characteristics, and pre-calibrated and operationally used item parameter values were provided by Skills Norway. For the formal test sample, reading test responses from May 2019 were used. For the pilot test, a sample of reading test responses from an anonymous year was selected for several reasons; first of all, this had the greatest number of common items between this pilot test and the formal test from May 2019, which gives the most possible items for comparison (the year is not given for security reasons). Another reason is that when only one specific sample of pilot test is used, it makes the data more stable rather than using all gathered pilot data across several years. There were, in total, 8,050 participants who took the formal test in May 2019 for 172 items. The total observation for the pilot test data, on the other hand, was 4,984 for 301 items. In total, 56 common dichotomous reading items between the formal and pilot tests were analyzed for further investigation.

Analysis procedure

All the analyses procedures were done in R program language (R Development Core Team, 2019). Several packages such as Psych and Mirt were used (Chalmers et al., 2019; Revelle, 2019). The current study investigated DIF in a reading comprehension test by using several sequences in the procedure. The response data sets from formal and pilot tests were analyzed for DIF using the two-step purification method. To clean the anchor items, the purification procedure detects DIF items in a preliminary DIF analysis and removes them from the list of anchor items in the main DIF analysis (Lee & Geisinger, 2016). The purification procedure is beneficial for large-scale analysis due to power improvement, but it is rarely used in the examination of language testing because of its highly technical procedure (Jodoin & Gierl, 2001; Lee & Geisinger, 2016). The item response theory-based likelihood ratio test was used to detect the invariant anchor items and significant DIF items model (de Ayala, 2009; Meade & Wright, 2012). Effect size was considered to find the most relevant DIF items, using the expected score standardized difference (ESSD) (Meade, 2010). Finally, the item contents and contexts were examined and discussed for the DIF items.

Results and discussion

Results revealed 17 anchor items and 39 DIF items at the second step of the purification method. Among the DIF items, 10 were found to have an absolute value of ESSD larger than 0.8. This result directly answered RQ1a. The Item characteristic curves (ICCs) and information curves showed that there was non-uniform DIF across most of the items, which means that items function differently in both difficulty and discrimination when conditioning on the same theta level. Although the difficulty was similar between the Formal and Pilot groups, the discrimination difference was large. Items showed generally higher discrimination power for the Formal group than the Pilot group, as shown through the steeper curves. This was also shown for the information functions. Nine out of ten DIF items with a large effect in the formal test had remarkably taller peaks of information functions than those conditional on the same theta level in the pilot test. These results addressed RQ1b.

The difference in information between the formal and pilot tests might be related to the systematic noise in low-stake situations. Low motivation in pilot tests might be reflected in the systematic noise of disengaged behavior, which leads to increased random error in pilot tests and decreased information. However, we do not have crucial evidence that motivation is the main issue here. Disengaged behavior is only suggested as a potential factor in causing DIF between the formal and pilot tests.

Item characteristics and features were also analyzed to examine and discover the potential factors that might be related to the DIF. The methodology used was descriptive, examining 10 DIF items with large effect size and comparing them to 17 anchor items. Item format, count of words, levels in the formal test, levels in the pilot test, proportion of correct responses, and item parameter estimates were compared between DIF and anchor items. The most striking finding was that eight out of ten DIF items with large effects were only piloted at one level. In comparison, most of the 17 anchor items were piloted at two levels. Only three out of seventeen anchor items were piloted at one level, which were Item 1, 2 and 7. Items that were piloted at two levels, were piloted either at A1-A2 and A2-B1 levels, or at A2-B1 and B1-B2 levels. This leads to the conclusion that items piloted in those levels are estimated in wider scales of ability levels: A2 items in A1, A2 and B1 levels, and B1 items in A2, B1 and B2 levels. The IRT parameters give us information about how the item functions at all ability levels (de Ayala, 2009). An item piloted on a larger scale of proficiency naturally gives more valid and stable parameter values than items piloted in a narrower ability range. Uncertainty would be greater for the scales that do not have many candidates. One would therefore get the most valid item parameters from the calibration with the most candidates in all the theta levels.

However, other item features seemed to be a minor concern related to DIF. There was no clear evidence in item formats and count of words that were clearly related only to DIF items. Based on these findings, we can conclude that item shift can be affected by the test administration rather than item format and count of words. More specifically, items being calibrated in several different levels in the pilot test are tested in a broader aspect in the theta scale, and therefore have more precise estimation. This responded to RQ2.

Limitation and implication of the study

Some limitations of the study are worth noting. One is that the single characteristics of items do not explain the DIF results adequately enough, and additional research may help to identify whether combinations of variables, such as those related to one specific type of item format, will correlate consistently with DIF. Furthermore, a tentative attempt at generalization will require further experimental confirmation. Our study does not focus on the elimination of DIF-related factors, but rather explores the quality of DIF-exhibiting items. Therefore, a further direction of investigation is suggested for the qualitative contents of DIF items and the possible mitigation of the DIF factor. A closer qualitative follow-up is also suggested for the items that were shown to fall prone to DIF to identify item features that contribute to item drift from low-stakes to high-stakes test administration. Another limitation is that we only investigated the reading test of Norskprøven. The DIF examination for the listening test between the formal and pilot groups is worthy of exploration in future studies to improve the quality of the item bank in Norskprøven. Writing and oral communication tests in Norskprøven do not operate the pilot administration, and therefore are not appropriate for this kind of DIF analysis.

However, several implications of the study can be discussed. Firstly, this study confirms that when obtaining parameters in the item bank, it is necessary to update them based on parameter estimation from the formal test. Otherwise, several items might exhibit DIF because of the different situation. Secondly, it can be recommended that Norskprøven should eliminate or update the parameters of detected DIF items from this study. Otherwise, ability estimation in future formal tests may be biased. Finally, as we have found 17 clean anchor items in this study, Skills Norway might consider using these items as a set of sample items for developing new items for the future administration of Norskprøven.

References

Birkeland, P., Midtbø, T., & Ulven, C. H. (2019). *Resultater på Norskprøven for voksne innvandrere 2014–2017*. Retrieved from: www.kompetansenorge.no/contentassets/fcb0d13d2d81485f92f9e4f96dc36767/resultater_pa_norskproven_for_voksne_innvandrere.pdf

Chalmers, P., Pritikin, J., Robitzsch, A., Zoltak, M., Kim, K., Falk, C. F., Meade, A., Schneider, L., King, D., Liu, C.-W., & Oguzhan, O. (2019). *mirt: Multidimensional Item Response Theory* (Version 1.31) [Computer software]. Retrieved from: CRAN.R-project.org/package=mirt

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

Davey, T., & Lee, Y.-H. (2011). *Potential Impact of Context Effects on the Scoring and Equating of the Multistage Gre® Revised General Test*. ETS Research Report Series. Princeton: Educational Testing Service.

de Ayala, R. J. (2009). *The Theory and Practice of Item Response Theory*. New York: The Guilford Press.

Jodoin, M. G., & Gierl, M. J. (2001). Evaluating Type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. *Applied Measurement in Education*, *14*(4), 329–349.

Kim, M. (2001). Detecting DIF across the different language groups in a speaking test. *Language Testing*, *18*(1), 89–114.

Kingston, N. M., & Kramer, L. B. (2013). High-stakes test construction and use. In Little, T. D. (Eds.), *The Oxford Handbook of Quantitative Methods Volume 1: Foundations* (pp. 189–205). Oxford: Oxford University Press.

Leary, L. F., & Dorans, N. J. (1985). Implications for altering the context in which test items appear: A historical perspective on an immediate concern. *Review of Educational Research*, *55*(3), 387–413.

Lee, H., & Geisinger, K. F. (2016). The matching criterion purification for differential item functioning analyses in a large-scale assessment. *Educational and Psychological Measurement*, *76*(1), 141–163.

Meade, A. W. (2010). A taxonomy of effect size measures for the differential functioning of items and scales. *Journal of Applied Psychology*, *95*(4), 728–743.

Meade, A. W., & Wright, N. A. (2012). Solving the measurement invariance anchor item problem in item response theory. *Journal of Applied Psychology*, *97*(5), 1,016–1,031.

Moe, E., & Verhelst, N. (2017). Setting standards for multistage tests of Norwegian for adult immigrants. In S. Blömeke & J.-E. Gustafsson (Eds.), *Standard Setting in Education* (pp. 181–204). New York: Springer.

Pokropek, A. (2016). Grade of membership response time model for detecting guessing behaviors. *Journal of Educational and Behavioral Statistics*, *41*(3), 300–325.

R Development Core Team. (2019). *R: A language and environment for statistical computing.* Retrieved from: www.r-project.org

Revelle, W. (2019). *psych: Procedures for Psychological, Psychometric, and Personality Research* [Computer software]. Retrieved from: CRAN.R-project.org/package=psych

Sadeghi, K., & Abolfazli Khonbi, Z. (2017). An overview of differential item functioning in multistage computer adaptive testing using three-parameter logistic item response theory. *Language Testing in Asia*, *7*.

Skills Norway. (2017). *Norskprøven ved opptak til høyere utdanning—Kompetanse Norge*. Retrieved from: www.kompetansenorge.no/prover/norskproven-ved-opptak-til-hoyere-utdanning/

Skills Norway. (2019). *Leseforståelse nivå A2–B1—Kompetanse Norge*. Retrieved from: www.kompetansenorge.no/prover/norskprove/ove-til-proven/leseforstaelse-niva-a2-b1/

Ulitzsch, E., von Davier, M., & Pohl, S. (2019). A hierarchical latent response model for inferences about examinee engagement in terms of guessing and item-level non-response. *British Journal of Mathematical and Statistical Psychology*, *73*(1), 83–112.

van der Linden, W. J. (Ed.). (2017). *Handbook of Item Response Theory Volume 3: Applications*. Boca Raton: Chapman and Hall/CRC.

Wang, C., & Xu, G. (2015). A mixture hierarchical model for response times and response accuracy. *British Journal of Mathematical and Statistical Psychology*, *68*(3), 456–477.

Yan, D., von Davier, A. A., & Lewis, C. (Eds.) (2014). *Computerized Multistage Testing; Theory and Applications*. London: Chapman and Hall/CRC.

Yen, W. M. (1980). The extent, causes and importance of context effects on item parameters for two latent trait models. *Journal of Educational Measurement*, *17*(4), 297–311.

Mono- and multilingual testing – common standards towards validity and fairness in testing language proficiency for aviation

Neil Bullock
English Plus LTS, Switzerland

Abstract

The mandatory testing of language proficiency for pilots and air traffic controllers introduced worldwide in 2008, with recommendations for the use of common international testing standards, was clear evidence of the need for minimum standards of communication in this domain. However, many language proficiency tests worldwide developed for this purpose – mostly in the *de facto* use of English, although all languages used via the radiotelephone must be assessed – have been done in isolation, leading to differing testing standards and practices where valid and reliable test results are not always assured.

This paper will show recent research that continues to show a high variance in quality and test usefulness in this domain and then go on to examine the development of one language proficiency test for pilots, demonstrating that good testing practice is possible and highlighting the need for test-takers to engage in test tasks replicating real-life communication.

Introduction: Towards validity and usefulness – an overview of international testing in aviation

Despite a rather obvious link between effective communication and safety, and recommendations from the International Civil Aviation Organisation (ICAO) that language proficiency testing (introduced in 2008) should follow good testing practices (International Civil Aviation Organisation, 2010), early research showed that, in the aviation world, heeding such advice was often the exception rather than the rule and that oversight of such tests be improved in order to adhere to recommended testing practices (Alderson, 2009; 2010).

One attempt to address these issues was made by ICAO, in 2011, when a method of endorsing language proficiency tests was set up. Test Service Providers (TSPs) could (pay and) apply to have their test endorsed as fulfilling the recommendations of the Language Proficiency Requirements (LPRs) system. Following the introduction of this service, over 20 TSPs applied for test endorsement, yet only one test (*English Language Proficiency for Aeronautical Communication*, ELPAC) is currently endorsed. Research and evidence as to why so many tests did not succeed in the endorsement process, which would be of particular value to all stakeholders, does not seem to be available (Bullock & Westbrook, 2021).

Another project aiming to tackle some of the concerns mentioned above, was the *Test Design Guidelines* (TDGs), developed by the International Civil Aviation English Association (ICAEA)[1], and supported by ICAO. Eight testing criteria were established that civil aviation authorities (CAAs) and Air Navigation Service Providers (ANSPs) could use to cross-check LPR tests under their jurisdiction. The criteria were based on the requirements of ICAO Doc9835 as well as recognised testing literature. Four three-day workshops were delivered worldwide in 2019 and a master document is due to be published as an official circular by ICAO in 2021.

Despite the measures mentioned above, key research in this domain has been conspicuous by its absence, which does appear unusual and little evidence is available to say why (Bullock & Kay, 2017; Bullock & Westbrook, 2021; Emery, 2014; Kim, 2018). There is also no current list of either privately or institutionally developed test instruments approved or certified by national CAAs worldwide. In order, therefore, to source some information from TSPs worldwide about what tests were currently being used at the moment, including having some idea of their quality and usefulness, a doctoral research project was carried out in late 2020 by the author. A random internet search was carried out using terms such as: 'aviation language proficiency test' and 'Test

[1] See Appendices 1 and 2.

of English for aviation'. From this search, and some already known tests, a list of 30 tests was compiled. In order to source specific evidence about the tests, a list of eight categories (A–H[2] – see Table 1) was drawn up relating to the typical sort of information that could be expected to be found on a TSP website according to that recommended in testing documentation and literature (Alderson, Clapham, & Wall, 1995; Association of Language Testers Europe, 2020; Fulcher, 2010).

Once this had been completed for the 30 tests, each one was evaluated against the eight criteria[3] of the ICAEA TDGs using the evaluation toolkit provided during the workshops.

Table 1: Percentage of TSPs that included information expected to be found on a TSP website relating to the test content and operation

No.	*Information recommended to be on website*	*% of tests where such information is given on website*
A	Regulatory approval of state CAA claimed	57%
B	Information given on test development	20%
C	Information given on who developed the test	17%
D	TT population known (pilots and/or ATCOs)	63%
E	Test includes both listening comprehension and speaking ability	47%
F	Specified ICAO levels tested	53%
G	Sample test and prep info available	27%
H	Location of test (online/face-to-face) stated	50%

Initial observations

Many of the 30 TSPs, based in 20 countries, included information on the websites which was often written in a register more appropriate for a sales platform using short punchy phrases such as: 'Approved by most EU CAAs'; 'See our success rate'; 'Certificate issued in only 30 minutes'; and the rather curious and unsubstantiated 'question bank workflow engine ensures quality'. The six institutional organisations (two CAAs, three universities and the European ATC agency, Eurocontrol), were, unsurprisingly, more pragmatic and formal in their language. 11 (37%) of the TSPs had no more than one page on their website dedicated to the test and therefore detailed information about the test was limited. 17 (56%) of the TSPs offered declarations about their test with little supporting evidence, including various claims about compliancy with ICAO.

Even though much of this data could not be corroborated against further evidence, the figures are nonetheless quite striking. Less than 20% of TSPs surveyed offered information on test development, and only 27% included evidence of a sample test. Of these, only a few provided video or audio files. Some simply offered PDF documents.

Further investigation of the websites clearly demonstrated testing practices that are not appropriate for the purpose designed in the ICAO LPRs. Comments noted included:

- *voice-only interaction* justified as two test-takers (TTs) sitting back-to-back in a room discussing a publicly available video
- *interaction* assessed through short decontextualised utterances between TT and computer
- very short tasks requiring minimal use of standard phraseology and no plain language
  - decontextualised routine weather information transmissions, labelled a *listening* task
  - statement that the test assesses ability to 'understand recorded messages'
  - inclusion of writing and/or theory tests
  - questions that can be answered by technical knowledge
  - single face-to-face interview *or* 'informal chat' tasks.

When matching the 30 tests according to the ICAEA TDG toolkit, only three (10%) surveyed provided sufficient evidence to demonstrate that they could be effective in at least all of the first three criteria, and notably in Criteria 1 which states that: 'Test instruments need to include appropriate tasks that directly assess how test takers use language in radiotelephony communication contexts' (International Civil Aviation English Association, 2019).

[2] These criteria were labelled A–H to avoid confusion with the eight criteria of the ICAEA TDGs labelled 1–8.

[3] The rationale of the toolkit, in simple terms, is that the test is evaluated against eight criteria in descending order of importance. If it does not include any one or more of the first three criteria, the test is seen as not 'effective' (International Civil Aviation English Association, 2019). The first three criteria are seen as the three most critical for inclusion in a test to assess the language proficiency of pilots and air traffic controllers (ATCOs). If it does include the first three but not any one of the remaining five then it *may* be effective but needs more work or supporting evidence.

This means that 90% of the tests that did provide sufficient information to be evaluated did not demonstrate test tasks that reflect how their TTs communicate in real-world contexts, i.e. in voice-only person to person co-constructed radiotelephone communication as a pilot or controller during routine and non-routine situations.

The evidence in this survey is clearly short of comprehensive and supporting data in order to make more specific conclusions. However, from the evidence shown so far, what is clear is that, not only do many ELP tests not follow the recommendations of the ICAO LPRs, despite wide and ambiguous claims on the website, but most failed to demonstrate in their assessment process the central role played by real-world communication between a pilot and a controller.

Demonstrating usefulness in practice – evidence through test development

One of the three tests surveyed, which was specifically developed to include assessment of language proficiency in real-life pilot/controller radiotelephony, and that would have included the first three items on the TDG checklist, was one developed in 2011, by the Swiss Federal Office of Civil Aviation[4] (FOCA), in which the author was involved as a test developer. Switzerland has four national languages, and therefore the proficient use of English as a means of communication between pilots and ATCOs is a critical factor in ensuring flight safety within a multicultural and multilingual domain. The test was developed to assess English language proficiency of pilots and replace an older version, sourced from an external supplier. The older test did not include any tasks that directly assessed language as used in real-world radiotelephony situations, leading to criticism from test-takers about the inappropriateness of tasks. Data, taken during the new test development and in subsequent feedback from TTs shown in the development team documentation, clearly showed the positive impact that inclusion of real-life communication had on the TTs, and therefore the usefulness of the test.

The test development team included people who all had both second language learning/teaching and testing experience, as well as operational experience as pilots or ATCOs. A bank of additional operational experts – typically pilots and ATCOs – and language experts was also used during the development process where required. Key theory in test development was evaluated for use before determining the specific requirements of the TT population's real-world communicative context (Bachman & Palmer, 1996; Brown, 2000; Field, 2013; Hughes, 2003; Luoma, 2004). From this evaluation, the developers adopted Fulcher's (2010) test development plan with additional cautionary stages for what Weir (2005) refers to as *a priori* and *a posteriori* validity but inserting these stages before and during the trialling stage (See Figure 1). The rationale was to have a certain indication of validity before any form of trialling takes place, so that the trialling provides tangible evidence for continued development of the test.

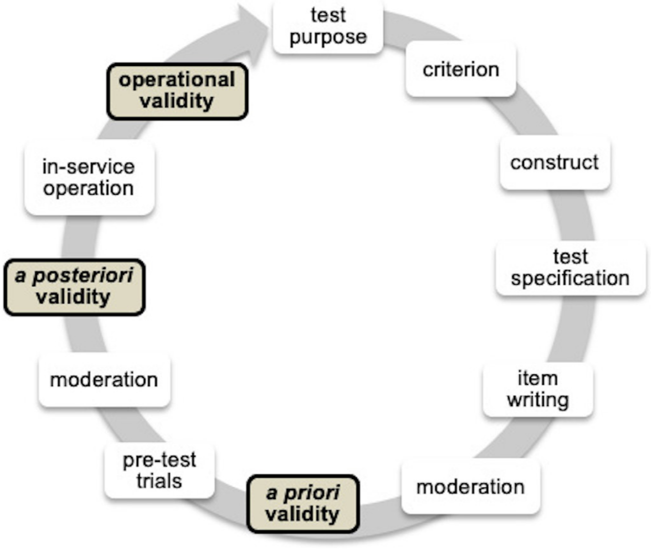


Figure 1 Test development cycle shown as a circle, starting from test purpose with three subsequent key stages: a priori validity, a posteriori validity and operational validity

Listening comprehension was assessed by means of a stand-alone test in which candidates listen to simulated pilot/ATCO exchanges using authentic task-based interactive language. Pilots and ATCOs were employed to record the scripts adding increased authenticity and cognitive validity.

Speaking ability was assessed via two tasks that matched the ICAO recommended criteria (ICAO, 2010), to include:

1. A voice-only interaction role-play (VOI) of a simulated flight.
2. A face-to-face task (F2F) where TTs were asked questions related to non-routine situations and events in their operational aeronautical environment.

The whole test development, including trialling, took approximately two years. Data from all test results was sourced and analysed for item performance, and for the first two years of the test operation feedback from all TTs was received and analysed. All of this data was intended to investigate whether the test was achieving levels of validity hitherto missing on the older versions.

[4] The national civil aviation authority for Switzerland.

In the first three years of test operation, 551 candidates had taken the exam and 280 (51%) returned a feedback response using the SurveyMonkey online tool. Questions were asked on a variety of elements related to the test, particularly on how much the material and questions reflected real-life operational aeronautical situations and to what extent they thought they were appropriate and useful. A 5-part Likert scale was used ranging from 'fully agree' to 'fully disagree'. The agree/fully agree coefficients show a very high level of positivity (average 0.91) in the opinions of testees (see Table 2).

Table 2: TT responses (280 from 551 total, 51%) on key test components from 14.11.2014–31.12.17

Exam stage	*Positivity co-efficient*
Voice-only interaction (role play)	0.84
Face-to-face interview	0.90
Listening test	0.91
Exam organisation	0.96
After the exam	0.94
Overall average	**0.91**

Conclusion

Whilst data sourced from the Swiss test results gave evidence of operational validity and consistent performance, the candidate feedback provided additional evidence of contextual, cognitive, face and consequential validity. High-stakes testing demands that development and operation not only adhere to good testing practices and recognised testing theory, but also include the actual real-life communication that allows the required language to be assessed. Thus, it can be said that despite the misgivings noted from a sample of tests worldwide earlier, including inappropriate testing practices, it is possible to develop an appropriate and useful language proficiency test for aeronautical communication. That the issues raised by authors over 10 years ago have not yet been resolved is still, however, a concern for air safety and these should be addressed as a matter of urgency.

References

Alderson, C. J. (2009). Air safety, language assessment policy, and policy implementation: The case of Aviation English. *Annual Review of Applied Linguistics*, *29*, 168–187.

Alderson, C. J. (2010). A survey of Aviation English tests. *Language Testing, 27*(1), 51–72.

Alderson, C. J., Clapham, C., & Wall, D. (1995). *Language Test Construction and Evaluation.* Cambridge: Cambridge University Press.

Association of Language Testers Europe (2020). *Principles of Good Practice*. Retrieved from: www.alte.org/resources/Documents/ALTE%20Principles%20of%20Good%20Practice%20Online%20(Final).pdf

Bachman, L. F., & Palmer, S. (1996). *Language Testing in Practice*. Oxford: Oxford University Press.

Brown, H. D. (2000). *Teaching by Principles*. New Jersey: Longman.

Bullock, N., & Kay, M. (2017, April 24). *Reviewing 10+ years of the ICAO LPRs* [Conference presentation]. ICAEA 2017 Conference, Dubrovnik, Croatia. Retrieved from: commons.erau.edu/icaea-workshop/2017/monday/8/

Bullock, N., & Westbrook, C. (2021). Testing in ESP: Approaches and challenges in aviation and maritime English. In B. Lanteigne, C. Coombe, & J. D. Brown (Eds), *Challenges in Language Testing Around the World - Insights for Language Test Users* (pp. 67–77). Singapore: Springer.

Emery, H. (2014). Developments in LSP testing 30 years on? The case of Aviation English. *Language Assessment Quarterly*, *11*(2), 198–215.

Field, J. (2013). Cognitive validity. In A. Geranpayeh & L. Taylor (Eds.) *Examining Listening: Research and Practice in Assessing Second Language Listening* (pp. 77–151). Studies in Language Testing volume 35. Cambridge: UCLES/Cambridge University Press.

Fulcher, G. (2010). *Practical Language Testing.* London: Hodder Education.

Hughes, A. (2003). *Testing for Language Teachers* (2nd ed.). Cambridge: Cambridge University Press.

International Civil Aviation English Association. (2019). *Using the ICAO LPR Test Design Guidelines: New Criteria for Effective ICAO LPR Test Design. Workshop Handbook*. Retrieved from: www.icaea.aero/wp-content/uploads/2019/12/ICAEA-TDG-Workshop-Handbook.pdf

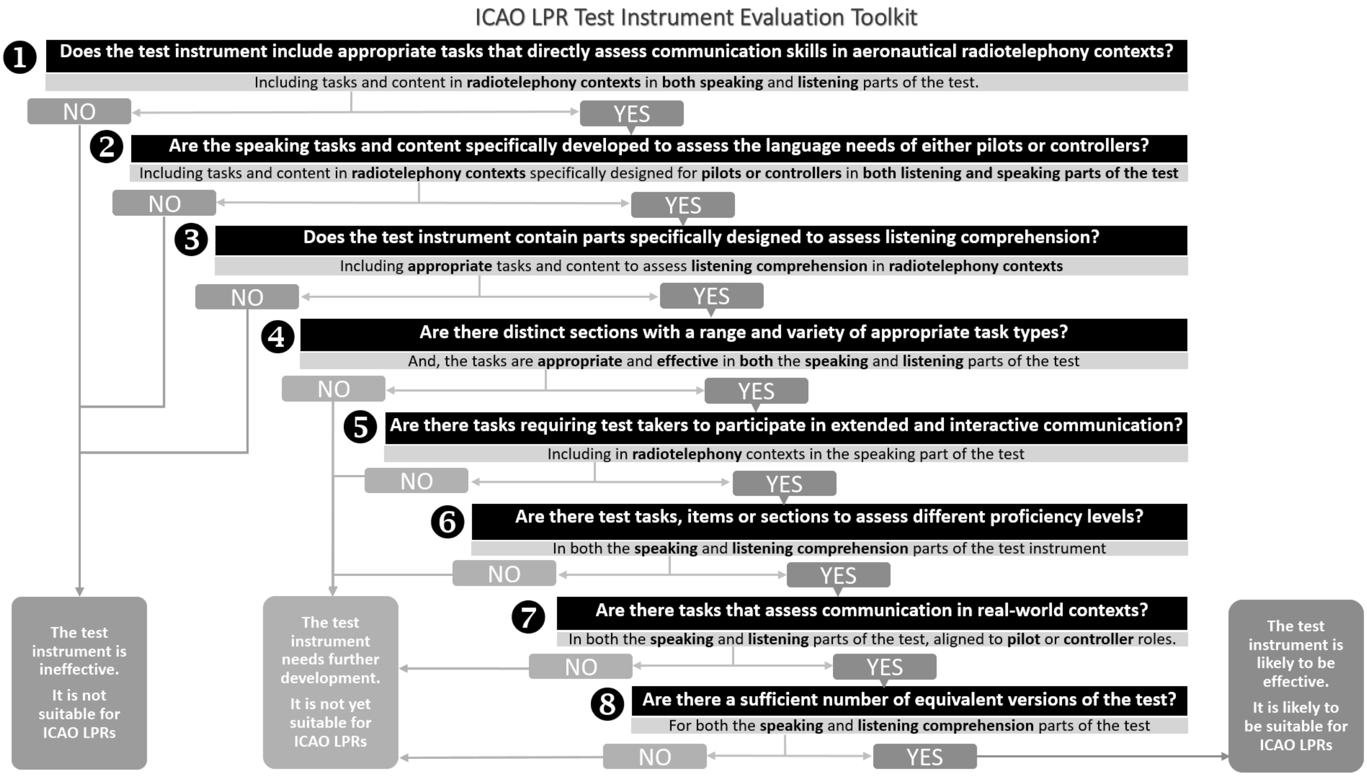
International Civil Aviation Organisation. (2010). *Doc 9835. Manual on the Implementation of ICAO Language Proficiency Requirements* (2nd ed.). Montreal: International Civil Aviation Organisation.

Kim, H. (2018). What constitutes professional communication in aviation: Is language proficiency enough for testing purposes?. *Language Testing*, *35*(3), 403–426.

Luoma, S. (2004). *Assessing Speaking*. Cambridge: Cambridge University Press.

Weir, C. J. (2005). *Language Testing and Validation – An Evidence-based Approach.* London: Palgrave Macmillan.

Appendix 1: ICAEA Test Design Guidelines Toolkit


ICAO LPR Test Instrument Evaluation Toolkit
1 Does the test instrument include appropriate tasks that directly assess communication skills in aeronautical radiotelephony contexts?
Including tasks and content in radiotelephony contexts in both speaking and listening parts of the test.
NO
YES
2 Are the speaking tasks and content specifically developed to assess the language needs of either pilots or controllers?
Including tasks and content in radiotelephony contexts specifically designed for pilots or controllers in both listening and speaking parts of the test
NO
YES
3 Does the test instrument contain parts specifically designed to assess listening comprehension?
Including appropriate tasks and content to assess listening comprehension in radiotelephony contexts
NO
YES
4 Are there distinct sections with a range and variety of appropriate task types?
And, the tasks are appropriate and effective in both the speaking and listening parts of the test
NO
YES
5 Are there tasks requiring test takers to participate in extended and interactive communication?
Including in radiotelephony contexts in the speaking part of the test
NO
YES
6 Are there test tasks, items or sections to assess different proficiency levels?
In both the speaking and listening comprehension parts of the test instrument
NO
YES
7 Are there tasks that assess communication in real-world contexts?
In both the speaking and listening parts of the test, aligned to pilot or controller roles.
NO
YES
8 Are there a sufficient number of equivalent versions of the test?
For both the speaking and listening comprehension parts of the test
NO
YES
The test instrument is ineffective. It is not suitable for ICAO LPRs
The test instrument needs further development. It is not yet suitable for ICAO LPRs
The test instrument is likely to be effective. It is likely to be suitable for ICAO LPRs

Appendix 2: ICAO LPR test design criteria: Overview

1. Test instruments need to include appropriate tasks that directly assess how test-takers use language in radiotelephony communication contexts.

Why is this important?

The ICAO LPRs refer to communication between pilots and air traffic controllers. Assessing proficiency for radiotelephony communication is central to the ICAO LPRs. This is the primary focus of the safety-critical communication contexts which the ICAO LPRs aim to address.

What does this mean for test design?

Speaking and listening skills in air-ground communication contexts need to be directly assessed through dedicated test tasks that reflect this content, context and type of communication. A substantial component of the test needs to contain content and task types which assess how test-takers communicate in radiotelephony communication contexts – in listening (understanding what pilots or controllers are saying over the radio) and speaking (being able to communicate as pilots or air traffic controllers over the radio) components of the test. The tasks used to assess language proficiency for radiotelephony communication need to be appropriate and reflect the communicative contexts in which pilots or controllers communicate in real-world situations.

2. Separate test instruments need to be designed for pilots and air traffic controllers.

Why is this important?

The communication contexts and language needs of pilots and air traffic controllers differ and therefore these need to be reflected in different forms of the test, catering to the needs of each profession.

What does this mean for test design?

The structure of test instruments for pilots/air traffic controllers may look similar; however, the content and task requirements need to differ and reflect the language needs and communicative contexts associated with each profession.

3. Test instruments need to contain tasks dedicated to assessing listening comprehension, separate from tasks designed to assess speaking performance.

Why is this important?

Listening comprehension represents at least half the communicative load in aeronautical communication. Proficiency in comprehension is determined by a range of different cognitive skills, language skills and knowledge. All of these attributes are internal and cannot be directly observed for assessment purposes. In contrast, speaking skills are more observable and can be assessed directly by observing speaking performance. Therefore, proficiency in listening comprehension is best assessed in contexts which are not affected by speaking ability because basing decisions on what test-takers say may be more of a result of their speaking skills rather than their internal comprehension proficiency.

What does this mean for test design?

Tests need to contain sections and parts which are designed to only assess listening comprehension. This means test-takers are required to listen to prescribed recordings and then complete follow-up comprehension tasks. Such tasks could be on paper, require test-takers to summarise information or answer prescribed written questions asked orally or provided on a test paper/computer screen.

It is possible for tests to also evaluate comprehension subjectively in an interactive context in addition to having a dedicated listening test section, but not to the exclusion of including dedicated listening comprehension test sections. In such situations the subjective ratings should be used to support the results of the dedicated listening sections.

4. Test instruments need to comprise distinct sections with a range of appropriate task types.

Why is this important?

Tests need to comprise different sections with different assessment purposes – assessing different skills/levels in a range of communication contexts. Tests which contain a range of different test tasks provide more opportunities to effectively sample the range, complexity and type of communication pilots or air traffic controllers may face. This improves both the fairness and the effectiveness of the test, including the validity of the interpretations made of test results and how these are used. Tests which do not include enough variety are not able to effectively sample test-takers' abilities to engage with the language or communicate in different situations. This undermines the test's overall validity. Such tests can also unfairly disadvantage test-takers who are less familiar or comfortable with certain test tasks which may dominate a test.

What does this mean for test design?

A variety of different task types, items, situations and content needs to be included throughout the test instrument to ensure the domain and range of language proficiency levels are effectively sampled.

5. Test instruments need to include tasks that allow test-takers to engage in interactive and extended communication.

Why is this important?

Aeronautical radio communication involves pilots and controllers communicating in interactive situations – responding to issues, enquiring, solving problems, providing advice etc. In all such communication, each participant is required to engage in topics, negotiate meaning and participate in a collective and shared communicative context which develops as a result of the interaction.

What does this mean for test design?

At least some speaking tasks need to provide opportunities for test-takers to participate in interactive communication with a trained interlocutor, i.e., tasks which require the test-taker to contribute to a co-constructed dialogue in the same way that communication occurs in real-world aeronautical contexts. Test tasks which are limited to test-takers responding to isolated questions or disconnected prompts do not allow interactive skills to be evaluated and do not reflect real-world communications. They are therefore not authentic. Authenticity is a key requirement of proficiency testing. Note that in interactive speaking tasks, comprehension should not be assessed, or only be limited to supporting the results of a part of the test dedicated to assessing comprehension separately. Comprehension could be rated and included as a subjective impression to confirm or support the results of a dedicated listening part of the test (see Item 6).

6. Test instruments need to include tasks and items which allow the assessment to differentiate between ICAO language proficiency levels.

Why is this important?

Content and tasks types which target the proficiency levels and skills associated with each of the ICAO levels the test aims to assess need to be incorporated into the test.

What does this mean for test design?

Parts and components of tests need to be accessible and achievable for test-takers at different levels, with some parts/tasks/items catering to test-takers with lower-level proficiency levels (below ICAO Level 4), and other parts for ICAO Level 4 (and above). If the test claims to be able to assess ICAO Level 5 or Level 6, different tasks/items or test parts need to be dedicated to assessing these higher levels and their associated competencies (as reflected in the ICAO LPR rating scale).

7. Test instruments need to contain appropriate tasks that assess test-takers' abilities to understand and communicate in real-world contexts.

Why is this important?

In order for the test to allow valid assessment decisions to be made about how well test-takers are able to communicate in their job as either pilots or air traffic controllers, the test needs to ensure the content and task requirements allow for this evaluation to be effective. The closer the test reflects the communicative requirements associated with real-world communication contexts which pilots or air traffic controllers face, the more meaningful the test results are. Test tasks which require test-takers to communicate or use language that is not directly associated with how they communicate in real-world situations are not able to allow meaningful assessment decisions to be made about how well the test-takers can communicate in their jobs as pilots or air traffic controllers.

In high-stakes testing, test-takers respect tests and the results of such tests when the tests mirror real-world communication needs of the test-takers.

What does this mean for test design?

Components of the test need to contain tasks and content that mirror the kind of communication settings and contexts associated with real-world situations that pilots and air traffic controllers may face, both in radiotelephony communication contexts and other job-related communication contexts. The more directly the test tasks mirror real-world communication contexts, including the type of language and how this is used, the more effective a test instrument is in allowing valid interpretations to be made about test performances and test scores.

8. Test instruments need to have a sufficient number of equivalent versions, with each version of the test representing the test instrument in the same way.

Why is this important?

Tests which do not draw on a sufficient bank of test versions lack security. Test-takers in a target population can become familiar with test content and therefore prepare and rehearse answers and responses for the test. Obviously, in such cases, the test is not able to accurately assess test-takers' real overall proficiency as test-takers may appear to perform on the test at levels above their 'real' proficiency level.

Test banks also need to comprise equivalent test versions. This means that test-takers receive similar results on whichever version of the test they take. If the test bank includes versions which are easier than other versions, the test and the results are not reliable and therefore the overall testing system is not effective.

What does this mean for test development?

Tests need to comprise a test bank where each version of the test has aspects which are unique to that version of the test. Test developers need to ensure that each version of the test is written to a set of specifications so that all test versions are parallel and more or less equivalent in their level of difficulty and the range of language and communicative contexts that they assess.

The larger the test-taker population and the more often they need to be tested, the larger the test bank needs to be.

Interviews to raters as part of the qualitative validation of the B1 writing rating scale designed for the CertAcles-UJA test

Carmen Álvarez García
Clie Culture And Language International Experience S.c. And - Jaén, Spain

Abstract

Personal interviews with raters about their experience analysing and using the scale provide valuable information on key aspects for the scale validation, for example, about its representativeness of the construct, its usefulness, difficulties in its application, and whether it can be improved (East, 2009; Harsch & Martin, 2012). Even though interviews are not commonly used in the literature to validate scales, its use was found relevant mainly for two reasons. First of all, raters are the main users of the scale and the decisions they make about the candidates' writing ability have deep consequential validity (Taylor, 2011), and, secondly, raters are qualified and trained professionals, so they can also be considered experts, and as such their contributions are certainly of great relevance for the scale improvement.

Introduction

When assessing writing, raters need to make a judgement about the candidates' writing ability on the basis of their performance in a writing task. Through assessment tools they ascertain and make decisions (possibly high stakes) on the person's level of proficiency. Accordingly, it seems logical to think that the better the assessment tool is the fairer and more reliable the assessment will be, since it will help reduce the unwanted variability in scores. On this basis, the study on which this article is based aimed at reproducing the Protocol developed by Cruz (*A Protocol to Design a CEFR-linked Proficiency Rating Scale for Oral Production and its App Implementation*, 2016) for the design and validation of oral production rating scales, but in this case for the design and validation of a B1 writing scale. Additionally, the study added *ex novo* one new stage within the qualitative validation phase in which interviews with raters are performed in order to know the opinion of the actual users of the scale in real conditions, and therefore to contribute to its validation.

The personal interviews with raters about their experience analysing and using the scale provided valuable information on key aspects for the scale validation, for example, as claimed by East (2009) and Harsch & Martin (2012), about its representativeness of the construct, its usefulness, difficulties in its application, and whether it can be improved. Even though interviews are more commonly used in the literature to validate test results rather than scales, for the purpose of this study the use of interviews was found important mainly for two reasons. First of all, raters are the main users of the scale and the decisions they make about the candidates' writing ability have deep consequential validity (Taylor, 2011), and, secondly, raters are qualified and trained professionals, so they can also be considered experts, and as such their contributions are certainly of great relevance for the scale improvement. According to Knoch (2009, p. 297), '[r]aters' perceptions of the scale usefulness are important as they provide one perspective on the construct validity of the scale. As language experts they are well qualified to judge whether the writing construct is adequately represented by the scale.'

Theoretical background

According to Knoch (2016), the validation work on writing assessment has traditionally been based on the statistical analysis of scales' characteristics and rater reliability, mainly by the analysis of essays scores. However, there is currently an increasing recognition of mixed methods research (Jang, Wagner, & Park, 2014), which combines the use of qualitative and quantitative approaches for a better understanding of the phenomenon in question. According to them, '[r]esearchers seek evidence supporting the validity of assessment practice by integrating the nuanced understanding of the human experiences and relations in LTA contexts into the traditional validity discourse' (Jang et al., 2014, p. 125). Examples of this are the work of Janssen, Meier and Trace (2015) in which a mixed methods study assessing the functioning of an existing rubric is presented, and that of Harsch and Martin (2012) who involve raters in the scale revision as part of the validation process. Moreover, the Common European Framework of Reference for Languages (CEFR) (Council of Europe, 2001) also recommends three validation methods to be combined: intuitive, qualitative and quantitative. 'Qualitative methods require the intuitive preparation and selection of material and intuitive interpretation of results. Quantitative methods should quantify qualitatively pre-tested material, and will require intuitive interpretation of results' (Council of Europe, 2001, p. 207). The *Reference Supplement to the Preliminary Pilot Version of the Manual for Relating Language Examinations to the CEFR* (Council of Europe, 2004) also refers to qualitative analysis methods as a way 'to look more closely at how a test is working', or in this case, how the scale is working, and within the feedback methods it describes, interviews with raters would provide thoughtful and informed judgement of the scale.

On this basis, the CEFR (Council of Europe, 2001) as well as the Protocol by Cruz (2016) were used as theoretical frameworks for our validation. Since this is a CEFR-linked scale, the adherence to CEFR methods will strenghen the linkage and will contribute to its validation. The Protocol (2016) also provides a step-by-step guide for validation (see Table 1).

Table 1: Scale validation phase: Protocol to design a CEFR-linked proficiency rating scale (Cruz, 2016, p. 154)

Stages		*Actions to be carried out*
Stage 3	**Validation 1 (qualitative)**	3.a Validate the rubric qualitatively by consulting other experts. 3.b Fine-tune the rubric following feedback from 3.a.
Stage 4	**Validation 2 (quantitative)**	4.a Analyse through Facets the scores that two raters give to 30 candidates. 4.b Fine-tune the rubric following feedback from 4.a. 4.c Analyse through Facets the scores that five to eight raters give to 40 to 50 candidates. 4.d Fine-tune the rubric following feedback from 4.c.

Methodology

The validation this study suggested slightly differed from the one carried out in the Protocol (Cruz, 2016) because, as mentioned above, it adds a new stage within the qualitative validation in which interviews with raters are performed in order to know the opinion of the actual users of the scale in real conditions who ultimately decide on candidates' level of ability.

The nine subjects involved in the validation of the study can be divided into two groups according to the validation method they contributed to. Firstly, a group of four experts in foreign language teaching, with three of them holding a PhD, participated in the correction and revision of the scale (intuitive validation; see Table 2). Three of the participants are experts in teaching English as a foreign language and one of them in Spanish as a foreing language. All of them are fully familiar with the CEFR (Council of Europe, 2001) and they all have broad experience of designing, administering and/or marking proficiency tests. In addition, three of the four experts are in charge of the CertAcles-UJA test design.

Secondly, the five remaining raters participated in the qualitative validation (see Table 2). All of them are experts in English as a foreign language teaching, with similar levels of qualifications (the minimum being a Master's degree) and they all hold a C2 certification in English proficiency. Additionally, they are all familiar with the CEFR (Council of Europe, 2001) and have received specific training on

Table 2: Scale validation phase, adapted from Cruz (2016)

Stages		*Actions to be carried out*
Stage 3	**Validation 1 (intuitive)**	3.a Validate the rubric by consulting other experts. 3.b Fine-tune the rubric following feedback from 3.a.
Stage 4	**Validation 2 (qualitative)**	4.a Conduct rater training and benchmarking sessions. 4.b Use the scale with real exam samples. 4.c Prepare for the interview. 4.d Interview the raters and analyse the data collected. 4.e Fine-tune the rubric following feedback from 4.d.
Stage 5	**Validation 2 (quantitative)**	5.a Analyse through Facets the scores that two raters give to 30 candidates. 5.b Fine-tune the rubric following feedback from 4.a. 5.c Analyse through Facets the scores that five to eight raters give to 40 to 50 candidates. 5.d Fine-tune the rubric following feedback from 4.c.

language testing. Finally, they all have experience administering and marking proficiency tests since they have participated in previous certification processes in the same context.

As can be seen in Table 2, this new phase was carefully defined and subsequently conducted.

Raters attended the personal inteview after a benchmarking session and after having marked 31 scripts individually with the new scale, so that they could get used to it, identify the strengths and weaknesses it presents, and prepare for the interview they would attend later on.

Since the information elicited from those interviews was intended to be analysed in order to support the validation, a semi-structured interview was scheduled and an interview guide was designed in order to ensure that key topics were covered, that results could be comparable, and that conclusions could be drawn. The interview guide contained a list of relevant aspects about the scale and some probes in order to clarify, extend or check respondents' answers. More specifically, it included questions about the writing construct and scale content (i.e. dimensions taken into account or descriptors), scoring (i.e. differences between bands), usefulness and practicality (i.e. ease in their understanding and use), and general questions about the scale (i.e. usefulness and fairness as a tool for writing assessment).

Finally, the interviews were performed as soon as possible after the benchmarking session and the individual marking period. It must be noted that raters had their markings and the scale in front of them while the interview was conducted. The interview was individual so that each rater could freely express themselves and was recorded for later transcription and analysis of the results.

Results and conclusions

The data collected from the interviews by comparing the raters' opinions and impressions of the scale after having used it provided the following insights.

Regarding the raters' own definition of writing ability, all raters stated that the scale was in general terms a good representation of what writing means for them; that is, of which aspects a piece of writing from candidates should contain.

Focusing on the dimensions as well as the components they include, the raters unanimously agreed that the dimensions were appropriate and well defined, and that they covered all the aspects they would consider when marking a writing task. Nevertheless, some aspects were suggested for revision. When further asked about those aspects, the need for more training in which the dimensions and the components they include are further defined was evidenced.

Regarding the descriptors, two out of the five raters referred to them as clear and elaborated, but two of them also considered it advisable to include examples or clarify some descriptors in further training.

Another relevant aspect of the scale is the difference between the bands; that is, whether raters found it difficult to choose between one band or another taking into account the performance they were marking and the descriptors definitions for each band. Three out of the five raters highlighted the gradual progression and modulation of the bands and they all stated that the bands were appropriately fine-tuned, the difference between each band being very clear.

When asked about their impressions on the scale practicality, all raters highlighted that the scale was easy to use and to understand. Three of the five raters mentioned that the fact that each descriptor contained three components to be considered made the scale more practical since it helped make the decisions, and that the high degree of detail generally facilitated moving along it.

Finally, three raters highlighted the fact that the scale provided them with informed criteria to deliver fairer judgements and that in the majority of the cases their general impressions of the writing after a first reading were quite in accordance with the mark obtained after applying the scale strictly.

Hence, the results of the interviews seemed to suggest that the scale was working considerably well, so it seemed to be a valid tool for the writing assessment of the CertAcles-UJA test. It is true that the raters have reported some difficulties in the use of the scale, mainly related to the dimensions and descriptors definition, but this turned into something positive as having information about those difficulties allowed the addressing of them. Therefore, depending on the nature of the raters' judgements of the scale, at this point it may need to be modified and thus the validation should start from the beginning; or, as in this case, they might spot the main difficulties to be taken into account for future training sessions.

Therefore, the analysis of data collected from the interviews contributed to construct validity, since according to raters, the previously defined construct is operationalised through the scale. It also contributed to content validity as raters' assessment of the scale determined that it covers all the required areas and the criteria are appropriate. Finally, it provided valuable information about the scale usefulness and practicality. Obtaining such valuable information justifies the departure from the original Protocol (Cruz, 2016) and proves to contribute to the scale validation, so it could be easily incorporated to the original Protocol.

This step forward into qualitative validation has proved to contribute significantly not only to the research results but also to the research process, as the preparation steps have also informed the scale behaviour while allowing the identification of strengths and weaknesses, reinforcing the argument of Harsch and Martin (2012) that involving raters has highly beneficial consequences, whether this is in the revision or in the validation phase.

References

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

Council of Europe. (2004). *Reference Supplement to the Preliminary Pilot Version of the Manual for Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching Assessment.* Strasbourg: Language Policy Division.

Cruz, J. (2016). *A protocol to design a CEFR-linked proficiency rating scale for oral production and its App implementation* [Unpublished doctoral dissertation]. University of Jaén, Spain.

East, M. (2009). Evaluating the reliability of a detailed analytic scoring rubric for foreign language writing. *Assessing Writing*, *14*(2), 88–115.

Harsch, C., & Martin, G. (2012). Adapting CEF-descriptors for rating purposes: Validation by a combined rater training and scale revision approach. *Assessing Writing, 17*, 228–250.

Jang, E. E., Wagner, M., & Park, G. (2014). Mixed methods research in language testing and assessment. *Annual Review of Applied Linguistics, 34*, 123–153.

Janssen, G., Meier, V., & Trace, J. (2015). Building a better rubric: Mixed methods rubric revision. *Assessing Writing*, *26*, 51–66.

Knoch, U. (2009). Diagnostic assessment of writing: A comparison of two rating scales. *Language Testing*, *26*(2), 275–304.

Knoch, U. (2016). Validation of writing assessment. In C. A. Chapelle (Ed.), *The Encyclopedia of Applied Linguistics*. Retrieved from: doi.org/10.1002/9781405198431.wbeal1480

Taylor, L. (Ed.). (2011). *Examining Speaking: Research and Practice in Assessing Second Language Speaking*. Studies in Language Testing volume 30. Cambridge: UCLES/Cambridge University Press.

Die Rolle von Lesen-um zu-Schreiben-Aktivitäten im Griechischen Staatlichen Sprachzertifikat

Dafni Wiedenmayer
Fachbereich für Deutsche Sprache und Literatur, Nationale und Kapodistrische Universität Athen

Evangelia (Evelyn) Vovou
Fachbereich für Deutsche Sprache und Literatur, Nationale und Kapodistrische Universität Athen

Abstract

Der vorliegende Beitrag gibt die Kerngedanken sowie die methodologischen Schritte eines laufenden Forschungsteilprojekts wieder, das im Rahmen der wissenschaftlichen Betreuung des Staatlichen Sprachzertifikats (KPG) für die Deutsche Sprache[1] an der Anpassung der beim Modul der schriftlichen Produktion eingesetzten Aufgabentypen arbeitet, und zwar was einen spezifischen Aufgabentyp betrifft, nämlich die integrierten *Lesen-um zu-Schreiben* (LuzS) Sprachtestaktivitäten. Ziel des Beitrags ist zu zeigen, was *LuzS*-Aktivitäten im Rahmen der KPG-Prüfung sind, welche Rolle sie bei der Evaluation kommunikativer Kompetenz spielen, welche Implikationen bei der Gestaltung sowie der Leistungsbeurteilung anhand eines solchen Aufgabentyps zu sehen sind und wie diese im Rahmen des Forschungsprojekts weiter untersucht werden.

Kurze Einführung in die Typologie ‚Lesen-um zu-Schreiben'

Obwohl der Aufgabentyp *Lesen-um zu-Schreiben* sowohl im Fremdsprachenunterricht (i.F. FSU) als auch bei Sprachtests sehr häufig – oder sogar ausschließlich, was das Einüben und die Beurteilung der Schreibfertigkeit angeht – gebraucht wird, ist der Terminus *Lesen-um zu-Schreiben* zumindest in der entsprechenden griechischsprachigen Literatur nicht weit verbreitet. Dies liegt daran, dass sich der Fokus von Forschern und Lehrenden gleichermaßen vielmehr auf das *Schreiben*, d.h. auf die Produktion der Lernenden bzw. Kandidaten, die die eigentliche Leistung verrät, und weniger auf das *Lesen*, nämlich auf den Schreibanlass, richtet. Delaney (2008) rechtfertigt eine solche eher einschränkende Auslegung von *LuzS*-Aktivitäten: „Das Lesen-um zu-Schreiben-Konstrukt kann aus der Perspektive des Lesens, des Schreibens oder des konstruktivistischen Ansatzes untersucht werden, abhängig von der Bedeutung, die den beteiligten Fertigkeiten beigemessen wird"[2] (Delaney, 2008, S. 141). Im vorliegenden Beitrag wird der Standpunkt dargelegt, dass rezeptive und produktive Fertigkeiten in Wechselwirkung zueinander stehen. Demzufolge wäre eine konstruktivistische Annäherung sowohl für die Entwicklung als auch für die Klärung von Forschungsfragen in Richtung wechselwirkender Fertigkeiten bzw. aufeinander aufbauender kognitiver Prozesse vorteilhaft, wie im weiteren Verlauf diskutiert wird.

LuzS-Aktivitäten fallen unter die Definition der *integrierten Aktivitäten* bzw. *Aufgaben*, die im Gegensatz zu *isolierten Aufgaben* „[. . .] die sprachlichen Situationen zu replizieren versuchen, in denen sich die Lerner oft [. . .] befinden"[3] (Lewkowicz, 1997, zitiert in Gebril, 2009, S. 508). *LuzS*-Aktivitäten dienen also dazu, Lernenden bzw. Kandidaten einen Kontext zu verschaffen, der a. auf eine bestimmte kommunikative Situation und ihre soziolinguistischen Normen hindeutet, b. meist eine bestimmte Textsorte hervorruft, und c. bestimmte Gegebenheiten voraussetzt, denen alle Lerner bzw. Kandidaten folgen sollten (Plakans, 2007), um die Aktivität kommunikativ angemessen durchführen zu können. In diesem Sinne sollen also *LuzS*-Aktivitäten so gestaltet sein,

[1] Das Staatliche Sprachzertifikat (KPG) ist eine nationalweite, gestaffelte Sprachprüfung, in deren Rahmen Sprachkenntnisse in folgenden Fremdsprachen auf den Sprachniveaus des Gemeinsamen Europäischen Referenzrahmens für Sprachen (A1–C2) evaluiert und zertifiziert werden: Englisch, Deutsch, Französisch, Italienisch, Spanisch und Türkisch. Die Prüfung für jede Sprache besteht aus vier Modulen: Schriftlicher Ausdruck (und Sprachmittlung bei den Niveaus B1–C2), Mündlicher Ausdruck (und Sprachmittlung bei den Niveaus B1–C2), Leseverstehen und Hörverstehen. Für die wissenschaftliche Betreuung des Staatlichen Sprachzertifikats für die Deutsche Sprache ist Assoc. Prof. Dafni Wiedenmayer zuständig.

[2] „The reading-to-write construct can be examined from a reading, writing, or constructivist approach depending on the importance given to the literacy skills involved" (Delaney, 2008, S. 141).

[3] „[. . .] intended to replicate the language situations that students often encounter [. . .]" (Lewkowczi, 1997 zitiert in Gebril, 2009, S. 508).

dass sie die Lerner- bzw. Kandidatenproduktion auf bestimmte Inhalte lenken, und zwar so effektiv wie möglich, denn auf diese Weise könnten kulturelle Schemata zwischen Lernern/Kandidaten unterschiedlicher Kulturen gemildert oder sogar abgeschafft werden. Der Einsatz also von *LuzS*-Aktivitäten in der didaktischen Praxis und besonders in der Evaluation fremdsprachlicher Kenntnisse dient der inhaltlichen Einheitlichkeit der Lerner- bzw. Kandidatenproduktionen und somit dem Gütekriterium der *Fairness*.

Im Rahmen der KPG-Sprachprüfung für die Deutsche Sprache ist Folgendes in Bezug auf *LuzS*-Aktivitäten festzustellen[4]: In der bisherigen KPG-bezogenen Literatur wird der Schwerpunkt auf das *Schreiben* gelegt und nicht auf die Wechselwirkung zwischen der Lesefertigkeit und der schriftlichen Produktion als aufeinander aufbauende kognitive Prozesse. Dies ist auch begrifflich zu erkennen: Statt auf den Terminus *Lesen-um zu-Schreiben*-Aktivitäten wird in Bezug auf diesen Aufgabentyp auf Testaufgaben für den Schriftlichen Ausdruck, die von multimodalen Stimulus-Texten begleitet werden (Karatza, 2017) oder auf die „[. . .] Vermittlung einer multimodalen Textquelle zu einem monomodalen Ziel(text)"[5] (Dendrinos, 2006, S. 18) hingewiesen (u.a. Dendrinos, 2006, 2009, 2010, 2019; Dendrinos & Mitsikopoulou, 2013). Es wird also bisher Wert auf a. die schriftliche Produktion per se, b. die Multimodalität der Anfangsquelle, c. die Umstellung von Modus (vom Multimodalen zum Monomodalen) bzw. die Behandlung semiotischer Informationen, und nicht auf die Fertigkeit *Lesen* gelegen, die zur Bedeutungserschließung einer (multimodalen) Anfangsquelle erforderlich ist. Man kann behaupten, dass die KPG-bezogene Literatur zwar der Forschung einen Schritt voraus ist, indem sie sich mit Multimodalität im Bereich der Evaluation beschäftigt, die Fertigkeit *Lesen* aber geht dem Modus, in dem eine Aufgabe dem Kandidaten präsentiert wird, voraus. Die Wechselwirkung zwischen den unterschiedlichen kognitiven Prozessen, die einerseits das *Lesen* und andererseits das *Schreiben* evozieren, sowie das Zusammenlegen dieser unterschiedlichen Prozesse in Form schriftlicher Produktion bleiben unerforscht. Das Forschungsprojekt, um das es in diesem Beitrag geht, versucht die obige Beobachtung systematisch zu erfassen.

Zuletzt ist eine zentrale Frage des Forschungsprojekts zu erwähnen, die eine präzisere Definition von *LuzS*-Aktivitäten sowohl bezüglich der Fremdsprachendidaktik als auch der Evaluation angeht. Da in beiden instruktiven Bereichen Schreibaufgaben selten ohne den Anlass einer textuellen Anfangsquelle gestaltet sind (Cumming, Kantor, Powers, Santos, & Taylor, 2000; Hamp-Lyons & Kroll, 1996; Weigle, 2002), bleibt der Unterschied zwischen *integrierten* und *isolierten* Schreibaufgaben eher unklar. Oft wird besonders unter Fremdsprachenlehrenden diskutiert, inwiefern die Aufgabenstellung als textuelle Anfangsquelle zu gelten habe. Nach der kommunikativen Wende und besonders nach dem Einsatz des Gemeinsamen Europäischen Referenzrahmens für Sprachen (GER) beinhalten Aufgabenstellungen oft kommunikative und nicht nur instruktive Elemente, d.h. sie legen den Kontext fest, an dem sich die Lerner bzw. Kandidaten orientieren sollten. Nimmt man also an, dass die Aufgabenstellung oft als textuelle Anfangsquelle dient, dann sind die quantitativen und qualitativen Charakteristika einer Aufgabenstellung, d.h. die Länge, die lexikalische und semantische Dichte usw., in Bezug auf die *Lesefertigkeit* zu berücksichtigen. Welche Charakteristika machen also eine *LuzS*-Aktivität zu einer *LuzS*-Aktivität?

Überblick der Forschungsfragen und des methodischen Vorgehens

Außer der oben beschriebenen Forschungsfrage fokussiert das Forschungsprojekt zzt und in Bezug auf die KPG-Prüfung vorwiegend auf folgende Aspekte: (i) die Charakteristika der untersuchten *LuzS*-Aktivitäten in Bezug auf den Zusammenhang zwischen der Lese- und Schreibkompetenz der Prüflinge, (ii) die Charakteristika der *LuzS*-Aktivitäten in Bezug auf die Formulierung von Items (Fragepunkten), (iii) die Funktion von multimodalen Schreibstimuli als Teil von *LuzS*-Aktivitäten, (iv) den Zusammenhang zwischen den oben erwähnten Charakteristika und dem verwendeten Bewertungsraster, und (v) die potentielle Diskrepanzen zwischen unterschiedlichen Bewertern angesichts der Natur von *LuzS*-Aktivitäten. Die dargestellten Forschungsziele setzen an einzelnen Phasen eines umfangreicheren methodologischen Vorgehens an, das vornehmlich anhand einer textlinguistischen Datenanalyse von a. Aufgabenstellungen der KPG-Prüfungen auf B-Niveau (B1&B2) im Zeitraum zwischen 2001 und 2019 und b. Kandidatenproduktionen derselben Prüfungsperioden empirisch zu beantworten versucht (Wiedenmayer & Vovou, Forthcoming 2021). Sowohl Aufgabenstellungen als auch Kandidatenproduktionen wurden vorerst – nach den textanalytischen Kriterien von Hausendorf und Kesselheim (2008) – anhand ihrer Organisationsformen und der daraus abgeleiteten Muster analysiert. Zu den anfänglichen Ergebnissen gehört die Feststellung, dass eine bestimmte Textsortenkompetenz vonseiten der Kandidaten erforderlich ist bzw. bei den Kandidaten entwickelt wird, die mit den Charakteristika der *LuzS*-Aktivitäten korreliert.

[4] Eine detaillierte Analyse der Merkmale von *LuzS*-Aktivitäten in der KPG-Prüfung für die Deutsche Sprache sowie eine Beschreibung deren Entwicklung im Zeitrahmen zwischen 2011 und 2019 ist in Wiedenmayer/Vovou (2021, in print) zu finden.

[5] „[. . .] mediating multimodal source text to monomodal target" (Dendrinos, 2006, S. 18).

Fazit und künftige Forschungsschwerpunkte

In Anlehnung an die oben beschriebenen Forschungsfragen ist für den Zeitrahmen 2021–2022 eine Erweiterung des Forschungsprojekts geplant, die sich auf folgende Schwerpunkte stützt:

- Untersuchung der Korrelation zwischen Lese- und *LuzS*-Sprachtestaktivitäten in Bezug sowohl auf die Kandidatenperformanz als auch auf die in Kraft tretenden kognitiven Prozesse.
- Untersuchung der Korrelation zwischen Schreib- und *LuzS*-Sprachtestaktivitäten in Bezug sowohl auf die Kandidatenperformanz als auch auf die in Kraft tretenden kognitiven Prozesse.
- Begutachtung und Evaluation der *LuzS*-Aufgabenstellungen anhand eines horizontalen Vergleichs der Kandidatenleistung in unterschiedlichen *LuzS*-Aktivitäten.
- Einschätzung der Rolle von Querschnittkompetenzen, wie z.B. kritischem Denken, bei der Auseinandersetzung mit *LuzS*-Sprachtestaktivitäten.
- Untersuchung des Phänomens des Plagiats anhand von *LuzS*-Aufgabenstellungen vonseiten der Sprachtestkandidaten und sein Einfluss auf die Validität der Aufgabenstellung sowie auf die Produktion.
- Einschätzung der Performanz von Fremdsprachenlernern bei *LuzS*-Aktivitäten in Vergleich zu ihrer Performanz bei *isolierten* Schreibaktivitäten, wie z.B. Tagebucheinträgen usw.

Die Untersuchung von *LuzS*-Aktivitäten im Rahmen der Evaluation schriftlicher Kompetenz, und zwar aus kognitiver Sicht, bietet einen Einblick in die Prozesse, die dem Lesen und dem Schreiben unterliegen. Die Lesefertigkeit könnte z.B. die organisatorischen und bedeutungstransformierenden Operationen, die bei der Verwirklichung einer schriftlichen Aufgabe vorhanden sind, beeinflussen. Die empirische Untersuchung zielt darauf ab, Belege für die obige Hypothese zu liefern und die Folgen für die Beurteilung schriftlicher Produktion ans Licht zu bringen. *LuzS*-Aktivitäten schaffen also eine fruchtbare Grundlage für zukünftige Forschung.

Literatur

Cumming, A., Kantor, R., Powers, D., Santos, T., & Taylor, C. (2000). *TOEFL 2000 Writing Framework: A Working Paper*. TOEFL Monograph Series, Report no. 18. Princeton: Educational Testing Service.

Delaney, Y. A. (2008). Investigating the reading-to-write construct. *Journal of English for Academic Purposes, 7*, 140–150.

Dendrinos, B. (2006). Mediation in communication, language teaching and testing. *JAL, 22*, 9–35.

Dendrinos, B. (2009). *Rationale and ideology of the KPG exams*. Retrieved from: rcel2.enl.uoa.gr/kpg/gr_kpgcorner_sep2009.htm.

Dendrinos, B. (2010). Genre-based writing in the KPG exams. *The KPG Corner*. Retrieved from: rcel.enl.uoa.gr/directions/issue1_2d.htm

Dendrinos, B. (2019, May). *Multilingual testing and assessment in different educational contexts* [Conference presentation]. ICC 26th Annual Conference, Berlin, 2019.

Dendrinos, B., & Mitsikopoulou, B. (2013). *The KPG Writing Test in English: A Handbook*. Athens: University of Athens/RCeL Publications.

Gebril, A. (2009). *Score Generalizability in Writing Assessment: The Interface between Applied Linguistics and Psychometrics Research.* Saarbrücken: VDM Verlag Dr. Müller.

Hamp-Lyons, L., & Kroll, B. (1996). Issues in ESL writing assessment: An overview. *College ESL, 6*(1), 52–72.

Hausendorf, H., & Kesselheim, W. (2008). *Textlinguistik fürs Examen*. Göttingen: Vandenhoeck & Ruprecht.

Karatza, S. (2017). *Analysing multimodal texts and test tasks for reading comprehension in the KPG exams in English* [Doctoral dissertation]. National and Kapodistrian University of Athens, School of Philosophy, Faculty of English Language and Literature.

Plakans, L. (2007). *Second language writing and reading-to-write assessment tasks: A process study* [Unpublished doctoral dissertation]. University of Iowa.

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

Wiedenmayer, D., & Vovou, E. (Forthcoming 2021). Validität und Reliabilität des Lesen-um zu-Schreiben-Konstrukt im Griechischen Staatlichen Sprachzertifikat. *Lexis, 2.*

Evaluating the fairness of a university English for Academic Purposes test

Gyula Tankó
Department of English Applied Linguistics, Eötvös Loránd University, Budapest

Abstract

In the large-scale testing of English for Academic Purposes (EAP), there has been a distinct shift to the use of integrated tasks to assess candidates' language ability by engaging them in target language use tasks relevant for the tertiary education domain. A similar shift is needed in small-scale EAP assessment because integrated tasks are considered to be more authentic and to have a positive washback effect on academic skills courses. This paper presents the results of an ongoing assessment justification project designed to investigate whether an EAP test measuring academic reading, summarisation, paraphrasing and academic prose writing developed for a university English language programme meets the fairness criteria specified in Bachman and Palmer (2010) and Bachman and Damböck (2018) based on the analysis of the test documentation, test administration procedures, and empirical evidence from a series of research studies that investigated various aspects of the EAP test.

Introduction

English for Academic Purposes (EAP) has become a mainstream discipline as a result of the internationalisation of tertiary education and the use of English as the language of science and technology. Written academic interaction through research articles (Tankó, 2017) or such educational genres as essays, reports or theses, has been thoroughly researched and described for teaching purposes (e.g., Hyland, 2000; Paltridge, 1997; Swales & Feak, 2004). Large-scale high-stakes international examinations (e.g. IELTS, TOEFL) started to use integrated tasks to assess academic writing, which can be defined as 'a writer-responsible, referential, often argumentative, transactional, and conventionalised type of writing' (Tankó, 2012, p. 11). Moreover, tertiary educational institutions with English language programmes have also started developing their own post-entry EAP examinations (e.g., Read, 2015). One such examination is the Academic Skills Test (AST) developed at a large university in Budapest, Hungary. The high-stakes test on which later course enrolment depends is administered annually at the end of an academic skills course to about 450 first year English majors. It consists of a guided summary task: the test-takers read a 700-word-long self-contained academic text in English with a selective reading goal; summarise and paraphrase the relevant propositional content; and write it up as a short academic text. The AST operationalises four constructs: (1) academic reading (i.e., reading selectively with a pre-set goal using academic reading behaviours) (cf. Khalifa & Weir, 2009); (2) summarisation (condensing ideational content by using the construction, generalisation, deletion and zero macro rules) (cf. Van Dijk, 1980); (3) paraphrasing (giving a new form to the summarised content while retaining specialised terminology) (cf. Tankó, 2019); and (4) academic writing (producing a short text that matches the schematic structure, organisation, register, mechanics, and APA style in-text citation use norms of academic prose). The scripts are double rated with an analytical rating scale. The raters are trained, and their work is monitored at each administration.

Fairness must be the concern of test developers and users whenever an advancement decision like the one based on the AST is made (Bachman & Palmer, 2010). This article presents the results of an ongoing assessment justification project also investigating whether the AST meets the criteria of fairness (cf. Bachman & Palmer, 2010; Bachman & Damböck, 2018). The investigation was guided by the Assessment Use Argument (AUA) (Bachman & Palmer, 2010) and relied on data obtained from the test documentation (i.e., the academic skills course syllabus, test specification, and item writer's guidelines), test administration procedures, and empirical evidence from research studies on the AST test.

Fairness in language testing

Fairness is an essential characteristic of an assessment (Bachman & Palmer, 2010) as the trust of stakeholders hinges on it (Bachman & Damböck, 2018). The concept is complex because 'fairness is not a single quality, but is a function of many different aspects of not only the assessment process itself, but also the use of assessments' (Bachman & Palmer, 2010, p. 128). Its facets are at the core of the language assessment approach proposed by Bachman and Palmer (2010) and are operationalised in the AUA, which is a conceptual framework for language test development and use.

The AUA has a primary, subordinative argumentation structure (Van Eemeren, Grootendorst, & Snoeck Henkemans, 2002). It consists of a chain of argument in which four main arguments connect performance with assessment use consequences. The four main claims are inferential conceptual links that formulate assertions about the intended consequences of using an assessment, decisions to be made to bring about those consequences, interpretations that inform the decisions, and assessment records arrived at through the observance and recording of test-taker performance. The claims have an outcome (i.e., consequences, decisions, interpretations, and assessment records) and qualities (e.g., beneficence or consistency) elaborated by warrants and their backings. As each main argument has a number of warrants and each claim-warrant-backing triad constitutes a separate sub-argument, the sub-arguments pertaining to each claim are instances of coordinative argumentation and therefore constitute the secondary argumentation structure of the AUA.

The AUA has two fundamental functions. With a top-down approach (i.e., consequences-to-performance), it serves as a tool for assessment development. Conversely, with a bottom-up approach, it describes the process of assessment use and functions as a comprehensive framework for assessment justification. In either of its uses, it is to be adapted to the specific assessment situation, for example, with the selection or addition of applicable warrants.

The AUA operationalises fairness in two ways. Bachman and Palmer (2010) conceptualised fairness as a multifaceted sub-argument level quality articulated by warrants. This may be regarded as the primarily assessment development interpretation of fairness. Bachman and Damböck (2018) elucidated its second dimension when they considered it primarily from an assessment use perspective as an overall AUA-level quality. As such, it denotes 'how well each link from students' performance on assessment tasks, to assessment records, to decisions, and to consequences can be supported or justified to stakeholders. If any link cannot be justified, then the assessment *use* is not likely to be fair' (p. 28). Given that the AUA has a complex argumentation structure which combines local coordinative argumentation embedded into a global subordinative argumentation structure, the presence and cogency of the coordinative arguments entail the presence and cogency of the inferential links that connect the four claims in the AUA and that make an assessment fair.

In the next section, the fairness of the AST is analysed through the evaluation of the extent to which it meets the criteria of fairness operationalised in the AUA. The discussion follows Bachman and Palmer's (2010) organisation of the aspects of fairness.

Fairness in the Academic Skills Test

Equality of treatment

Equality of treatment criteria are operationalised in each main AUA argument. Evidence about whether the AST is administered under the same conditions across different occasions is provided by the test administration procedures, which regulate the test administration and are reviewed yearly. Additionally, a study was conducted on the context validity (cf. Weir, 2005) of the AST (Tankó, 2015). Questionnaire data and test results collected from 234 first year English majors showed that the fairness criteria stated in the consistency warrants were partially met because the test administration was uniform and followed the specification of the procedures, but the study also revealed some issues to be addressed (e.g., students reported to have been disturbed by invigilators talking among themselves during the test). The study suggested some anxiety-reducing improvements (e.g., lengthening of test time, more simplified seating protocol, more efficient collection of test papers).

Evidence that the AST meets the fairness criteria stated in the impartiality warrants can be found in the item writer's guidelines, which regulate response format and content (e.g., the reading passage cannot be offensive and must be on a general academic topic students may encounter in their content courses); the online test specification describing the assessment content and procedures; and the unified course syllabus, which ensures that the students have equal opportunities to prepare for the assessment as they receive uniform input in the academic skills courses. Students have equal access to the AST and the accommodation of special needs students is regulated by the administration procedures.

A follow-up study to the initial task piloting conducted on the AST response format investigated the optimal length constraints set for the guided summary writing task (Szőcs, 2019). A discourse analysis of student scripts confirmed the originally set length specification by showing that it allowed for the inclusion of the highest number of accurately reproduced, summarised, and paraphrased content points.

The AST also meets the fairness criteria specified in the equitability warrants. By means of the annual rater trainings, the tutors rating the scripts internalise and practise using the rating scale, and in the process learn to classify students according to set criteria. Furthermore, the first reader of each script must be a rater other than the test taker's academic skills tutor to ensure that the classification is less influenced by course work. The academic skills course syllabus stipulates that students must be familiarised with the test specification to understand the way the scripts are assessed and decisions are made. Additionally, tutors follow a unified syllabus focusing on the abilities assessed in the AST and provide extensive summary writing practice, so students are given equal opportunities for development.

The fairness of the AST could be improved with blind double rating or by using raters who have not taught the student, but these are impractical options. What needs to be added to the syllabus is a compulsory mock test so that students can experience completing the integrated task under time constraints.

Finally, the AST administration procedures ensure that the assessment reports of individual students are treated confidentially as only the test-taker, the raters, and the course coordinator have access to the scripts and scores. The procedures also state that students must be informed about their results in a clearly understandable manner within two weeks after the test. This indicates that the fairness criteria of the beneficence warrants are met.

Absence of bias

The fairness criteria about the absence of bias are stated in the impartiality warrants. Measures are taken to eliminate the potential bias in the scoring method during rating scale revisions, task design sessions, and rater trainings. The key (i.e., the content points to be included in the summary) is checked by the item writer team and the academic skills tutors. The scoring procedures and decision rules are discussed and practiced during rater trainings. Inconsistent scoring concerns are addressed before and after each test administration. Before the rater training, raters submit their scores for benchmarked scripts. Their scores are analysed and the assessment criteria with notable heterogeneous variance are discussed in more detail in the training. After the test, the academic skills coordinator selects and rates a set of scripts and compares their scores to the raters' final and individual scores. The monitoring revealed that inter-rater consistency is still an occasional problem as regards tutors who have just started teaching the academic skills course or who are less familiar with language assessment. Occasionally, the differences in assessment performance across groups were found not to be due to differences in the ability assessed by the AST. Therefore, some academic skills tutors need input on the constructs and abilities developed by the course and on language assessment in general, and the fluctuation of tutors must be reduced, taking into account that EAP courses are not mere 'skills' courses that anyone can convey.

A study (Tankó, 2021a) conducted with 445 English majors on the AST at the time when it used to consist of two tasks, the current integrated and an argumentative essay task, found gender-related differences in the written production scores of the students: although syntactic and lexical analyses showed that male writers significantly outperformed them, their female peers received significantly higher total scores for their essays. However, it is reassuring that a follow-up study investigating differential rater behaviour in the case of 221 guided summaries elicited with the new version of the AST found no evidence of such bias (Tankó, 2021b in press).

Assessment use

Two related studies investigating the guided summary writing processes of first-year English majors using think-aloud as a research method (Szűcs, 2015, 2020) confirmed that the AST task engages the abilities intended to be measured. Two additional studies motivated by practicality concerns explored which of two academic writing tasks, an essay or a summary, elicited academic prose more efficiently and whether keeping only one of the two tasks provided sufficient information for EAP proficiency assessment (Tankó, 2016, 2020). Linguistic complexity analyses indicated that the summary was the significantly better task for academic prose elicitation, and that it provided more relevant information for the decision to be made based on the AST. These four studies also provided evidence in support of the meaningfulness warrant.

An interview study (Stemmer, 2019) conducted with content course teachers from the same English programme in which the AST is administered demonstrated that the interpretations made based on the AST were generalizable beyond the assessment itself. Teachers from five different departments (i.e., literature, history, linguistics, applied linguistics, and language pedagogy) reported that in their courses they used summarisation as an instructional and assessment tool.

Conclusion

The results available from the ongoing AST justification project have shown that the AST meets the majority of the fairness criteria articulated in the warrants in the AUA. Given the relationship between the sub- and main arguments in the AUA and their contributions to fairness, the evidence of the presence of fairness as a sub-argument-level quality can be assumed to indicate the presence of fairness as an overall AUA-level quality, which is reassuring.

Some of the weaknesses found have already been addressed (e.g., Tankó, 2015), some cannot be resolved due to the unavailability of resources, and others need to be dealt with in the future. The AUA-informed systematic investigation of the fairness of the AST has already resulted in and is expected to lead to further improvements of the AST.

References

Bachman, L. F., & Damböck, B. (2018). *Language Assessment for Classroom Teachers*. Oxford: Oxford University Press.

Bachman, L. F., & Palmer, A. S. (2010). *Language Assessment in Practice: Developing Language Assessments and Justifying Their Use in the Real World*. Oxford: Oxford University Press.

Hyland, K. (2000). *Disciplinary Discourses: Social Interactions in Academic Writing*. London: Longman.

Khalifa, H., & Weir, C. (2009). *Examining Reading: Research and Practice in Assessing Second Language Reading*. Studies in Language Testing volume 29. Cambridge: UCLES/Cambridge University Press.

Paltridge, B. (1997). *Genre, Frames and Writing in Research Settings*. Amsterdam: John Benjamins.

Read, J. (2015). *Assessing English Proficiency for University Study*. Basingstoke: Palgrave Macmillan.

Stemmer, G. (2019). *The use of summarisation as an educational tool in university content classes* [Unpublished master's thesis]. Eötvös Loránd University, Budapest.

Swales, J. M., & Feak, C. B. (2004). *Academic Writing for Graduate Students* (2nd ed.). Michigan: University of Michigan.

Szőcs, E. (2019). *The effect of limited and unlimited length conditions on the quality of guided summaries written by English major BA university students* [Unpublished master's thesis]. Eötvös Loránd University, Budapest.

Szűcs, Á. (2015). *The summary writing processes of first year EFL learner BA students* [Unpublished master's thesis]. Eötvös Loránd University, Budapest.

Szűcs, Á. (2020). *Reading English for academic purposes: The reading processes of first year EFL learner BA students* [Unpublished doctoral dissertation]. Eötvös Loránd University, Budapest.

Tankó, Gy. (2012). *Professional Writing: The Academic Context* (Rev. ed.). Budapest: Eötvös University Press.

Tankó, Gy. (2015). Investigating the context validity of an English academic writing test. In D. Holló & K. Károly (Eds.), *Inspirations in Foreign Language Teaching: Studies in Language Pedagogy and Applied Linguistics in Honour of Péter Medgyes* (pp. 128–150). New York: Pearson.

Tankó, Gy. (2016). Written summarisation for academic writing skills development: A corpus-based contrastive investigation of EFL student writing. In N. DobriĐ, E.-M. Graf, & A. Onysko (Eds.), *Corpora in Applied Linguistics: Current Approaches* (pp. 53–78). Cambridge: Cambridge Scholars Publishing.

Tankó, Gy. (2017). Literary research article abstracts: An analysis of rhetorical moves and their linguistic realizations. *Journal of English for Academic Purposes*, *27*, 42–55.

Tankó, Gy. (2019). *Paraphrasing, Summarising and Synthesising Skills for Academic Writers: Theory and Practice* (Rev. ed.). Budapest: Eötvös University Press.

Tankó, Gy. (2020). Eliciting written academic performance: An investigation of task effect on the written production of EFL students. In Cs. Kálmán (Ed.), *DEAL 2020: A Snapshot of Diversity in English Applied Linguistics* (pp. 1–37). Eötvös Loránd University.

Tankó, Gy. (2021a). Gender-related differences in performing a test-task in academic writing: Insights from performance data on an argumentative essay task. In É. Illés & J. Kenyeres (Eds.). *Changing Perspectives: Studies in English at Eötvös Loránd University* (pp. 76–94). Budapest: Eötvös Loránd University.

Tankó, Gy. (2021b in press). Fairness in integrated academic writing ability assessment: An empirical investigation of the effect of test-taker gender on task performance. In Gy. Tankó & K. Csizér (Eds.), *DEAL 2021: Current Explorations in English Applied Linguistics* (pp. 1–31). Budapest: Eötvös Loránd University.

Van Dijk, T. A. (1980). *Macrostructures*. Mahwah: Lawrence Erlbaum.

Van Eemeren, F. H., Grootendorst, R., & Snoeck Henkemans, A. F. (2002). *Argumentation: Analysis, Evaluation, Presentation*. Mahwah: Lawrence Erlbaum.

Weir, C. J. (2005). *Language Testing and Validation: An Evidence-based Approach.* Basingstoke: Palgrave Macmillan.

Needs analysis for the examination and course of language for medical academic purposes

Dina Vîlcu
Babeş-Bolyai University, Romania

Abstract

The linguistic premises for the prospective students following higher education in a foreign language are a matter more serious than many stakeholders are prepared to admit. Needs analysis (NA), which is normally at the core of any valid, fair and reliable language for specific purposes (LSP) examination, is rarely performed for LSP courses which are part of academic language programmes (LSAP). Consequently, the logistic, material and didactic resources necessary for them are hard to be fully understood, accepted and met.

The present paper aims to prove that NA, even if rather difficult to perform at the level of exams and courses which are part of academic language programmes, is extremely relevant, its results contributing in a great measure to understanding target language use (TLU) and to tailoring the course and the examination for the students' needs.

Introduction

The relevance of NA for the courses and examinations of LSP was claimed decades ago. Since then, NA has been not only extensively implemented and researched, but also questioned and reshaped (Hutchinson & Waters, 1987; Dudley-Evans & St John, 1998; Fulcher, 1999; Douglas, 2001; Long, 2005; O'Sullivan, 2012; Brunfaut, 2014). The LSP exam developers have often been careful to base their construction and administration process on the results obtained from interviews and surveys targeting different categories of stakeholders, consultations with experts or observation at the workplace.

The construction and administration process of LSP examinations is particularly difficult. The target population of such examinations is usually lower in number than that of general language tests, and this impacts on the costs involved, the possibility of ensuring human and material resources, and on the possibility of determining the candidate profile (Van Gorp & Vîlcu, 2018). Other factors might complicate the process even further: a less widely tested language, a domain with very few candidates (compare, for instance, music or philosophy with medicine), the difficulty of obtaining texts/materials from the domain due to the confidential nature of the work, as well as the candidates' preparation for entering not only the work market, but also academic training programmes (Elder, 2016; Hyland, 2016; Ingham & Thighe, 2006).

The providers of courses and examinations of LSAP often experience such complications. Given that they sometimes struggle to ensure even the basic needs for developing the course and the examination, like qualified human resources, specific materials or sufficient teaching time (Krajka, 2018), they can rarely develop the stage of NA (Vîlcu & Van Gorp, 2018). This paper aims to show that the strenuous process of developing the NA stage may be rewarded with better quality LSAP courses and examinations. We had in view, with the NA process described here, the language for medical academic purposes (LMAP) examination. However, its impact on the LMAP courses is expected to be very consistent. Since new types of tasks and of texts will be introduced in the examination specifications and the weight among the components of the examination will be changed, these aspects will be reflected in the activities the teachers include in the course, resulting in a better preparation of the students for the medical studies they will pursue in the following years.

The present study was focused on the course of LMAP, which is part of the language year at Babeş-Bolyai University, attended by foreign students who will do their medical studies in Romania, some of them in Romanian, the others in French or English. Given the limited space, the results will be presented here only in part, and will focus on the findings concerning the students' needs in terms of listening comprehension, showing: the instruments used for NA; the identified needs of the subjects; the problems revealed by NA; and the solutions proposed by specialists.

The study was conducted between 2018 and 2020 and was developed in close collaboration with the faculty members teaching the course of LMAP, and with the assessment experts participating in the language for specific purposes special interest group (LSP SIG) of the Association of Language Testers in Europe (ALTE) (www.alte.org/SIGs).

The semi-structured interviews

The first stage of NA consisted of a series of semi-structured interviews with test-takers who had graduated from the LMAP course, currently studying general medicine or kinesiotherapy. At the time when the interviews were conducted, the participants were in different years of study (between the first and the third). They had graduated in the top 20% of their general language studies (Levels A1–B2).

The questions they were addressed during the interview were related to: the difference in experience as students in the language year and in the medical school; how helpful the LMAP course and examination were; suggestions concerning the LMAP course and examination; frequency of use of Romanian in their medical studies; their opinion on the level of difficulty of the LMAP examination; and receptive/productive activities they thought they should have practised more during the LMAP course.

In this paper I will report on the findings related to the listening aspect involved in the learning and assessment activities, findings which gave the exam developers a new perspective on the students' needs and problems once they begin their medical studies.

Starting with the answers to the very first question (How different was your experience as a student in the preparatory year and the one in the faculty where you study now?), the students said that the first semester was really difficult ('a nightmare', one of them said); not because of the course content, but because of the language, in particular understanding their colleagues and teachers. Being integrated into groups of Romanian students, they were faced with the inherent variations in native pronunciation, contrary to their previous experience with students in the language year and general Romanian teachers, who used mainly standard pronunciation. Furthermore, the teachers used a lot of idioms, abbreviations and changes in rhythm and intonation which were problematic for them. A surprising answer came from one of the really good students in general Romanian, who said that in the first semester he did not participate in the courses at all because he did not understand the teachers, learning instead at home from the textbooks and written course materials.

When asked what would be a good addition to the LMAP course, one of the main suggestions was more listening. They also suggested more interaction in class, their proposals revealing the pressing need for an increased number of listening/speaking activities.

The semi-structured interviews are a very useful first step in the NA process (Datko, 2015; Long, 2005), given the researcher's facility for guiding the discussion to obtain the most relevant information. In our particular case, this step proved extremely useful in designing the questions for the subsequent survey.

The survey

Based on the results from the semi-structured interviews and on the literature in the domain (de Leeuw, Hox, & Dillman, 2008; Douglas, 2000; Long, 2005), the questionnaires were designed and distributed to different categories of informants. Essential information came from the students in the medical school, related to their needs when they started studying in Romanian and the problems which hindered their learning process.

The listening-related questions were broken down into the following situations of communication:

How important are/were the following activities in Romanian language during your medical studies: listening to

- medical courses taught by teachers;
- presentations on medical topics given by fellow students;
- lectures and conferences on medical topics;
- audio and video medical sources;
- conversations on medical topics;
- conversations on general topics;
- patients?

The questions targeted the same situations for the students enrolled in the LMAP course, but phrased differently, so as to measure their expectations (How important do you think the following activities will be during your medical studies?). The same situations were included in the questions addressed to the LMAP teachers, with the purpose of understanding how these kinds of activities are part of their courses (How important are the following activities for the course of medical language?).

A comparative representation of the results obtained for these questions resulted in:

- a better understanding of the students' needs, in terms of listening comprehension

- a reshaping of the expectations of the students in the LMAP course (more realistic, better preparation for the activities they are to participate in)
- the possibility of presenting the teachers with the relevance of the listening activities to their courses, as well as suggestions of the type of input and tasks most beneficial for the students.

The results, needs/problems and solutions

The results in Table 1 indicate the percentage of respondents who chose the answers 'very important' and 'important' for the questions above.

Table 1: Answers to the questions related to the importance of listening activities

Activities	*Students in or graduates from the medical school*	*Students enrolled in the LMAP course at the time of the survey*	*LMAP teachers*
Listening to medical courses taught by teachers	90.5%	80%	100%
Listening to presentations on medical topics given by fellow students	71%	80%	85%
Listening to lectures and conferences on medical topics	76%	80%	85%
Listening to audio and video medical sources	80%	80%	71.5%
Listening to conversations on medical topics	85%	84%	85%
Listening to conversations on general topics	71%	76%	57%
Listening to patients	95%	80%	71.5%

Using the answers coming from the students in or graduates from the medical school as our point of reference and drawing on the other two perspectives, we can conclude that:

- the percentages do not reflect contradicting views on the listening aspect of the learning context in the medical school, but rather the necessity for a better understanding of the needs related to listening comprehension
- the teachers tend to focus on the types of listening input (and tasks) specific to courses, lectures, presentations – contexts in which the students' participation is rather passive; far from representing an unessential aspect of listening activities, they might be balanced with those involving more interaction
- the students in the LMAP course could be exposed more to real discourse, which can then be analysed and explained
- the point above relates in a revealing way to one of the lowest percentages we discover in the teachers' answers, in terms of coverage during the course: listening to online medical audio and video sources; more materials of this type could be included in the course, as the students need to work with them once they start their medical studies
- the lowest percentage in the teachers' appreciation of the importance of listening activities is listening to conversations on general topics; however, these situations seem to take up a large part in the students' regular interactions; the communication with the teachers and colleagues on general topics is essential for the students' integration in the learning process
- the greatest difference between the teachers' and the medical school students' answers concerns the activity of listening to the patients (rated the highest in importance by the students in the medical school); the understanding of the patients' speaking is often hindered by a number of factors (state of health, age, region of origin, etc.) and practising their understanding should have a larger part in courses and be reflected in examinations
- a more similar perspective is shared by the students in the medical school and the LMAP teachers when it comes to listening to the courses taught by teachers; however, the difficulties indicated by the students in this respect should result in more

practice with samples of real medical courses, with an emphasis on the analysis of the language aspects (not so much of content)

- the range of percentages coming from the students in the medical school varies between 71% and 95% when it comes to the importance of listening activities, which is much more nuanced than the one coming from the students enrolled in the LMAP course situated between 76% and 84%, indicating a rather logical assumption that all the listening activities included in the questionnaire are rather important for the medical studies and for their activity in hospital; the percentage given by the students in the LMAP course is 80% for five out of the seven questions related to listening in the questionnaire, indicating a normal presupposition about these, in the absence of real experience; however, their expectations, as well as their preparation for the different types of listening contexts can be adjusted to the real needs revealed by the students already in the medical school, resulting in a more adequate preparation for the academic studies in the medical field.

Conclusions

Facing a new programme of academic studies in a new language is a real challenge for any student. The language teachers make huge efforts in order to prepare their students for this step. Adequate instruments are necessary for this and NA proves to be one of the essential ones.

The goal of this research was that of redesigning the examination and consequently the course of LMAP at the Department of Romanian language, culture and civilisation, Babeș-Bolyai University. New specifications were necessary as was a new perspective on the LMAP course. The experts participating in the LSP SIG of ALTE suggested significant changes concerning the listening component of the examination, in terms of type of tasks included (more varied, integrated), of difficulty (higher level), and of proportion (more listening tasks included). The results also seem to have a great impact on the syllabus of the LMAP course.

The questionnaires we sent as part of the survey included a section of open questions, which returned extremely relevant answers from the students.

The process of NA brought, in the case of the LMAP examination at Babeș-Bolyai University, crucial information and insight related to the students' needs and the problems they face when starting their academic training in a new language. The providers of LSAP courses are thus strongly encouraged to perform the stage of NA, as much can be done for the students' psychological comfort, trust and better results during their studies. The effort of adapting our courses and examinations to their needs is entirely worth making.

References

Brunfaut, T. (2014). Language for Specific Purposes: Current and future issues. *Language Assessment Quarterly*, *11*(2), 216–225.

Datko, J. (2015). Semi-structured interview in language pedagogy research. *Journal of Language and Cultural Education*, *3*(2), 142–156.

de Leeuw, E., Hox, J., & Dillman, D. (Eds.). (2008). *International Handbook of Survey Methodology*. Abingdon: Routledge.

Douglas, D. (2000). *Assessing Language for Specific Purposes*. Cambridge: Cambridge University Press.

Douglas, D. (2001). Language for Specific Purposes assessment criteria: where do they come from?. *Language Testing, 18*(2), 171–185.

Dudley-Evans, T., & St John, M. J. (1998). *Developments in English for Specific Purposes: A Multi-Disciplinary Approach*. Cambridge: Cambridge University Press.

Elder, C. (2016). Exploring the limits of authenticity in LSP testing: The case of a specific-purpose language test for health professionals. *Language Testing, 33*(2), 147–152.

Fulcher, G. (1999). Assessment in English for Academic Purposes: Putting content validity in its place. *Applied Linguistics*, *20*(2), 221–236.

Hutchinson, T., & Waters, A. (1987). *English for Specific Purposes. A Learning-Centred Approach*. Cambridge: Cambridge University Press.

Hyland, K. (2016). General and specific EAP. In K. Hyland & P. Shaw (Eds.), *The Routledge Handbook of English for Academic Purposes* (pp. 17–29), Abingdon: Routledge.

Ingham, K., & Thighe, D. (2006). Issues with developing a test in LSP: the International Certificate in Financial English. *Research Notes*, *25*, 5–9.

Krajka, J. (2018). The ESP teacher as a researcher – From needs analysis to materials development. *Scripta Manent, 13*(1), 2–25.

Long, M. (2005). *Second Language Needs Analysis*. Cambridge: Cambridge University Press.

O'Sullivan, B. (2012). Assessment issues in Language for Specific Purposes. *The Modern Language Journal, 96*, 71–88.

Van Gorp, K., & Vîlcu, D. (Eds.). (2018). *Guidelines for the Development of Language for Specific Purposes Tests. A Supplement to the Manual for Language Test Development and Examining*. Produced by ALTE. Retrieved from: www.alte.org/resources/Documents/6093%20LSP%20Supplement%20-%20WEB.pdf

Vîlcu, D., & Van Gorp, K. (2018, April 13). *Developing resources for LSP tests. A reflection* [plenary presentation]. 51st ALTE Meeting and Conference Day, Cluj-Napoca, Romania.

Enabling interaction: A study of examiner behaviour in an interactive speaking test

Richard Harris
Trinity College London, United Kingdom

Abstract

Reliability is a primary concern for high-stakes speaking assessments though this can be at the expense of authenticity. Nevertheless, authentic interaction and representation of the complete construct in speaking assessments remains an ongoing aspiration. This article describes an investigation of the Interactive Task in Trinity College London's Graded Examinations in Spoken English (GESE). This task transfers responsibility to the candidate to take control of their interaction with the examiner. Using transcripts from the Trinity-Lancaster Corpus, grounded theory methodology was employed to identify patterns of examiner behaviour in Interactive Task performances. The data yielded 886 examples of performance that followed examiner strategies. Using non-parametric ANOVA, the study then quantitatively compared indices of candidate performance that had been elicited by each examiner strategy. The small and medium effect sizes discovered in the differences between some strategies indicate the authenticity achieved through the freedom afforded examiners to interact spontaneously does not bias performance.

Introduction

Trinity College London's GESE

Trinity College London's Graded Examinations in Spoken English (GESE) is a suite of speaking and listening exams of 12 grades in four stages placed at a user level from the Common European Framework of Reference for Languages (CEFR, Council of Europe, 2001): Initial (Grades 1–3; CEFR Basic User), Elementary (Grades 4–6; CEFR Basic User to Independent User), Intermediate (Grades 7–9; CEFR Independent User), Advanced (Grades 10–12; CEFR Proficient User). Depending on the stage, each exam is composed of one or more of the following sections: the Conversation, the Topic Presentation, the Topic Discussion, the Interactive Task and the Listening Task. In a GESE exam, there is one candidate and one examiner, who acts as both interlocutor and rater. This study investigates examiner behaviour in the Interactive Task, which appears at GESE Grades 7–12, where a 'prompt from the examiner initiates candidate-led speaking and listening interactions – requiring an authentic exchange of information and opinions' (Trinity College London, 2018, p. 7).

The Interactive Task

In the Interactive Task (IT), candidates lead the interaction. It is a phase of the exam in which the candidate can display their abilities in directing conversation through, for example, questioning and encouraging contributions. The candidate should 'initiate "turns" in the conversation and control the direction of the interaction' (Trinity College London, 2018, p. 7). The examiner's contributions are not scripted; however, in order to be able to respond in an authentic manner, they produce a test plan or 'back story' as part of their preparation (Boyd & Taylor, 2016). At the beginning of the IT, a short prompt, which presents a dilemma of some kind, is delivered verbatim from a script by the examiner and subsequently the examiner's test plan provides background from which the examiner draws in order to naturally respond to the candidate's lines of questioning. Therefore, the GESE IT is an example of 'recipient design', where the candidate must adjust and adapt a speech action to their interlocutor and the situation (Drew, 2013).

There are two immediately obvious and pressing considerations that the design of the IT presents. The first is to do with roles and the second is to do with unpredictability. van Lier (1989) questioned whether an oral proficiency 'interview' can fully represent conversation for assessment when the format and direction of an interview are entirely under the control of the interviewer. The IT phase of the GESE changes the role of the examiner from interviewer to conversational partner by shifting the responsibility for direction principally to the candidate. Whether the hierarchical roles of examiner and candidate can be properly set aside in favour of equal roles as conversation partners is debateable. However, van Lier (1989) asserts that these two relationships are not mutually exclusive. Therefore, even in a test situation it is possible for natural conversation between examiner and candidate to take place on an equal footing.

The second question of unpredictability is an important one because it speaks directly to the issue of authenticity. Irregularity of interaction is a part of authentic dialogue (Morrow, 1979) and the IT distributes the responsibility for that irregularity and unpredictability more fairly between examiner and candidate than more structured interview-type assessments. Wall and Taylor (2014) advance the argument that the GESE is indeed an example of authentic interaction precisely because of the irregularity of communication and the distribution of responsibility.

What do speakers *do* in interaction?

When analysing spoken interaction, it is tempting to look for a taxonomy or ethnography of types of interaction. However, Schegloff (2007), in describing the analytic stance related to speech act theory of John Austin and John Searle, states that one should not begin with categories of action but should start with observation of a piece of talk and ask what the speaker is doing. In this way, 'one can find new things, new actions, that we did not previously know people did' (Schegloff, 2007, p. 8).

An important concept that guided this research is that of preference organisation, which occurs as 'a structural bias toward affiliation and reciprocity of perspectives' (Seedhouse, 2004, p. 9). Preference and dispreference are not about liking or not liking something, they are about aligning or not aligning oneself with one's interlocutor. They might be termed affiliation and disaffiliation (Seedhouse, 2004, p. 23) or confirmations and discomfirmations (Pomerantz & Heritage, 2013, p. 213). The preference principle operates such that there is a strong tendency towards preference or alignment with one's conversational partner. This principle manifests in preference being signalled with readiness and without delay. Dispreference on the other hand is accompanied by delay or mitigation, or may include within it weak agreement.

This study sought to answer two research questions: What interaction strategies (hereafter, *strategies*) do examiners employ in the GESE IT in order to allow candidates to complete the task appropriately and to display rateable performance? And, do different examiner strategies elicit quantitatively different performance from candidates in nine performance measures in the categories of fluency, lexical sophistication, grammatical accuracy and syntactical sophistication?

Methodology

In the first instance, transcriptions of 54 different examples of the GESE interactive phase (Trinity College London, 2018) were obtained from the intermediate-level subset of the Trinity Lancaster Spoken Learner Corpus (Gablasova, Brezina, & McEnery, 2009; McEnery et al., n.d.). Each transcript represented up to four minutes of spoken dialogic interaction between examiner and candidate. All examiners in the sample were native speakers of English. The 54 candidates came from 10 countries, and there were 37 different ITs.

Coding examiner interventions

Following the approach of grounded theory, all transcripts were read, and the examiners' moves, or interaction strategies, coded using NVivo12 software. Since contributions to conversation may embody more than one action, a decision was made to select the one most apparent, or which seemed to be most apparent, to the candidate. The final level of axial coding involved grouping codes into similar categories to form the 'theory' in grounded theory.

Once the coding of examiner strategies had been finalised, candidate responses to those strategies needed to be formatted for quantitative analysis. Therefore, the next step was to extract all the responses that followed the strategies so that three text outputs of candidate response phase were generated. The first responses from the transcription had the examiner's backchannels removed so that it was a complete utterance of the candidate only. These responses held information about false starts, truncated words, empty pauses and filled pauses. A second response was created with non-lexical filled pauses removed. A third response was created following Iwashita, Brown, McNamara and O'Hagan, where 'features of repair', such as false starts and reformulations, were 'pruned' so that responses could 'show evidence of correct use' of language for analysis (2008, p. 32).

Quantitative analysis of candidates' responses

Once the three versions of each of the responses had been created, quantitative text analysis was conducted using Natural Language Processing (NLP) tools (Kyle & Crossley, 2015). Nine measures were selected to enable broad coverage of three areas of performance: fluency, lexis and grammar. The three measures of fluency were length of response in words, reformulation ratio and pause ratio; the three measures of lexis were type-token ratio (TTR), lexical range and lexical frequency; and the three measures of grammar were grammatical accuracy as measured by an error ratio and grammatical complexity as measured by a complex t-units to total t-units ratio and the number of complex nominals per t-unit.

Assessing statistical significance and effect

In order to determine whether any of the strategies identified through grounded theory had an impact on candidates' performance as quantitatively measured for fluency, lexical sophistication, grammatical accuracy or grammatical complexity, statistical analysis in SPSS V24 was used. An omnibus Kruskal-Wallis H test was used to determine whether one or other of the strategies has a significant effect on one or other of the measures of the responses, and post-hoc Mann-Whitney U Tests with Bonferroni adjustment to significance values were employed to assess the direction and size of any effects.

Results and discussion

In investigating the types of strategies examiners employ in the GESE IT through the inductive reasoning of grounded theory methodology, 65 codes rolled up into eight axial codes corresponding to eight types of examiner behaviour that were identified, of which there were 886 examples. The eight types of behaviour were termed *accepting*, *holding*, *inauthenticity*, *parsimony*, *reformulating*, *rejecting, resolving* and *steering*. Here, four of these strategies will be compared, the responses to which exhibited statistically significant differences in measures of fluency.

Examiner behaviour

The axial code *holding* described a small group of codes that describe conversational behaviour that appeared to make no demands on the direction of conversation. For example, three open codes that were combined into holding involved supplying some form of information intended to supply grist to the mill of the candidate's questioning. These codes tended to occur early in the interactive tasks as the background to the situation posed by the initial prompt was established and explored.

A strategy sometimes used by examiners to encourage contribution from candidates was to leave a gap, a strategy that was termed *parsimony*. The gap may have been silence, a filled pause or information given in a sparing or parsimonious manner. This feature appeared in the transcriptions as examiners being noticeably or unusually unforthcoming.

Steering often became apparent where the examiner appeared to be directing the conversation or manipulating the solution from a position of pseudo-passivity. Subjectively, this strategy often did not appear to be successful as candidates often did not 'take the bait'. Other more overt examples of steering included examiners offering their own solutions for candidates to take up, asking for opinions about options, or challenging a candidate's contribution or suggestion.

The axial code *reformulating* summarised a group of open codes that include 'clarifying', 'confirming', 'repeating' and 'restating the original problem'. The last of these codes was often an almost word-for-word restating of the task prompt. The most frequent of these codes in the data is 'confirming', which often occurred in response to a request for clarification.

Effect on candidates' fluency

Considering the two measures of fluency, reformulation ratio and pause ratio, the strategy of reformulating elicited more fluent responses than the strategies of holding, parsimony and steering. There was a medium effect size of the difference between parsimony and reformulating and the other differences had a small effect size. Reformulating may not make much conversational demand on one's conversational partner. Holding, meanwhile, may increase the cognitive load for the candidate as it may introduce an element of 'conceptualisation' into the speaking process (Field, 2011, p. 74) as the candidate considers what to do with some information, or what direction to take the conversation. It was hypothesised that parsimony produced greater disfluency in the candidates' responses because there is a possibility that this strategy may introduce feelings of uncertainty of the examiner's intentions. There is evidence that parsimony elicits less fluent responses than reformulating. Reformulating is a strategy that includes examples of behaviour that seeks to clarify and remove uncertainty. In this respect, it is behaviour that is opposite to that of parsimony. The third statistically different relationship within these two fluency measures is between steering and reformulating. It is possible to explain this difference by considering that steering may introduce disfluency because the examiner attempts, often without being overt, to take the conversation in a different direction to that in which the candidate may wish or expect.

Conclusion

This exploratory study used grounded theory as a structured qualitative research method to explore the interactions that take place in the GESE IT at CEFR B2 level with a particular focus on examiners' linguistic behaviour. The findings of statistically significant small to medium differences between types of examiner behaviour confirm the understanding in the literature that, given the co-constructed nature of interaction, examiners do inevitably have some effect on candidate performance in spoken interactive exams. However, the findings confirm there is a 'bias for best' (Swain, 1984), a stated Trinity goal (Boyd & Taylor, 2016), and add insight into the type of interactive elicitation behaviour Trinity examiners employ.

References

Boyd, E., & Taylor, C. (2016). Presenting validity evidence: The case of the GESE. In J. V. Banerjee & D. Tsagari (Eds.). *Contemporary Second Language Assessment : Contemporary Applied Linguistics* (pp. 37–59). London: Bloomsbury Academic.

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

Drew, P. (2013). Turn design. In J. Sidnell & T. Stivers (Eds.), *The Handbook of Conversation Analysis* (pp. 131–149). Hoboken: Wiley-Blackwell.

Field, J. (2011). Cognitive validity. In L. Taylor (Ed.), *Examining Speaking: Research and practice in Assessing Second Language Speaking* (pp. 65–11). Studies in Language Testing volume 30. Cambridge: UCLES/Cambridge University Press.

Gablasova, D., Brezina, V., & McEnery, A. (2009). The Trinity Lancaster Corpus: Development, description and application. *International Journal of Learner Corpus Research*, *5*(2), 126–160.

Iwashita, N., Brown, A., McNamara, T., & O'Hagan, S. (2008). Assessed levels of second language speaking proficiency: How distinct?. *Applied Linguistics*, *29*(1), 24–49.

Kyle, K., & Crossley, S. A. (2015). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly*, *49*(4), 757–786.

McEnery, T., Baker, P., Brezina, V., Hardie, A., Boyd, E., Taylor, C., & Ikeda-Wood, A. (n.d.). *Spoken Learner Corpus (SLC) Project*. Retrieved from: cass.lancs.ac.uk/cass-projects/the-spoken-learner-corpus-slc-project/

Morrow, K. (1979). Communicative language testing: Revolution or evolution?. In C.J. Brumfit and K. Johnson (Eds.), *The Communicative Approach to Language Teaching* (pp. 143–157). Oxford: Oxford University Press.

Pomerantz, A., & Heritage, J. (2013). Preference. In J. Sidnell & T. Stivers (Eds.), *The Handbook of Conversation Analysis* (pp. 210–229). Malden: Wiley-Blackwell.

Schegloff, E. A. (2007). *Sequence Organization in Interaction*. Cambridge: Cambridge University Press.

Seedhouse, P. (2004). Conversation analysis methodology. *Language Learning*, *54*(1), 1–54.

Swain, M. (1984). Large-scale communicative language testing: a case study. In S. Savignon & M. Berns (Eds.), *Initiatives in Communicative Language Teaching* (pp. 185–201). Reading: Addison-Wesley.

Trinity College London. (2018). *Exam Information: Graded Examinations in Spoken English (GESE)*. Trinity College London. Retrieved from: www.trinitycollege.com

van Lier, L. (1989). Reeling, writhing, drawling, stretching, and fainting in coils: Oral Proficiency Interviews as conversation. *TESOL Quarterly*, *23*(3), 489–508.

Wall, D., & Taylor, C. (2014). Communicative language testing (CLT): Reflections on the 'issues revisited' from the perspective of an examinations board. *Language Assessment Quarterly*, *11*(2), 170–185.

Testing writing: Washback on B2 First preparation courses in southern Spain

Victoria Peña Jaenes
Cambridge Assessment English, United Kingdom

Abstract

This paper starts with a brief overview of the close relationship between teaching and assessment and the role that assessment and more precisely, proficiency exams, have in society today, which serves as introduction to the discussion about washback. It then tries to determine the influence that Cambridge Assessment English B2 First, both the standard and for Schools versions, may have on learning and teaching by looking at the task types practised in more exam-oriented courses and in more general English courses offered in a sample of language schools in southern Spain, and by considering the aspects perceived as key for success in the test.

Introduction

'Testing and teaching are so closely interrelated that it is virtually impossible to work in either field without being constantly concerned with the other' (Heaton, 1990, p. 5, cited by Peña Jaenes, 2017, p. 76). In fact, the different approaches to language teaching have influenced the evolution of and current trends in testing, while the effect of tests on teaching and learning has been discussed by scholars such as Alderson and Wall (1993), Bailey (1996; 1999), Cheng (2005) or Green (2007), all cited by Peña Jaenes (2017, p. 76). Test-related aspects may have an impact at the macro-level, i.e. the educational system and society in general, but also at the micro-level, i.e. language courses, people's attitudes or course materials (Bachman and Palmer, 1996, p. 30). Washback has to do with the impact of tests on teaching and learning.

Washback

Research into washback started following Alderson and Wall's 'Does washback exist? (1993), and has since explored the complexity of this phenomenon (Messick, 1996, p. 242; Bachman and Palmer, 1996, p. 30; Collins, Reiss, & Stobart, 2010; Polesel, Rice, & Dulfer, 2014; Cheng, Sun & Ma, 2015) and the difficulty in identifying the factors affecting it. In terms of its complexity, Watanabe (2004) identifies five dimensions of washback: specificity, intensity, length, intentionality, and value. For our purposes, this paper will mainly focus on the last dimension i.e. the value of the washback. Although washback is frequently considered a neutral term (Alderson & Wall, 1993), hence referring to both positive and negative effects, Bailey and Masuhara (2013, p. 304; cited by Ha, 2019, p. 4) point out that the value of washback is not absolute, as it depends on our views regarding the desirable outcomes of language learning. Therefore, a test may be considered to exert negative washback when its content or format is based on a narrow definition of language ability and, as a consequence, has a detrimental effect on the breadth of, or the variety to be found within a curriculum, preventing students from learning real-life skills. Some teachers might associate standardised testing with negative washback, although experts such as Messick (1998) or Shohamy (2001), both cited by Xie and Andrews (2012, p. 50), point out that test design does not affect the nature of washback – or is not the only factor affecting it – and argue that it is the misuse or overuse of test results that produce negative washback. Positive washback occurs when a test results in good teaching practice (Taylor, 2005, p. 154) or enhances learning (Hakim & Tasikmalaya, 2018, p. 62).

Washback in context

Research into washback first focused on how assessment innovation affected teachers and how tests could shape teaching. However, more recent research has been conducted in areas where high-stakes tests have long been present in the education system. It has also paid greater attention to how tests can impact learning and learners. Despite the large number of studies carried out in countries and regions such as Canada, Central and Eastern Europe, China, Japan, UK and Sri Lanka (Tsagari, 2011, p. 432, cited by Peña Jaenes, 2017), and the number of Spanish candidates sitting accreditation exams every year, very few studies have been conducted in Spain on the washback of well-known language proficiency exams.

Study

The present study was conducted in six language schools located in the South of Spain – in Jaén, Málaga, Granada and Murcia – with a total of 136 students and 17 teachers taking part. Students were divided into a control group – those who attended more general English courses, and an experimental group – those who attended more exam-oriented courses. The instruments used were questionnaires and teaching diaries. There were two versions of the questionnaires, one for students and one for teachers. The two variables considered for this study were:

1. The text types practised in class.

2. The activities considered most effective – by participants – to successfully prepare for the B2 First Writing paper.

Results: Presentation and discussion

Text types

Students were asked about the tasks they did in class and the text types that they worked on. They were also asked to give information about the frequency with which this was the focus of their classes. The table the students were asked to complete focused mainly on the text types that appear in the B2 First Writing paper i.e. reports, letters, essays, reviews, articles, and short stories (University of Cambridge Local Examinations Syndicate, 2019, p. 27). However, students were also encouraged to add any other text type they practised in class. Answers were analysed taking into consideration which institution students were enrolled in. This area of the study was carried out in only three of the six schools involved overall.

Table 1: Student questionnaire, Question 18: Results

Text type	*Language School 1 (experimental group)*			*Language School 2 (experimental group)*			*Language School 3 (control group)*		
Frequency	1/m[1]	+ 1/m	I don´t know/ No answer	1/m	+ 1/m	I don´t know/ No answer	1 /m	+ 1/m	I don´t know/ No answer
Reports	29%	**64%**	7%	44%	**48%**	8%	**48%**	26%	26%
Letters	30%	**63%**	7%	43%	**51%**	6%	31%	**46%**	23%
Essays	36%	**64%**	0%	40%	**56%**	4%	33%	28%	**39%**
Reviews	29%	**49%**	22%	45%	**47%**	8%	33%	18%	**49%**
Articles	**58%**	35%	7%	**49%**	46%	5%	33%	23%	**44%**
Short stories	43%	**50%**	7%	**47%**	43%	10%	**44%**	26%	30%
Other	29%	**50%**	21%	33%	32%	**35%**	**41%**	31%	28%

[1] m – month

The questionnaires for teachers also included a very similar table on text types and other writing-related activities. In this case, the differences were not very significant as most professionals – 43 % in the case of reports and letters, 38% in the case of essays, 40% in the case of reviews, and 46% in the case of articles and short stories – reported working on each text type on a monthly basis. However, differences appeared when it came to the amount of time devoted to the different text types every month, as 40% of the teachers from Language School 2 claimed that they worked on text types for 1 to 2 hours a month, while none of the other teachers reported devoting as much time every month to any text type.

When teachers were asked about whether they worked on text types that were not included in the B2 First exams, all teachers delivering more general English courses – Language School 3 – answered yes. They justified this by explaining that students should have preparation in all genres and as much writing practice as possible. As for teachers in charge of more exam-oriented courses, only 20% of those working at Language School 2 answered that they looked at other text types with their students, while none of the professionals working at Language School 1 did so. They justified their answers on the basis of workload and time. As for the additional text types and tasks practised, teachers mentioned writing formal letters and descriptions, and also working on content by asking students to read texts and guess the title, to rephrase sentences or to fill in the gaps to work on linking devices.

The analysis of the data suggests that the B2 First exam influences the choice of activities and texts used in class. The results show that the three language schools mainly work on text types tested in the B2 First exam. Although students' results show that there is also time for other text types in the three schools, the teachers' answers evidence that the washback of B2 First is more pronounced on more exam-oriented courses. This is because while all the teachers delivering general English courses said they practised other text types in class, none of the teachers in Language School 1 and only 20% of those in Language School 2 mentioned doing so. Therefore, the washback of the exam is strong when it comes to text and task selection. This finding is in line with Green (2007, p. 75, cited by Peña Jaenes, 2017, p. 95), who found that 'teachers, for their part, also reported that IELTS influenced their choice of activities', with Cheng (2010, p. 49, cited by Peña Jaenes, 2017, p. 95), who obtained similar evidence in her interviews and classroom observation, and by Tsagari (2011, p. 437, cited by Peña Jaenes, 2017, p. 95), who reported that 'the exam encouraged teachers and students to place more value on the skills and activities that were assessed on the exam'.

When teachers were asked about what other text types they used in their lessons, they mentioned formal letters and descriptions. However, letters do appear in Part 2 of the B2 First Writing paper, and it could be argued that description is a key feature of both reviews and short stories, which are also options. Therefore, all of the professionals in this study, regardless of whether they were delivering general English courses or more exam-oriented courses, focused on virtually the same text types. In this sense, it could be argued that the washback of B2 First is positive, as it does not appear to limit the variety of text types that students engage with in class. This conclusion is in line with Shohamy (1992, p. 514, cited by Peña Jaenes, 2017, p. 77).

Most effective task types

To analyse washback from a different perspective, it was considered relevant to collect feedback from both students and teachers on activities that, in their opinion, helped in the successful preparation for the B2 First Writing paper.

Results show that most students identify working on grammar and vocabulary on the one hand, and active writing practice on the other, as key activities for success. They also consider familiarisation with different text types and activities that focus on text structure as being useful. Finally, students place considerable value on the feedback they receive from their teachers.

The teachers involved in this study ranked active writing tasks as the most helpful for their students, although they consider that these are more effective when carried out under exam conditions. In general, teachers seem to consider activities related to assessment, such as giving a mark, explaining test-taking techniques, and explaining marking criteria as very effective for their students' success. It is interesting to note that a similar percentage of students and teachers consider working on text structure as beneficial.

When it comes to identifying which activities are the most effective when preparing students for the B2 First Writing paper, the washback differs to some degree between students and teachers. This agrees with the findings reported by Green (2007, p. 86, cited by Peña Jaenes, 2017, p. 96). Students give more weight to active writing tasks, and while teachers recognise their importance, they give priority to writing practice carried out under exam conditions. This emphasis on active writing is beneficial and can be seen as an example of positive washback as it motivates learners to practise and thus develop their writing skills. The fact that grammar and vocabulary were also identified as being very important is not surprising if we take into account that when students were asked about their expectations regarding the lessons they attended, they mentioned that grammar and vocabulary both played an important role, almost as important as practising all four skills (Peña Jaenes, 2017, p. 96). Similar findings were reported by Tsagari (2011, p. 438, cited by Peña Jaenes, 2017, p. 96). This interest in grammar and vocabulary could be explained by the fact that language is one of the assessment criteria for writing used by Cambridge Assessment English (University of Cambridge Local Examinations Syndicate, 2019) but this probably has more to do with the influence of local teaching practices and beliefs, which also appeared in Tsagari (2009, p. 8, cited by Peña Jaenes, 2017, p. 96). Finally, the value that teachers give to assessment-related activities reflects students' expectations regarding English courses, since 58% of them were interested in doing mock tests regularly.

Conclusion

This paper aimed to gain a better understanding of the washback that a well-known exam may have on language courses. For that purpose, two variables were considered i.e. choice of text types and key strategies for success in the B2 First Writing paper.

The washback of the test is observed in the text types that are practised in class. This washback can be said to be positive because preparing for B2 First does not limit the variety of genres students engage with, and in turn fosters the development of a range of writing skills.

Washback on teachers is slightly different because while both teachers and learners identify writing practice as the key activity for success in the B2 First Writing paper, they disagree when choosing the other key activities.

References

Alderson, J.C., & Wall, D. (1993). Does washback exist? *Applied Linguistics, 14*, 115–129.

Bachman, L., & Palmer, A. (1996). *Language Testing in Practice*. Oxford: Oxford University Press.

Bailey, K. M. (1996). Working for washback: a review of the washback concept in language testing. *Language Testing 13*(3), 257–279.

Bailey, K. M. (1999). *Washback in Language Testing*. Princeton: Educational Testing Service.

Bailey, K. M., & Masuhara, H. (2013). Language testing washback: The role of materials. In B. Tomlinson (Ed.), *Applied Linguistics and Materials Development* (1st ed.) (pp. 303–318). London/New York: Bloomsbury Academic.

Cheng, L. (2005). *Changing Language Teaching Through Language Testing: A Washback Study*. Studies in Language Testing volume 21. Cambridge: UCLES/Cambridge University Press.

Cheng, L. (2010). How does washback influence teaching? Implications for Hong Kong. *Language and Education*, *11*(1), 38–54.

Cheng, L., Sun, Y., & Ma, J. (2015). Review of washback research literature within Kane's argument-based validation framework. *Language Teaching, 48*(4), 436–470.

Collins, S., Reiss, M., & Stobart, G. (2010). What happens when high-stakes testing stops? Teachers' perceptions of the impact of compulsory national testing in science of 11-year-olds in England and its abolition in Wales. *Assessment in Education: Principles, Policy and Practice*, *17*(3), 273–286.

Green, A. (2007). *IELTS Washback in Context: Preparation for Academic Writing in Higher Education*. Studies in Language Testing volume 21. Cambridge: UCLES/Cambridge University Press.

Ha, N. T. T. (2019). A literature review of washback effects of assessment on language learning. *Journal of Science Ho Chi Minh City Open University, 9*(5), 3–15.

Hakim, L. N., & Tasikmalaya, U. P. (2018). Washback effect in language testing: What do we know and what is its effect?. *Jurnal Forum Didaktik, 2*(1), 59–68.

Heaton. J. B. (1990). *Classroom Testing*. Harlow: Longman.

Messick, S. (1996). Validity and washback in language testing. *Language Testing*, *13*(3), 241–256.

Messick, S. (1998). Test validity: A matter of consequence. *Social Indicators Research, 45*(1), 35–44.

Peña Jaenes, V. (2017). Testing writing: The washback on Cambridge English: First preparation courses in southern Spain. *The Grove, 24,* 75–106.

Polesel, J., Rice, S., & Dulfer, N. (2014). The impact of high-stakes testing on curriculum and pedagogy: A teacher perspective from Australia. *Journal of Education Policy*, *29*(5), 640–657.

Shohamy, E. (1992). Beyond proficiency testing: a diagnostic feedback testing model for assessing foreign language learning*. The Modern Language Journal*, *76*(4), 513– 521.

Shohamy, E. (2001). *The Power of Tests: A Critical Perspective on the Uses of Language Tests*. New York: Pearson Education.

Taylor, L. (2005). Washback and impact. *ELT Journal, 59*(2), 154–155.

Tsagari, D. (2009). Revisiting the concept of test washback: investigating FCE in Greek language schools. *Research Notes*, *35*, 5–9.

Tsagari, D. (2011). *Washback of a high-stakes English exam on teachers' perceptions and practices*. Retrieved from: core.ac.uk/download/pdf/267932175.pdf

University of Cambridge Local Examinations Syndicate. (2019). *B2 First for Schools Handbook for Teachers*. Cambridge: Cambridge Assessment English.

Watanabe, Y. (2004). Methodology in washback studies. In L.Y. Cheng, Y. Watanabe & A. Curtis (Eds.), *Washback in Language Testing: Research Contents and Methods* (pp. 19–36). Mahwah: Lawrence Erlbaum.

Xie, Q., & Andrews, S. (2012). Do test design and uses influence test preparation? Testing a model of washback with Structural Equation Modeling. *Language Testing, 30*(1), 49–70.

The future of self-assessment: understanding students' and teachers' attitudes

Nawal Kadri
University of Bejaia, Algeria

Abstract

With the new paradigm shift witnessed in EFL teaching and learning, there has been an increasing concern for self-assessment. Ample evidence shows that self-assessment is an effective learning and pedagogical tool that enhances students' learning and improves classroom instruction. Nevertheless, the development of this process depends on many factors among which students' and teachers' attitudes play a prominent role. The present study attempted to address this issue by exploring EFL Algerian students and teachers' attitudes about self-assessment. Specifically, 109 students and 20 teachers at the University of Bejaia, Algeria completed an online survey. Findings revealed that although students and teachers developed positive attitudes towards self-assessment and recognized its potential benefits on students' learning, there seems to be a discrepancy between students' and teachers' beliefs and classroom practice. As a result, the study came up with significant pedagogical implications on how to integrate self-assessment into EFL instruction and enhance its effectiveness.

Introduction

A generally held view among researchers investigating formative assessment is that learning is effective when students become self-reliant and assess their own learning progress; this is because assessment is part of classroom instruction. According to Paris and Ayres (1994, p. 7), 'students need to be active participants in assessment of their own learning rather than passive respondents to a series of tests.' Several studies have proven that self-assessment is an effective learning and pedagogical tool that enhances students' learning and improves classroom instruction. Nevertheless, the development of this process depends on many factors among which attitudes and beliefs play a prominent role. In this concern, little research has been conducted to investigate students' and teachers' attitudes about self-assessment, notably in Algeria. The current study attempts to address this gap in research by exploring Algerian students' and teachers' attitudes about self-assessment and evaluating the degree to which these attitudes are in alignment with classroom assessment practice. To this end, five important questions are addressed:

1. What are Algerian EFL teachers' attitude towards self-assessment as an alternative method of assessment?
2. What are Algerian EFL students' attitudes towards self-assessment?
3. Is there a connection between the teachers' attitudes about self-assessment and their classroom assessment practices?
4. Is there a discrepancy between the teachers' and the students' attitudes towards self-assessment?
5. What are the self-assessment literacy needs of the students and the teachers?

Self-assessment potential

Self-assessment is considered a type of formative assessment. There are three major elements that condition the effective implementation of self-assessment in EFL classrooms: reflection, criteria and revision. These are captured in Andrade and Valtcheva's (2009, p. 13) definition of self-assessment as 'a process of formative assessment during which students reflect on the quality of their work, judge the degree to which it reflects explicitly stated goals or criteria, and revise their work accordingly.' First, unlike traditional assessment, self-assessment is an ongoing process that engages students in constant review of their work and reflection on its quality. Harris (1997, p. 17) supports this view by stating: 'self-assessment can help to compensate by providing a continuous, personalized, and formative element of assessment, in settings where the only practical assessment measures may be periodic tests.' The second important condition in the self-assessment process is access to criteria. Students need to be aware of the criteria of assessment in order to compare their work against target performance. Self-assessment is a goal-oriented process in the classroom. Last but not least, self-assessment has no value if students do not revise their work.

Students should be given opportunities to use their own self-assessment results in accordance with the teacher feedback in order to take action and improve their performance. Revision is important for students to make their own decisions.

Ample evidence has shown that self-assessment has potential benefits on students' learning if implemented effectively. Researchers stress the role of self-assessment in developing reflection and critical thinking. In this regard, Kurnaz and Çimer (2010, p. 3,667) state that through the process of self-assessment, students gain insights into their own understandings. By getting insight into their strengths and weaknesses, students become aware of their learning progress. Gottlieb (2000, p. 97) goes further by explaining that students take an active role in their learning process which allows them to develop as independent learners. This conveys the idea that learning is a constructive and manageable process. Moreover, self-assessment is assumed to increase students' internal control and motivation and make learning meaningful (Elliott & Higgins, 2005; Panadero & Alonso-Tapia, 2013; Rolheiser & Ross, 2001). Another benefit is on instruction. In this regard, Russell and Airasian (2012, p. 8) explain that self-assessment provides useful data for teachers to interpret students' results by getting insight into their thinking, their learning strategies and emotions and therefore understand why they perform as they do. Subsequently, they can identify students' learning difficulties and address their needs. Other conditions of self-assessment use are: awareness, direct instruction, the use of self-assessment models and practice (Goodrich, 1996, pp.31–32).

Attitudes towards self-assessment

The successful implementation of self-assessment in EFL classrooms depends on various factors among which both teachers' and students' attitudes play a determinant role. There is relatively a limited body of research that addressed the issue. Drawing on the available limited sources, teachers and students appear to view self-assessment in two different ways. Some teachers view self-assessment positively and believe it is useful in EFL classrooms (Kadri & Amziane, 2017; Tshabalala & Ndimande, 2016), while others develop negative attitudes towards the process (Gardner, 2000; Sadler, 1989). Likewise, empirical research on students' attitudes shows both positive (Andrade & Du, 2005, 2007; Chelli, 2013; Hanrahan & Isaacs, 2001; Mok, Lung, Cheng, Cheung, & Ng, 2006) and negative (Butler & Lee, 2010; McMullan, 2006) attitudes.

The study

The present paper aims to give a comprehensive account of Algerian EFL students' attitudes towards self-assessment in contrast with those of teachers. To this end, 109 EFL students and 20 EFL teachers from Bejaia University participated in the study. The sample of students consisted of 74 third year undergraduate students and 35 second year master students ranging from 20 to 28 years old; they were 96 females and 13 males. Teacher participants were 17 females and 3 males holding an MA or PhD degree with different teaching experiences. To collect data about the participants' attitudes and classroom practices, we used an online survey.

The survey consisted of questions related to three main aspects: classroom practice, awareness and attitudes about self-assessment, and literacy needs. Different types of questions were used: multiple choice, yes/no and open-ended questions, and a Likert scale that ranges from 1) strongly disagree to 5) strongly agree.

Findings

Classroom practice

The analysis of the students' and teachers' answers made it clear that the dominant mode of assessment in EFL classrooms in the department of English in Bejaia University is teacher assessment. According to 100% of teachers' answers and 88% of students', teachers assess their students' performance themselves. These results are similar to those found by Benettayeb-Ouahiani in her small scale study conducted at the department of English at the University of Chlef (Algeria) in 2016 and the study of Kadri and Amziane conducted in 2017 at the University of Bejaia. What seems interesting in the findings is the use of both summative and formative assessments. In addition to scores obtained through testing, teachers also provide feedback in the form of written comments or correction of students' errors.

Regarding self-assessment, there appears to be a disagreement between students' and teachers' answers. There are students (47.7%) who revealed that no teacher engages them in self-assessment and others (32.1%) reported that few of them do. However, responses in the teacher survey indicate that 70% of teachers use self-assessment in their EFL classrooms. What is sure is that both students (39%) and teachers (65%) agree on the fact that teachers encourage their students to self-assess their work. By taking a look at teachers' comments on how they integrate self-assessment it seems that there is a lack of understanding of what self-assessment is or how it is implemented. Just asking students to reflect on their learning without self-assessment techniques or asking students to correct their peers is not self-assessment. As explained in the review of literature, there are conditions for effective implementation of self-assessment. As far as self-assessment instruction and modeling are concerned, statistics show similar results: 46% of students reported that they receive no instruction on self-assessment and 40% of teachers confirmed this. This means that one important condition in the process of self-assessment is absent.

To understand the principles underlying these practices, there is a need to get into the participants' attitudes.

Awareness and attitudes

Students' and teachers' answers provided insight into their beliefs about self-assessment. The vast majority of the participants reported positive attitudes towards students' self-assessment and recognised that this process is important in EFL classrooms as it improves learning (68% of students and 95% of teachers) and promotes critical thinking (83% of students and 100% of teachers). These benefits have been reported in other studies (Bullock, 2011; Elliott & Higgins, 2005; Harris, 1997; Panadero & Alonso-Tapia, 2013).

Moreover, half of the students (50%) seem to place high value on their ability to assess their own progress accurately. Nevertheless, the same number of teachers (50%) believes that students are unable to accurately assess the quality of their work. In reference to teachers' comments, some potential reasons which inevitably hinder students from identifying their strengths and weaknesses include students' low proficiency, lack of motivation and autonomy.

Regarding knowledge and skills in self-assessment, more than half of the teachers (60%) reported, in the Likert scale, they possess the necessary skills for self-assessment and are aware of the conditions required for its implementation in EFL classrooms and 45% stated they can design self-assessment techniques. However, many teachers made it clear in their comments that although they find self-assessment useful, their main problem is lack of awareness on how to implement the process in EFL classrooms. This is confirmed by the findings obtained in the next section.

Needs and expectations

One interesting finding is that both students (83%) and teachers (95%) believe that self-assessment should be integrated into EFL classrooms. Another finding is that the same number of participants previously provided voiced their ultimate need for instruction and training in the skills required for self-assessment to be implemented successfully.

Conclusion

Based on our findings, there appears to be a discrepancy between students' and teachers' responses and a lack of alignment between teachers' attitudes and classroom practices at the University of Bejaia. In spite of their positive attitudes about students' self-assessment, results presented in the three sections confirm that self-assessment is not integrated appropriately into EFL classroom instruction; teacher assessment is the dominant mode of assessment. Nevertheless, teachers and students seem willing to use self-assessment as they recognised the potential benefits of the process and expressed their need for instruction and training.

Teachers' classroom practices might be related to teachers' and students' belief that assessment is the teacher's responsibility, a lack of knowledge due to a lack of training, time constraints, large classes, students' low language proficiency and lack of autonomy. This may explain why we have got controversial results. As a matter of fact, desirability and positive attitudes about self-assessment are not enough to ensure its implementation and practice. For this reason, a number of suggestions for EFL classrooms might include the following:

- Raising both teachers' and students' awareness of the importance of self-assessment.
- Training teachers and providing support in developing the skills necessary for effective implementation of self-assessment in EFL classrooms.
- Providing classroom instruction on accurate self-assessment.
- Adopting models of self-assessment suggested in the literature.
- Allowing constant practice with provision of constructive feedback.

References

Andrade, H., & Du, Y. (2005). Student perspectives on rubric-referenced assessment. *Practical Assessment, Research & Evaluation*, *10*(3), 1–11.

Andrade, H., & Du, Y. (2007). Student responses to criteria-referenced self-assessment. *Assessment and Evaluation in Higher Education*, *32*(2), 159–181.

Andrade, H., & Valtcheva, A. (2009). Promoting learning and achievement through self-assessment. *Theory Into Practice*, *48*(1), 12–19.

Benettayeb-Ouahiani, A. (2016). Assessment in the EFL university classroom: Between tradition and innovation. *Revue des études humaines et sociales-B/Littérature et Philosophie*, *15*, 3–10.

Bullock, D. (2011). Learner self-assessment: An investigation into teachers' beliefs. *ELT Journal*, *65*(2), 114–125.

Butler, Y. G., & Lee, J. (2010). The effects of self-assessment among young learners of English. *Language Testing*, *27*(1), 5–31.

Chelli, S. (2013). Developing students' writing abilities by the use of self-assessment through portfolios. *Arab World English Journal*, *4*(2), 220–234.

Elliott, N., & Higgins, A. (2005). Self and peer assessment – Does it make a difference to student group work?. *Nurse Education in Practice*, *5*, 40–48.

Gardner, D. (2000). Self-assessment for autonomous language learners. *Links & Letters*, *7*, 49–60.

Goodrich, H. (1996). *Student Self-assessment: At the Intersection of Metacognition and Authentic Assessment* [Doctorial dissertation]. Harvard University.

Gottlieb, M. (2000). Portfolio practices in elementary and secondary schools: Toward learner-directed assessment. In G. Ekbatani & H. Pierson (Eds.), *Learner-directed Assessment in ESL* (pp. 89–104). New York: Routledge.

Hanrahan, S. J., & Isaacs, G. (2001). Assessing self- and peer-assessment: The students' views. *Higher Education Research & Development*, *20*(1), 53–70.

Harris, M. (1997). Self-assessment of language learning in formal settings. *ELT Journal*, *51*(1), 12–20.

Kadri, N., & Amziane, H. (2017).Teachers' attitudes about students' self-assessment: what research says and what classrooms reveal. الممارسات اللغوية, *42*(8), 167–189.

Kurnaz, M. A., & Çimer, S. O. (2010). How do students know that they have learned? An investigation of students' strategies. *Procedia-Social and Behavioral Sciences*, *2*(2), 3,666–3,672.

McMullan, M. (2006). Students' perceptions on the use of portfolios in pre-registration nursing education: A questionnaire survey. *International Journal of Nursing Studies*, *43*(3), 333–343.

Mok, M. M. C., Lung, C. L., Cheng, D. P. W., Cheung, R. H. P., & Ng, M. L. (2006). Self-assessment in higher education: Experience in using a metacognitive approach in five case studies. *Assessment & Evaluation in Higher Education*, *31*(4), 415–433.

Panadero, E., & Alonso-Tapia, J. (2013). Self-assessment: Theoretical and practical connotations. When it happens, how it is acquired, and what to do to develop it in our students. *Educational Journal of Research in Educational Psychology, 11*(2), 551–576.

Paris, S., G., & Ayres, L., R. (1994). *Becoming Reflective Students and Teachers: With Portfolios and Authentic Assessment*. Washington, DC: American Psychological Association.

Rolheiser, C., & Ross, J. A. (2001). Student self-evaluation: What research says and what practice shows. In R. D. Small & A. Thomas (Eds.), *Plain Talk about Kids* (pp. 43–57). Covington: Center for Development and Learning.

Russell, M., K., & Airasian, P., W. (2012). *Classroom Assessment: Concepts and Applications* (7th ed.). New York: McGraw Hill.

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science*, *18*(2), 119–144.

Tshabalala, N. G., & Ndimande, A. (2016). The perceptions of students and staff towards portfolio assessments: The case of Mangosuthu Technikon in Kwa Zulu-Natal Province South Africa. *Mediterranean Journal of Social Sciences*, *7*(3), 319–330.

Assessment literacy in a transnational teacher education context

Samar Yakoob Almossa
English Language Centre, Umm Al-Qura University, Makkah, Saudi Arabia

Abstract

This paper is concerned with seven teachers' conceptualisation and practice of assessment in relation to their background, education, learning and teaching experience in Saudi Arabia and in their homelands. The research found that a common theme among the participating transnational teachers was a lack of pre-service and in-service training in assessment. Their prior learning and teaching experiences and beliefs contributed to their conceptualisation and practices with regard to assessment. This paper argues for the urgent need for a re-examination of the theoretical and practical issues involved in developing teachers' assessment literacy (AL). Given the need to consider AL, there are implications for an assessment training and development repertoire which can be adopted in transnational AL contexts.

Assessment literacy in a transnational teacher education context

In Saudi Arabia today, English language centres (ELCs) are sites of diversity that employ educators with various affiliations who, as a result of transnational movement, have brought with them distinct identities, experiences and cultural and linguistic backgrounds. As part of a prevailing wave of change in the Saudi education system to meet Saudi Vision 2030, teachers' assessment literacy (AL) in the age of transnationalism is an important point of discussion. There are several interrelated issues surrounding classroom-based assessment (CBA) and how it is defined and understood, who defines it and how it is perceived and practiced. This ethnographic study was conducted in an ELC at a Saudi university, with a curriculum and professional training adopted from a UK-based institution. An ethnographic approach was employed to observe and interview seven teachers who worked in the centre. A transnationalism lens and sociocultural perspective were utilised as a theoretical framework to capture recent changes in the higher education scene in Saudi Arabia. *Transnationalism* has been defined as the 'movement of people, media, language, and goods between distinct nation states, particularly that which flows in both directions and is sustained over time' (Jiménez, Smith, & Teague, 2009, p. 17).

Assessment literacy

The position of assessment as a high-stakes accountability tool, has increased academic interest in teachers' AL (Popham, 2013). Given the role assessment plays in teaching and learning, teachers are expected to have an adequate understanding of several aspects of assessment to effectively develop their teaching, support their students, respond to students' needs and meet the expectations of stakeholder groups e.g., policymakers, administrators, students and parents (Herrera Mosquera & Macías, 2015). There has been considerable discussion regarding teachers' AL (Fulcher, 2012; Inbar-Lourie, 2012; Malone, 2017), with great emphasis placed on the importance of teachers' knowledge, skills, principles and expertise in using assessment results, and the impact of that knowledge on students in general (Malone, 2013; Vogt & Tsagari, 2014) and on students' AL (Smith, Worsfold, Davies, Fisher, & McPhail, 2013).

Various concepts associated with teachers' AL and several models and frameworks have been proposed, each with its own potentials and limitations (see Fulcher, 2012; Pill & Harding, 2013; Taylor, 2013). For instance, Taylor (2013) introduced language AL profiles for test writers, classroom teachers, university administrators and professional language testers. In her model, she addressed eight dimensions and their levels to differentiate between stakeholders' needs. In contrast, Fulcher (2012, p. 125) viewed AL as:

> The knowledge, skills and abilities required to design, develop, maintain or evaluate large-scale standardized and/or classroom-based tests, familiarity with test processes, and awareness of principles and concepts that guide and underpin practice, including ethics and codes of practice. The ability to place knowledge, skills, processes, principles and concepts

> within wider historical, social, political and philosophical frameworks in order to understand why practices have arisen as they have, and to evaluate the role and impact of testing on society, institutions, and individuals.

Some aspects from Fulcher's conceptualization were adopted in the current study.

Methodology

This research was based on an ethnographic approach and used a sociocultural perspective to explore individual teachers' experiences with CBA. The participants were seven teachers who taught in ELCs at English language institutes (ELIs) in Saudi universities. I interviewed the participating teachers twice and observed some of their classes, and we had informal chats and conversations both face to face and over the phone. The teachers came from different countries and had lived in several countries before moving to Saudi Arabia. They each had between 10 and 20 years of teaching experience in Saudi Arabia. The teachers' accounts of their knowledge and practices of assessment and training needs provided rich data that support transnational teachers' AL. The research questions that guided this study were:

- What are transnational teachers' understanding and practices of assessment in a Saudi TESOL context?
- What AL opportunities are available to teachers?

Findings

What are transnational teachers' understanding and practices of assessment in a Saudi TESOL context?

This study addresses transnational teachers' understanding and practices of CBA in a higher education context. A unified approach to teaching and assessment was a noticeable phenomenon in the ELCs/ELIs, which resulted in restrictions with regard to what teachers could do in the classroom and limited their roles in assessment. The research findings suggest that a common theme in terms of the participating transnational teachers was a lack of pre-service and in-service teacher education in assessment. Additionally, their prior learning and teaching experiences and beliefs contributed to their conceptualisation and practices with regard to assessment.

Participants were first asked about their learning and teaching experiences to build a profile. One of the key findings is that while the participants had different learning and teaching experiences in different countries, their views, education about assessment and daily CBA practices were similar. The participant teachers were expected to follow fixed assessment tasks with minor variations, but their roles in summative testing were limited. Summative testing accounted for up to 80% of the course grades, and it was organised by an exam committee; the teachers had no role in it.

The participants referred to all their daily practices as CBA practices. These activities included inviting students to participate, quizzing students, asking them questions and leading discussions. Their assessment activities, even those they referred to as formative assessments, were graded, and the weight given to graded tasks was linked purely to the high-stakes nature of the course in the foundation year programme. Thus, there was increased emphasis on students' assessments, which were high stakes. This resulted in mapping their classroom activities and CBA around summative assessments, given that these were important to students' grade point averages.

What AL opportunities are available to teachers?

In response to the question of whether they participated in professional development (PD) sessions that were focused on AL, the teachers said that they had no previous training in assessment. They noted that opportunities for in-house PD were available but never focused on assessment. They were not offered nor had received pre- or in-service assessment-related PD sessions. One reason for this could be that the summative assessments for the course, which accounted for up to 80% of the students' grades, were standardised and not prepared by the teachers. Thus, the teachers' roles in summative assessment were limited. The admins in the ELCs/ELIs focused more on developing teaching practices and neglected assessment in their PD priorities. When the teachers were asked about their PD needs, only one teacher signalled that assessment was an essential part of her PD needs. This finding suggests that not all the participant teachers valued assessment as an important part of their professional growth. This might be for several reasons; one obvious reason is their role in summative assessment was limited, which impacted their CBA practices and PD needs.

Conclusion

In this study, seven participants shared their accounts and experiences as transnational teachers working in the Saudi context. It became very clear they were excluded from participating in making decisions about their PD needs. The paper amply displays the urgent need for a re-examination of the theoretical and practical issues involved in developing teachers' AL. Given that the existing research has exposed the need to consider assessment illiteracy, there are implications for an assessment training and development repertoire that can be adopted in transnational AL communities. A transnational teacher's AL should not be assumed based on the years of experience they have accumulated in different countries. In-house seminars and workshops as well as sponsorships for training and conferences should focus on ongoing PD support to develop different assessment aspects.

This small case study sheds light on the experiences of transnational teachers with assessment in the Saudi context. Future studies should explore the experiences of transnational teachers by comparing teachers from different institutions in Saudi Arabia, transnational teachers in different countries, and transnational teachers with different backgrounds/countries of origin.

References

Fulcher, G. (2012). Assessment literacy for the language classroom. *Language Assessment Quarterly, 9*(2), 113–132.

Herrera Mosquera, L., & Macías, V. D. F. (2015). A call for language assessment literacy in the education and development of teachers of English as a foreign language. *Colombian Applied Linguistics Journal, 17*(2), 302–312.

Inbar-Lourie, O. (2012). Language assessment literacy. In C. A. Chapelle (Ed.), *The Encyclopedia of Applied Linguistics* (pp. 1–9). Oxford: John Wiley & Sons.

Jiménez, R. T., Smith, P. H., & Teague, B. L. (2009). Transnational and community literacies for teachers. *Journal of Adolescent & Adult Literacy*, *53*(1), 16–26.

Malone, M. (2013). The essentials of assessment literacy: Contrasts between testers and users. *Language Testing, 30*(3), 329–344.

Malone, M. (2017). Training in language assessment. In E. Shohamy, I. G. Or, & S. May (Eds.), *Language Testing and Assessment: Encyclopedia of Language Education* (3rd ed., pp. 225–240). Cham: Springer.

Pill, J., & Harding, L. (2013). Defining the language assessment literacy gap: Evidence from a parliamentary inquiry. *Language Testing*, 30(3), 381–402.

Popham, W. J. (2013). Tough teacher evaluation and formative assessment: Oil and water?. *Voices from the Middle, 21*(2), 10–14.

Smith, C. D., Worsfold, K., Davies, L., Fisher, R., & McPhail, R. (2013). Assessment literacy and student learning: The case for explicitly developing students 'assessment literacy'. *Assessment & Evaluation in Higher Education, 38*(1), 44–60.

Taylor, L. (2013). Communicating the theory, practice and principles of language testing to test stakeholders: Some reflections. *Language Testing, 30*(3), 403–412.

Vogt, K., & Tsagari, D. (2014). Assessment literacy of foreign language teachers: Findings of a European study. *Language Assessment Quarterly, 11*(4), 374–402.

Developing classroom practice: The use of a MOOC to increase teachers' language assessment literacy

Richard Spiby
British Council, United Kingdom

Carolyn Westbrook
British Council, United Kingdom

Abstract

Teachers spend large amounts of time on assessment-related activities, yet they may have received limited training in assessment during their teacher training (Stoynoff & Coombe, 2012). Teachers therefore need opportunities for assessment literacy training through Continuing Professional Development but these can be expensive and difficult to access (Odden, Archibald, Fermanich, & Gallagher, 2002).

This paper presents a free language assessment, Massive Open Online Course (MOOC). MOOCs are flexible, low cost and have the capacity to reach large numbers of teachers worldwide. Pre- and post-course questionnaires were administered to investigate participants' perceptions of course effectiveness and the potential impact on classroom practice. Findings suggest that many participants initially had low levels of assessment literacy but later displayed greater understanding of assessment and its link to teaching. The benefits and drawbacks of MOOCs are also considered.

Introduction

Assessment plays a significant role in the work of any teacher as they spend up to a third of their time on assessment-related work despite not having the requisite training to do it well (Stiggins, 2014; Stoynoff & Coombe, 2012). However, assessment could be considered the most important thing teachers do for students because 'the results . . . influence students for the rest of their lives' (Race, Brown, & Smith, 2005, p. xi). Consequently, assessment literacy should be 'a pivotal content area' in teacher training courses (Popham, 2009, p. 4).

Although more courses include training in assessment as part of pre-service teacher training, many teachers are still left to develop their knowledge through in-service training. However, access to professional development can be challenging due to lack of access to resources and materials, costs for course fees and travel, and time pressure (Odden et al., 2002). MOOCs therefore provide an ideal alternative since they are often free of charge and offer flexibility in terms of time and location.

This paper reports on a free language assessment MOOC and presents the findings of pre- and post-course questionnaires aimed at investigating participants' background, expectations at the beginning of the MOOC, their perceptions of course effectiveness, and the potential impact on classroom practice upon completion of the course.

Literature review

Language assessment literacy

Inbar-Lourie (2008) argues that being assessment literate involves 'having the capacity to ask and answer critical questions about the purpose for assessment, about the fitness of the tool being used, about testing conditions, and about what is going to happen on the basis of the results' (p. 389). Taylor (2012) adds to this definition. She argues that being assessment literate also involves the ability to select an appropriate assessment for a given purpose and to use assessment results to feed into improvements in teaching, assessment and decision-making.

Thus, in order to develop these capabilities, teachers need training. The components of language assessment literacy include training in the skills, knowledge and principles of assessment (Davies, 2008) as well as an understanding of the contextual framework of assessment (Fulcher, 2012). This involves developing the skills necessary to design good assessments, write test items and analyse the results as well as an understanding of language teaching, testing, and measurement while also taking into account issues of fairness, ethics and the correct use of language tests (Davies, 2008, p. 328). Fulcher (2012) expands on Davies (2008) to include a contextual framework which takes into account the origins, reasons and impact of assessment in terms of historical, social, political and philosophical aspects. With regard to teachers' assessment literacy needs, in addition to skills and knowledge, the teaching context is also a crucial component (Crusan, Plakans, & Gebril, 2016; Giraldo, 2019; Levi & Inbar-Lourie, 2020; Scarino, 2013; Vogt, Tsagari, & Spanoudis, 2020).

A good deal of research has been done in the last decade to look into definitions and components of language assessment literacy and stakeholder needs. However, a more in-depth discussion of the literature is beyond the scope of this article. Suffice to say, though, that assessment literacy must include other stakeholders, not only teachers (Taylor, 2009), and several studies have investigated the differing assessment literacy needs of various stakeholders including teachers (Kremmel & Harding, 2020; Taylor, 2009; Vogt & Tsagari, 2014). While the authors endorse this wider view, it should be noted that the MOOC presented here was aimed specifically at teachers who need to be able to use assessment results to feed into improvements in teaching, assessment and decision-making (Taylor, 2012) (albeit not all participants were teachers) so that we could focus explicitly on the context of classroom-based language assessment.

MOOCs as a vehicle for professional development

There are different types of MOOCs but the most common by far is the type that is provided as an online space for university courses (Kay, Reimann, Diebold, & Kummerfeld, 2013, p. 70) in which the instructor plays a central role, providing input in the form of texts and videos for participants to engage with (Lowenthal, Snelson, & Perkins, 2018, p. 3) and it is this type of MOOC that is presented here.

Aside from the different types of MOOCs, research has looked at the motivational factors behind MOOC participation (Hakami, White, & Chakaveh, 2017), learner perceptions of trust and credibility (Chu, Ma, Feng, & Lai, 2015; Costello, Brunton, Brown, & Daly, 2018), and retention on MOOCs (Alraimi, Zo, & Ciganek, 2015).

Methodology

The data collection tools were two surveys. One questionnaire (N=2,031) provided through a link embedded in the first page of the course content, consisted of 12 items: 10 selected response and two open-ended. The second questionnaire (N=385), administered through a link at the end of the course, comprised 25 items: 21 selected response and four open-ended. Participation was anonymous and completely voluntary. Population-level demographic data from the registration process was also available. Selected responses were summarised descriptively, while open-ended responses were coded for emerging themes.

Results

A total of 10,348 individuals from 158 different countries registered for the course. Of these, 56% went on to participate in the course, referred to as 'learners'. Just over 10% of these learners completed 90% of the course. The most notable aspects of the survey results are described below.

Pre-course questionnaire

In terms of background characteristics, respondents were working across a range of institutions, but most were in secondary education (33%). A slim majority (55%) had done online professional development courses before. Some 69% of respondents had developed language tests as part of their work. However, as Table 1 shows, 36% of all respondents had never received any assessment training. When viewed more closely, over a quarter (27%) of those involved in test development had not undergone any form of training.

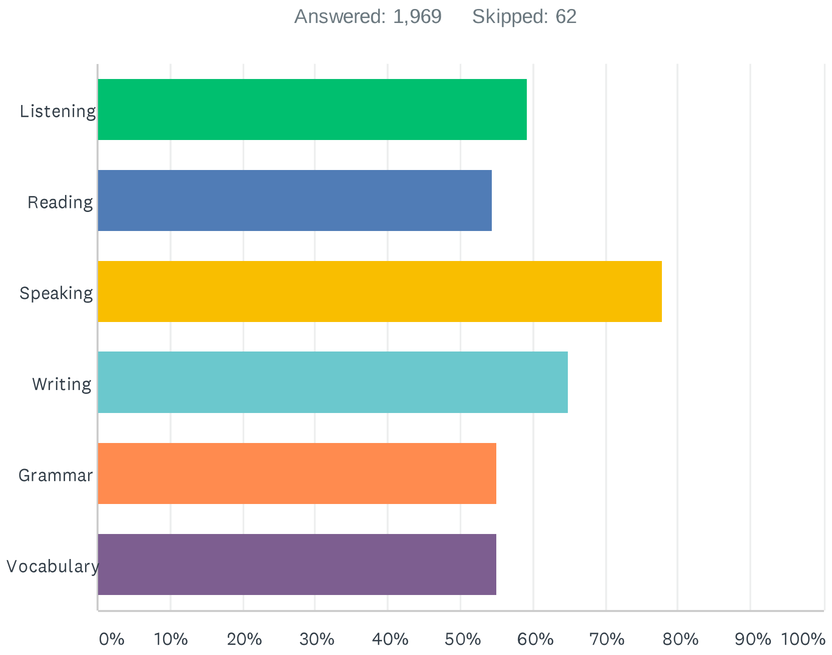
Subjects were asked which skill areas they were most interested in learning about (multiple responses were possible). While all skill areas were chosen by over half of respondents, the productive skills were the most popular (Figure 1).

Survey respondents stated their intentions of how many hours per week they would spend on the course. Around two-thirds (65%) selected from within the range 2–4 hours, 3 hours (27%) being the most popular (Figure 2).

In terms of general course expectations, there was a wide variety of responses, but some key topics emerged. First, respondents expressed a wish to produce more effective assessments. Second, there was recognition of the potential for positive impact on both teaching skills and student learning. The other important theme to emerge was the need to continue their own professional growth and keep up to date with developments in the field.

Table 1: Have you received previous assessment training?

Responses	%
No, I haven't received any training on assessment.	35.8
Yes, in a teaching course.	28.7
Yes, in a workshop.	16.4
Yes, by self-study.	13.7
Yes, in an assessment course.	5.6



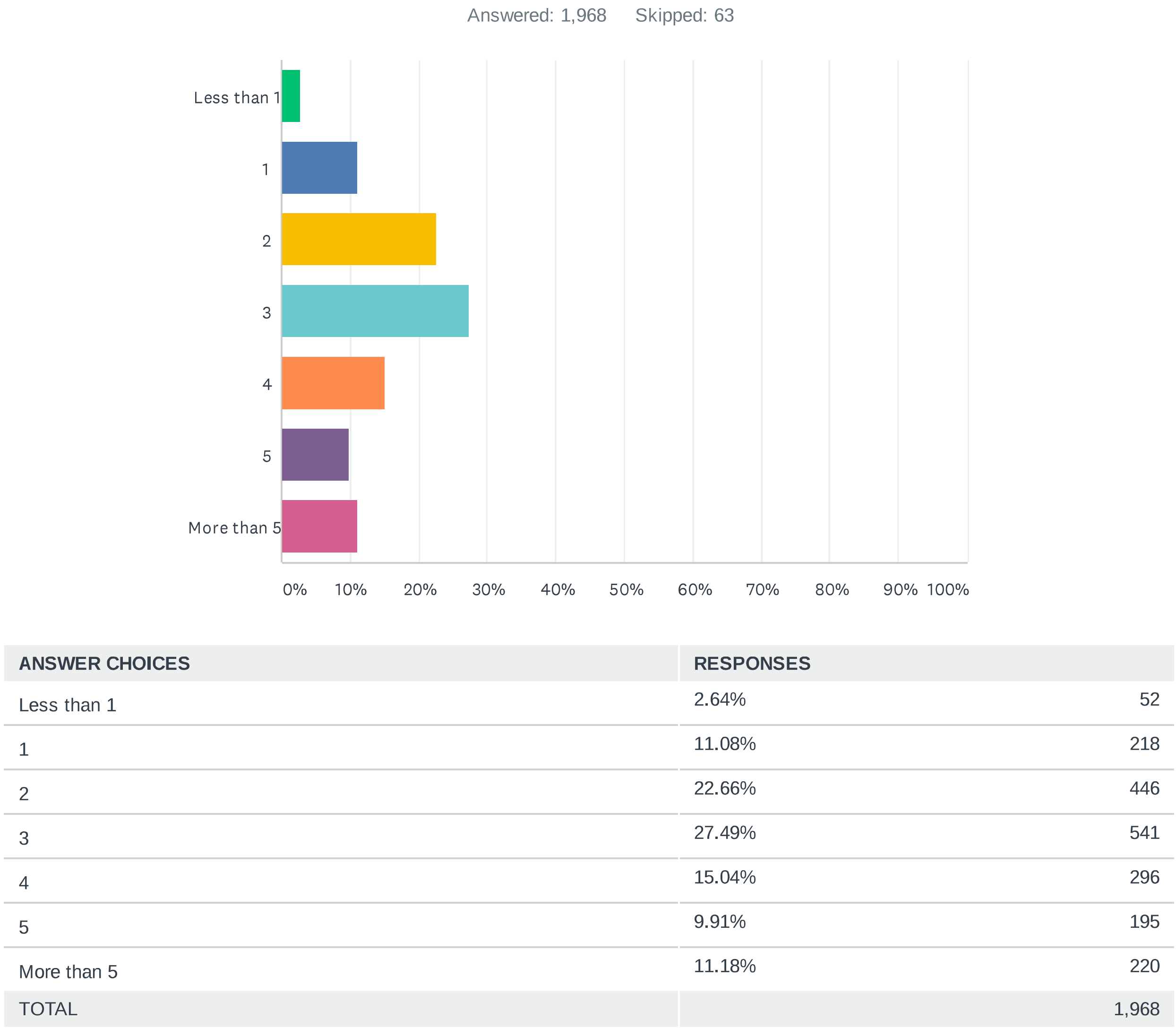
ANSWER CHOICES	RESPONSES	
Listening	59.17%	1,165
Reading	54.39%	1,071
Speaking	77.86%	1,533
Writing	64.75%	1,275
Grammar	55.00%	1,083
Vocabulary	54.95%	1,082
Total Respondents: 1,969		

Figure 1 Bar graph and table breakdown of responses to questionnaire question 'Which skill area(s) are you most interested in learning about?'

Post-course questionnaire

Of those who completed the post-course questionnaire, 85% were teachers. The rest were spread across a range of working areas such as teacher training, materials production and administration. Only 5% claimed to be 'assessment specialists'.

When asked about the course itself, 98% of respondents agreed that the language of the course was easy to understand, while a further 98% agreed that the content of the course was easy to understand. While 38% claimed that they had already known a lot of the content, 94% still said that they learned a lot from the course and 91% stated it was either 'very useful' or 'extremely useful' (Figure 3).



ANSWER CHOICES	RESPONSES	
Less than 1	2.64%	52
1	11.08%	218
2	22.66%	446
3	27.49%	541
4	15.04%	296
5	9.91%	195
More than 5	11.18%	220
TOTAL		1,968

Figure 2 Bar graph and table breakdown of responses to questionnaire question 'How many hours per week do you intend to spend on this course?'

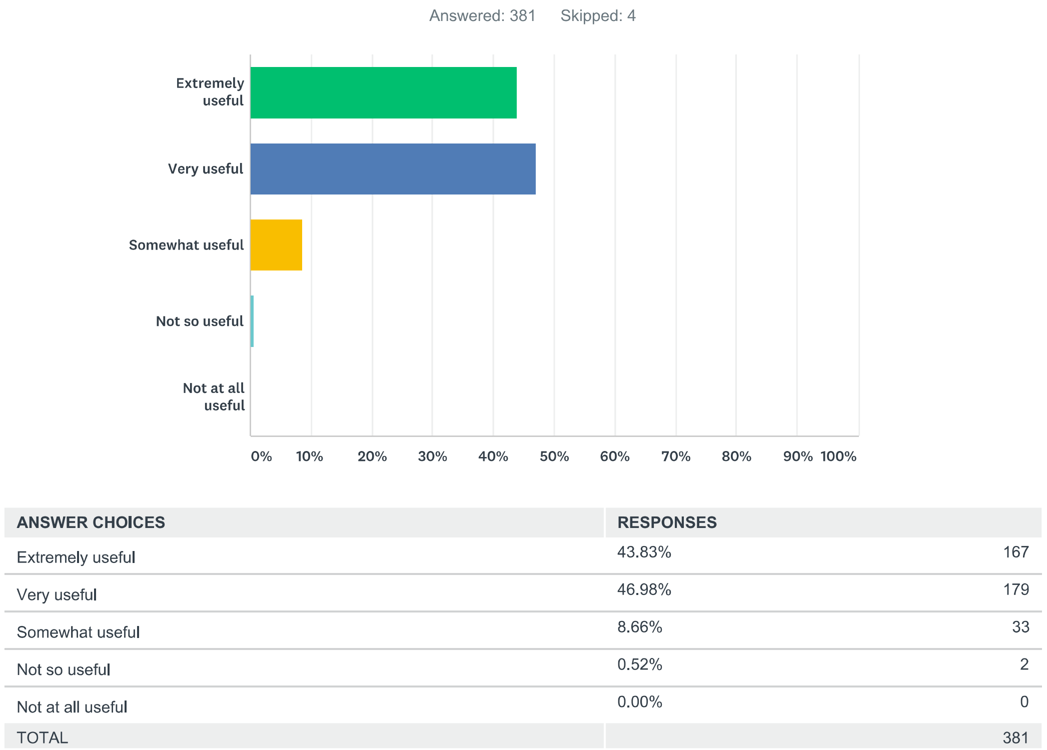
Asked why they had chosen to participate in the MOOC, the most popular answer was interest in the topic followed by the ability of learners to work at their own pace. In line with other research (Hakami et al., 2017), positive previous experience of online courses and the reputation of the British Council as a content provider were also cited as major reasons for participation.

Three questions related to the amount of time spent on the course. About two-thirds of respondents claimed to have spent 2–4 hours per week working through the course (Figure 4) and a very high proportion (93%) agreed or strongly agreed that this was acceptable, with a slightly lower number (78%) agreeing that they had enough time to work on the course activities. However, time was also indicated to be a major challenge (58%) as was the amount of course content (21%).

More detailed questions were asked about learners' interest and motivation. With very few exceptions, respondents found the course interesting overall (99%). Interest in the course content was the main motivational factor in continuing learning (86%), followed by belief that the course would be useful for their career (75%).

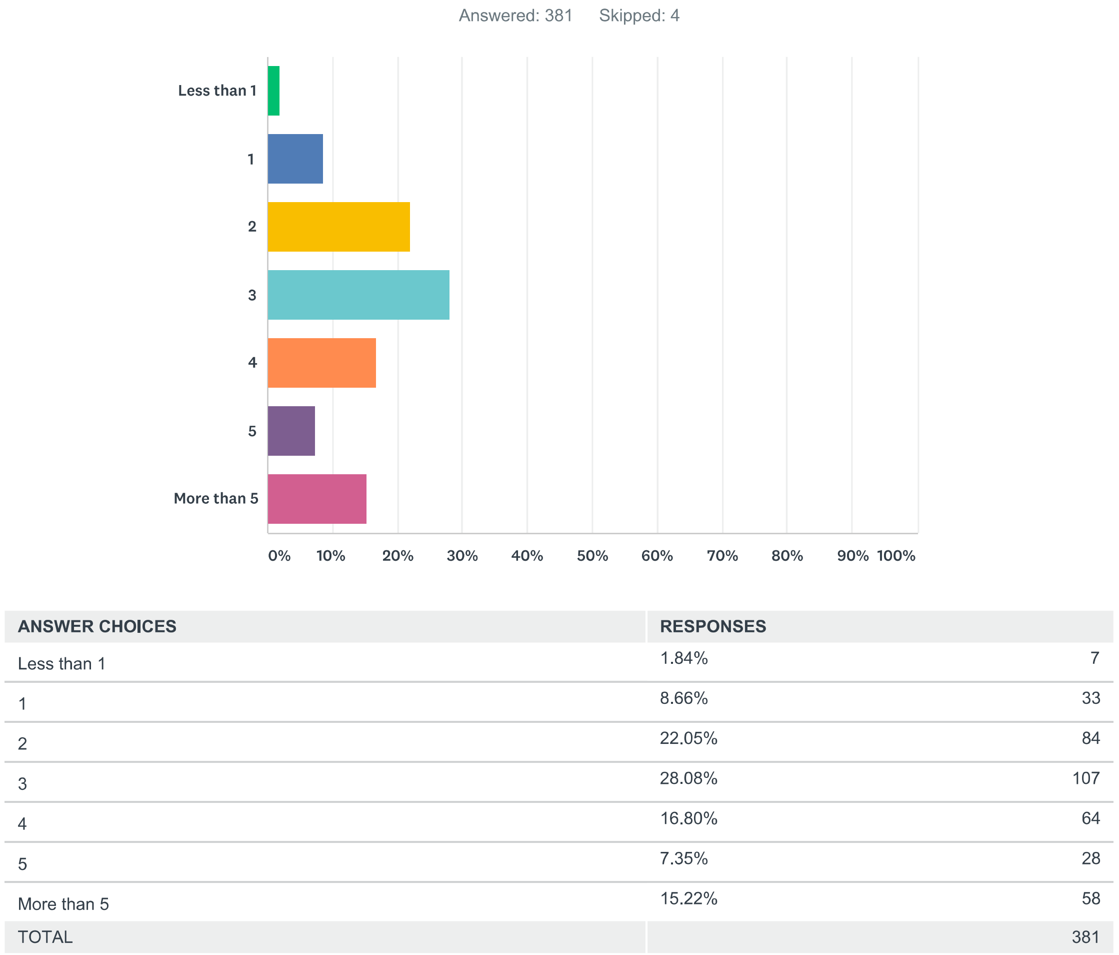
Participants were generally very positive about the course delivery but the most favoured aspects were short articles, interviews with assessment experts and video animations.

To understand potential impact on classroom practice, participants were asked whether the course would help them understand how to use their knowledge of assessment in the classroom. As shown in Figure 5, 97% agreed that it would, and 95% planned to use their learning in the classroom. When asked for specifics, a variety of areas were mentioned, yet major themes were the planning and design of assessments, the creation of more authentic tests, and ways of evaluating tests after administration,



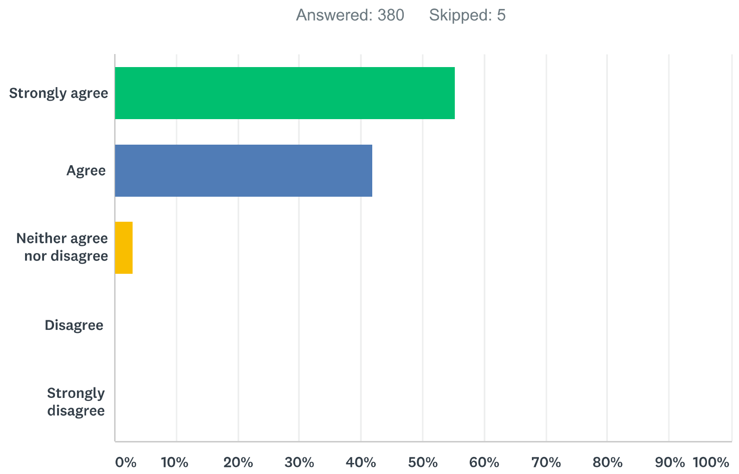
ANSWER CHOICES	RESPONSES	
Extremely useful	43.83%	167
Very useful	46.98%	179
Somewhat useful	8.66%	33
Not so useful	0.52%	2
Not at all useful	0.00%	0
TOTAL		381

Figure 3 Bar graph and table breakdown of responses to questionnaire question 'How useful was the overall content of the course?'



ANSWER CHOICES	RESPONSES	
Less than 1	1.84%	7
1	8.66%	33
2	22.05%	84
3	28.08%	107
4	16.80%	64
5	7.35%	28
More than 5	15.22%	58
TOTAL		381

Figure 4 Bar graph and table breakdown of responses to questionnaire question 'On average, how many hours per week did you spend on this course?'

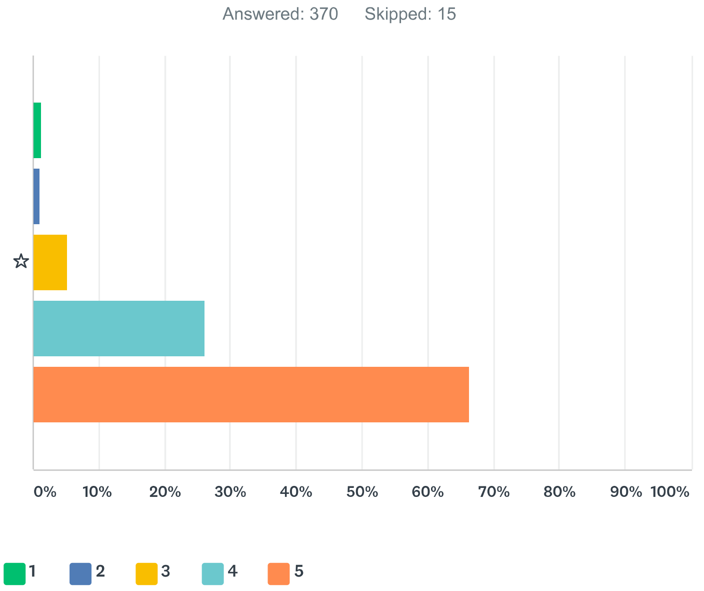


ANSWER CHOICES	RESPONSES	
Strongly agree	55.26%	210
Agree	41.84%	159
Neither agree nor disagree	2.89%	11
Disagree	0.00%	0
Strongly disagree	0.00%	0
TOTAL		380

Figure 5 Bar graph and table breakdown of responses to questionnaire statement 'The course helped me understand how to use my knowledge of assessment in the classroom.'

which appeared to be a new area for many: 'To pay more attention to the right type of test I should use with my young learners. To keep seeking for feedback from other colleagues and use the new method of calculating facility values.'

Finally, participants were asked to provide an overall rating for the course out of a maximum of 5. Two-thirds awarded the course a 5-star rating with four stars awarded by a further 26% (Figure 6).



	1	2	3	4	5	TOTAL	WEIGHTED AVERAGE
☆	1.35% 5	1.08% 4	5.14% 19	26.22% 97	66.22% 245	370	4.55

Figure 6 Bar graph and table breakdown of responses to questionnaire question 'What overall rating would you give the course?'

Discussion

It should be noted that, since both surveys were conducted online and were completely voluntary, care needs to be taken with interpretation of the results. Nevertheless, some tentative conclusions can be drawn.

Firstly, from the demographic and post-course data, it appears that the course managed to reach the target audience of predominantly secondary school teachers. Although there was a high dropout rate from registration onwards, this was lower than might be expected from a typical MOOC (Alraimi et al., 2015). The survey data revealed that many respondents had not received any assessment training. This is a concern as it includes many who have been required to actively produce assessments on top of other assessment-related activities such as interpretation and use of results, suggesting that more assessment literacy support for teachers is needed. In addition, almost half had not done any online professional development before. Thus, one implication is that the MOOC needs to accommodate users with limited experience of relevant training.

The broad expectations of the course participants were identified at the beginning of the course. Unsurprisingly, these were mainly related to teaching and the classroom environment, emphasising that it is important for courses of this nature to have a strong practical element. In this respect, the course appears to have fulfilled participants' needs. At the end of the course, participants reported high levels of satisfaction with the course overall, the accessibility of language and content, and rated the usefulness very highly. In terms of the amount of time which participants were prepared to invest, findings pre-course were consistent with those post-course. Participants seemed satisfied with this, but did indicate time pressure as a challenge, which needs to be factored into future courses.

According to the post-course survey, participants acquired an understanding of how they would implement their learning from the MOOC. They also had clear ideas about which aspects they planned to try, although ultimately further research would be needed to determine actual impact in the classroom.

Implications and conclusion

The results clearly demonstrate that the participants benefitted from the course; however, there are several benefits, drawbacks, opportunities and challenges to running this (and other) MOOCs which should be considered. For participants, the benefits are access to quality courses with world-renowned experts in the field (Evans & Myrick, 2015, p. 304) at no cost to them and flexibility to study when convenient. For the provider, it raises the profile of the organisation and moderators (Hakami et al., 2017, p. 326).

However, the drawbacks of MOOCs should not be overlooked. They require a huge time commitment by developers and moderators. Similarly, managing the time commitment from participants is challenging. While it is relatively straightforward to estimate the time required to work through the input, engaged learners will also want to follow the discussions and the more interaction there is, the longer this takes. Promoting informed interaction between participants is also challenging. While interaction and exchanges of ideas are certainly to be encouraged, participants' comments can sometimes have the effect of misleading others. Nonetheless, MOOCs provide opportunities for participants to engage in an intercultural and educational exchange with colleagues in different contexts around the world.

In terms of lessons learned, it is necessary to find a balance between the demand for the course and the resources available to moderate it. Another point is to try to gauge the participants' level of knowledge. While the content cannot be amended mid-course, moderators can tailor questions to the participants, consider where issues may arise, and focus on providing clarification for those issues. The pre-course questionnaire helped to some extent but not all participants will complete such questionnaires so, to 'level the playing field', a glossary of terms can be provided at the beginning of the course, along with a task which involves discussing the meaning of key terms from the glossary. End-of-week videos and Facebook Live sessions also help to provide ongoing clarification in a quasi-face-to-face situation. Finally, access to technical assistance is important.

To conclude, we would advocate using MOOCs for professional development because, despite the time and effort involved on the part of the provider, participants can benefit from low-cost, flexible, high-quality education, and this increases trust in the institution and raises the profile of the provider and moderators.

References

Alraimi, K.M., Zo, H., & Ciganek, A.P. (2015). Understanding the MOOCs continuance: the role of openness and reputation. *Computers and Education, 80,* 28–38.

Chu, R., Ma, E., Feng, Y., & Lai, K.W. (2015). Understanding learners' intension toward Massive Open Online Courses. In K. S. Cheung, L. F. Kwok, H. Yang, J. Fong, & R. Kwan (Eds.), *International Conference on Hybrid Learning: and Continuing Education* (pp. 302–312). New York: Springer.

Costello, E., Brunton, J., Brown, M., & Daly, L. (2018). In MOOCs we trust: learner perceptions of MOOC quality via trust and credibility. *International Journal of Emerging Technologies in Learning, 13*(6), 214–222.

Crusan, D., Plakans, L., & Gebril, A. (2016). Writing assessment literacy: surveying second language teachers' knowledge, beliefs, and practices. *Assessing Writing, 28,* 43–56.

Davies, A. (2008). Textbook trends in teaching language testing. *Language Testing, 25*(3), 327–347.

Evans, S., & Myrick, J. G. (2015). How MOOC instructors view the pedagogy and purposes of massive open online courses. *Distance Education, 36*(3), 295–311.

Fulcher, G. (2012). Assessment literacy for the language classroom. *Language Assessment Quarterly, 9*(2),113–132.

Giraldo, F. (2018). Language assessment literacy: implications for language teachers. *Profile: Issues in Teachers' Professional Development, 20*(1), 179–195.

Hakami, N., White, S., & Chakaveh, S. (2017). Motivational factors that influence the use of a MOOC: learners' perspectives. Retrieved from: www.scitepress.org/Papers/2017/62595/62595.pdf

Inbar-Lourie, O. (2008). Constructing a language assessment knowledge base: a focus on language assessment courses. *Language Testing, 25*(3), 385–402.

Kay, J., Reimann, P., Diebold, E., & Kummerfeld, B. (2013). MOOCs: so many learners, so much potential . . . . *IEEE Intelligent Systems, 28*(3), 70–77.

Kremmel, B., & Harding, L. (2020). Towards a comprehensive, empirical model of language assessment literacy across stakeholder groups: Developing the language assessment literacy survey. *Language Assessment Quarterly, 17*(1), 100–120.

Levi, T., & Inbar-Lourie, O. (2020). Assessment literacy or language assessment literacy: learning from the teachers. *Language Assessment Quarterly, 17*(2), 168–182.

Lowenthal, P., Snelson, C., & Perkins, R. (2018). Teaching Massive, Open, Online, Courses (MOOCs): tales from the front line. *International Review of Research in Open and Distributed Learning, 19*(3). Retrieved from: www.irrodl.org/index.php/irrodl/article/view/3505

Odden, A., Archibald, S., Fermanich, M., & Gallagher, H. A. (2002). A cost framework for professional development. *Journal of Education Finance, 28*(1), 51–74.

Popham, W. J. (2009). Assessment literacy for teachers: faddish or fundamental?. *Theory into Practice, 48*(1), 4–11.

Race, P., Brown, S., & Smith, B. (2005). *500 Tips on Assessment* (2nd ed). London: Routledge.

Scarino, A. (2013). Language assessment literacy as self-awareness: Understanding the role of interpretation in assessment and in teacher learning. *Language Testing, 30*(3), 309–327.

Stiggins, R. J. (2014). Improve assessment literacy outside of schools too. *Kappan, 96*(2), 67–72.

Stoynoff, S., & Coombe, C. (2012). Professional development in language assessment. In C. Coombe, P. Davidson, B. O'Sullivan, & S. Stoynoff (Eds.), *The Cambridge Guide to Language Assessment* (pp. 122–129). Cambridge: Cambridge University Press.

Taylor, L. (2009). Developing assessment literacy. *Annual Review of Applied Linguistics*, *29*, 21–36.

Taylor, L. (2012, February). *Developing assessment literacy* [Conference presentation]. ProSET Project Group, University of Bedfordshire.

Vogt, K., & Tsagari, D. (2014). Assessment literacy of foreign language teachers: Findings of a European study. *Language Assessment Quarterly, 11*(4), 374–402.

Vogt, K., Tsagari, D., & Spanoudis, G. (2020). What do teachers think they want? A comparative study of in-service language teachers' beliefs on LAL training needs. *Language Assessment Quarterly, 17*(4), 386–409.

Noch mehr Chancengleichheit für Prüfungsteilnehmende mit spezifischem Bedarf: Ein gemeinsames Vorhaben der Ungarischen Akkreditierungsbehörde für Sprachprüfungen und des ECL Sprachprüfungssystems

Hrisztalina Hrisztova-Gotthardt
Fremdsprachenzentrum der Universität Pécs, Ungarn

Réka Werner
Fremdsprachenzentrum der Universität Pécs, Ungarn

Abstract

Im Sinne der Regierungsverordnung Nr. 137/2008 (V. 16.) sind alle in Ungarn akkreditierten Prüfungsanbieter dazu verpflichtet, geeignete Maßnahmen zu ergreifen, um Prüfungsteilnehmenden mit spezifischem Bedarf einen Nachteilsausgleich zu gewähren. Da es im Fall von Prüfungsteilnehmenden mit Lese- und/oder Rechtschreibschwäche diesbezüglich keine einheitlichen Richtlinien gibt, hat die Ungarische Akkreditierungsbehörde für Sprachprüfungen im Frühling 2019 ein aus Vertretern mehrerer Prüfungsanbieter bestehendes Expertenteam einberufen. Die Experten hatten die Aufgabe, konkrete Vorschläge für Lösungsmöglichkeiten zur Prüfungsdurchführung bei Teilnehmenden mit spezifischem Bedarf zu formulieren.

Das ECL Sprachprüfungssystem ist seit Jahren darauf bedacht, Prüfungsteilnehmende mit spezifischem Bedarf beim Ausgleich ihrer Beeinträchtigung zu unterstützen. Daher wurden auch die Mitarbeiter des Internationalen ECL Prüfungszentrums dazu eingeladen, an den Beratungsgesprächen teilzunehmen und über ihre Erfahrungen auf diesem Gebiet zu berichten.

In diesem Zusammenhang werden im Rahmen des vorliegenden Beitrags jene Regeln für Good Practice vorgestellt, die vom ECL Sprachprüfungssystem bei Prüfungsteilnehmenden mit Lese- und/oder Rechtschreibschwäche angewendet werden und deren Angemessenheit und Wirksamkeit durch diverse Expertenempfehlungen und -meinungen bekräftigt wurden.

Einleitung

In seinem Werk *Fundamental Considerations in Language Testing* bezeichnete Bachman (1990, S. 24) Reliabilität und Validität als „die zwei grundlegenden Gütekriterien in Bezug auf Sprachtests“.[1] Dementsprechend waren Reliabilität und Validität jahrzehntelang die zwei Testcharakteristika, denen Aufgabenersteller und Testentwickler die größte Aufmerksamkeit geschenkt haben. Seit Beginn der 2000er Jahre rückt jedoch der Begriff *Testfairness* immer häufiger in den Fokus wissenschaftlicher Studien, Untersuchungen und Präsentationen zum Thema „Testen und Bewerten von fremdsprachlichen Kompetenzen“ (vgl. Kane, 2010; Kremmel, 2019; Kunnan, 2000; 2004; 2007; 2014; Stoynoff, 2012, etc.). Auch Handbücher und Richtlinien für Testersteller wie der *Code for Fair Testing Practices in Education* (Joint Committee on Testing Practices, 1988; 2005), die *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999; 2014); die *ETS International Principles for the Fairness in Assessments* (Educational Testing Service, 2016) und die *ALTE Principles of Good Practice* (Association of Language Testers in Europe, 2020) heben explizit hervor, dass Test- und Prüfungsanbieter bestrebt sein sollen, ihre Tests für Prüfungsteilnehmende unterschiedlichen Geschlechts, Alters, kulturellen und sprachlichen Hintergrunds sowie für solche mit spezifischem Bedarf (z.B. mit einer Körperbehinderung oder Lese- und/oder Rechtschreibschwäche) so fair wie möglich zu gestalten. Demnach sollen auch Sprachtests ausschließlich diejenige(n) Sprachfähigkeit(en) der Prüfungsteilnehmenden messen, die sie ursprünglich zu

[1] Die Übersetzung aller nicht-deutschsprachigen Zitate erfolgte durch die Autorinnen.

messen bezwecken. Sie dürfen keineswegs „einzelne Individuen aufgrund von Merkmalen bevorteilen oder benachteiligen, die für das zu messende Konstrukt irrelevant sind" American Educational Research Association et al., 2014, p. 50).

Das von Kunnan entwickelte Testfairness-Modell[2] (Kunnan, 2000; 2004), das grundsätzlich auf den im *Code for Fair Testing Practices in Education* und in den *Standards for Educational and Psychological Testing* formulierten Richtlinien basiert, behandelt fünf Aspekte, die berücksichtigt werden müssen, um die Fairness von Sprachtests und somit die Chancengleichheit für alle Teilnehmenden zu gewährleisten: (1) Validität, (2) die Vermeidung von Verzerrungsfaktoren (Bias) bei der Bewertung, (3) Zugänglichkeit, (4) Durchführung und (5) gesellschaftliche Auswirkungen. Unter Zugänglichkeit (personal access) versteht Kunnan jene Maßnahmen, die Test- und Prüfungsanbieter ergreifen können bzw. müssen, um Prüfungsteilnehmenden mit spezifischem Bedarf individuelle Lösungen zur Prüfungsdurchführung und somit einen Nachteilsausgleich zu gewähren (vgl. Kunnan, 2004, p. 37).

Die Association of Language Testers in Europe (ALTE) richtet an ihre Mitglieder ebenfalls die dringende Empfehlung, im Rahmen ihrer Möglichkeiten geeignete Prüfungsbedingungen für Teilnehmende mit spezifischem Bedarf zu schaffen (s. ALTE, 2020, p. 13). Es gilt nämlich mittlerweile als unbestrittene Tatsache, dass Tests, die sich systematisch als schwieriger für Personen mit spezifischem Bedarf erweisen, sehr wahrscheinlich eine konstrukt-irrelevante Varianz in den Ergebnissen zur Folge haben werden. Dadurch wäre auch die Validität der Rückschlüsse über die tatsächlichen fremdsprachlichen Kompetenzen der einzelnen Prüfungsteilnehmenden gefährdet, die auf der Grundlage ihrer Testergebnisse gezogen werden (vgl. Stoynoff, 2012, p. 4).

Das Konzept von Chancengleichheit für Prüfungsteilnehmende mit spezifischem Bedarf im ungarischen Kontext

Das Recht von Prüfungsteilnehmenden mit spezifischem Bedarf auf alternative Lösungen zur Prüfungsdurchführung und somit auf einen Nachteilsausgleich im Rahmen ihrer Sprachprüfung ist auch in der ungarischen Gesetzgebung verankert. Um von der Ungarischen Akkreditierungsbehörde für Sprachprüfungen akkreditiert zu werden, müssen etliche Prüfungsanbieter laut Regierungsverordnung: „[. . .] geeignete Maßnahmen ergreifen, um Prüfungsteilnehmenden mit spezifischem Bedarf im Rahmen ihrer Sprachprüfung einen Nachteilsausgleich zu gewähren. (s. § 8 Absatz 1 Satz f der Regierungsverordnung Nr. 137/2008. (V. 16.)).

Das sog. Akkreditierungshandbuch, das von der Akkreditierungsbehörde für Sprachprüfungen herausgegeben und jährlich aktualisiert wird, schreibt vor, dass etliche in Ungarn akkreditiere Prüfungsanbieter in ihrer Geschäftsordnung jene konkreten Maßnahmen darlegen müssen, die ergriffen werden, um Prüfungsteilnehmenden mit spezifischem Bedarf einen Nachteilsausgleich zu gewähren:

> VII. Wie stellt der Prüfungsanbieter [. . .] sicher, dass die vom Gesetzgeber vorgeschriebenen Bedingungen für Personen mit spezifischem Bedarf (z.B. Barrierefreiheit für Teilnehmende mit einer motorischen Beeinträchtigung, angepasste Prüfungsbedingungen für Teilnehmende mit Seh- oder Hörschwäche bzw. mit Lese- und/oder Schreibschwäche und für Teilnehmende mit sonstigen Behinderungen) gegeben sind und dass zugleich die Gleichwertigkeit der Prüfung gewahrt wird? (Accreditation Centre for Foreign Language Examination, 2021, p. 10)

Dementsprechend sind die in Ungarn akkreditierten Prüfungsanbieter bemüht, die für sie relevanten Empfehlungen und Richtlinien von Experten auf dem Gebiet der Leistungsmessung zu internalisieren und in die Praxis umzusetzen.

Konkrete Empfehlungen von Experten auf dem Gebiet der Leistungsmessung

In ihren *Standards for Educational and Psychological Testing* haben die American Educational Research Association, die American Psychological Association und der National Council on Measurement in Education konkrete Hinweise in Bezug auf die Gestaltung der Prüfungsdurführung für Teilnehmende mit spezifischem Bedarf formuliert. Folgende Hinweise können (und müssen) auch beim Testen und Bewerten von fremdsprachlichen Kompetenzen berücksichtigt werden:

- Jegliche Änderungen der Prüfungsmaterialien sollten gezielt auf die spezifischen Bedürfnisse der Teilnehmenden zugeschnitten sein.
- Je nach Bedarf können eine oder mehr Änderungen vorgenommen werden.
- Es können Änderungen in der Präsentation der Prüfungsmaterialien vorgenommen werden. So z. B. können Testhefte für Teilnehmende mit Sehschwäche in einer größeren Schriftgröße gedruckt werden.

[2] Die englische Originalbezeichnung für das Modell lautet *Test Fairness Framework*.

- Möglich sind auch Änderungen in Bezug auf die Art und Weise, wie Prüfungsteilnehmende ihre Antworten kommunizieren. Teilnehmende, die nicht in der Lage sind, ihre Lösungen per Hand aufzuschreiben, können eine Computertastatur benutzen.
- Die Zeit zum Durchführen der einzelnen Testteile kann verlängert werden.
- Die Prüfungen von Teilnehmenden mit spezifischem Bedarf können individuell, in einem getrennten Prüfungsraum durchgeführt werden.

(American Educational Research Association et al., 1999, SS. 102–103; 2014, SS. 49–62).

Sicherstellung einer entsprechend angepassten Prüfungsdurchführung bei Prüfungsteilnehmenden mit spezifischem Bedarf im Rahmen des ECL Sprachprüfungssystems

Im Rahmen des ECL Sprachprüfungssystems[3] haben Prüfungsteilnehmende mit spezifischem Bedarf, die diesen Status durch ein ärztliches Attest belegen können, einen Anspruch auf eine entsprechend angepasste Prüfungsdurchführung. Alle ECL Prüfungsorte müssen diese Möglichkeit in folgenden Fällen gewährleisten: bei Teilnehmenden mit Seh- oder Hörschwäche, bei blinden Teilnehmenden, bei Teilnehmenden mit einer motorischen Beeinträchtigung sowie bei Teilnehmenden mit Lese- und/oder Rechtschreibschwäche (s. *Richtlinien zur Durchführung von ECL Prüfungen bei Prüfungskandidaten mit spezifischem Bedarf*, eclexam.eu/deutsch/informationen-fur-kandidatenmit-spezifischem-bedarf/).

Offene Fragen in Bezug auf Prüfungsteilnehmende mit Lese- und/oder Rechtschreibschwäche

Die Maßnahmen, die im Fall von Personen mit einer diagnostizierten Seh- oder Hörschwäche bzw. bei Personen mit einer motorischen Beeinträchtigung eingeleitet werden können, sind verhältnismäßig eindeutig zu definieren und festzulegen (vgl. *Richtlinien zur Durchführung von ECL Prüfungen bei Prüfungskandidaten mit spezifischem Bedarf*).

Im Gegensatz dazu gibt es bei Prüfungsteilnehmenden mit einer ärztlich attestierten Lese- und/oder Rechtschreibschwäche mehr offene Fragen als zufriedenstellende Antworten. Aus diesem Grund hat die Ungarische Akkreditierungsbehörde für Sprachprüfungen 2019 ein Expertenteam einberufen, das sich u.a. mit folgenden Fragen auseinandersetzen sollte:

- Auf was für Maßnahmen greifen die in Ungarn akkreditierten Prüfungsanbieter zurück, um einen angemessenen Nachteilsausgleich für Prüfungsteilnehmende mit einer Lese- und/oder Rechtschreibschwäche zu gewährleisten?
- Ist eine Normierung dieser Regelungen und Maßnahmen möglich?

Vertreter des ECL Sprachprüfungssystem wurden ebenfalls dazu eingeladen, an den Beratungsgesprächen teilzunehmen und über ihre Erfahrungen auf diesem Gebiet zu berichten.

Ergebnisse der Beratungsgespräche und einheitliche Empfehlungen für die in Ungarn akkreditierten Sprachprüfungssysteme

Es wurden folgende einheitliche, jedoch unverbindliche Empfehlungen formuliert, die im Einklang mit der bisherigen Praxis des Internationalen ECL Prüfungszentrums stehen:

- Prüfungsteilnehmende mit einer diagnostizierten Lese- und/oder Rechtschreibschwäche müssen einen Antrag zur Gewährung einer individuellen Prüfungsdurchführung stellen. Dem Antrag ist in jedem Fall ein ärztliches Attest beizulegen.
- Über die Gewährleistung einer individuellen Prüfungsdurchführung anhand des eingereichten Antrags und des beigefügen ärztlichen Attests entscheidet der Referent für Chancengleichheit des jeweiligen nationalen (oder ggf. des Internationalen) Prüfungszentrums.

[3] Die ECL Sprachprüfungen wurden von den Mitgliedern des *European Consortium for the Certificate of Attainment in Modern Languages* nach einheitlichen Standards konzipiert und werden seit 1999 vom Internationalen ECL Prüfungszentrum an der Universität Pécs in Ungarn verwaltet. Derzeit können ECL Sprachprüfungen auf vier GER Niveaustufen (A2, B1, B2 und C1) und in 15 Sprachen (Bulgarisch, Deutsch, Englisch, Französisch, Hebräisch, Italienisch, Kroatisch, Polnisch, Rumänisch, Russisch, Serbisch, Slowakisch, Spanisch und Tschechisch) abgelegt werden. Mehr zu den ECL Prüfungen auf https://eclexam.eu.

- Bei der individuell angepassten Prüfungsdurchführung dürfen keine inhaltlichen Änderungen des Prüfungsmaterials vorgenommen werden; die Abweichungen dürfen ausschließlich die Präsentation der Materialien und die Kommunikation der Antworten durch die Prüfungsteilnehmenden betreffen.

Konkrete Maßnahmen zum Nachteilsausgleich für Prüfungsteilnehmende mit Lese- und/oder Rechtschreibschwäche im Rahmen des ECL Sprachprüfungssystems

Beim Anpassen der Prüfungsmaterialien an die Bedürfnisse der Teilnehmenden mit Lese- und/oder Rechtschreibschwäche orientiert sich das ECL Sprachprüfungssystem u.a. an den Empfehlungen der British Dyslexia Association (2018). Dementsprechend werden die Tests:

- auf hellgelbem Papier
- in einer verhältnismäßig großen Schriftgröße (16 Pt.)
- mit 1,5, fachem Zeilenabstand
- in einer speziellen Schriftart (OpenDyslexic) und
- ohne Blocksatzformat gedruckt.

Außerdem wird die Zeit für die Prüfungsteile Lesen, Schreiben und Hören in der Regel um 30 Prozent[4] verlängert.

Ferner dürfen Prüfungsteilnehmende im Falle einer ärztlich attestierten Rechtschreibschwäche ihre Lösungen mithilfe eines Laptops oder PCs mit integriertem Wörterbuch – jedoch ohne Internetzugang – aufzeichnen.

Bis Oktober 2018 wurde in den ECL Tests der auch von der British Dyslexia Association empfohlene Buchstabentyp Verdana verwendet. Seit Dezember 2018 wird die speziell für Personen mit Leseschwäche entwickelte Schriftart OpenDyslexic benutzt. Im Unterschied zu herkömmlichen Schriftarten sind die Buchstaben bei OpenDyslexic unten etwas dicker. Durch dieses „Gewicht" können auch legasthene Menschen die Richtung der Buchstaben leichter erkennen, es kommt zu keinem Umdrehen oder Verwechseln der Buchstaben mehr (s. OpenDyslexic).

2019 wurde auch das erste Vorbereitungsbuch (Szabó, Barefield & Papp, 2019) veröffentlicht, das ECL Mustertests für Englisch als Fremdsprache beinhaltet, deren typografische Gestaltung auf die speziellen Bedürfnisse von Prüfungsteilnehmenden mit Leseschwäche zugeschnitten ist.

Erste Ergebnisse

Die Einführung der neuen Schriftart hat bereits erste positive Auswirkungen auf die Anzahl der ECL Prüfungsteilnehmenden mit Lese- und Rechtschreibschwäche und auf ihre Prüfungsergebnisse gezeigt. Wie den folgenden Abbildungen zu entnehmen ist, sind ein spürbarer Anstieg in Bezug auf die Anzahl der Prüfungsteilnehmenden sowie eine leichte Verbesserung ihrer Prüfungsergebnisse zu vermerken.

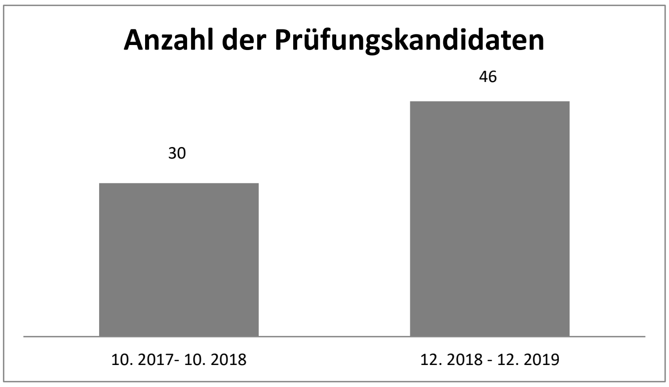


Abbildung 1: Anzahl der Prüfungsteilnehmenden mit Lese- und/oder Rechtschreibschwäche

Fazit und Ausblick

Wie aus den obigen Ausführungen hervorgeht, besteht immer noch Nachholbedarf, was die Gewährleistung von Chancengleichheit und die Schaffung von entsprechenden

[4] In der einschlägigen Literatur finden sich keine konkreten Empfehlungen bzw. Hinweise bezüglich einer eventuellen Verlängerung der Prüfungszeit zum Zwecke des Nachteilsausgleichs. Demzufolge haben sich die in Ungarn akkreditierten Prüfungssysteme an den für Abiturprüfungen geltenden Richtlinien orientiert und gemeinsam den Vorschlag unterbreitet, im Falle von Teilnehmenden mit spezifischem Bedarf die ursprüngliche Prüfungszeit um 30 Prozent zu verlängern. Dem Vorschlag wurde von der Ungarischen Akkreditierungsbehörde für Sprachprüfungen zugestimmt.

Prüfungsbedingungen für Teilnehmende mit einer Lese- und/oder Rechtschreibschwäche angeht. Einige erste Maßnahmen, die u.a. vom ECL Prüfungssystem ergriffen werden, haben bereits positive Auswirkungen auf die Prüfungsergebnisse der betroffenen Personen gezeigt. In Expertenkreisen in Ungarn wird seit Kurzem darüber gemutmaßt, dass bestimmte objektive Aufgabentypen wie z.B. Lückentexte – die auch in den ECL Tests zum Leseverstehen vorkommen – eine zusätzliche kognitive Herausforderung für Prüfungsteilnehmende mit Lese- und/oder Rechtschreibschwäche darstellen könnten und daher durch alternative Aufgabentypen wie z.B. Zuordnungsaufgaben ersetzt werden müssten. Diese Hypothese wurde jedoch bisher nicht verifiziert und soll im Rahmen künftiger empirischer Untersuchungen überprüft werden.

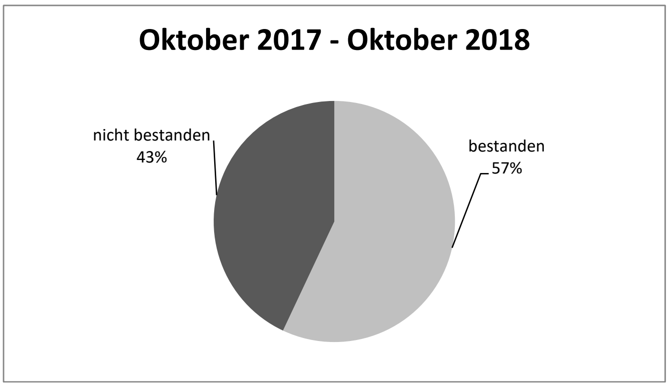


Abbildung 2: Erfolgreich abgelegte Prüfungen im Zeitraum zwischen Oktober 2017 und Oktober 2018

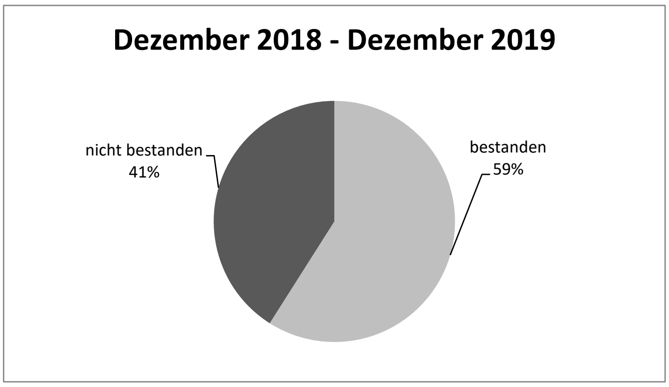


Abbildung 3: Erfolgreich abgelegte Prüfungen im Zeitraum zwischen Dezember 2018 und Dezember 2019

Literatur

Accreditation Centre for Foreign Language Examination. (2021). *Accreditation Handbook. Budapest: Educational Authority*. Retrieved from: nyak.oh.gov.hu/nyat/doc/AH2021-eng/EN_Accreditation_Handbook_2021.pdf

Association of Language Testers in Europe. (2020). *ALTE Principles of Good Practice*. Cambridge: ALTE.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for Education and Psychological Testing*. Washington, D. C.: American Educational Research Association.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for Education and Psychological Testing*. Washington, D. C.: American Educational Research Association.

Bachman, L. F. (1990). *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press.

British Dyslexia Association. (2018). *Dyslexia Style Guide 2018: Creating Dyslexia Friendly Content*. Retrieved from: cdn.bdadyslexia.org.uk/documents/Advice/style-guide/Dyslexia_Style_Guide_2018-final-1.pdf?mtime=20190409173949

Educational Testing Service. (2016). *ETS International Principles for the Fairness of Assessments. A Manual for Developing Locally Appropriate Fairness Guidelines for Various Countries*. Retrieved from: www.ets.org/s/about/pdf/fairness_review_international.pdf

Joint Committee on Testing Practices. (1988). *Code of Fair Testing Practices in Education*. Washington, D.C.: Joint Committee on Testing Practices.

Joint Committee on Testing Practices. (2005). *Code of Fair Testing Practices in Education* (revised). Washington, D.C.: Joint Committee on Testing Practices.

Kane, M. (2010). Validity and fairness. *Language Testing, 27*, 177-182.

Kremmel, B. (2019, November). *Avoiding bias in language test development* [Conference presentation]. ATLE 54th Meeting and Conference, Ljubljana.

Kunnan, A. J. (2000). Fairness and justice for all. In A. J. Kunnan (Ed.), *Fairness and Validation in Language Assessment: Selected Papers from the 19th Language Testing Research Colloquium, Orlando, Florida* (pp. 1–13). Studies in Language Testing volume 9. Cambridge: UCLES/Cambridge University Press.

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

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

Kunnan, A. J. (2014). Fairness and Justice in Language Assessment. In A. J. Kunnan (Ed.), *The Companion to Language Assessment* (pp. 1–17). Hoboken: John Wiley & Sons.

OpenDyslexic. (n.d.). Retrieved from: opendyslexic.org/

Regierungsverordnung Nr. 137/2008. (V. 16.). [137/2008. (V. 16.) *Korm. Rendelet az idegennyelv-tudást igazoló államilag elismert nyelvvizsgáztatásról és a külföldön kiállított, idegennyelv-tudást igazoló nyelvvizsga-bizonyítványok Magyarországon történő honosításáról*. (2008). Retrieved from: net.jogtar.hu/getpdf?docid=a0800137.kor&targetdate=20180101&printTitle=137/2008.+%28V.+16.%29+Korm.+rendelet

Stoynoff, S. (2012). Fairness in language assessment. In C. A. Chapelle (Ed.), *The Encyclopedia of Applied Linguistics* (Online edition). Retrieved from: onlinelibrary.wiley.com/doi/abs/10.1002/9781405198431.wbeal0409

Szabó, Sz., Barefield, J., & Papp, E. (2019). *The most hated topics at the ECL exam for individuals with dyslexia – Level B2*. Nyíregyháza: Szabó Nyelviskola Kft.

Big wrong data: Insights from developing and using an automatic translation revision tool with error memories

Bert Wylin
KU Leuven and Televic Education, Belgium

Gys-Walt van Egdom
Universiteit Utrecht, The Netherlands

Abstract

Since August 2017, more than 40 institutions (and over 800 translators) piloted the translationQ revision platform. This paper will not focus on the development of but shows some lessons learned from translationQ's error/revision memories: how to recycle the errors ('Big Wrong Data') into useful insights for translation, translation evaluation and translation teaching.

The translationQ project (a joint initiative of KU Leuven and Televic Education) was developed to automate and speed up evaluation processes in both translator education and the translation profession. The tool works for bilingual or monolingual text productions. The program's core is an error or revision 'memory': it allows the system to recognize errors in new translations and to suggest corrections and feedback automatically; the program leaves room for human intervention. Still, the bulk workload of retyping the same corrections and feedback time and again is now done automatically by the tool, leading to more rapid and more consistent revision feedback and scoring. Revision memories can be shared and reused with new texts and with new trainees.

The reporting module of the platform allows the profiling of translators (strengths and weaknesses) in an objective way.

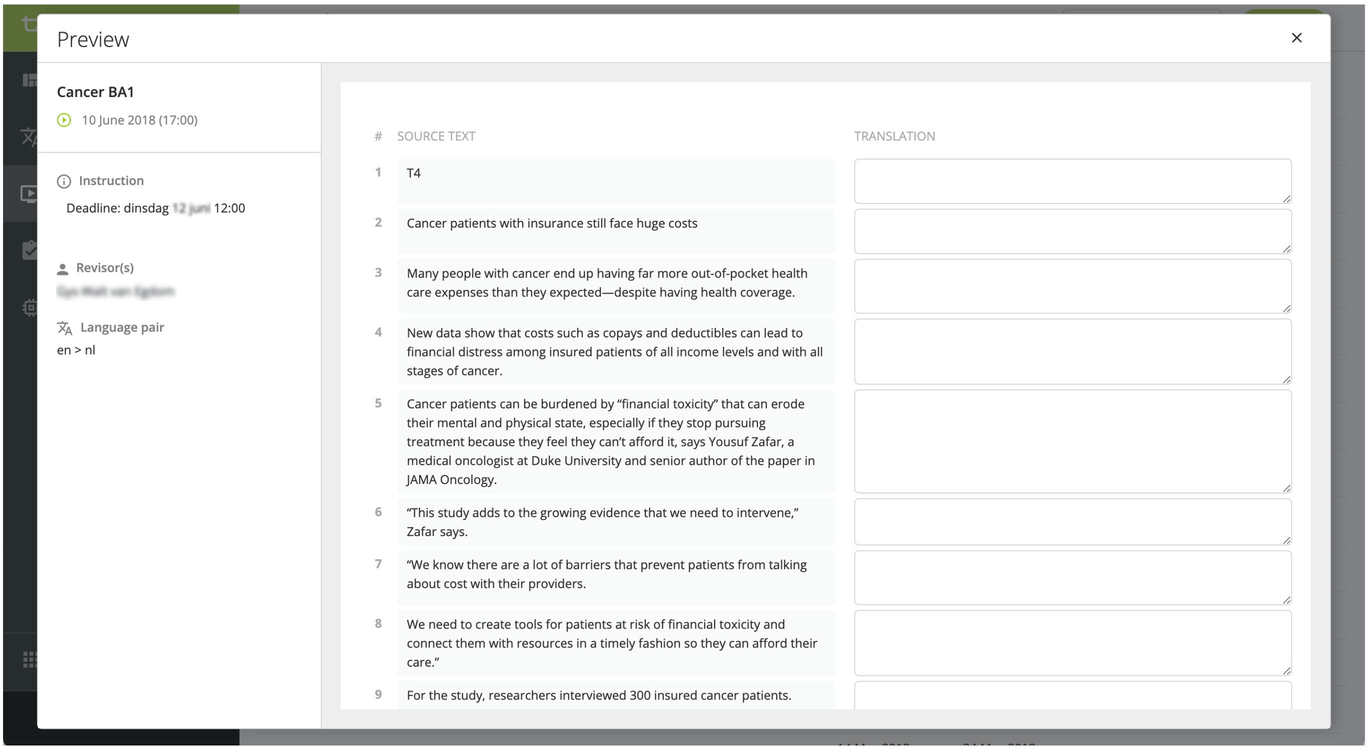
Introduction

The domain of translation technology can be described as a hive of activity. In a survey, Translation Automation User Society (TAUS) canvassed no less than eighty 'technological profiles' for Language for Specific Purposes (LSP) (2016). We believe that, with the latest advances in machine learning, this number is likely to continue to increase. A survey amongst translator trainers would yield different results. Although modern translation environment tools (TEnTs) can serve to aid some of the ergonomic and general qualitative problems in translator training, one will be hard-pressed to name but one tool that is geared to the specific needs of translator trainers. By ergonomics, we understand the combination of psychological and technical principles in the product design, with the goal being to reduce human error, to increase productivity and to enhance the wellbeing of the people involved. At first blush, it is quite remarkable that little to no attention has been paid to technological means to remove the drudgeries from tasks like trainer-to-trainee revision and student translation evaluation, and, thereby, speed up these processes, and boost quality.

With its official release in April 2018, the cloud-based tool translationQ has come to occupy an exceptional position in translator training. The tool, designed by Televic, was developed in close collaboration with KU Leuven, and, as a result of this co-creation, is primarily attuned to translator trainers' and trainees' needs. When designing the product, researchers and developers concluded that what is needed in trainer-to-trainee revision is 1) authenticity, 2) objectivity, and 3) efficiency.

Authenticity

First of all, the tool offers the potential to enhance authentic experiential learning, as it bears resemblance to the virtual working environment of TEnTs, not only in the translation mode but also in the revision and the post-revision mode. A good case in point is the lay-out of a translation assignment: the source text and the target text are segmented and appear in two columns (Figure 1).



Cancer BA1
10 June 2018 (17:00)

Instruction
Deadline: dinsdag [redacted] 12:00

Revisor(s)
[redacted]

Language pair
en > nl

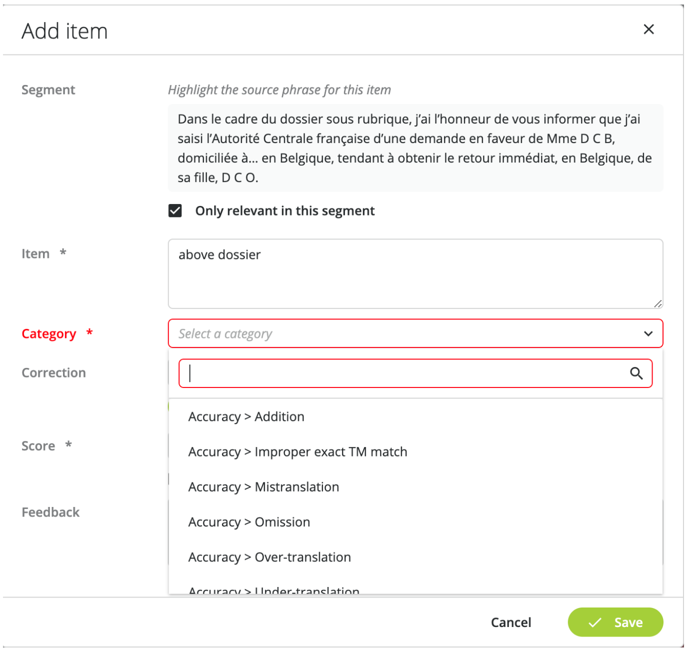
#	SOURCE TEXT	TRANSLATION
1	T4	
2	Cancer patients with insurance still face huge costs	
3	Many people with cancer end up having far more out-of-pocket health care expenses than they expected—despite having health coverage.	
4	New data show that costs such as copays and deductibles can lead to financial distress among insured patients of all income levels and with all stages of cancer.	
5	Cancer patients can be burdened by “financial toxicity” that can erode their mental and physical state, especially if they stop pursuing treatment because they feel they can’t afford it, says Yousuf Zafar, a medical oncologist at Duke University and senior author of the paper in JAMA Oncology.	
6	“This study adds to the growing evidence that we need to intervene,” Zafar says.	
7	“We know there are a lot of barriers that prevent patients from talking about cost with their providers.	
8	We need to create tools for patients at risk of financial toxicity and connect them with resources in a timely fashion so they can afford their care.”	
9	For the study, researchers interviewed 300 insured cancer patients.	

Figure 1 Screenshot of traditional two column source text target text layout for trainees to enter their translation

Trainees and trainers benefit from this layout: there is no need for them to divide attention between a source file (or even a paper sheet with the source text) and a target file. Furthermore, revision processes are authenticated through the use of error categories that have become common currency in the language industry, more specifically, the Dynamic Quality Framework or DQF categories (TAUS, 2015), as well as through error weighting, which means that a distinction can be made between minor and major errors (Figure 2).

Objectivity

By employing the analytical categories of DQF, translationQ developers have not only sought to authenticate trainer-to-trainee revision processes, but also to provide a fillip to objectivity in revision. The tool allows for trainers to rename the DQF categories and use their own preferential error labels. This renaming can be beneficial to individual trainers but pose a threat for cross-trainer portability of the categories or error memories. For that reason, renamed categories are still (silently) linked to the original DQF categories, thus permanently allowing the sharing and comparing of error memories of all trainers and institutions. However, we must not be too hasty to assume that the common use of DQF categories allows for complete objectivity in revision. From the criticism garnered against analytical testing, one can infer that, despite clear definitions, ʻ it is often a matter of subjective judgement [. . .] whether an error falls into one category or another’ (Saldanha & O’Brien, 2014, pp. 101–102). Furthermore, subjectivity looms large in error weighting: it is up to the reviser to decide whether an error is minor, major or even critical. By sharing and exchanging revision memories, revisers can make sure they use the exact same penalties



Segment
Highlight the source phrase for this item
Dans le cadre du dossier sous rubrique, j’ai l’honneur de vous informer que j’ai saisi l’Autorité Centrale française d’une demande en faveur de Mme D C B, domiciliée à... en Belgique, tendant à obtenir le retour immédiat, en Belgique, de sa fille, D C O.

☑ Only relevant in this segment

Item * above dossier

Category * *Select a category*

Correction

Score *

Feedback

Accuracy > Addition
Accuracy > Improper exact TM match
Accuracy > Mistranslation
Accuracy > Omission
Accuracy > Over-translation
Accuracy > Under-translation

Figure 2 Screenshot of how to add and categorise a translation error in translationQ

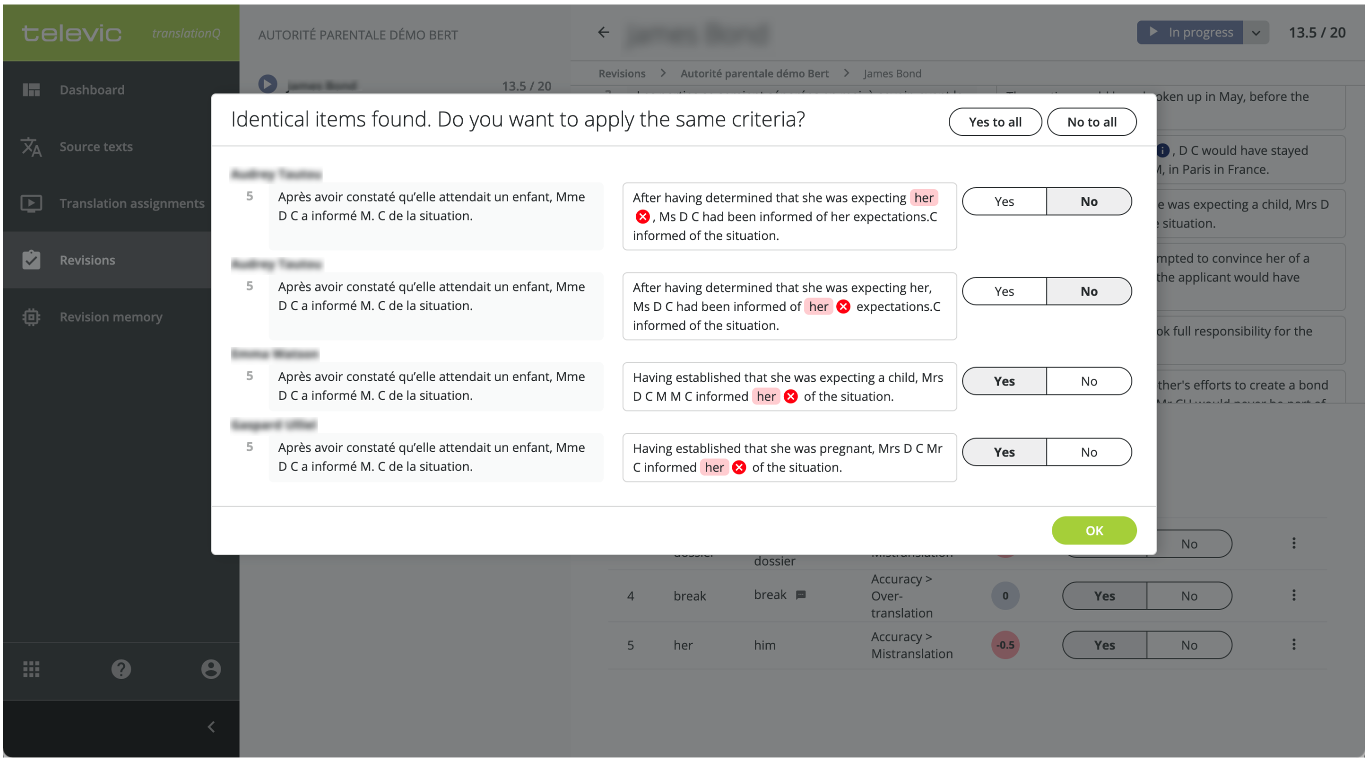
and weights for identical errors, which will improve the inter-rater consistency. Of course, the lack of inter-rater reliability cannot be *fully* remediated in the tool; as has been pointed out by a great many specialists, there will probably never be a panacea for subjectivity in translation revision and evaluation (e.g. Bowker, 2001; Holmes, 1988; Thelen, 2016). Still, inter-rater evaluation is made more consistent, as the error categories are institution-dependent and serve as a basis for an intersubjective error memory. What is more, the tool does seem to provide a solution to problems with intra-rater reliability, problems that are not very often admitted in revision literature but that are perhaps the most salient problems in trainer-to-trainee revision processes. In translator training, it is incredibly difficult to keep tabs on all revisions of a single assignment. Common human reviser errors that cause low intra-rater reliability are:

- identical errors are sanctioned in one student version and overlooked in the other
- identical errors are sanctioned by the reviser, but the error made by one student is classified differently from that of another student
- identical errors are sanctioned by the reviser, but the error weighting differs from one version to another.

In translator training, these inconsistencies are frequently brought to light in translation courses, since students not only have to complete similar tasks, they are often in close contact and tend to compare their versions. The consequences of low intra-rater reliability are far-reaching, as blatant inconsistencies undermine the trainer's expert position and might even discourage students (Hönig, 1998, p. 15). Drawing upon basic insights in corpus-based approaches to translation and computational linguistics, the developers of translationQ have provided a way for the translator trainer to revise in a consistent manner: the algorithm of the tool goes in search of, detects and flags identical errors (in the same segment of other students' versions) and similar errors (in different segments); and the trainer can decide to accept or reject computed revision suggestions (Figure 3).

Efficiency

This brings us to the final issue that the translationQ project members have sought to address: efficiency. In keeping with current technological trends in the language industry, the tool has been developed with a view to making lighter work of professional practices, in this case: translation revision. This means that a great deal of time has been dedicated to enhancing user experience and, ultimately, meeting the ergonomic threshold set by potential users. The tool has been made accessible for the less IT-literate as well as challenging for trainers with a knack for technology. By this, we mean that it allows for project creation and correction in a few simple steps; however, it also provides functionalities for trainers who wish to get the most out of a revision tool (authentication, elaborate revision, revision data ('rich wrong data')).



Identical items found. Do you want to apply the same criteria?

[redacted]
5 Après avoir constaté qu'elle attendait un enfant, Mme D C a informé M. C de la situation.
After having determined that she was expecting her ✕, Ms D C had been informed of her expectations.C informed of the situation.
Yes [✓ No]

[redacted]
5 Après avoir constaté qu'elle attendait un enfant, Mme D C a informé M. C de la situation.
After having determined that she was expecting her, Ms D C had been informed of her ✕ expectations.C informed of the situation.
Yes [✓ No]

[redacted]
5 Après avoir constaté qu'elle attendait un enfant, Mme D C a informé M. C de la situation.
Having established that she was expecting a child, Mrs D C M M C informed her ✕ of the situation.
[✓ Yes] No

[redacted]
5 Après avoir constaté qu'elle attendait un enfant, Mme D C a informé M. C de la situation.
Having established that she was pregnant, Mrs D C Mr C informed her ✕ of the situation.
[✓ Yes] No

4	break	break	Accuracy > Over-translation	0	[✓ Yes] No
5	her	him	Accuracy > Mistranslation	-0.5	[✓ Yes] No

Figure 3 Screenshot of translationQ error algorithm

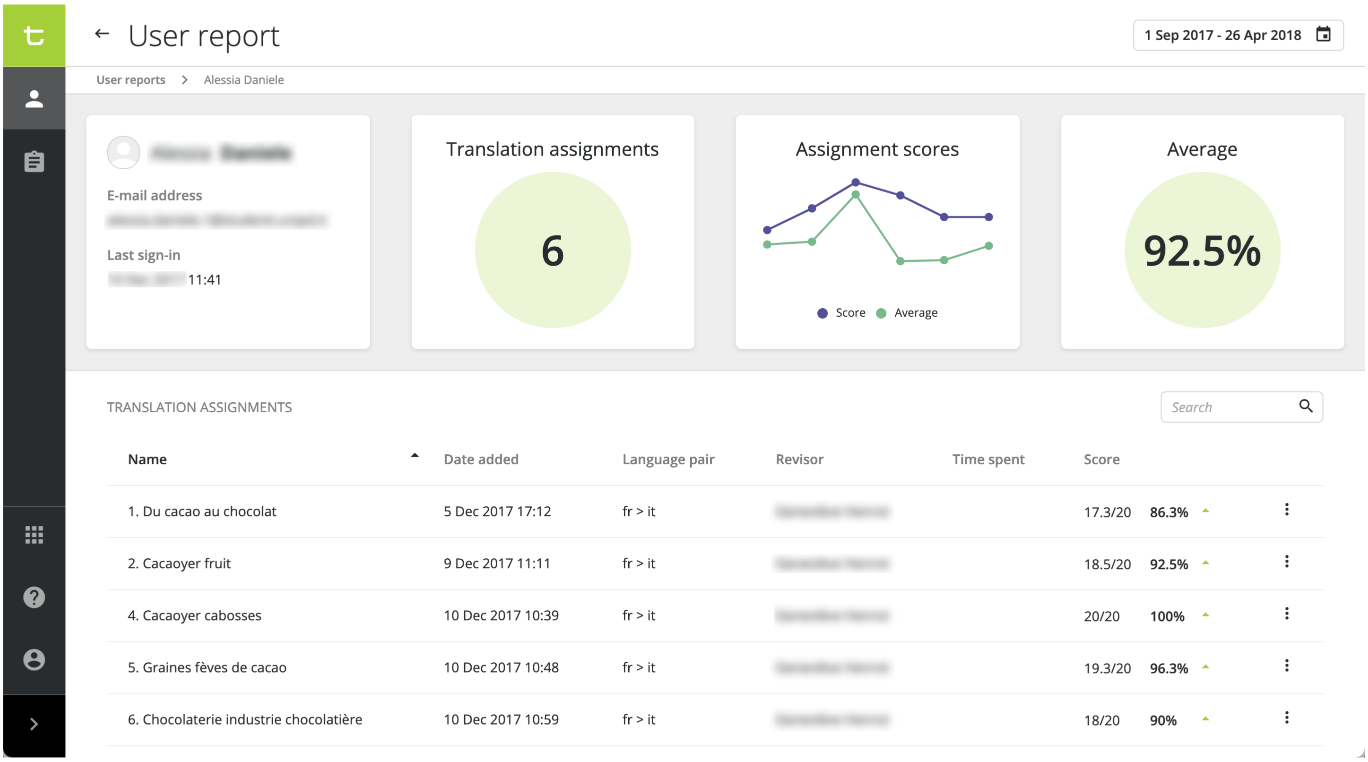
In order to be able to meet the ergonomic threshold, special heed was to be taken of factors that make trainer-to-trainee revision a tedious and tiresome task. Trainer-to-trainee revision can be aptly described as donkey work, primarily because of its repetitive nature. In trainer-to-trainee revision, it is imperative that the reviser work his way through every student version and examines the text in detail for its suitability to the purpose as laid out in the assignment. In practice, this means that the reviser gets caught up in a vicious cycle, in which identical mistakes are detected, labelled and, ideally, corrected time and again. As mentioned earlier on, translationQ removes this drudgery from the revision process: the algorithm flags identical (and similar) errors in the same student version and in others' versions and copies the comments and suggestions for improvement onto the flagged fragments that are deemed relevant by the reviser.

The ergonomic quality is further increased through the revision memory function. The functioning of the algorithm is not limited to one sole project; translationQ automatically stores identified errors (along with the error types, comments and suggestions) and incrementally builds up a revision memory. By gradually building up a revision memory, translationQ reduces the reviser's cognitive load and speeds up revision practices considerably.

This is not all there is to say about the added ergonomic value of translationQ. The software also provides automated solutions for the following tasks:

- calculating scores for formative assessment
- publishing feedback
- archiving student translations
- adapting didactic strategies to student needs.

The feedback provided by trainers in a training context is intended to give students an idea of (the types of) errors made. However, students seem to feel somewhat in the dark when interpreting the number of errors; they often wonder whether the quality of their translation would be sufficient to obtain a passing grade in a summative context or, perhaps even, how far they are removed from a professional level of translation competence. As a solution to this problem, trainers can treat the assignment as if there is something at stake; they count and weigh errors and calculate a score for each translation. Unfortunately, the calculation of scores usually takes quite a bit of time. In translationQ, points are automatically subtracted from the total score of the assignment. Trainers can publish the automatically calculated scores along with the feedback for each student with a simple click of a button. Published revisions are then stored in the 'User Reports' tab (which can be consulted via the Dashboard, another feature that is built in to foster authentic experiential learning). Through the user reports, the trainer can glean an idea of student performance on specific tasks, and get an overview of student and group progress and error frequency over time (Figure 4).



← User report

1 Sep 2017 - 26 Apr 2018

User reports > Alessia Daniele

[redacted]

E-mail address

[redacted]

Last sign-in

[redacted] 11:41

Translation assignments

6

Assignment scores

Average

92.5%

TRANSLATION ASSIGNMENTS

Search

Name	Date added	Language pair	Revisor	Time spent	Score		
1. Du cacao au chocolat	5 Dec 2017 17:12	fr > it	[redacted]		17.3/20	86.3%	⋮
2. Cacaoyer fruit	9 Dec 2017 11:11	fr > it	[redacted]		18.5/20	92.5%	⋮
4. Cacaoyer cabosses	10 Dec 2017 10:39	fr > it	[redacted]		20/20	100%	⋮
5. Graines fèves de cacao	10 Dec 2017 10:48	fr > it	[redacted]		19.3/20	96.3%	⋮
6. Chocolaterie industrie chocolatière	10 Dec 2017 10:59	fr > it	[redacted]		18/20	90%	⋮

Figure 4 Screenshot of user report

The trainer can also delve deeper into the 'big wrong data', by exporting the data in an Excel sheet and experimenting with pivot tables. By pivoting the data in the tables, insights can be gained and visualised that seem outside the trainer's grasp in traditional revision practices. The tables offer insight into: score distribution per error category (answering the question: does the trainer give the same or other weights to errors within the same category?); general error frequency (answering the question: which are the most frequent errors incurred in training?); group profiles (answering questions like: which error categories are more frequent in a group and require the trainer's special attention?); and trainee profiles (answering key questions like: which are the specific weaknesses of a student?). In the same 'User report' tab, the student too gets a good idea of their performance, as the tool displays error frequency, tasks scores over time and it sets individual scores in opposition to group means. It stands to reason that this formative feature is likely to add a powerful impetus to social-organisational ergonomics. Information yielded through user reports can be used not only to better address the needs, abilities and limitations of the student group, but they also serve to identify the specific needs, abilities and limitations of the individual, steering didactic practices in the desired direction. In terms of social and organisational ergonomics, this effect is highly desirable, as didactics not only inform assessment, but assessment also come to inform didactics.

References

Bowker, L. (2001). Towards a methodology for a corpus-based approach to translation evaluation. *Meta: Journal des Traducteurs*, *46*(2), 345–364.

Holmes, J. S. (1988). *Translated! Papers on Literary Translation and Translation Studies*. Amsterdam: Rodopi.

Hönig, H. G. (1998). Positions, power and practice: Functionalist approaches and translation quality assessment. In C. Schäffner (Ed.), *Translation and Quality* (pp. 6–34). Clevedon: Multilingual Matters.

Saldanha, G., & O'Brien, S. (2014). *Research Methodologies in Translation Studies*. London: Routledge.

TAUS. (2015). *Quality Dashboard White Paper June 2015*. Retrieved from: www.taus.net/insights/reports/taus-quality-dashboard-white-paper

Thelen, M. (2016). Quality and objectivity of testing and assessing in translator training: Dilemmas and suggestions. In M. Thelen, G. van Egdom, D. Verbeeck, & B. Lewandowska-Tomaszczyk (Eds.), *Translation and Meaning* (pp. 79–98). New Series volume 1. Frankfurt am Main: Peter Lang.

Innovations in Language Assessment

Foreign language assessment in virtual exchange – The ASSESSnet project

Anna Czura
Autonomous University of Barcelona

Melinda Dooly
Autonomous University of Barcelona

Abstract

In this article we provide an overview of the ASSESSnet project, funded through a Marie Skłodowska Curie Actions grant by the European Commission Research Executive Agency (H2020-MSCA-IF-2018 845783). The project aims to analyse assessment practices in foreign language courses that involve virtual exchange (VE) and to explore teacher beliefs about the shape and the content of assessment in the virtual contexts where these exchanges take place. The research findings of this in-progress project, illustrated with relevant examples of promising practices, will be used to formulate practical recommendations for improving the type, content and implementation of assessment in VE projects across diverse educational contexts. In the following, we outline the detailed objectives of the project, the methodology used and the projected outcomes. We conclude with a presentation of the project website and an invitation to participate in the project by sharing further examples of good practices with other practitioners.

Introduction

Despite the undeniable benefits it offers and the efforts on the part of the European Union and higher education institutions (HEIs), face-to-face mobility is still available only to a small percentage of students. This trend became even more pronounced in the time of the current pandemic, during which mobility programmes were dramatically reduced or came to a halt, thus depriving students of the opportunity to experience a different educational context, engage in an intercultural dialogue and improve foreign language (FL) skills.

Although it cannot be treated as a substitute for face-to-face mobility, virtual exchange (VE) offers a valuable alternative and presents itself as a valid strategy of promoting internationalisation at home and internationalisation of the curriculum (Jager, Nissen, Helm, Baroni, & Rousset, 2019). VE is understood here as an integral part of a course during which students interact and cooperate, and thus engage in intercultural exchange with peers from partner institutions, by means of online communication tools under the guidance and with the support of teachers or facilitators (Belz, 2003; O'Dowd, 2020; O'Dowd & O'Rourke, 2019). VE is often referred to as telecollaboration, online intercultural exchange or teletandem; however, as Dooly and Vinagre (Forthcoming 2022) point out, these terms are not always considered to be synonyms and each term has emerged from different epistemologies and contexts, and preferences in terminology are often linked to an individual's dynamics and background references. Regardless of the different connotations each of these terms carry, O'Dowd and Dooly (2020, p. 362) indicate that 'they all highlight both the medium (virtual, online, digital, distance, global, networked) and the underlying purpose (exchange, intercultural, collaboration, learning).' Nonetheless, VE has emerged as the preferred term in the EU, especially as both governmental and private institutions increasingly adopt this term.

So far the research has tended to focus on investigating the impact of VE on foreign language competences (Sauro, 2009), intercultural skills (Vogt, 2006), learner autonomy (Fuchs, Hauck, & Müller-Hartmann, 2012) and selected transversal skills (Vinagre, 2010). Given the benefits of VE, a steady interest in implementing such programmes in teacher education courses can be observed (Dooly & Sadler, 2020; Nissen & Kurek, 2020; Pfingsthorn, Czura, Kramer, & Stefl, Forthcoming 2021). However, there is an acute shortage of studies and case studies that explore and elaborate on the topic of assessment in this mode of learning. The 'ASSESSnet: Language assessment in virtual mobility initiatives at tertiary level – teachers' beliefs, practices and perceptions' project aims to fill in this gap in the subject literature by exploring tertiary-level teachers' assessment-related beliefs, perceptions and practices, and formulate practical resources to facilitate the assessment process in VE courses. In this article, before we present the project objectives and the methodology used, we elaborate on the difficulties commonly associated with assessment in VE courses. We conclude with the presentation of the expected outputs, preliminary research results and the overview of the project website.

The ASSESSnet project

The rationale behind the ASSESSnet project

Assessment in VE remains a largely underexplored topic in research despite the fact that it is perceived by many practitioners as one of the most difficult aspects of running such a course (O'Dowd, 2013). The difficulties arise from a number of factors. Firstly, at the institutional level, the success of assessment may depend on whether VE constitutes an integral part of the curriculum, and whether students are granted credit points for their involvement in VE (Cloke, 2010). Assessment is also sociocultural and context-dependent and, thus, largely determined by an educational and institutional context in which it is situated. The perceived importance – by both teacher and learners – of foreign language learning, attitudes to learner autonomy, and perceptions of the importance of assessment and grading may come into play here. Finally, the dynamic and interactive nature of VE projects makes assessment of student learning highly unpredictable (Akiyama, 2014). It is therefore necessary that assessment in VE does not merely replicate the assessment strategies applied in a traditional language classroom, but that its objectives, form and content reflect the intricate and interactive nature of VE projects, which in most cases take part outside the classroom and involve an intercultural component. The research findings of the ASSESSnet project are also of relevance to teachers involved in different forms of distance education, which was abruptly integrated into mainstream education in the time of the Covid-19 pandemic. VE and distance education are not tantamount as they differ, for instance, in terms of objectives, mode and teacher presence; however, they share such distinctive features as the medium of communication and the autonomous nature of the learning process, which essentially necessitate the emphasis on self-reflection, formative assessment and the mediating role of assessment.

Research objectives

The ASSESSnet project is a two-year (2019–21) project carried out as a part of Marie Skłodowska Curie Actions Individual Fellowship (MSCA IF) at the Autonomous University of Barcelona (UAB) by Anna Czura (researcher) and Melinda Dooly (supervisor), with cooperation of the GREIP (Research Center for Plurilingual Teaching & Interaction) research group. The project sets out to explore the planning, design and implementation of the assessment process in VE in FL education, through the compilation of data regarding the assessment objectives, tools and criteria used in existing VE projects. In particular, we focus on the relationship between summative and formative approaches and the use of specific assessment tools (e.g. portfolios, projects, peer-assessment, etc.). Given the interdisciplinary nature of a VE project, the study also aims to identify the content of assessment; establishing which elements of learners' activity and performance are typically subject to assessment (e.g. FL competence, multimedia literacy, intercultural competence, transversal competences). The study also explores the implementation of VE projects with the objective of identifying possible promising practices as well as the challenges the teachers experience at both institutional and classroom levels. At this point, we have collected both quantitative and qualitative data and are now in the process of analysis.

Methodology

In order to collect the quantitative data in this mixed methods study, a questionnaire containing both Likert-type rating scales and open-ended questions was used. This tool focused mainly on teacher assessment beliefs as regards assessment objectives, tools, content and implementation. The questionnaire was available in four languages (English, Spanish, Catalan and Polish) and was completed by 60 participants. The qualitative data was collected by means of in-depth oral interviews conducted with 25 FL teachers in tertiary education, which focused on their assessment practices, instruments and strategies in courses involving elements of VE. The interviews additionally touched upon the factors that affected the implantation of assessment in VE and the evolution of assessment strategies over time. The interviews were transcribed and content analysed by means of NVivo software. This data is supplemented with the analysis of assessment-related resources and documents (e.g. syllabi, assessment rubrics, descriptions of assessment tools) provided by research participants. Due to the international nature of VE projects, the practitioners who took part in the study came from a wide range of educational context in Europe and beyond, which enabled us to explore the diverse approaches to assessment in VE in different institutions and in different countries.

Project outputs

Teaching resources and project website

In further steps, the research findings will be used to formulate practical recommendations concerning the scope and form of assessment in VE in FL courses. These will be available in an open access handbook that will focus on the aspects of good assessment and will present an array of assessment tools and criteria that can be implemented to attend to student learning in VE. The guidebook will be supplemented with a collection of good practices illustrating actual use of assessment across a variety of educational contexts. Depending on the VE project, the good practices delineate the assessment process in the entire course or focus on a specific tool. What is most important is that these examples are produced *by* practitioners *for* practitioners with the objective of facilitating future assessment practices in VE projects. The assessment strategies the teachers use are characterised by a high degree of diversity that is shaped by the local educational context, institutional demands, the set-up of the VE partnership as well as the teachers' personal preferences. Consequently, the examples of good practices are applicable to synchronous, asynchronous or a mixture of these two communication approaches, in settings in which the participation in a VE component is either an integral part of the syllabus or offered as a voluntary activity. Although addressed primarily to tertiary-level teachers, the guidelines will find application also in other educational contexts. The results and the practical implications of the study may be of interest to school authorities and policy makers interested in improving the quality of VE or in introducing this form of learning as an element of internationalisation at home. All the publications presenting research results, the handbook and the examples of good practice will be available open access at the UAB Repository and project website: www.assessnet.site. The website also contains an annotated bibliography of publications and projects that touch upon the topic of assessment in VE and distance learning.

Preliminary results

This is research in-progress and the data analysis is still underway; however, on the basis of the questionnaire and interview data collected and analysed so far, the following general conclusions can be formulated:

- The approaches to assessment are highly diversified across educational contexts and depend to a large extent on whether or not VE is formally incorporated in the curriculum on an institutional level.
- Teacher beliefs are principally oriented towards formative objectives of assessment, with the aim of improving student learning and informing the planning of the teaching process.
- Formative assessment that offers informative feedback on both the process and the product of learning tends to prevail; however, its implementation is to a large extent context-dependent.
- Teachers tend to consider reflective and collaborative approaches as key to assessing students' VE experience.
- The assessment of the intercultural component usually entails students' reflective practice on the VE experience and formative feedback.
- There is a shortage of training opportunities and resources aimed specifically at assessment-related teaching competences.
- Parallel approaches to assessment in all partner institutions facilitate, but are not essential to, the success of the assessment process.

Conclusions

Although assessment is generally recognised as key to the success of VE, some teachers acknowledge their lack of efficient strategies, which is coupled with the shortage of relevant research findings and practical resources that would guide them in the process. It was also observed that the planning and implementation of assessment are characterized by a high degree of variability not only from project to project, but also within one project. To address the urgent need for hands-on resources and training opportunities aimed at these aspects of VE, the ASSESSnet project will include practical resources that are research-based and draw from the experiences from practising teachers across many educational contexts. We would like to take this opportunity to invite teachers experienced in VE to contact us and share the examples of their assessment practices on the project website.

Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 845783.

We would like to express our thanks to our research participants, who in the midst of the pandemic, found the time to complete the questionnaire, answer our questions and share their assessment experiences with us.

References

Akiyama, Y. (2014). Review of issues and potential solutions of Japan-U.S. telecollaboration: From the program coordinator's viewpoint. *Studies in Japanese Language Education*, *11*, 3–14.

Belz, J. A. (2003). Linguistic perspectives on the development of intercultural competence in telecollaboration. *Language Learning & Technology*, *7*(2), 68–99.

Cloke, S. (2010). The Italia-Australia Intercultural Project. In S. Guth & F. Helm (Eds.), *Telecollaboration 2.0: Language, Literacies and Intercultural Learning in the 21st Century* (pp. 375–384). Bern: Peter Lang.

Dooly, M., & Sadler, R. (2020). "If you don't improve, what's the point?" Investigating the impact of a "flipped" online exchange in teacher education. *ReCALL, 32*(1), 4-–24.

Dooly, M., & Vinagre, M. (Forthcoming 2022). Research into practice: Virtual exchange in language teaching and learning. *Language Teaching*.

Fuchs, C., Hauck, M., & Müller-Hartmann, A. (2012). Promoting learner autonomy through multiliteracy skills development in cross-institutional exchanges. *Language Learning & Technology*, *16*(3), 82–102.

Jager, S., Nissen, E., Helm, F., Baroni, A., & Rousset, I. (2019). *Virtual Exchange as Innovative Practice across Europe. Awareness and Use in Higher Education*. Retrieved from: evolve-erasmus.eu/wp-content/uploads/2019/03/Baseline-study-report-Final_Published_Incl_Survey.pdf

Nissen, E., & Kurek, M. (2020). *The Impact of Virtual Exchange on Teachers' Pedagogical Competences and Pedagogical Approach in Higher Education.* Retrieved from: hdl.handle.net/11370/bb89998b-c08b-41f4-aee6-08faf1208433

O'Dowd, R. (2013). Telecollaborative networks in university higher education: Overcoming barriers to integration. *Internet and Higher Education*, *18*, 47–53.

O'Dowd, R., & Dooly, M. (2020). Intercultural communicative competence through telecollaboration and virtual exchange. In J. Jackson (Ed.), *The Routledge Handbook of Language and Intercultural Communication* (2nd ed.) (pp. 361–375). Abingdon: Routledge.

O'Dowd, R., & O'Rourke, B. (2019). New developments in virtual exchange for foreign language education. *Language Learning & Technology*, *23*(3), 1–7.

Pfingsthorn J., Czura A., Kramer, C., & Stefl, M. (Forthcoming 2021). Interculturality and professional identity: Exploring the potential of telecollaboration in foreign language teacher education. In M. Victoria, & C. Sangiamchit (Eds.), *Interculturality and the English Language Classroom.* Basingstoke: Palgrave Macmillan.

Sauro, S. (2009). Computer-mediated corrective feedback and the development of L2 grammar. *Language Learning & Technology, 13*(1), 96–120.

Vinagre, M. (2010). *Teoría y práctica del aprendizaje colaborativo asistido por ordenador*. Madrid: Síntesis.

Vogt, K. (2006). Can you measure attitudinal factors in intercultural communication? Tracing the development of attitudes in e-mail projects. *ReCALL, 18*(2), 153–173.

Digital text types in a computer-based foreign-language test

Katharina Karges
University of Fribourg, Switzerland

Malgorzata Barras
University of Fribourg, Switzerland

Peter Lenz
University of Fribourg, Switzerland

Abstract

With the advent of new technologies and social media, language use has changed dramatically in the last few years. To account for this, we developed scenario-based assessment tasks which use digital text types such as smartphone chats, websites, audio messages and blogs in an elementary-level foreign language test. The present article provides evidence on how adolescent test takers perceived the use of these digital text types and discusses whether the inclusion of these text types may improve the test-takers' motivation and, as a consequence, more accurately reflect their true ability.

Introduction

In the digital age, reading and listening comprehension have gained new dimensions. The nature of reading has changed through hypertexts, the use of search engines, 280-character tweets or online chats. Listening comprehension has become much more present in everyday life with the use of smartphones. Having long since arrived in the everyday lives of many students, these changes have an impact on foreign language teaching in schools and, thus, should be reflected in the assessment of foreign language skills. In the following paper, we present evidence from a research project in which we developed test items using digital text types and investigated the effect this had on the students' perception of the test items.

Context

Reading and listening comprehension in the digital age

Text comprehension, in particular reading comprehension, plays a key role in academic success and social participation (OECD, 2019). As a result, reading comprehension assessment has been researched extensively (Alderson, 2000, p. 1; see also Khalifa & Weir, 2009, p. 35 ff.; Sabatini & O'Reilly, 2013) and is continually evolving. Recent approaches such as the *Reading for Understanding* framework define reading literacy as a purposeful, complex activity which involves 'cognitive, language, and social reasoning skills, knowledge, strategies, and dispositions, directed towards achieving specific reading purposes' (Sabatini, O'Reilly, & Deane, 2013, p. 7).

Within this framework, the authors developed scenario-based assessment (SBA): test items are embedded in scenarios, which are designed to provide test-takers with relevant communicative goals that motivate them to use their skills more completely (Sabatini et al., 2013, p. 29 f.).

The project

In our research project 'Innovative Forms of Assessment' (IFB), we developed computer-based tasks suitable for use in a large-scale foreign language assessment targeting Levels A2 and B1 of the Common European Framework of Reference for Languages (CEFR) (Council of Europe, 2001). These tasks were embedded in scenarios intended to simulate real-life reading and listening comprehension in a foreign language – an adaptation of the SBA concept mentioned above. We investigated the functioning of these tasks in a mixed-methods design using introspective methods (think-aloud protocols and stimulated recall interviews), a series of tests of separate language skills and competences (including vocabulary, speed of language processing, and grammar) as well as questionnaires.

Research questions

The use of digital text types in tests is generally assumed to be authentic and motivating for test-takers (e.g. OECD, 2019; Sabatini et al., 2013). A test which is assumed to be more motivating, in turn, 'may be more reflective of [the students'] true ability than an assessment that is not as engaging' (Sabatini et al., 2013, p. 28 ff.).

To explore the relevance of digital text types for the students, two research questions (RQs) are pursued below:

1. To what extent are digital text types present in students' everyday life?
2. How do students respond to the digital text types in the test items?

RQ1 explores to what extent the use of digital text types is actually 'authentic' for the test-takers. We draw on evidence from two questionnaires that were used as part of the quantitative data collection. To address RQ2, we will present findings from the qualitative part of the research project, which used introspective methods to explore the students' individual perspectives (Heine & Schramm, 2016).

Data collection

Quantitative study

The quantitative study, which plays only a minor role in this article, represented the core of the IFB project. In addition to the scenario-based tasks, 'traditional' reading and listening comprehension tasks as well as vocabulary and grammar tests were developed or adapted to our purposes. In addition, questionnaires on language learning motivation, test-taking strategies and computer use, as well as on the learners' social and linguistic background, were used.

Prior to the main study, all tasks and scenarios were piloted in an extensive pre-operational testing phase (Kenyon & MacGregor, 2012), which included both individual and group interviews, and field tests in classrooms. A total of 631 learners in 39 classes later participated in the main study. At the time of testing, the students attended a Swiss-German lower secondary school (average age: 15 years) and were learning French and English in non-intensive foreign language classes. Each learner completed the entire test battery, either in English or French.

The responses from two questionnaires are of primary interest for this paper. In a background questionnaire, the students were asked about their use of computers, smartphones, and tablets at school and at home, among other things. In a post-test questionnaire, they were asked to comment on what they found particularly easy, difficult or interesting about the test items they had encountered.

Qualitative study

For the qualitative study, 30 students completed a selection of four to five scenario-based tasks while verbalising their thoughts aloud (Bowles, 2010). Following this, they discussed their test responses, their reasoning and their thoughts with a researcher in a stimulated-recall interview (Gass & Mackey, 2017). The individual sessions with the learners were recorded, transcribed and analysed according to the principles of structuring qualitative content analysis as described by Kuckartz (2018). For the purpose of this paper, all passages from the stimulated-recall interviews in which the students commented on the digital text types were identified and coded according to three main categories:

- negative assessment of the digital text type(s)
- positive assessment of the digital text type(s)
- indifference to the (digital) text type(s).

Results

Familiarity with computers and smartphones

Responses to our background questionnaire clearly show that students in our context (n=563) are familiar with smartphones, tablets and computers: only 38 learners (7%) reported that they never used these devices, while a majority of 59% used computers or tablets at least once a day, and in the case of smartphones 91% of the students reported daily use. Interestingly, all devices are mainly used at home: while computers or tablets are often used occasionally (i.e. 'not daily') in school, the smartphone clearly had not found its way into the classrooms at the time of the study: 512 learners (91%) indicated that they never or almost never used it at school.

The students were also asked how much they liked using computers, tablets and smartphones. As expected, many young people agree with the statement that 'using these devices is fun' (494 out of 562 learners, i.e. 88%). About two-thirds also stated that they used these devices with ease. Not nearly all learners, on the other hand, indicated a particular curiosity towards modern technologies: 215 students (38%) agree somewhat or completely with the statement that '[they] are very interested in technology' and only 168 (30%) think they know more about technology than other people their age.

Statements about the test items

Based on the test-takers' responses to the open-ended questions 'What did you find particularly interesting?' and 'What did you find particularly difficult?', we can conclude that certain scenarios were positively remembered by the learners (e.g., the one about dog-walking robots). The same is true for the digital text types, especially the smartphone chats, which were perceived as particularly interesting.

Results of the qualitative study

The data from the qualitative interviews show that the digital texts and test items developed in the IFB project were overall well received by the students. Some criticism arose when individual learners encountered unfamiliar item formats, otherwise unexpected features, or topics they found particularly boring.

Many learners praised the realistic design of the input texts, although Sheldon (a self-chosen pseudonym) considered it 'annoying':

> [. . .] it looks like [it] is on the internet but that's just not really (-) like a test- [. . .] somehow I think it doesn't work that well [626–639][1].

Similarly, students generally liked the smartphone chats serving as input texts, but Bonnie also said:

> partly it is (-) just more difficult to understand because [. . .] in a normal text you have () a clear structure (-) main part final part and so on [318–326].

Nevertheless, most learners responded very positively to the digital text types. Several indicated that these were familiar to them, 'easier' to read (Hector) and that they made you feel more 'at home' (Hector) and 'comfortable' (Howard) during the test. The learner Picard pointed out:

> that's something you have in your own life and so you could (-) imagine (-) [something] with it [889–890].

Some students also mentioned that the digital text types and the attractive, realistic design could help increase their motivation. For example, Arya thought it was 'really cool' that the reading comprehension stimulus was presented in the form of a group chat. In her opinion,

> it just comes across much more em (-) realistic than if there was just a boring text and it also motivates much more if [. . .] you [have] a nice design and you know it from everyday life [. . .] [565–571]

> I have more motivation when I see something and I find that maybe I can still use it later in life [1,374–1,375]

An interesting statement was made by Omega. When the researcher inquired if digital texts that the student had described as 'funny' and 'cool' would motivate him a little more in a testing situation, he responded:

[1] All transcripts were translated from the German originals into English. The numbers at the end of the quotes indicate the lines in the respective transcripts. The pseudonym of the student speaking is either indicated there as well or in the text. Inserts in parentheses indicate pauses. Inserts in square brackets indicate that a part of the transcript was omitted [. . .] or that something was inserted to make the meaning clearer.

> [yes as/ yes] eh no it makes (-) so it's motivation first and then somehow it occurs to you that it's a test ((laughs)) then you think yes it's funny but now I have to focus again. [473–477]

Some students also pointed out that in the end, understanding the text was more important to them than whether or not it looked like a text that might interest them. Ezra, for his part, did not even notice that one of the listening tasks was presented in the form of a podcast, while Steve was 'indifferent' not only to the form of the task but also to French in general.

Finally, the student Sofia insightfully explains the conflict she experiences in a testing situation:

Sofia: I don't think there has ever been a topic that would interest me in a test because then I don't even really think about it [. . .] it's just so (-) stressful [. . .] that I have to find the solutions [. . .].

Researcher: um (-) that means that during (-) a test you do not really pay attention to the topic uh.

Sofia: *yes yes but I can't enjoy it [. . .] even if it would be a topic I would like I am stressed and [. . .] I can't concentrate on the test [. . .] the thing I think about is that I have to find the solutions and then I don't think oh cool [1,504–1,517].

Discussion and conclusion

The primary purpose of this paper was to show how learners perceived the digital text types used in a computer-based foreign language test. With regard to RQ1, it can be assumed that the learners were familiar with the text types used and the handling of the computer since almost all learners regularly use computers, tablets and smartphones.

With regard to RQ2, we could observe that the learners reacted quite positively to the digital text types – both because they are part of their daily life and because they are a novelty in their school life. However, some statements made by the students also suggest that in the end, successfully completing a task is more important to them than the task itself.

Our observations suggest that motivation in a testing situation comes not only from the choice of tasks and texts, but also from the consequences the test will have for a student's life: if the test results are important enough to the test-takers, they may attentively complete any task, even those they would not describe as particularly motivating. Nevertheless, if test items intended to simulate real-life communication allow a more accurate reflection of the test-takers' skills and competences, then the use of authentic text types and the development of scenario-based tasks may be worth the effort.

References

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

Bowles, M. A. (2010). *The Think-aloud Controversy in Second Language Research*. New York: Routledge.

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

Gass, S. M., & Mackey, A. (2017). *Stimulated Recall Methodology in Applied Linguistics and L2 Research* (2nd ed.). New York: Routledge.

Heine, L., & Schramm, K. (2016). Introspektion. In D. Caspari, F. Klippel, M. Legutke, & K. Schramm (Eds.), *Forschungsmethoden in der Fremdsprachendidaktik. Ein Handbuch* (pp. 173–181). Tübingen: Narr Francke Attempto.

Kenyon, D. M., & MacGregor, D. (2012). Pre-operational testing. In G. Fulcher & F. Davidson (Eds.), *The Routledge Handbook of Language Testing* (pp. 295–306). Abingdon: Routledge.

Khalifa, H., & Weir, C. J. (2009). *Examining Reading: Research and Practice in Assessing Second Language Reading*. Studies in Language Testing volume 29. Cambridge: UCLES/Cambridge University Press.

Kuckartz, U. (2018). *Qualitative Inhaltsanalyse: Methoden, Praxis, Computerunterstützung* (4th ed.). Weinheim: Beltz Juventa.

OECD (2019). *PISA 2018 Assessment and Analytical Framework*. Paris: OECD Publishing.

Sabatini, J. P., & O'Reilly, T. (2013). Rationale for a new generation of reading comprehension assessments. In B. Miller, P. McCardle, & R. Long (Eds.), *Teaching Reading and Writing: Improving Instruction and Student Achievement* (pp. 100–111). Baltimore: Brookes Publishing.

Sabatini, J. P., O'Reilly, T., & Deane, P. (2013). *Preliminary Reading Literacy Assessment Framework: Foundation and Rationale for Assessment and System Design*. Research Report RR-13-30. ETS. Retrieved from: www.ets.org/research/policy_research_reports/publications/report/2013/jrmh

The development and delivery of online-proctored speaking exams: The case of LanguageCert International ESOL

Leda Lampropoulou
LanguageCert, United Kingdom

Yiannis Papargyris
LanguageCert, United Kingdom

Abstract

Responding to commercial requests for the online delivery of its exams, LanguageCert embarked on an attempt to develop an online-proctored (OLP) equivalent to its established International English for Speakers of Other Languages (IESOL) Speaking exam suite. To this extent, we considered the practical aspects of replicating the exam in an online environment, the potential need to adjust the content and format of the test, the applicable variants of the exam registration and exam administration processes, security and test integrity issues, as well as any potential impact on the assessment methodology the exam employs. Most importantly, we needed – and still do – to ensure that the accuracy of assessment outcomes is not compromised by the mode of delivery in any way.

Background

In mid-2018, LanguageCert began developing an online equivalent of the face-to-face International English for Speakers of Other Languages (IESOL) Speaking exam. Several factors had led to this decision, the implementation of which eventually coincided with the severe impact the Covid-19 pandemic had on paper-based exam delivery globally; among others (Marshall, Shannon, & Love, 2020), the growing dominance of online communication and the popularity of online learning (Lim & Wang, 2016), both of which increasingly dictate the need for online assessment solutions. Furthermore, LanguageCert, a member of the PeopleCert Group of Companies, already had access to an electronic delivery system, through which PeopleCert has successfully and securely delivered over 390,000 exams in an online-proctored environment, mainly in the fields of Business and IT. The online-proctoring platform, offered by PeopleCert, is manageable in terms of technical competence and equipment required, and familiar and straightforward in terms of navigation and candidate expectations. It is worth noting that LanguageCert had also been delivering its IESOL (Listening, Reading, Writing) exam suite through PeopleCert's proprietary online-proctoring system since the beginning of 2017. On that occasion, it was only a system check and a test integrity check that were required for the effective implementation of the computer-based exam delivery, using the same, already tested and successfully used, online-proctored environment. The IESOL (Speaking) exam suite, though, was a different case, as it both relied on and effectively assessed live human interaction.

Objective

The main objective of this paper is to outline our rationale for the development of the online version of the IESOL Speaking exam, to identify the potential implications for the comparability between the different modes of delivery, and to explain how these were addressed in the context of LanguageCert's test development processes.

The LanguageCert IESOL Speaking exam

The LanguageCert IESOL Speaking exam suite has been available since 2015 in a paper-based format. Each test consists of a short, spoken interview between one candidate and an interlocutor, who manages the interaction and is responsible for recording the session, but does not assess the candidate. Test format is consistent across all CEFR levels, comprising four parts: personal information, situational roleplay, interactive task, and long turn. Prescribed exam duration varies depending on the level of the exam from six (A1) to 17 (C2) minutes.

The transition from face-to-face to online

The majority of studies conducted during the 1980s and 1990s, when computer-based testing was gradually introduced, supported the view that the computer-based testing process was comparable to paper-administered exams in terms of candidate experience, construct validity, scoring, etc. (Burke & Normand, 1987; Levin & Gordon, 1989; Taylor, Kirsch, Jamieson, & Eignor, 1999; Vincino & Moreno, 1988; Wise, Boettcher Barnes, Harvey, & Plake, 1989). LanguageCert's experience of the delivery of the online-proctored IESOL (Listening, Reading, Writing) exam echos findings reported in research such as the above.

The study reported in this paper extends research into online-proctored delivery of tests in that the focus is on the skill of Speaking and on the implications which online communication may bring to the test.

It is important to point out that all of LanguageCert's claims to comparability presented below are based on the decision that both exam versions – face-to-face and online – are delivered by a human interlocutor; the automated delivery of the spoken exam would pose an undisputed limitation to the capturing of any measurable extent of interaction (Galaczi, 2010).

In preparation for the transition from a face-to-face to an online environment, we needed to ensure that the mode of delivery did not exert an impact on the existing assessment arrangements. To this extent, a team of experts reviewed all assessment descriptors and criteria for references to paralinguistic elements and/or other features or interactional competences which might presuppose or rely upon face-to-face communication (e.g., eye contact, body language, gesture, etc.). We eventually concluded that the criteria used in the LanguageCert scales (i.e., Task Fulfilment and Coherence, Accuracy and Range of Grammar, Accuracy and Range of Vocabulary, Pronunciation, Intonation and Fluency) could be applied equally effectively – regardless of the mode of delivery.

In terms of exam content, a review of LanguageCert's Item Bank was conducted, to identify items which might be adversely affected by the mode of delivery. With that in mind, a panel of experts reviewed references to places (e.g., *We are sitting in a restaurant and you want to order a drink*.) and references to actions which presuppose proximity (e.g., *Can you pass me this magazine, please?*). In the same sense, we decided to amend certain instances of the latter, especially certain actions reflecting the practicality of the exam – i.e., instructions to the interlocutor – which may include passing on paper and pencil for notetaking during certain parts of the test. In terms of exam content, very few and minor amendments were undertaken specifically for the online delivery of certain items which denoted proximity between interlocutor and candidate (e.g., *How did you get here?* became *How did you get there?*).

Following the initial stages of the qualification review with the objective of administering it online, we proceeded with a pretesting phase, which would give us the opportunity to investigate a series of issues pertaining to the comparability of the two delivery streams. More specifically, we needed to ensure the comparability of assessment outcomes and to collect feedback from candidates, exam personnel and stakeholders as to the suitability of the proposed solution. The pilot began in mid-March 2019 and lasted two months. Approximately 180 spoken interviews were conducted at CEFR Levels B1–C2. We decided to exclude lower levels (A1, A2) from online delivery, as we felt that the English-medium onboarding process (a series of brief exchanges between the interlocutor and the prospective candidate to ensure that test-takers have set up their systems appropriately and are ready to start the exam, e.g. 'Please show me your desk with your camera') would possibly disadvantage candidates at those levels. In brief, the pilot objectives were the following: i) to ensure error-free test administration (systems check); ii) to identify any potential linguistic limitations during the onboarding process for candidates at B1; iii) to identify any irregular trends in terms of assessment or any indication of irreconcilable lack of comparability.

It was anticipated that online delivery would affect interlocutor conduct in one respect, namely the management of potential sound delays and/or minor connection issues during the interview. As opposed to cases where actual technical support is needed, we are referring here to very minor connection and/or bandwidth issues which do not necessarily interrupt the interview but may, nonetheless, cause a momentary disconnection or delay. On such occasions, interlocutors were advised to actively encourage candidates to ask for repetition, whereas accommodating short delays, pauses and break-ups became part of standard interlocutor training. For instance, to accommodate minor technical issues, interlocutors were advised to repeat a question as many times as requested if there was a break-up in the signal and the candidate requested repetition. In contrast, for a similar request during a face-to-face interview, the interlocutor would be advised to change the question instead of repeating it a third time. Backchanneling is also a technique which interlocutors have been trained to perform differently. In Part Three of the test, for example, where a dialogue takes place between interlocutor and candidate, the former would be instructed to encourage the candidate with backchanneling sounds, such us 'um-hum'. It was noticed that such interjections might, however, reach the candidate a bit later than intended, or cause them to think that the interlocutor might want to start speaking, and interlocutors are no longer encouraged to interact in such a way, but to prioritise clearer turn-taking instead.

Observations and next steps

Upon completion of the pretest analysis, we found convincing evidence that scores were comparable between the face-to-face and the online delivery of the exam. This came as the result of analysing the average scores from candidate performances on the online pretest and comparing them to the averages of candidates participating in the paper-based speaking exams at the respective level. As a next step, the comparability study aims to investigate the effect of the mode of delivery on score accuracy and on interrater reliability.

In parallel to the – ongoing – quantitative analysis of candidate performance, we also conducted a qualitative study with the view to identifying candidate reactions to the online test-taking experience. This consisted of a candidate survey and one focus group consisting of assessment experts. Feedback was overwhelmingly positive both from candidates and stakeholders, who highlighted aspects of practicality and accessibility. A comment which stood out but seemed to be widely shared among survey participants is that the online interface bestows a 'more democratic' power relationship between the candidate and the interlocutor. Another described the online exchange between candidate and interlocutor as 'less intimidating than actually visiting a test centre'.

Candidates – especially those at B1 level – were able to follow the onboarding instructions without any noticeable difficulty. 88% of candidates described the experience as good (51.28%) or excellent (36.90%).

No higher levels of malpractice or cases of compromised exam delivery have been detected since the online delivery of Speaking exams commenced, either. Unsurprisingly, though, as online testing remains something of a predominantly solitary activity for now, malpractice activities seem to revolve around individual test-takers rather than centres, making the occurrence of malpractice qualitatively different to centre-based malpractice that affects several test-takers during a single administration (Draaijer, Jefferies, & Somers, 2017).

Overall, the digitisation and online delivery of existing assessments contribute to the modernisation of the assessment landscape and to its being in step with the developments in communication, technology and learning. It is imperative, nonetheless, that assessment development and certification principles be adhered to, and that the electronic delivery of traditional assessments is comparable and consistent with standard delivery methodologies.

References

Burke, M., & Normand, J. (1987). Computerized psychological testing: Overview and critique. *Professional Psychology: Research and Practice*, *18*, 42–51.

Draaijer, S., Jefferies, A., & Somers, G. (2017). Online proctoring for remote examination: a state of play in higher education in the EU. In E. Ras & Roldan, A. E. G. (Eds.), *Technology Enhanced Assessment - 20th International Conference, TEA 2017, Revised Selected Papers* (pp. 96–108). New York: Springer.

Galaczi, E. D. (2010). Face-to-face and computer-based assessment of speaking: Challenges and opportunities. In L. Araújo (Ed.), *Proceedings of the Computer-based Assessment (CBA) of Foreign Language Speaking Skills* (pp. 29–51). Strasbourg: Publications Office of the European Union.

Levin, T., & Gordon, C. (1989). Effect of gender and computer experience on attitudes toward computers. *Journal of Educational Computing Research, 5*(1), 69–88.

Lim, C. P., & Wang, L. (Eds.). (2016). *Blended Learning for Quality Higher Education: Selected Case Studies on Implementation from Asia-Pacific*. Bangkok: UNESCO Bangkok Office.

Marshall, D. T., Shannon, D. M., & Love, S. M. (2020). How teachers experienced the COVID-19 transition to remote instruction. *Phi Delta Kappan, 102*(3), 46–50.

Taylor, C., Kirsch, I., Jamieson, J. & Eignor, D. (1999). Examining the relationship between computer familiarity and performance on computer-based language tasks. *Language Learning*, *49*(2), 219–274.

Vincino, F., & Moreno, K. (1988). *Test-takers' attitudes toward and acceptance of a computerised adaptive test* [Conference presentation]. Annual meeting of the American Educational Research Association, New Orleans, USA.

Wise, S. L., Boettcher Barnes, L., Harvey A. L. & Plake, B.S. (1989). Effects of computer anxiety and computer experience on the computer-based achievement test performance of college students. *Applied Measurement in Education*, *2*(3), 235–241.

Towards error-based strategies for automatically assessing ESL learners' proficiency

Stefano Bannò
Digital Society/Department of Cognitive Science/Fondazione Bruno Kessler/University of Trento, Italy

Marco Matassoni
Digital Society/Fondazione Bruno Kessler, Italy

Sofia Simakova
University of Trento, Italy

Abstract

In this paper we propose potential strategies for automatically assessing second language proficiency based on the presence of errors only. We used an open-source grammar and spelling check tool to extract errors from the answers of the written section of an Italian English as a second language (ESL) learners' corpus annotated with human scores and we automatically generated the respective correct versions. We found a moderate correlation between the presence of errors and the scores assigned by human experts. As such, we believe that error-rate may be particularly suitable for automatic assessment tools. Therefore, we envisage the use of various state-of-the-art machine learning approaches, aiming at developing useful techniques for both ESL learners and teachers.

Introduction

Automatic evaluation of language proficiency is gaining growing attention and importance since an increasing number of students start to learn English as a second language (ESL) worldwide (Howson, 2013) and therefore this requires the development of advanced techniques for assessing linguistic proficiency objectively.

A common issue in this field is the lack of publicly available data specifically designed and annotated for automatic assessment. Another typical problem is the lack of consistency and coherence in human evaluation, as it frequently relies on proficiency indicators that often have biases and are not clearly generalizable, and therefore not easily transferable into automatic scoring systems (Engelhard, 2002; Zhang, 2013). Although second language proficiency cannot be assessed on the mere basis of the presence of errors in learners' texts, this aspect is highly consistent and plays a major role in language assessment by human experts (James, 2013).

This paper proposes potential strategies for assessing linguistic proficiency based on errors made by native Italian ESL speakers starting from the written section of a corpus (TLT-school corpus) collected in 2016, 2017 and 2018 from students aged between 9 and 16 years in Trentino, an autonomous region in Northern Italy. The errors have been automatically extracted using the spelling and grammar check LanguageTool.[1]

Related work

Theoretical framework

The origins of the field of second language assessment date back to the influential work of Lado (1961), who believed that the problems of learning a new language could 'be predicted as described in most cases by a systematic linguistic comparison of the two language structures' (p. 24), i.e. the native language (L1) and the second language (L2), which is consistent with his

[1] languagetool.org

structuralist perspective of language and contrastive linguistics. Language was taught and assessed as a set of distinct elements, starting from a contrastive analysis of sounds, grammar and finally – but only partially – vocabulary. As a result, errors play an important role in this construct, especially those errors traceable to specific contrasts between a learner's native language and English (Lado, 1957). In response to and in continuation of contrastive analysis, at the end of the 1960s the seminal work of Corder (1967) set the foundation for error analysis and considered the concept of error from a developmental perspective, i.e. errors typical of any learner, independently of their L1, at a particular stage in learning English. In the 1970s, the subsequent fundamental step in language testing and assessment was inspired by the forward-looking work on communicative competence by Hymes (1972). This seminal work was then refined and framed in the so-called communicative approach by Canale (1983) and Canale and Swain (1980) in the 1980s and further by Bachman and Palmer (1996) in the 1990s. According to this approach, language is used to communicate meaning, which encompasses: grammatical knowledge, sociolinguistic competence, and strategic competence. Around the same time, an approach theoretically rooted in the communicative approach started to be developed and was later fixed in the Common European Framework of Reference for Languages (CEFR) (Council of Europe, 2001), which was meant to provide 'a common basis for the elaboration of language syllabuses, curriculum guidelines, examinations, textbooks, etc., across Europe'. Although it might seem that this approach privileges communication at the expense of formal correctness, errors still play a major role in assessing language proficiency, especially in the field of morphosyntax and lexis (Pfingsthorn, 2013). Furthermore, Thewissen (2013) has shown that learner errors can be connected to CEFR proficiency levels and they can be considered as criterial features for each level, together with other linguistic features, as illustrated in Hawkins and Buttery (2010).

State of the art

The roots of the field of automated scoring of language proficiency can be traced back to the work of Page (1968) on automatic essay scoring. His Project Essay Grader (PEG) was a system that evaluated writing skills based only on proxy traits: handwritten texts had to be manually entered into a computer and a scoring algorithm then quantified superficial linguistic features, i.e. essay length, average word length, count of punctuation, count of pronouns and prepositions, etc.

Across the following decades, the field of automated scoring of writing has expanded and improved, and more significant studies have been conducted from the 1990s and early 2000s, as computational techniques and software technology have increased their power (Landauer, 2003). At the same time, the shift from rule-based models to data-driven methods started requiring the use of learner corpora to train models (see Higgins, Ramineni, and Zechner (2015) for an exhaustive overview on the subject). A corpus-based approach is a common feature to two of the most efficient automated scoring programs for essays, i.e. the *erater®*, developed at Educational Testing Service (ETS) (Attali & Burstein, 2006; Burstein, 2003), and the *Intelligent Essay Assessor™*, built at Pearson Knowledge Technologies (Landauer, Laham, & Foltz, 2003).

Despite the growing use of learner corpora, the use of errors as features in the field of automatic assessment of second language proficiency has been sporadic so far. Error-rate is one of the features employed in Yannakoudakis, Briscoe and Medlock (2011) along with lexical, part-of-speech (POS) and syntactic features for automatically assessing English as a Second or Other Language (ESOL) examination scripts, and it was found to be significant for enhancing the overall correlation between true scores and predicted ones. Similarly, errors are a feature investigated in the work of Vajjala (2018), in which spelling and grammar errors are extracted by LanguageTool. In this case, the error-rate feature considered individually was found to have little impact on the classification performance, reaching an accuracy of 51%. Similar experiments were conducted again by Vajjala and Rama (2018) with German, Czech and Italian, including errors as a feature. This work was reproduced by Caines and Buttery (2020) who applied such experiments also to English and Spanish corpora. Recently, the work described by Ballier et al. (2019) has investigated the possibility of predicting CEFR proficiency levels based on manually annotated errors in the EF-Cambridge Open Language Database (EFCAMDAT) corpus. Their classifier based on a random forest model achieved 70% accuracy. Moreover, their study identified certain types of errors, such as punctuation, spelling and verb tense, are characteristic of specific CEFR proficiency levels.

Experimental setting

The corpus: Trentino Language Testing

In Trentino, an autonomous region in northern Italy, large-scale evaluation efforts have been underway for testing L2 linguistic competence of Italian students taking written and spoken proficiency tests in both English and German (Gretter, Matassoni, Banno`, & Falavigna, 2020); two campaigns have been completed in 2016 and 2018. Each campaign involved about 3,000 students ranging from 9 to 16 years old, belonging to four different school grade levels (5, 8, 10, 11) and three CEFR levels (A1, A2, B1). Since we are planning on conducting our experiments only on the English written part of the corpus, we will not describe the section concerning the spoken utterances and the German section, as their analysis goes beyond the scope of this paper.

Written production and scores

Each answer received scores assigned by human experts, according to different indicators. For our initial experiments, we focus on the data collected in 2016, considering as total score the sum of six indicators (see Table 1).

Table 1: List of the indicators used by human experts to evaluate specific linguistic competences

Lexical richness
Pronunciation and fluency
Syntactical correctness: morpho-syntactical correctness, orthography and punctuation
Fulfillment on delivery: relevancy of the answer with respect to the prompt
Coherence and cohesion
Communicative, descriptive, narrative skills

The considered data consists of about 4,000 answers to five question prompts provided in written form and divided according to CEFR levels. Four question prompts are formulated with slight differences (e.g. changes in setting, place, objects, people) in order to avoid cheating. As for the A1 level, the question prompt asks the test-taker to send a postcard to a friend and tell them about the place they are visiting and about what they like or do not like on their holidays. In the A2 level test, students are asked two questions: one requiring a reply to an email from a friend who is inviting them to their house, the other one asking to write an email to a pen pal who lives abroad and to describe their classmates. In the B2 section, test-takers are asked two questions. The first one requires them to write a blog entry in which they have to describe what happened during the day and to talk about their plans for the rest of the week, while the second one asks them to write an email to a friend who broke an object (a pair of headphones, a hard disk, a game console or a sportcam) borrowed from them.

It is worth mentioning that some written answers are characterized by a number of issues: presence of words belonging to multiple languages, or off-topic answers (e.g. containing jokes, comments not related to the questions, etc.). Nonetheless, we decided not to eliminate these answers from the data.

Error extraction and analysis

We extracted errors from the data using the open-source spelling and grammar check LanguageTool, thus generating more than 200 error labels, mostly ascribable to grammar and spelling and partly to style and collocation. We found a moderate correlation between the number of errors per words and the human scores, especially as regards A1 level (see Figure 1), reaching a Pearson correlation coefficient of about 0.50. In other words, the error-rate tends to be higher in the answers scored with lower grades.

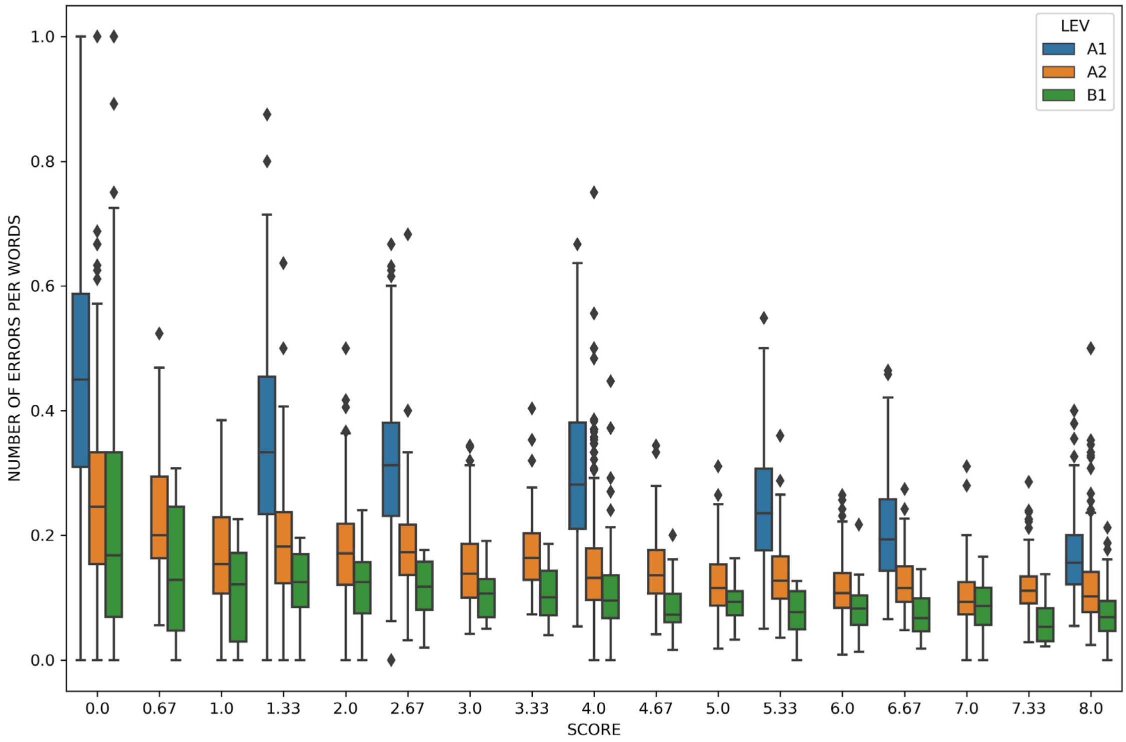


Figure 1 Scatterplot graph of correlation between scores and number of errors per words

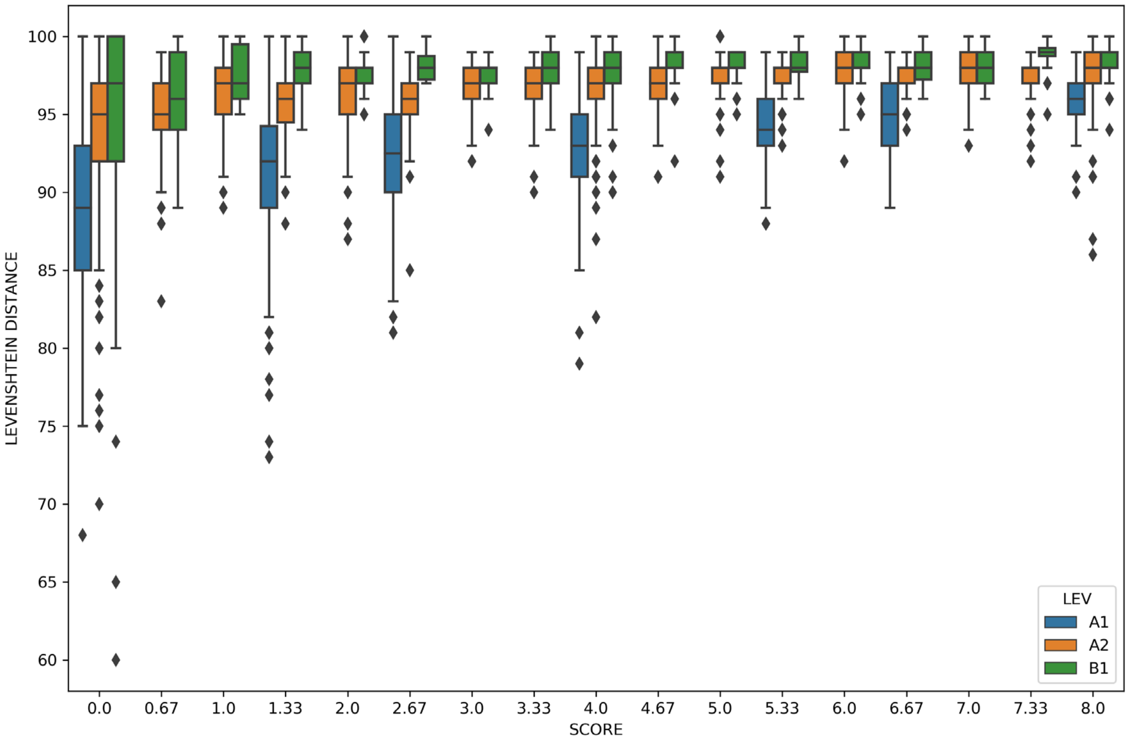


Figure 2 Scatterplot graph of correlation between scores and Levenshtein distance ratio between original and corrected answers

Using the tool, we also obtained the corrected version of each answer. Subsequently, we calculated the Levenshtein distance between the original answers and the ones corrected by LanguageTool in order to quantify the difference between the two strings, and plotted it against the scores. Also in this case we found a rather interesting correlation, as shown in Figure 2, obtaining a Pearson correlation coefficient of about 0.50.

Discussion

In light of the above analyses, although errors are not the only criteria taken into account in the phase of assessment, we confirm the insight that errors show a moderate correlation with the scores assigned by human experts, especially for lower proficiency levels reaching a Pearson correlation coefficient of about 0.50, which might increase if errors are evaluated in a weighted rather than linear manner. Due to its consistency, we believe that error-rate lends itself particularly well to automatic tools. Hence, we propose the employment of machine learning techniques in order to automatically learn such correlation. We envisage various strategies for automatically assessing ESL learners' written proficiency. Firstly, the use of error-rate as a feature may give rather significant results both if used as the only feature, as illustrated by Ballier et al. (2019), or if used in combination with other linguistic features, as shown in Yannakoudakis et al. (2011), but unlike such studies – in which shallow learning algorithms were employed – we propose to investigate the application of deep neural networks. Secondly, the use of two input texts – one containing answers with errors, the other one containing the corrected versions of the answers – may also result in an efficient approach. With regard to this strategy, we have carried preliminary experiments testing different pre-trained models involving BERT (Devlin, Chang, Lee, & Toutanova, 2018) and other BERT-derived models i.e. DistilBERT (Sanh, Debut, Chaumond, & Wolf, 2020) and ALBERT (Lan et al., 2019), and we obtained promising results on the TLT-school corpus. Specifically, our models take two sequences of token embeddings, i.e. of the answers provided by the test-takers and of the answers corrected by LanguageTool, as inputs and predict the total score of each answer normalized on a scale from 0 to 1. We look forward to applying this approach to other publicly available datasets.

Despite the goodness of these preliminary results, we are aware of the limitations of these error-based strategies, with regards specifically to higher-level aspects, such as semantic and pragmatic errors. Other issues may arise when applying such approaches to spoken texts. In this case, the selection of errors should be properly filtered and tailored to the specificities of speech.

Conclusions and future works

In the present paper, we have focused on the correlation between the presence of errors and the scores assigned by human experts in the written section of the TLT-school, a corpus of young Italian ESL learners. We have automatically extracted the errors and generated the correct version of the learners' answers by means of an open-source grammar and spell check. Since we have found a moderate correlation between errors and human scores, we have proposed the use of error-based automatic techniques for assessing proficiency, specifically some state-of-the-art deep learning approaches.

Further work should be undertaken in terms of error classification in order to test the impact of certain categories of errors on assessment (e.g. errors of spelling, grammar, collocation, style, etc.), also considering typical errors due to interference from learners' L1 and to their CEFR proficiency level. Moreover, we envisage the potential application of an error-based scoring system to spoken answers, starting from a preliminary work based on speech transcriptions.

Finally, we acknowledge that the presence of errors cannot be the only feature to be taken into account when evaluating second language proficiency but, if properly weighted and balanced with other proficiency indicators, it might improve consistency and objectivity in assessment.

References

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater v.2. *Journal of Technology, Learning and Assessment*, *4*(3), 1–31.

Bachman, L. F., & Palmer, A. (1996). *Language Testing in Practice: Designing and Developing Useful Language Tests*. Oxford: Oxford University Press.

Ballier, N., Gaillat, T., Simpkin, A., Stearns, B., Bouye´, M., & Zarrouk, M. (2019). *A supervised learning model for the automatic assessment of language levels based on learner errors*. Retrieved from: hal.univ-rennes2.fr/hal-02496688/document

Burstein, J. (2003). The e-rater scoring engine: Automated essay scoring with natural language processing. In M. Shermis & J. Burstein (Eds.), *Automated Essay Scoring: A Cross-disciplinary Perspective* (pp. 113–122). New York: Routledge.

Caines, A., & Buttery, P. (2020). REPROLANG 2020: Automatic proficiency scoring of Czech, English, German, Italian, and Spanish learner essays. In European Language Association (Ed.), *Proceedings of the 12th Conference on Language Resources and Evaluation* (pp. 5,614–5,623). Marseille: European Language Association.

Canale, M. (1983). On some dimensions of language proficiency. In J. Oller (Ed.), *Issues in Language Testing Research* (pp. 33–42). New York: Newbury House Publishers.

Canale, M., & Swain, M. (1980). Theoretical bases for communicative approaches to second language teaching and testing. *Applied Linguistics*, *1*(1), 1–47.

Corder, S. P. (1967). The significance of learners' errors. *International Review of Applied Linguistics in Language Teaching*, *V*(1), 161–170.

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

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. Retrieved from: arXiv:1810.04805

Engelhard, G. (2002). Monitoring raters in performance assessments. In G. Tindal & T. M. Haladyna (Eds.), *Large-scale Assessment Programs for All Students: Validity, Technical Adequacy and Implementation* (pp. 261–287). Abingdon: Routledge.

Gretter, R., Matassoni, M., Banno`, S., & Falavigna, D. (2020). *TLT-school: a corpus of non native children speech*. Retrieved from: arxiv.org/pdf/2001.08051v1.pdf

Hawkins, J., & Buttery, P. (2010). Criterial features in learner corpora: Theories and illustrations. *English Profile Journal*, *1*(1), 1–23.

Higgins, D., Ramineni, C., & Zechner, K. (2015). Learner corpora and automated scoring. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge Handbook of Learner Corpus Research* (pp. 587–604). Cambridge: Cambridge University Press.

Howson, P. (2013). *The English Effect*. London: British Council.

Hymes, D. (1972). On communicative competence. In J. Pride & J. Holmes (Eds.), *Sociolinguistics: Selected Readings* (pp. 269–293). Harmondsworth: Penguin.

James, C. (2013). *Errors in Language Learning and Use: Exploring Error Analysis*. London: Routledge.

Lado, R. (1957). *Linguistics Across Cultures*. Ann Arbor: University of Michigan Press.

Lado, R. (1961). *Language Testing: The Construction and Use of Foreign Language Tests*. Bristol: Longman.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2019). *ALBERT: A Lite BERT for Self-supervised Learning of Language Representations*. Retrieved from: arXiv:1909.11942

Landauer, T. K. (2003). Automatic essay assessment. *Assessment in Education: Principles, Policy and Practice*, *10*(3), 295–308.

Landauer, T. K., Laham, D., & Foltz, P. W. (2003). Automated scoring and annotation of essays with the Intelligent Essay Assessor™. In M. Shermis & J. Burstein (Eds.), *Automated Essay Scoring: A Cross-disciplinary Perspective* (pp. 87–112). New York: Routledge.

Page, E. (1968). The use of the computer in analyzing student essays. *International Review of Education*, *14*(2), 210–225.

Pfingsthorn, J. (2013). *Variability in Learner Errors as a Reflection of the CLT Paradigm Shift*. Frankfurt am Main: Peter Lang.

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2020). *DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter*. Retrieved from: arXiv:1910.01108

Thewissen, J. (2013). Capturing L2 accuracy developmental patterns: Insights from an error-tagged EFL learner corpus. *The Modern Language Journal*, *97*(1), 77–101.

Vajjala, S. (2018). Automated assessment of non-native learner essays: Investigating the role of linguistic features. *International Journal of Artificial Intelligence in Education*, *28*, 79–105.

Vajjala, S., & Rama, T. (2018). *Experiments with universal CEFR classification*. Retrieved from: core.ac.uk/download/pdf/212816462.pdf

Yannakoudakis, H., Briscoe, T., & Medlock, B. (2011). A new dataset and method for automatically grading ESOL texts. In Association for Computational Linguistics (Ed.), *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (pp. 180–189) Portland: Association for Computational Linguistics.

Zhang, M. (2013). Contrasting automated and human scoring of essays. *R&D Connections*, *21*, 1–11.

Piloting an automatic clustering algorithm to supplement placement test results for large groups of students

Marie-Pierre Jouannaud
Université Grenoble Alpes, France

Sylvain Coulange
Université Grenoble Alpes, France

Anne-Cécile Perret
Université Grenoble Alpes, France

Abstract

We report on the initial piloting of an online application using a co-clustering algorithm to supplement the results of a placement test (SELF) developed at Université Grenoble Alpes in six languages, and used at a number of partner universities in France. Automatic clustering models aim to group together similar objects (in our case, test-takers) according to certain variables (test items), and do this without supervision. Our co-clustering algorithm groups test-takers and items simultaneously by identifying groups of test-takers who answered similarly to groups of items (a first step toward learner profiles). The application, developed in R, provides a graphic interface enabling users to visually compare SELF and co-clustering results. It is possible to set the number of groups desired, which might be useful when many students receive the same test placement results but institutions want to make finer-grained groupings.

Introduction

The SELF placement test is a semi-adaptive multi-stage test developed at Université Grenoble Alpes in six languages (English, French as a Foreign Language, Italian, Japanese, Mandarin Chinese and Spanish) and used at a number of partner universities in France. The first stage of the test (the initial testlet) is common to all test-takers, but the items in the second stage depend on test-takers' results in the first. Results in the second stage are used to refine the estimation of learners' level and arrive at placement results expressed in Common European Framework of Reference for Languages (CEFR, Council of Europe, 2001) levels that are as reliable as possible. SELF was designed and developed following Association of Language Testers in Europe (ALTE) guidelines (2011): needs analysis, construct definition, choice of reference level descriptors, test and item specifications, item writing, reviewing and piloting, pretesting, standard setting, final test assembly, and post-administration analyses. Currently, SELF only provides test users with a CEFR 'aggregate level' (corresponding to proposed course level enrollment), and a CEFR level in three macro skills: listening comprehension (L), reading comprehension (R), and 'limited' writing (W), but does not provide further diagnostic information about test-takers. Our goal is to explore the use of automatic clustering to identify subgroups of learners, or to discover learner profiles, which could then be used to enrich the feedback given to learners and help them (and their teachers) decide what skills or areas they need to work on.

Co-clustering models

Automatic clustering models aim to group together similar objects (in our case, test-takers) according to certain variables (test items), and do this without supervision. The co-clustering algorithm we are using, derived from Latent Block Modeling or LBM (Brault & Mariadassou, 2015), groups test-takers and items simultaneously by identifying groups of test-takers who answered similarly to groups of items (right or wrong, since our items are dichotomous). LBM is especially suited to our needs because, being derived from mixture models, it creates homogeneous, non-overlapping groups (Brault & Lomet, 2015). Our application, developed in R, provides a graphic interface enabling users to visually compare SELF and co-clustering results. It is possible to set the number of groups desired, which might be useful when many students receive the same test placement results

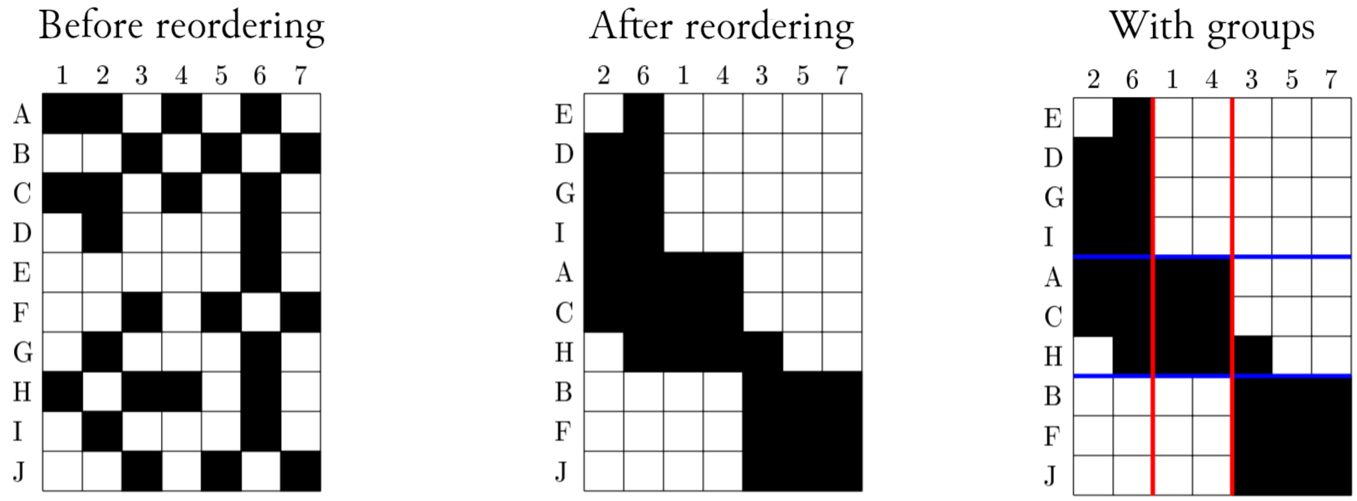


Figure 1 Three matrixes divided into 7 columns and 10 rows to group students according to item response patterns

but institutions want to make finer-grained groupings. Lastly, the algorithm identifies items whose results do not contribute significantly to the classification of test-takers (a so-called 'noise cluster').

In SELF, item levels were determined by a panel of experts in a standard-setting session convened before the test was originally assembled, which means that the results are interpretable in terms of CEFR levels and are thus directly useful to stakeholders (students, teachers, as well as administrative staff, for managing groups, scheduling, etc.). However, the result is ultimately based on total score, and does not distinguish between two students who received the same score but had different patterns of responses (i.e. did not answer the same items correctly). The co-clustering algorithm works in the opposite way: it only looks at patterns of responses and tries to identify groupings based on these patterns (Figure 1). Since it is unsupervised, no meaning is attached to items beforehand. Our objective is to see whether it is possible to make sense of these automatic groupings and whether they can be meaningfully interpreted in terms of learner profiles for diagnostic or placement purposes. In our example (the far right matrix in Figure 1), students A, C and H (lines) did well on items 1 and 4 (columns), but not on 3, 5 and 7, and students B, J and F responded in the opposite way. The question is whether we can identify what items 1 and 4 (or 3, 5, and 7) have in common, and whether we can use them to characterize the difference between students A, C, H on the one hand, and B, J, F on the other.

Initial results

Comparison of placement test and co-clustering results for student groupings

Because the algorithm has not been used with language test results before, the initial step in our analysis is a simple comparison of SELF placement results and the co-clustering algorithm in two groups of students (studying English or Japanese). We only used SELF results in the initial testlet, which separates test-takers into three groups (for the English test, A1/A2, B1 and B2/C1, B1 being the most common level observed in incoming students; for the Japanese test, A1/A2, A2/B1 and B1/B2), and we set the number of groups desired to three in the application.

Figure 2 shows (part of) the output of the application, with SELF results (here, for Japanese students) on the left-hand side, and co-clustering reordering on the right. The white squares correspond to right answers, and the dark ones to incorrect answers (all of the students answered all of the questions). The red/orange/green color coding corresponds to student groups according to SELF results in the initial testlet. We observe that, according to the co-clustering algorithm (on the right), some of the 'intermediate' (orange) students in SELF are more similar to 'advanced' (green) students than to other intermediate students, and are thus placed in the same group. The group of 'beginner' (red) students is essentially the same in SELF and with co-clustering.

For the Japanese as a foreign language cohort (n=101), the correlation between the two grouping methods (SELF and co-clustering) is 0,67 (Kendall's tau rank correlation coefficient), and in English (n=228), $\tau = 0,82$ (in both cases, $p \leftarrow .000$). We find high correlations for both languages, which is not surprising, given that the application uses patterns of successful responses and thus indirectly takes the total score into account. We feel that this validates the use of the co-clustering algorithm, which is capable of arriving at similar results as the placement test when the number of groups is the same, i.e. it can classify students into groups that are interpretable in terms of language level. We find similar results when we increase the number of groups desired.

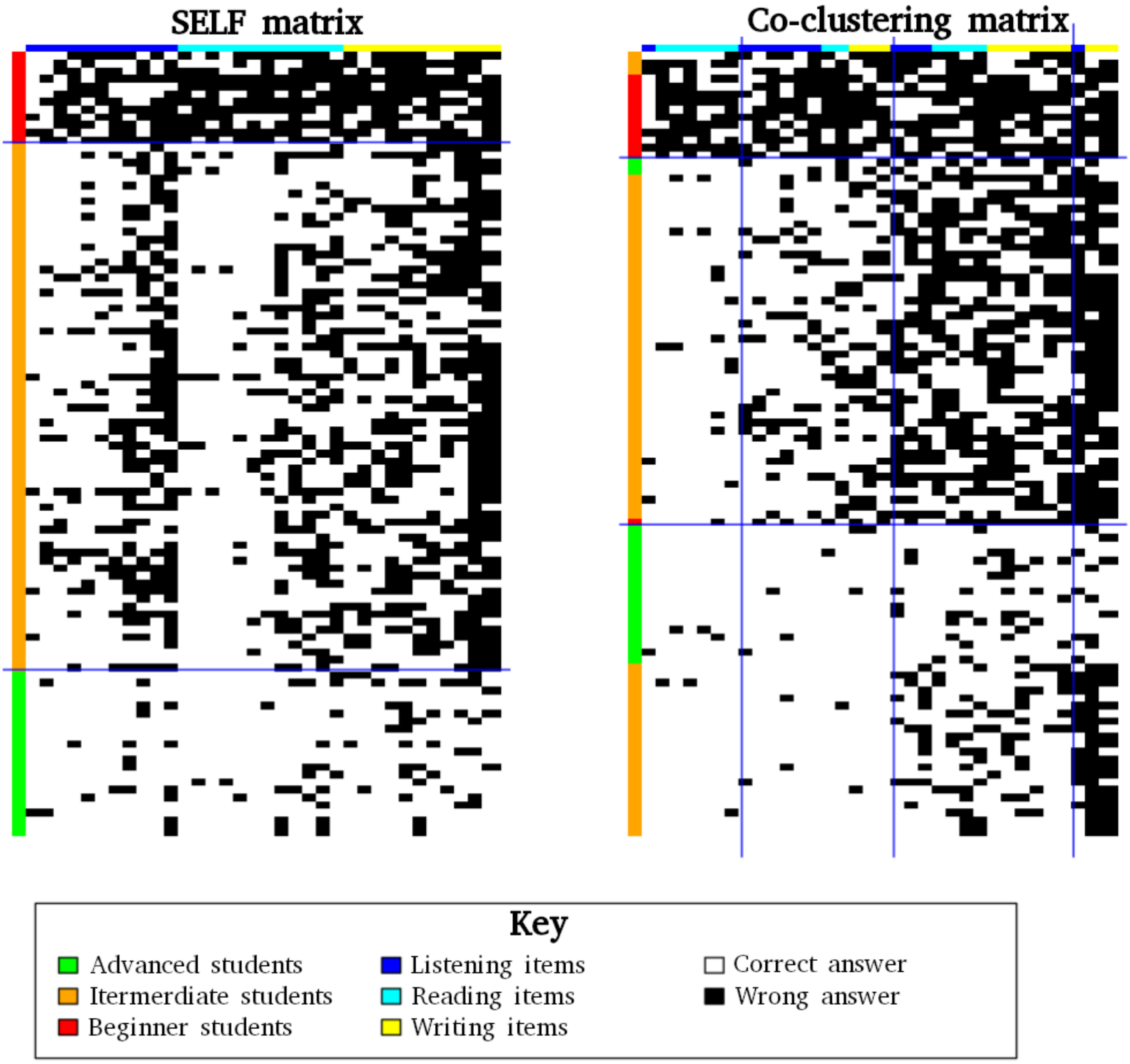


Figure 2 Screenshot of output of the application, with SELF results on the left, and co-clustering reordering on the right

Interpretability of item subgroups

Our original goal is not simply to use the co-clustering algorithm to group students according to levels (since we can already do that with our placement test), but to create more subgroups and/or to gain deeper insight into the characteristics of students placed in the same subgroup. In order to do this, we can analyze the item subgroups with the algorithm used to create the student groupings.

In the example above (Figure 2, right hand side), it is difficult to identify commonalities between items in each item subgroup. The first subgroup (first group of columns) is mostly composed of items targeting reading (light blue), with one listening item (dark blue). This first group might be said to contain 'comprehension' items, and could be used to characterize students according to their receptive skills (regardless of their results in productive skills). The next two item subgroups, however, are composed of items targeting all three language skills included in the test (listening, reading, and writing, in dark blue, light blue and yellow, respectively). The main difference between these two subgroups of items seems to be item difficulty, with the third group containing more difficult items than the second (as can be seen by the greater number of dark squares in the columns of the third item subgroup). The last group of items contains listening and writing items (two skills that do not necessarily have much in common) which are all difficult, as can be seen by the large number of dark squares in the last column, indicating that most students failed to answer correctly.

Thus, targeted language skill does not seem to be a relevant variable (over and above item difficulty) to interpret item groupings and define learner profiles according to these groupings. We are exploring the role of other item characteristics such as language focus (the critical information that item writers believe test-takers need to understand in order to answer the question correctly, which can be lexical, morphosyntactic or pragmatic), discourse type (the prevalent genre of the text the item bears on: narrative, informative, argumentative, etc.), and other characteristics laid out in item specifications, to see if these play a larger role in determining item groupings and students' response to them. Although the method is very different, the goal is similar to what cognitive diagnosis assessment (CDA) approaches have tried to accomplish (Liu, 2015): using results of large-scale tests, and characteristics of the items included in these tests, to provide diagnostic information to learners beyond general language results.

Conclusion and further study

SELF is currently used by more than 25 French universities and language centers, and more than 150,000 students have taken the test in one (or more) of its six foreign languages since it became operational in 2016. Data from administration to two groups of students (tested in Japanese and English) were used to explore the use of unsupervised co-clustering models to automatically create student groups based on their patterns of responses to groups of items, in an effort to automatically uncover learner profiles. We have shown that test-taker subgroupings by the co-clustering algorithm are interpretable in terms of language level and are very similar to SELF results based on CEFR levels. Item groupings, on the other hand, are interpretable in terms of item difficulty, but cannot at present be easily used to give finer-grained information about learner profiles.

These results are only preliminary, and we are exploring avenues for further study. One is the use of more item characteristics to try to interpret item subgroups created by the algorithm (test-taker characteristics could also be used to enrich the interpretation of test-taker subgroups). Another avenue is the analysis of items identified as 'noise' by the co-clustering algorithm (items that do not help in the definition of learner groups): are these items also characterized by lower discrimination in more traditional analyses (classical test theory)? Lastly, the algorithm in its present form does not respond well to missing data, which is why we have only used results to the first stage of our multistage test (the initial testlet), completed by all test-takers. We are working on integrating results to the second stage to enrich the data the co-clustering algorithm has access to.

References

Association of Language Testers in Europe. (2011). *Manual for Language Test Development and Examining*. Strasbourg: Language Policy Division, Council of Europe.

Brault, V., & Lomet, A. (2015). Revue des méthodes pour la classification jointe des lignes et des colonnes d'un tableau. *Journal de la Société Française de Statistique, 156*(3), 27–51.

Brault, V., & Mariadassou, M. (2015). Co-clustering through Latent Bloc Model: A review. *Journal de la Société Française de Statistique*, 156(3), 95–119.

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

Liu, H. H-T. (2015). The conceptualization and operationalization of diagnostic testing in second and foreign language assessment. *Working Papers in TESOL and Applied Linguistics*, *14*(1), 1–12.

Enhancing the classification of group-level diagnostic results from cognitive diagnostic assessment: Application of multi-CDM and cluster analysis

Du Wenbo
School of Foreign Studies, Xi'an Jiaotong University, Xi'an, Shaanxi, China

Ma Xiaomei
School of Foreign Studies, Xi'an Jiaotong University, Xi'an, Shaanxi, China

Abstract

Cognitive diagnostic assessment (CDA) has recently aroused an increasing interest in various domains of language assessment. The focus, however, lies too much in individual-level diagnosis with very little concern about learners' group-level performance. Additionally, the classification of group-level results obtained from most CDA studies with a single cognitive diagnostic model (CDM) typically results in a 'flat pattern', which does not bring much more benefit than the traditional total score. The present study attempts to apply multi-CDM along with cluster analysis to enhance the group-level classification results extracted from a sample of 740 EFL learners' reading performances on a diagnostic reading test. The results show that the multi-CDM demonstrates the best model-data fit over single CDMs. The subsequent diagnostic information was clustered by simple K-means and Expectation Maximization (EM) algorithms, with the latter yielding a more appropriate and interpretable classification. The findings provide a constructive and practical basis for instructors to refine EFL reading curricula and instructional plans based on the enhanced group-level results.

Introduction

As an innovative measurement theory, cognitive diagnostic assessment (CDA) aims at measuring specific knowledge structures and processing skills in learners in order to provide formative diagnostic feedback pertaining to their strengths and weaknesses in a variety of disciplines (Leighton & Gierl, 2007). CDA breaks down the knowledge and skills of a particular domain into various subskills and strategies that learners might need to complete a given task, referred to as *cognitive attributes.* With reference to theories of cognition and test construct, these attributes are embedded in test items represented in a *Q-matrix*, which can be expressed with 0/1, indicating an item not requiring/requiring an attribute (Rupp, Templin, & Henson, 2010). To extract diagnostic information, a series of multivariate discrete latent variable models, generally referred to as *cognitive diagnostic models* (*CDMs*), have been developed (de la Torre, 2011). These psychometric models can identify learners' mastery status of the tested cognitive attributes by generating fine-grained skill mastery profiles (SMP) for each individual learner. Due to its advantage over traditional tests, CDA has been widely applied in the field of education, psychometrics, and language assessment such as reading (e.g., Du & Ma, 2021; Kim, 2015, etc.), listening (e.g., Lee & Sawaki, 2009), and writing (Kim, 2011; Xie, 2017).

However, most of the extant CDA studies (e.g., Jang, 2009; Lee & Sawaki, 2009; Ravand, 2016; etc.) put much emphasis on individual-level diagnostic purpose. Only a few studies (e.g., Du & Ma, 2021; Kim, 2015) sought to investigate learners' group-level performance. Even though personalized diagnostic feedback is the unique feature that differentiates CDA from other assessment tools, group-level diagnostic results cannot be overlooked as they may be more practical in a classroom setting. Furthermore, the single CDM method in most previous CDA studies (e.g., Li, Hunter, & Lei, 2016; Ravand & Robitzsch, 2018, etc.) tends to classify learners into a 'flat pattern', that is, the mastery/non-mastery of all attributes accounts for the two highest proportions. In that case, the group-level classification results do not bring much more benefit than the total score. Therefore, new methods of group classification in CDA are definitely required.

Literature review

Representative CDMs

As a core component of CDA, CDMs have different assumptions regarding the probability of a correct response to an item with the joint impact of different parameters, including guessing, slipping and the main and/or interaction effects of the required attributes for an item (Du & Ma, 2021). In terms of model complexity, CDMs can be classified as saturated and reduced models. Saturated models, such as Generalized deterministic-input, noisy-and-gate model (G-DINA), encompass more parameters than reduced models, taking both main and interaction effects into consideration, and thus are complex to interpret and require a larger sample size to yield accurate estimates (Li et al., 2016). Reduced models normally require a smaller sample size and are easy to interpret (Rojas, de la Torre, & Olea, 2012). CDMs could be further categorized into compensatory or non-compensatory models. For compensatory models, such as Deterministic-input, noisy-or-gate model (DINO) and Additive CDM (ACDM), mastery of one attribute can compensate for non-mastery of other attributes required to answer an item correctly. Conversely, for non-compensatory models, such as Deterministic-input, noisy-and-gate model (DINA) and Reduced reparameterized unified model (RRUM), non-mastery of certain attributes cannot be compensated for by other mastered attributes. Multi-CDM is a combination of different single CDMs (either saturated or reduced models) which aims at choosing the best-fitting model at item level. Compared to single CDMs, multi-CDM has the potential to enhance the accuracy of the diagnostic results and the interpretation of the inter-skill relationship, which might be more suitable for a language domain with a complex nature (Du & Ma, 2021; Ravand & Robitzsch, 2018).

The 'flat pattern' issue in CDA-based reading studies

As stated earlier, most previous CDA-based reading studies applied only one predetermined CDM. The 'flat pattern' issue is prevalent in some of these applications. For instance, Chen and Chen (2015) adopted G-DINA to diagnose 1,029 British L2 learners' performance on Programme for International Student Assessment (PISA) English reading items and successfully classified students into 32 skill mastery patterns. Yet the pattern '00000', i.e., non-mastery of all attributes, took up 15.69% of the classification, and the pattern '11111', i.e., mastery of all attributes, accounts for 24.99%, the two highest among all. Li et al. (2016) obtained similar classification results through the estimation of five different CDMs (G-DINA, RRUM, ACDM, DINA and DINO) on the Michigan English Language Assessment Battery (MELAB) reading test. Over 40% of the test-takers had been classified into non-mastery of all attributes by four of the five models, while the masters of all attributes ranged from 16.06% to 29.83%. Li et al. (2016) claimed that DINA and DINO might be too restrictive for a reading comprehension test, while ACDM had a close affinity to the performance of G-DINA. However, the practicality of the classification results is worth further elaboration. Not surprisingly, the flat pattern can also be found in other single CDM studies (e.g., Javidanmehr & Anani Sarab, 2019; Ravand & Robitzsch, 2018). The question arises as to whether a single CDM could yield accurate diagnostic results even with good model-data fit. Is there any other classification method which could be implemented to make up for the deficiency of the current CDA approach?

Research purpose and questions

To fill the above gap, the present study turns to the multi-CDM method which has been addressed in Ravand and Robitzsch (2018) and found its practical use in Du and Ma (2021). As an extension of Du and Ma (2021), this study also manages to involve cluster analysis in the hope of achieving more appropriate and interpretable group-level classification results.

Two research questions are addressed:

1. To what extent does multi-CDM better estimate group-level diagnostic results?
2. To what extent does cluster analysis enhance group-level classification results?

Method

Data description

The test response data was from Du and Ma (2021), which contained a sample of 740 college freshmen's performance on a diagnostic reading test designed by the PELDiaG (Personalized English Learning: Diagnosis & Guidance) research team. The online diagnostic system and the reading comprehension test are available at: 202.117.216.242/english2.0/jsp/. Eight reading attributes were identified and embedded in the 40 test items, including A1, Understanding Sentence Literal Meaning; A2, Understanding Discourse Literal Meaning; A3, Deducing Word Meaning; A4, Contextual Inference; A5, Elaborative Inference; A6, Synthesizing and Summarizing; A7, Locating Relevant Information; and A8, Eliminating Alternative Choices. They were applied to construct two Q-matrices for this study, one with seven experts' coding (E-Qmat), and the other with nine students' verbal reporting (S-Qmat). The two provisional Q-matrices were then estimated by the real dataset and modified into a best-fitting Q-matrix tagged as R-Qmat for the present study. The detailed construction and modification process can be seen in Du and Ma (2021).

Data analysis

The data analysis of CDM estimation was conducted with R-package CDM Version 6.3-45 (Robitzsch, Kiefer, George, & Uenlue, 2018) and the cluster analysis was completed in Weka (Version 3.8.5, sourced from waikato.github.io/weka-wiki/). First, for the construction of the multi-CDM, we fitted the R-Qmat to the data with the Wald test implemented to objectively choose the best-fitting model for each item (de la Torre & Lee, 2013). Then, the models selected for each item constituted the multi-CDM and were fitted to the respective items. By doing so, the multi-CDM in this study consisted of five different single CDMs, including G-DINA, DINO, RRUM, DINA and ACDM. To make a comparison, the R-Qmat was also fitted with five single CDMs, whose fit statistics were compared with that of multi-CDM. If the multi-CDM demonstrated best model-data fit, it would be applied to extract group-level diagnostic results. Once the diagnostic results were obtained, two algorithms, simple K-means and Expectation Maximization (EM) method, were adopted to classify test-takers into different clusters. Based on a parsimony principle, the algorithm with more interpretable results theoretically and practically was retained to be analyzed. The characteristics of the selected group-level classification results from cluster analysis were then compared with that of the CDA approach.

Results

Group-level diagnostic results by multi-CDM

The fit statistics shows that the multi-CDM performs best in contrast to its rival models. For absolute fit indices, it has a non-significant value of p-MX^2(.210), indicating good model-data fit. The values of other absolute fit indices all demonstrate the lowest ones (i.e., MADcor, .033; SRMSR, .041; MADQ3, .035), proving its enhanced estimation accuracy. The relative fit indices of the multi-CDM also corroborate its better performance than the other five models. Hence, the group-level diagnostic results were extracted by the multi-CDM. Since we involved eight attributes, the total number of SMPs would thus be 256. We selected the top 10 patterns and found that 'mastery of all attributes' still accounted for the highest proportion, with 31.30%. Its counterpart, 'non-mastery of all attributes', however, did not show a very high proportion, with only 3.69% of all the patterns. This provides evidence that the multi-CDM might solve the flat pattern issue to some degree. However, too many SMPs are difficult to handle in a classroom setting. Hence, a cluster analysis was conducted to refine the classification results.

Group-level classification results from cluster analysis

With Simple K-means and EM method applied, the group diagnostic results were successfully classified into different clusters. For Simple K-means method, it only classified test-takers into two groups, i.e., masters and non-masters. For A1, Understanding Sentence Literal Meaning, the mastery probability of 61.46% was still sorted into non-master, whose classification validity should be questioned. On the other hand, the EM method yielded more appropriate classification results. All test-takers were classified into 12 different groups, with more interpretable characteristics in each group. For example, 68 test-takers were categorized into Group 1. They all have a solid mastery of attributes A1, A2, A3, A4, A7 and A8, while the two most difficult attributes, A5 and A6, are not mastered well. In the same vein, test-takers in Group 3 demonstrated a poor mastery of all attributes, with only 5% mastering all. The unique features in different groups are helpful for instructors to adjust their teaching plans to focus on the identified weaknesses of different groups. Meanwhile, the scale of the classification results (i.e., 12 groups) by the EM method is more controllable than that of CDA approach (i.e., 256 groups), which suits the purpose of classroom teaching and assessment.

Conclusion

The present study applied the multi-CDM method along with cluster analysis to enhance the classification of group-level diagnostic results extracted from the CDA approach. The multi-CDM was found to estimate more accurate model-data fit than any single CDMs did, demonstrating its practicality in reading comprehension tests. The group-level diagnostic results obtained by the multi-CDM also ameliorated the flat pattern issue in prior CDA studies to some extent. With the addition of cluster analysis, enhanced group-level classification results were generated with the EM method. A manageable size of groups with distinct strengths and weaknesses on the tested eight reading attributes are more practical for the classroom setting. Some pedagogical implications could be drawn from the findings of the present study. First, L2 instructors could design personalized EFL reading materials or teaching activities for different groups of students. Second, the cluster analysis method could also be implemented in other classroom assessment, which may assist instructors to get a whole picture of students' current learning outcomes. Finally, regarding diagnostic purpose, the multi-CDM method could be more appropriate in language assessment with intricate cognitive processing. It is worth noting that the present study is an initial investigation of new methods for group-level classification in CDA. The effectiveness and applicability of such incorporated methods could be further testified in other educational settings.

References

Chen, H., & Chen, J. (2015). Exploring reading comprehension skill relationships through the G-DINA model. Educational Psychology, *36*(6), 1–20.

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika*, *76*(2), 179–199.

de la Torre, J., & Lee, Y-S. (2013). Evaluating the Wald test for item-level comparison of saturated and reduced models in cognitive diagnosis. *Journal of Educational Measurement*, *50*(4), 355–373.

Du, W., & Ma, X. (2021). Probing what's behind the test score: application of multi-CDM to diagnose EFL learners' reading performance. *Reading and Writing*. Advance online publication. doi.org/10.1007/s11145-021-10124-x

Jang, E. E. (2009). Cognitive diagnostic assessment of L2 reading comprehension ability: Validity arguments for Fusion Model application to *LanguEdge* assessment. *Language Testing*, *26*(1), 31–73.

Javidanmehr, Z., & Anani Sarab, M. R. (2019). Retrofitting non-diagnostic reading comprehension assessment: Application of the G-DINA model to a high stakes reading comprehension test. *Language Assessment Quarterly*, *16*(3), 294–311.

Kim, A. (2015). Exploring ways to provide diagnostic feedback with an ESL placement test: Cognitive diagnostic assessment of L2 reading ability. *Language Testing*, *32*(2), 227–258.

Kim, H-Y. (2011). Diagnosing EAP writing ability using the reduced reparameterized unified model. *Language Testing, 28*(4), 509–541.

Lee, Y. W., & Sawaki, Y. (2009). Application of three cognitive diagnosis models to ESL reading and listening assessments. *Language Assessment Quarterly*, *6*(3), 239–263.

Leighton, J. P., & Gierl, M. J. (2007). *Cognitive Diagnostic Assessment for Education: Theory and Applications*. New York: Cambridge University Press.

Li, H., Hunter, C. V., & Lei, P-W. (2016). The selection of cognitive diagnostic models for a reading test. *Language Testing, 33*(3), 391–409.

Ravand, H. (2016). Application of a cognitive diagnostic model to a high-stakes reading comprehension test. *Journal of Psychoeducational Assessment*, *34*(8), 782–799.

Ravand, H., & Robitzsch, A. (2018). Cognitive diagnostic model of best choice: a study of reading comprehension. *Educational Psychology*, *38*(10), 1,255–1,277.

Robitzsch, A., Kiefer, T., George, A. C., & Uenlue, A. (2018). *CDM: Cognitive diagnosis modeling. R package version 6.3-45*. Retrived from: CRAN.R-project.org/package=CDM

Rojas, G., de la Torre, J., & Olea, J. (2012, April). *Choosing between general and specific cognitive diagnosis models when the sample size is small* [Conference presentation]. National Council on Measurement in Education, Vancouver, Canada.

Rupp, A. A., Templin, J., & Henson, R. (2010). *Diagnostic Measurement: Theory, Methods, and Applications*. New York: The Guildford Press.

Xie Q. (2017). Diagnosing university students' academic writing in English: Is cognitive diagnostic modelling the way forward?. *Educational Psychology*, *37*(1), 26–47.

Funding details

This paper was supported by the National Social Science Fund of China under Grant No.17BYY015, and the National Education Examinations Authority–British Council English Assessment Research under Grant No. EARG2020004.

Making formative assessment learner-friendly

Rubina Gasparyan
American University of Armenia

Abstract

While recognizing the importance of summative assessment as a way of reporting student learning at the end of a course or program (Suskie, 2009), it is hard to disagree with Coombe, Folse and Hubley (2010), who mention that assessment goes beyond tests to include a variety of tasks and activities that teachers use to evaluate students' daily progress.

To make this process enjoyable and interesting, as well as informative and useful for learners, different formative assessment tools can be employed, including role play, negotiation, giving feedback or asking and answering questions. However, these activities are only informative in terms of assessment if they are accompanied with tools, such as performance tables, evaluation grids or checklists. This paper offers a brief discussion of several formative assessment activities and looks into some ways of giving effective and useful feedback in order to promote further learning and teaching.

Introduction

The concept of formative assessment is not new in the field of language teaching and assessment. Together with summative assessment, it is an integral part of every school curriculum and a major focus in every classroom. While summative assessment refers to end-of-the-course tests and/or tasks administered 'to determine if students have achieved the objectives set out in the curriculum' (Coombe et al., 2010, p. xiv) and to make 'a judgment about student competence or program effectiveness' (Chappuis, 2015, p. 5), formative assessment is conducted during the course through feedback provided to students to improve instruction (Coombe et al., 2010, p. xiv) and, what is more, 'to help students guide their own subsequent learning or for helping teachers modify their teaching methods and materials so as to make them more appropriate for their students' needs, interests, and capabilities' (Bachman & Palmer, 2000, p. 98).

Needless to say that formative assessment serves as a cornerstone for classroom teaching and assessment since, as Suskie (2009) mentions, it is done midstream and provides a faculty with an opportunity to improve student learning 'by making immediate changes to classroom activities and assignments' (p. 24). Fulcher and Davidson (2007) claim that in the classroom the teacher 'can draw on a much wider range of evidence' (p. 27) based on the fact that they know the learners very well and thus can make better judgments about the learners' abilities. They further explain that the assessment of the learner's current abilities helps teachers decide what their subsequent steps should be for further learning to take place.

To further elucidate the purpose of formative assessment as part of overall assessment, it is worth mentioning that both assessment of learning and assessment for learning 'must be of high quality, yielding accurate results' (Stiggins, Arter, Chappuis, & Chappuis, 2007, p. 36) since important decisions are made based on assessment results. Chappuis (2015) seconds this idea, mentioning that formative assessment accounts for the use of information which can be gathered through formal and informal assessment, and based on which, decisions are made to help both teaching and learning.

With these features of formative assessment in mind, this paper aims to focus on several formative assessment practices by providing not only activities suitable for formative assessment (from here on formative assessment activities) but also, most importantly, assessment hints and tools for teachers.

Formative assessment activities

Student performance summary chart

In Chapters 9 and 10 of her book *Assessing Student Learning*, Suskie (2009) suggests a number of reflection checklists and provides hints for effective assignments. These suggestions are the inspiration behind the activities described below.

The first activity discussed here features a simple checklist to be completed by students based on a reading or a listening text. The idea for this particular reflection checklist was inspired by the Strategy Inventory for Language Learning (Oxford, 1990). This type of checklist can provide valuable information about what students have learnt and what they still need guidance on. Table 1 is an example of a checklist based on an adapted text from Mark Twain's *The Adventures of Tom Sawyer*. Such checklists are often referred to as self-assessment checklists or checklist rubrics (Suskie, 2009). The teacher can include as many key points and supporting details as they want; however, to make the checklist user-friendly, it should be simple and clear. One important feature of similar checklists is to keep statements affirmative and avoid using negative adjectives and/or forms.

Table 1: Reflection checklist based on an excerpt from Mark Twain's *The Adventures of Tom Sawyer*

Statement	*Put a tick if the statement is true for you. (Add comments if necessary).*
I know who Mark Twain is.	
I know who Tom Sawyer is.	
I can talk about Tom Sawyer.	
I understand all the words.	
I need some help with the new words.	
I need more explanation to understand the text.	

By including questions or statements derived from the learning outcomes of a given lesson or course, the teacher can gather information about the amount of learning that has taken place and the challenges that their students face. It is advised that the teacher gets the students' agreement to use the information in the checklist and explain to them that the purpose is to understand how to build further instruction for the students to improve.

To visualize students' answers and to summarize the information gathered for further decision-making related to learning outcomes, the teacher can create a simple table including the following sections: statement, students' responses, and further action (see Table 2).

Table 2: Student performance summary chart for teacher

Statement	*Students' responses*	*Further action*
I know who Mark Twain is.	All the students have answered positively.	Additional information about the writer can be provided.
I need some help with the new words.	Several students need help.	Vocabulary practice Making up sentences Role-play (encouraging students to use the new words)

Teacher evaluation grid

The importance of a rubric, a scoring guide describing the criteria necessary to evaluate an assignment (Suskie, 2009), cannot be overestimated. Among the advantages of using rubrics, Suskie mentions improving feedback to students, making scoring accurate and unbiased, inspiring better student performance, and clarifying vague goals. While some rubrics may be loaded with descriptors, simple checklists or grids reflecting the expectations of student performance and learning outcomes can be created based on those rubrics for quick classroom use. The following activity on creating reading or listening comprehension questions explains how teacher evaluation grids can be used. This and other similar activities can provide the teacher with ample time to observe their students and to take notes in the grid for further reference.

First, the teacher explains literal and inferential questions to their students by stating that the answers to literal questions can be found in the text directly, while the answers to inferential questions need to be inferred from other details in the text.

Furthermore, students work in small groups and create questions based on the text assigned. The next step is to exchange the questions formed with another group and receive/give feedback on question types and accuracy thus getting involved in active negotiations. Furthermore, the teacher can ask either the reviewing group or a third group to answer the questions.

There are a number of variations on how to conduct this activity, depending on the purpose and the learning outcomes. However, what makes it especially useful for formative assessment and giving feedback is the information that can be gathered during this activity.

Table 3 is a sample evaluation grid. This grid is useful for documenting general information rather than details in relation to each student individually. The information gathered will further help the teacher to reflect on the need for further explanation regarding comprehension questions and the comprehension of the text itself.

Table 3: Teacher's evaluation grid for question types

	Literal questions	*Inferential questions*
Easy to create		
Difficult to create		
Accurate grammar		
Clear content		
Elicited accurate answers		
Inaccurate answers due to question formulation		
Helped address text content effectively		
Notes		

Student performance table

Suskie (2009) claims that a better approach towards accomplishing learning goals are real-life tasks since such assignments 'engage students and help them see that they are learning something worthwhile' (p. 157). Students learn to implement their passive knowledge into real-life situations, which makes language learning very meaningful for them. Stiggins et al. (2007) mention that assessment can have a motivational aspect, and, in this respect, 'we use it to show students both what they have learned and what they need to learn next' (p. 41). Following these considerations, a student performance table can be created, and this time, as opposed to the teacher evaluation grid described above, the table can be used to observe individual student performance behaviour. Depending on such conditions as time allotted for the task and the number of students, the teacher should decide whether evidence is to be collected from all students or from several.

Table 4: Student performance table

Student	*Understanding/ Challenges*	*Feedback and follow-up*
Student 1	Has difficulty differentiating between logical fallacies and argumentation.	Teacher-student conferences. Assign additional exercises.
Student 2	Inadequate reasoning. Arguments are weak. The information in the reading is used inadequately.	Assign summaries (this will make the student read the text more attentively).

An example of a real-life task can be a role-play including negotiation where students are asked to focus on argumentation skills. Their task is to avoid using logical fallacies in negotiation. While students are engaged in the interesting and yet demanding task of negotiating according to the rules set, the teacher completes the student performance table. Table 4 is an example of a student performance table.

Conclusion

As can be seen, the proposed activities are discussed together with teacher tools that help collect and document information which can also be used to make 'the necessary instructional adjustments' (Marsh, 2007, p. 26). Marsh explains that formative assessment also helps students understand to what extent they have acquired their desired goals or knowledge. Furthermore, formative assessment empowers students to see their strengths and weaknesses and helps teachers with planning. However, Marsh claims that with all these advantages, teachers fail to effectively use formative assessment tools for various reasons, such as teachers' own experiences as students or the growing importance given to high-stakes standardized tests. To encourage teachers and other stakeholders to employ formative assessment, the author refers to research results claiming that techniques that include higher-order thinking, problem solutions, oral feedback and use of assessment criteria can be very beneficial for overall progress and achievement.

To conclude, I would like to agree with Stiggins et al. (2007), who believe that assessment for learning is about getting better. Formative assessment becomes the 'innovation that causes change in student achievement' (Stiggins et al., 2007, p. 37) and from an index of change becomes the change itself.

References

Bachman, L. F., & Palmer, A. S. (2000). *Language Testing in Practice.* Oxford: Oxford University Press.

Chappuis, J. (2015). *Seven Strategies of Assessment for Learning*. Boston: Allyn & Bacon.

Coombe, C., Folse, K., & Hubley, N. (2010). *A Practical Guide to Assessing English Language Learners.* Ann Arbor: University of Michigan Press.

Fulcher, G., & Davidson, F. (2007). *Language Testing and Assessment: An Advanced Resource Book.* London/New York: Routledge/Taylor & Francis Group.

Marsh, C. J. (2007). *A Critical Analysis of the Use of Formative Assessment in School.* Berlin: Springer Science + Business Media.

Oxford, R. (1990). *Language Learning Strategies: What Every Teacher Should Know*. Boston: Heinle.

Suskie, L. (2009). *Assessing Student Learning: A Common Sense Guide* (2nd ed.). California: Jossey-Bass Inc.

Stiggins, R. J., Arter, J. A., Chappuis, J. & Chappuis, S. (2007). *Classroom Assessment for Student Learning: Doing it Right – Using it Well.* New Jersey: Pearson Education, Inc.

Create your own wheel: Using evidence-centered design in the development of a workplace written communication assessment

Jennifer K. Morris
British Side Educational Services, United Kingdom

Ahmet Dursun
University of Chicago, USA

Abstract

CEPA® Written Communication Assessment™ is one of four stand-alone components in the overall Communicative English Proficiency Assessment (CEPA®), a computer-based, criterion-referenced and proficiency-oriented test designed to measure non-native speakers' workplace communicative English proficiency. It measures, specifically, 21st-century workplace written communication. In the creation of the CEPA® Written Communication Assessment, developers used the evidence-centered design (ECD) approach (Mislevy & Haertel, 2006) to make informed design choices through a systematic approach. In keeping with the notion that assessment as argument is a cornerstone of test validation (Kane, 2006; 2012) and that assessment practices are conceived as evidentiary arguments, this study reports on the practical application of ECD (with a focus on the first three phases) in the development of the CEPA® Written Communication Assessment™, demonstrating the unique process so it may be better understood and utilized by others in future testing development projects.

Introduction

On a daily basis, companies and organizations worldwide face similar challenges when it comes to decisions of selecting, hiring, promoting, and placing their employees based on a range of their abilities and skill sets. Foreign language skills are no exception, and the assessment process is not always as straightforward, nor as accurate, as simply reviewing formal education degrees, reference letters, institutional scores, and cover letters. This type of evaluation is further complicated by the oscillating nature of a second language, specifically its progression or attrition over time, its embeddedness in modern technology, and the sociolinguistic competencies (i.e., workplace *and* cross-cultural communication knowledge and strategies) which inevitably accompany it. Thus, a new type of assessment is needed to reveal the language capabilities and needs of the candidates as they pertain to their immediate level of English within the functioning, communicative environment of a real-life, modern workplace.

These challenges prompted the realization of the CEPA® (Communicative English Proficiency Assessment) research and development group in 2016, which would establish the foundation for a new and enriched type of online English language assessment in Turkey for adults in the business world (CEPA®, 2021, see cepatest.org/Web/). The current study is the documented process of this working group and attempts to explain and summarize how evidence-centered design (ECD), with a focus on the first three phases (domain analysis, domain modeling, and conceptual assessment framework) and an abridged overview of last two (assessment implementation and assessment delivery), guided the continuing design and structure of the computer-based proficiency-oriented writing performance tasks of CEPA®.

CEPA® assessment and mandate-driven evidence-centered design

CEPA® Written Communication Assessment™ was designed in response to the mandate handed down by British Side Educational Services, a private entity, to develop a screening assessment that could measure a non-native English speaker's workplace communicative English proficiency, and accordingly their ability to function in a real-world workplace. The need for this assessment was due to a niche market in Turkey with unmet demand. More specifically, the mandate required CEPA® to be a criterion-referenced, proficiency-oriented, and performance-based test reflecting the attributes of work for any potential employee in the real-world workplace.

In order to address these questions, a development path guided by the five layers of ECD, as described by Mislevy, Steinberg, and Almond (2003), Mislevy and Haertel (2006), and Mislevy (2011), was selected because it could provide tangible concepts (i.e., English language writing skills in real-life 21st century workplaces in Turkey) and a high level of detail that could be used to inform the design decisions (i.e., which skills were in more or less demand and how those could be represented in assessment tasks). More importantly, ECD had the potential to help create a solid argument for validation of test score interpretation and use by establishing explicit links between design decisions and scores obtained from the test (Chapelle et al., 2018). The first phase in this path, domain analysis, was the largest and most crucial undertaking.

Domain analysis: A systematic analysis of language use in workplace written communication

The British Side mandate revealed three goals for the domain analysis: (1) the need to screen candidates for corporations requiring workplace English proficiency for employability, (2) the need to define 21st-century workplace written communication knowledge, skills, and abilities, and (3) the need to design performance-based assessment tasks.

According to these needs, the domain analysis phase, which included 'a gathering of substantive information about the domain to be assessed, was initiated (Mislevy & Haertel, 2006, p. 7). Evidence was gathered and analyzed from three sources. The first was a survey of the stakeholders, including human resources staff, and other key personnel, both in domestic and international corporations in Turkey. These people were included because they were the ones who created the demand for an assessment, and therefore it was important to investigate their responses as to how and why English written communication is *actually* being used in their workplaces. The second component included professional workplace research reports from LinkedIn and prominent popular business magazines, which were used to determine the most recent and applicable discourse topics, trends, and values in workplace written communication. The third component was business English/English for workplace communication textbooks and curricula. This final component was included to capture the fundamental skills and knowledge being taught by, and the relevant learning materials provided in, major textbooks, which were in use in English language schools in Turkey and beyond.

For the sake of brevity, a detailed analysis of these resources (methodology, instruments, materials, and procedures, and results) can be found in a recent publication by Dursun, Morris and Ünaldı (2020). In the CEPA® project, the outcomes of the domain analysis helped us identify the central skill requirements and situations of written communication in global corporations as well as vital knowledge representations (e.g. topics and themes commonly found in the business world, relationships between participants in communication, directions and properties of communication practices, etc.), language functions, and key task features and performances needed to function within this domain, all of which led to a global definition of 21st-century workplace written communication:

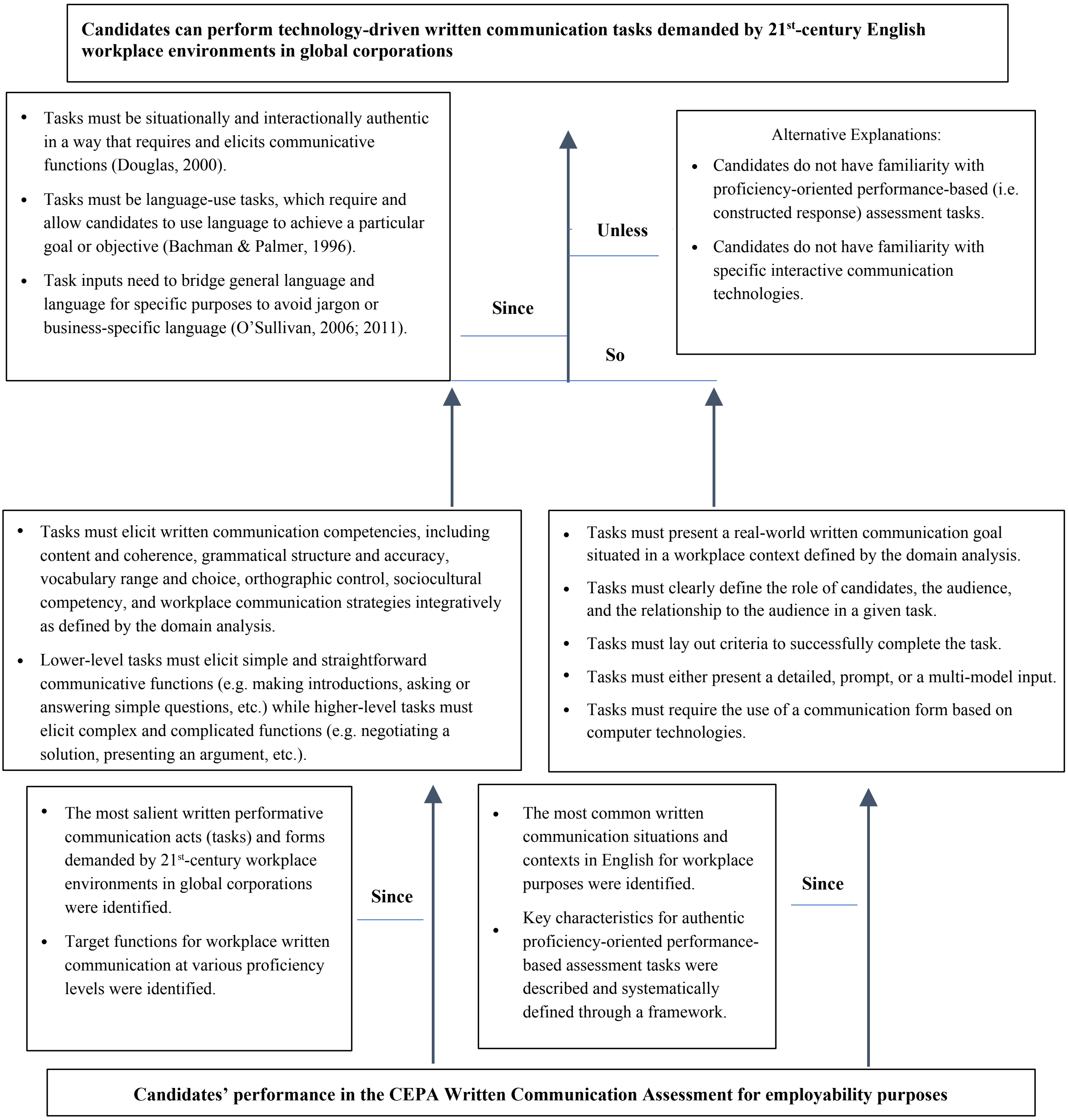
> Business correspondence involving both real-time and delayed unidirectional or bidirectional communicative acts either between internal members (same corporation, department, etc.) or other external members (business partners, potential clients, etc.), mainly through technology-mediated communication channels, in order to work on tasks situated within socioculturally dynamic contexts. (Dursun et al., 2020, p. 9)

In the following section, we further organize and describe these outcomes in conjunction with assessment arguments, thus forming a model of the domain.

Domain modeling: Developing evidentiary arguments for assessment design

The layer that follows the domain analysis in the ECD process is domain modeling in which the test developers 'organize the information and relationships discovered in domain analysis into the shape of assessment arguments' (Mislevy & Haertel, 2006, p. 8). These arguments help test designers to 'lay out what an assessment is meant to measure and how and why it will do so' (Mislevy & Haertel, 2006, p. 8). In this layer, the goal is to establish a connection between the knowledge, skills, and abilities discovered in the domain of interest, and the situations, activities, and tasks that engender these in the form of assessment arguments. Therefore, the findings from the domain analysis serve as input in the domain modeling process of defining design patterns.

Domain modeling can be developed through different supporting tools. One tool for domain modeling is to develop Toulmin diagrams for assessment arguments. A Toulmin diagram for assessment maps the assessment argument into a schema that lays out what assessments intend to measure and 'what behaviors or performances should reveal those constructs, and what tasks or situations should elicit those behaviors' (Messick, 1994, p. 16). Figure 1 presents the domain model for the CEPA® Written Communication Assessment using a Toulmin's argument diagram (the diagram is read from bottom to top). The warrants used in the assessment argument come from the above-mentioned domain analysis study (Dursun et al., 2020).



Candidates can perform technology-driven written communication tasks demanded by 21st-century English workplace environments in global corporations

- Tasks must be situationally and interactionally authentic in a way that requires and elicits communicative functions (Douglas, 2000).
- Tasks must be language-use tasks, which require and allow candidates to use language to achieve a particular goal or objective (Bachman & Palmer, 1996).
- Task inputs need to bridge general language and language for specific purposes to avoid jargon or business-specific language (O'Sullivan, 2006; 2011).

Alternative Explanations:

- Candidates do not have familiarity with proficiency-oriented performance-based (i.e. constructed response) assessment tasks.
- Candidates do not have familiarity with specific interactive communication technologies.

- Tasks must elicit written communication competencies, including content and coherence, grammatical structure and accuracy, vocabulary range and choice, orthographic control, sociocultural competency, and workplace communication strategies integratively as defined by the domain analysis.
- Lower-level tasks must elicit simple and straightforward communicative functions (e.g. making introductions, asking or answering simple questions, etc.) while higher-level tasks must elicit complex and complicated functions (e.g. negotiating a solution, presenting an argument, etc.).

- Tasks must present a real-world written communication goal situated in a workplace context defined by the domain analysis.
- Tasks must clearly define the role of candidates, the audience, and the relationship to the audience in a given task.
- Tasks must lay out criteria to successfully complete the task.
- Tasks must either present a detailed, prompt, or a multi-model input.
- Tasks must require the use of a communication form based on computer technologies.

- The most salient written performative communication acts (tasks) and forms demanded by 21st-century workplace environments in global corporations were identified.
- Target functions for workplace written communication at various proficiency levels were identified.

- The most common written communication situations and contexts in English for workplace purposes were identified.
- Key characteristics for authentic proficiency-oriented performance-based assessment tasks were described and systematically defined through a framework.

Candidates' performance in the CEPA Written Communication Assessment for employability purposes

Figure 1 CEPA® Written Communication Assessment domain modeling through Toulmin's argument diagram

Conceptual assessment framework: Creating test and task specifications

The next phase, the conceptual assessment framework (CAF), entails articulating assessment arguments in structures and specifications for tasks and tests (Mislevy & Haertel, 2006). According to the domain analysis and modeling, communication in global corporations were identified and defined, for the most part, as action-oriented or performance-based tasks. They also revealed the indispensable role of sociocultural competency (an awareness and sensitivity to a group/context) and workplace communication strategies (support of argument, providing examples, re-phrasing, etc.) in this domain. This highlighted a need to have a framework that defines the key characteristics of and establishes clear criteria for authentic proficiency-oriented performance-based tasks. Such tasks are distinguished from other types of assessments by their particular features in that they (a) present test-takers with a problem: a real-world goal, set within a realistic and relevant context of challenges and possibilities;

(b) push test-takers to develop a concrete product or performance for an intended audience (real or simulated); and (c) set evaluation criteria and performance standards that are appropriate for the task (Wiggins & McTighe, 2008).

For this purpose, we decided to adapt the *goal*, *role*, *audience*, *situation*, *product*, *standards* (GRASPS) model by Wiggins and McTighe (2008), which appeared in the context of the general education field. As the name indicates, the model sets six principles to ensure that the task is proficiency-oriented and performance-based so that it can elicit the target function(s) being measured in the task.

In addition to this, the framework outlines assessment tasks that are created to elicit a range of various language levels and competencies in order to adhere to the assessment arguments that require the use of measurable and demonstrable language functions and competencies. Furthermore, themes or topics are quite broad. Since each business has its own specific context and language that might necessitate the use of particular vocabulary, jargon or ideas, it was not practical or reasonable to assume universal knowledge in these areas. Thus, the framework includes a focus on identifying the overall purposes of the correspondence that were applicable across different business contexts. These include (a) information presentation, (b) information exchange or business inquiries, (c) sales and advertisements, (d) successfulness of customer relations, (e) negotiation, (f) resolving a conflict, and (g) networking. This outcome highlights the need for an assessment that bridges general language and language for specific purposes.

To this end, the Written Communication Assessment Framework in this domain outlines an interrelated integration of linguistic (as observed in the domain analysis of textbooks), sociolinguistic (as observed in the domain analysis of common global business world language content from reports and popular business magazines), and pragmatic competences (as observed in the domain analysis of situational contexts revealed by reports and surveys of talent acquisition managers) (Dursun et al., 2020).

Assessment implementation, delivery, and the future of CEPA®

CEPA® Written Communication Assessment has been implemented, meaning it has been 'constructed' and developed from 'all of the operational elements specified in the CAF' (Mislevy & Haertel, 2006, p. 16), using a series of three authentic written performance tasks. Each task represents a different range of proficiency levels, which progressively become more difficult in the subsequent tasks. For example, the first task has a low-high beginner-level range, the second task has a low-high intermediate-level range, and the third task has an advanced to mastery range. This assessment structure leveling is specified by the evidentiary argument highlighted in the framework. Likewise, these tasks are prefaced with or have embedded information (i.e., prompts) concerning the intended goals, role, audience(s), situation, product, and standards, which adhere to real-world and the most frequently used contexts and topics in the general business domain using information gathered in the domain analysis. For example, in the first task, set in a low-high beginner-level range, a test taker is shown an email and asked to write a reply to a colleague answering a few questions concerning their upcoming visit to another office located in an English-speaking country.

Another distinctive aspect of the CEPA® Written Communication Assessment is in its final delivery phase where assessment takers 'interact with tasks' (Mislevy & Haertel, 2006, p. 16). The act of incorporating 21st-century workplace written communication knowledge, skills, and abilities is partially actualized through the process of taking the assessment via a computer, an online platform, and remote proctoring. However, this assessment has gone even further as test developers worked with IT program engineers to simulate authentic materials and functions found in current computer communication technologies. For example, the email in the first task resembles a *real* email found in an inbox and the reply space mimics many of the functions of a word processor.

It is important in the implementation and delivery of assessments that the first three phases of the ECD process are consistently referenced and the framework kept in mind as the implementation and delivery phases represent *how* tasks and concepts will be received by assessment takers. Their reception (i.e., how authentic tasks are, how well they are understood through previously existing schema, etc.) is not always foreseeable; nonetheless, this can be mitigated by a sound consideration of the framework derived from the phases of ECD. A further step may also be taken to pilot the assessment or tasks with real assessment takers and revise task design and assessment delivery where needed, a step which the CEPA® developer group chose to employ.

Conclusion

From the final assessment platform and delivery of it, the CEPA® developer group has now begun to gather and review feedback, not just from users, but developers and IT engineers too, in order to reintegrate it back into the foundational definitions and findings first established by the domain analysis: a global definition of 21st-century workplace written communication, central skill requirements and situations of written communication in global corporations, and vital knowledge representations, language functions, and key task features and performances needed to function within the workplace domain.

Assessment development should always be a work-in-progress endeavor; however, this does not mean re-inventing the wheel every time. By using the phases of ECD, developers may create their *own* wheel by investigating a language domain, creating evidence-based arguments for their assessment design, and developing a framework whereby they can more easily integrate feedback and changes. Ultimately, they will be able to make informed, evidence-based choices concerning their assessment specifications thereafter.

References

Chapelle, C., Schmidgall, J., Lopez, A., Blood, I., Wain, J., Cho, Y., Hutchison, A., Lee, H., & Dursun, A. (2018). *Designing a prototype tablet-based learning-oriented assessment for middle school English learners: An evidence-centered design approach*. Research Report No. RR-18-46. Princeton: Educational Testing Service.

Dursun, A., Morris, J.K., & Ünaldı, A. (2020). Designing proficiency-oriented performance tasks for the 21st-century workplace written communication: An evidence-centered design approach. *Assessing Writing*, *46*, Article 100487.

Kane, M. (2006). Validation. In R. Brennen (Ed.), *Educational Measurement* (4th edn.) (pp. 17–64). Westport: Greenwood Publishing.

Kane, M. T. (2012). Validating score interpretations and uses. *Language Testing*, *29*(1), 3–17.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, *23*(2), 13–23.

Mislevy, R. J. (2011). *Evidence-centered design for simulation-based assessment*. CRESST Report 800. Los Angeles: University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Mislevy, R. J., & Haertel, G. (2006). Implications of evidence-centered design for educational testing. *Educational Measurement: Issues and Practice*, *25*(4), 6–20.

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

Wiggins, G., & McTighe, J. (2008). *Understanding by Design Professional Development Workbook*. Alexandria: Association for Supervision and Curriculum Development.

The development of linguistic profiles and online tests for Quebec health professionals working in English

Michel Laurier
University of Ottawa, Canada

Denise Lussier
McGill University, Canada

Hélène Riel-Salvatore
McGill University, Canada

Abstract

The research team first developed linguistic profiles which identify the linguistic tasks that health practitioners, such as nurses, use in their professional functions, mostly related to oral (expression and comprehension) interactions with patients. Three levels of complexity were considered: daily routine, professional healthcare, and complex procedures including tasks from the affective domain. The second phase of the research focuses on the development of a placement tests for oral skills. The test sections are based on authentic scenarios; each scenario consists of tasks that nurses have to perform in order to reach the required level of complexity. All the candidates have to go through the oral comprehension section if they reach the expected level; they are directed to the oral expression section which first asks for short answers to be analyzed by a voice recognition system, and then asks for longer answers to be corrected by an external evaluator.

Introduction

This paper reports on a language test for specific purposes in Canada that is under development. The content of the test is based on a set of linguistic profiles that describe language tasks that nurses may have to perform in their second language when they interact with English-speaking patients. The profiles are organized as a six-level scale which describes what nurses can actually do depending on their English as a Second Language (ESL) level.

A pilot version of the test is now under revision by experts, which will be implemented on an online platform. Nurses will then be able to access it freely to estimate their level of proficiency when interacting with patients in English. Until the test is fully validated, data will be collected to ensure that it allows us to draw appropriate inferences in connection with the linguistic profiles.

Access to health care for language minorities in Canada

At the federal level, Health Canada is responsible for allowing all Canadian citizens to receive the necessary assistance to maintain or improve their health, and must guarantee free medical care. Programs have been developed to ensure that language minorities in Canada can access health and social services in their own language. ESL courses are offered to health professionals. They are beneficial to a high percentage of patients whose mother tongue is English (Ouimet, Trempe, Vissandjée, & Hemlin, 2013). At province level, the Government of Quebec is committed in upholding the right for the English-speaking community to receive the same services in its own mother tongue as citizens from the French-speaking community. Although the majority of Quebec population speaks French as a mother tongue, the 2006 Canada Census indicates that 8% identify English as their mother tongue and 12% a language other than English or French.

Interacting with a health practitioner in a second language appears to be a difficult situation especially if the patient is not fluent, emotionally unstable or has to deal with technical concepts. Bowen (2001) shows how linguistic barriers may impede access to health and social services and even affect the quality of these services. The language barrier may make it difficult to establish a good caregiver-patient rapport. As a consequence, the patient may receive less attentive care and experience less satisfaction with the medical treatment than patients for whom there is no language barrier (Carrasquillo, Orav, Brennon, & Burstin, 1999; Jacobs, Chen, Karliner, Agger-Gupta, & Mutha, 2006). For emotional, ethical and legal reasons, good communication with the patient can be especially critical (Anderson, Scrimshaw, Fullilove, Fielding, & Norman, 2003; Chen, Youdelman, & Brooks, 2007). The delivery of bad news is a good example (Gillotti, Thompson, & McNeilis, 2002).

de Negri, Brown, Hernández, Rosenbaum and Roter (1997, p. 12) explain the importance of developing efficient interpersonal communication between healthcare provider and patients: 'Good interpersonal interaction between client and provider is, by definition, a two-way street where both speak and are listened to without interruption, both ask questions, express opinions and exchange information and both are able to fully understand what the other is trying to say.' Lussier (2009) argues that effective interactions rely on two essential dimensions: cognitive and affective. They also involve intercultural elements that impact the management of emotions and the expression of empathy (Robinson, 2002). Blaser (2009) reports that only 45% of the Quebec nursing staff can sustain a conversation in English, and that access to services in English is a major issue. In addition, as Moreno, Otero-Sabogal and Newman (2007) mention, being able to converse is not sufficient for a practitioner to provide adequate healthcare services.

The *Training and Retention of Health Professionals* program was launched at McGill University in 2005. The aim is to ensure that high-quality healthcare services are provided. Every year, it offers language training in English to more than 2,000 health practitioners. Learning activities focus on professional needs. The approach is based on the 'English for Specific Purposes' principles that are fully documented in the literature (Liu & Hu, 2021). Professional needs have been defined in collaboration with employers, professional bodies and practitioners. Learning activities are based on participants' professional interests and take into account various learning styles and conditions. The intent is not to develop a native-like competence but rather transform learners into efficient users of the target language to fulfill professional demands. Because of limited resources and time constraints, the language training is mainly intended for learners who have already reached intermediate proficiency level.

The program has already funded research to develop an online formative tool to help nurses assess their capacity to use English in a professional setting (Isaacs, Laurier, Turner, & Segalowitz, 2011). However, there is no formal assessment that could determine the proficiency level reached by a nurse at the end of a training session or determine need for further training. This is the reason why the program is now involved in the development of an *oral achievement test for English as a professional language for nurses* as health professionals.

Assessment framework

The first stage of this project was to design linguistic profiles for nurses in Quebec. The framework of reference includes competency descriptors, performance indicators, stages and levels of competency, and definition of language tasks.

The competency descriptors were categorized according to their complexity in relation with five performance elements:

- **types of discourse** from informative to narrative, to analytical to evaluative
- **professional contexts** from basic everyday to complex professional healthcare with interpersonal, interactional, and multicultural dimensions
- **situations** from everyday predictable to wide-ranging, often new and unpredictable
- **patients' needs** from primary basic care to complex and specialised nursing care in response to emotional situations
- **language tasks** from very simple to varied and specialised.

The descriptors were validated with the nursing community (Riel-Salvatore & Lussier, 2019).

> The competency descriptors are observable linguistic behaviours from which a competency level can be inferred. They are defined as types of discourse (relevance), performance (coherence) and language (grammar, vocabulary, pronunciation, intonation and fluency). The degree of complexity is determined by a performance indicator according to how easy or difficult it is to realise a language task. For example, getting information or describing a person are less demanding than analysing or making hypotheses. During the validation, practitioners were asked to assign a degree of complexity to each indicator.

Three stages are used to indicate the complexity of a performance: *Elementary, Intermediate* and *Advanced.* Each of them includes two levels of competence for a total of six levels of complexity. *Elementary* indicators usually refer to daily basic and routine needs/care, *Intermediate* indicators usually refer to professional healthcare requirements and *Advanced* indicators refer to specialized procedures and management of the sensitive and affective issues. The performance indicators that operationalise this level often refer to the affective domain. They generally involve a degree of intercultural awareness.

For example, a nurse who is positioned at the first level of the *Advanced* stage should be able to accomplish all the language tasks that are required in her interaction with patients. Oral Expression at this level is described as:

> **Highly functional and autonomous** as well as fluid in a specialized professional context often involving new situations. Pronunciation errors rarely impede communication. The nurse can link ideas while using a variety of sentence structures as well as specific and specialized vocabulary in order to provide information, explanations and recommendations, discuss choices, respond to embarrassing situations, propose solutions to problems and predict consequences. The nurse is generally at ease and responsive to the client's personal and emotional situation and shows empathy in complex and specialized nursing task situations. Can sustain a simple predictable phone conversation. (Riel-Salvatore & Lussier, 2019, p. 33)

Characteristics of the language test

The development of the language test for Quebec nurses follows the principles of testing for specific purposes as described by Douglas (2000). Since nurses work in a French milieu but have to communicate with the English-speaking population, written skills are not relevant to assess Quebec nurses' capacity to interact with patients in ESL. In this particular context, it has been decided to restrict the test to listening and speaking. Since authenticity is a major concern, the language tasks relate to situations that a nurse may encounter when interacting with patients who prefer to communicate in English. It is important to ensure that the test does not assess professional knowledge but rather focuses on the language competence as described by the language profiles.

The test is not intended for certification or licensure purposes. The aim is to position a candidate on a six-level scale and assign the most appropriate group. It may also eventually provide helpful information to allocate health resources to better serve the population. In this context, the test should be shorter than a high-stakes test and online delivery does not raise major security issues.

Both listening and speaking sections use assessment scenarios that relates to situations that may occur in nurses' professional life. For example, a nurse may have to understand a patient's request regarding a particular treatment or explain how to take a medication. Some scenarios are divided into several parts. Depending on the level they target, the scenarios are more or less complex. They are designed so that relevant performance indicators be observed in relation to the target level. In the listening section, the candidate hears excerpts (two to three minutes) and then answers multiple-choice questions (with one or two right answers) or completes a drag-and-drop exercise. In the speaking section, the candidate hears dialogues and is asked to provide a two-to-three-minute-answer based on a canvas.

The test is available online. It uses the *Learning Branch* platform (www.learningbranch.com). The platform allows conducting soft-skill assessment online and is used by several institutions in Canada to create language learning modules and assessment tools. It integrates some artificial intelligence techniques and allows various format of tasks. However, the test to be implemented cannot be considered as a fully adaptive test because no branching procedure is yet integrated in the platform code (Laurier, 1997). However, some adaptive strategies are used so that the test does not last too long if the candidate's level is clearly below the expected level. The test begins with the listening section. The first scenario is chosen among scenarios at the *Elementary* stage. If the candidate is successful, a more complex scenario is presented; otherwise, the candidate is invited to quit the environment. The process goes on until the candidate's actual level is reached. We consider that it is pointless to ask a candidate who is positioned at the *Elementary* levels in listening to enter the speaking section. However, if the score is higher than the *Intermediate* threshold, scenarios at the *Intermediate* levels and, if necessary, at the *Advanced* levels, will be presented. The candidate must record a two-to-three-minute answer. The answers are then concatenated and sent to a human rater who will establish the candidate's level in speaking. We are also considering the inclusion of short-answer items (one sentence) using the voice recognition device and natural language processor that are available on the platform in order to identify candidates who are not proficient enough to provide adequate long answers. This would reduce the number of recordings to be sent to the raters.

The raters can enter their judgement in the system. The candidates are then informed of the final result without undue delay. The raters refer to three criteria to determine the level: Message, Discourse and Language. The scoring sheet and guide, which are still under development, are key factors to maintain the quality of the test. Since the validity rests on the authenticity of the assessment scenarios and language tasks, the feedback from representatives of the nurse association is crucial. Validity is also ensured by the inclusion of all relevant performance indicators for a given level, both in the listening and speaking sections.

Conclusion

As soon as the scenarios and tasks are implemented on the platform, the test will be available online for any Quebec nurses. Data will be collected. The responses on the listening section will be analysed and modelled. Appropriate changes will be made. For the speaking section, we will look at the feasibility of the tasks and monitor the rater consistency.

We believe that this test will support the efforts of the nurses to maintain or develop their language skills. The test should then contribute to improving the quality of the healthcare services that are provided to minority language communities in Canada.

References

Anderson, L., Scrimshaw, S., Fullilove, M., Fielding, J., & Norman, J. (2003). Culturally competent healthcare systems: A systematic review. *American Journal of Preventative Medicine, 24,* 68–79.

Bowen, S. (2001). *Language Barriers in Access to Health Care*. Retrieved from: www.canada.ca/en/health-canada/services/health-care-system/reports-publications/health-care-accessibility/language-barriers.html

Blaser, C. (2009). *Health Care Professionals and Official-Language Minorities in Canada, 2001 and 2006*. Ottowa: Statistics Canada, catalogue no. 91-550-X-an.

Carrasquillo, O., Orav, J., Brennon, T., & Burstin, H. (1999). Impact of language barriers on patient satisfaction in an emergency department. *Journal of General Internal Medicine, 14,* 82–87.

Chen, A. H., Youdelman, M. K., & Brooks, J. (2007). The legal framework for language access in healthcare settings: Title VI and beyond. *Journal of General Internal Medicine, 22*, 362–367.

de Negri, B., Brown, D. L., Hernández, O., Rosenbaum, J., & Roter, D. (1997). *Improving interpersonal communication between health care providers and clients*. Bethesda: USAid. Retrieved from: pdf.usaid.gov/pdf_docs/pnace294.pdf

Douglas, D. (2000). *Assessing Languages for Specific Purposes*. Cambridge: Cambridge University Press.

Gillotti, C., Thompson, T., & McNeilis, K. (2002). Communicative competence in the delivery of bad news. *Social Science & Medicine, 54*, 1,011–1,023.

Isaacs, T., Laurier, M., Turner, C.E., & Segalowitz, N. (2011). Identifying second language speech tasks and ability levels for successful nurse oral interaction with patients in a linguistic minority setting: An instrument development project. *Health Communication*, *26*(6), 560–570.

Jacobs, E., Chen, A., Karliner, L. S., Agger-Gupta, N., & Mutha, S. (2006). The need for more research on language barriers in health care: A proposed research agenda. *The Milbank Quarterly, 84*, 111–133.

Laurier, M. (1997). Different models for different language tests. In D. Ajar & H.M. Kandarakis (Eds.), *New Horizons in Learning Assessment* (pp. 183–192). Montréal: Université de Montréal.

Liu Y., & Hu, G. (2021). Mapping the field of English for specific purposes (1980–2018): A co-citation analysis. *English for Specific Purposes*, *61*, 97–116.

Lussier, D. (2009). Common reference for the teaching and assessment of 'Intercultural Communicative Competence' (ICC). In L. Taylor & C. J. Weir (Eds.), *Language Testing Matters: Investigating the wider social and educational impact of assessment. Proceedings of the ALTE Cambridge Conference, April 2008* (pp. 234–244). Studies in Language Testing volume 31. Cambridge: UCLES/ Cambridge University Press.

Moreno, M. R., Otero-Sabogal, R., & Newman, J. (2007). Assessing dual-role staff-interpreter linguistic competency in an integrated healthcare system. *Journal of General Internal Medicine*, *22*(2), 331–335.

Ouimet, A-M., Trempe, N., Vissandjée, B., & Hemlin, I. (2013). *Language Adaptation in Health Care and Health Services: Issues and Strategies*. Québec: Institut National de Santé Publique du Québec, Gouvernement du Québec. Retrieved from: www.inspq.qc.ca/pdf/publications/1697_AdapLinguisSoinsServicesSante_VA.pdf

Riel-Salvatore. H., & Lussier, D. (2019). *Projet de développement de profils linguistiques pour les infirmières et infirmiers du Québec œuvrant auprès d'une clientèle d'expression anglaise en milieu francophone.* Montreal: Dialogue-McGill (Mimeo).

Robinson, M. (2002). *Communication and Health in a Multi-Ethnic Society*. Bristol: Policy Press.

Assessment practices in English language courses in Saudi public universities

Samar Yakoob Almossa
English Language Centre, Umm Al-Qura University, Makkah, Saudi Arabia

Abstract

This study examines current English language course assessment practices in Saudi higher education in order to provide insight into coordination between those practices and recent changes to Saudi higher education policy. The findings reveal that assessment practices are largely influenced by the principles of summative assessment and that the most common assessment instruments are written examinations. The research implications suggest that Saudi public universities should reconsider their assessment methods and instruments and should include more activities that help students practice and develop language skills. Urgent consideration should also be given to how assessment methods and instruments might impact development of 21st-century skills, such as critical thinking and problem solving.

Introduction

Assessment is an important aspect of learning because it helps to determine if the intended learning outcomes were achieved (Wiliam, 2011). In the context of higher education, students move to a range of classroom settings in which they are expected to navigate self-learning and to become autonomous learners who require less attention and less feedback than would be the case in directed learning (Panadero, Fraile, Fernández Ruiz, Castilla-Estévez, & Ruiz, 2019). The roles of teaching, learning, and assessment are interrelated steps in the process of encouraging students to become self-regulated. It is clear that the way in which students are taught and encouraged to learn, and the way they are assessed, has an impact on their learning and on the skills that they develop; the same is true for teachers (Brown, 2004).

Assessment researchers in the Middle East and North Africa (MENA) region have explored some aspects of the summative assessment washback effect on teaching and learning (Gebril & Brown, 2014; Gebril & Taha-Thomure, 2014; Troudi, Coombe, & Al-Hamly, 2009). However, there are still gaps in our knowledge regarding how assessment is perceived, reported, and used in higher education (Almossa, 2018). Several studies in the Saudi context have explored different aspects of assessment in higher education, including teacher education programmes (Alaudan, 2014; Almalki, 2014; Almansory, 2016; Almossa, 2018; Gaffas, 2019). Gebril and Hozayin (2014) stress the need for more studies in the MENA region that explore assessment practices, as previous studies have tended to focus on the perceptions and attitudes of teachers and students at one institution. While these studies are valuable sources of information, more data is needed to understand English language assessment practice trends in Saudi higher education. There are gaps in our understanding of assessment practices in the MENA region, and this study attempts to close these gaps.

This study examines syllabi from English language courses to determine the most common assessment instruments, the skills that are assessed, and the weight given to different assessment instruments. Investigation of assessment practices can help us identify the skills that are being evaluated. In addition, it helps us understand how language courses are contributing to student achievement of the courses' intended objectives. Beyond improving language skills in support of university studies and beyond, these courses should also impart 21st century skills. This paper is part of a research project that explores assessment practices across Saudi universities. It seeks to answer the following research questions (RQs):

- RQ 1: What are the most commonly used assessment instruments? Are there any variations in assessment types across streams?
- RQ 2: How many instruments are used to assess productive language skills? How much are the assessed skills (e.g., speaking and writing) weighed in the total grade?

Methodology

This study adopted syllabus analysis as a methodology for collecting data because the syllabus is an official document that can provide 'an interesting picture of assessment practices' and 'the instructional environment teachers create in their university courses' (Panadero et al., 2019, p. 382). Syllabi also have a clear format (and similar structures) and provide current information about the course and its assessment practices. It is an 'unobtrusive but powerful indicator of what takes place in classrooms' (Bers, Davis, & Taylor, 2000, p. 7).

Data collection and analysis procedures

To collect the research data, I visited the website of each public university and downloaded the syllabus for each English language course and stream. Twenty-one of the 29 Saudi public university syllabi were accessible; eight universities' English Language Institutes and English Language Centres (ELIs/ELCs) do not publish their syllabi. In total, 89 syllabi were downloaded and reviewed for analysis.

XLSTAT and SPSS were used to analyse the data, which were classified and coded into categories before they were entered into the SPSS software. The main categories were examination types and coursework types. Other categories were created according to the system operated by individual institutions, such as semester-based or quarterly-based or according to the age of the universities (new or established). Descriptive analysis involved the calculation of the frequency, the mean, and the minimum and maximum of all the tested variables.

Research findings

The results from analysis of assessment instruments in 21 English language courses in Saudi universities showed that, in total, 13 instruments were utilized. Written exams were the most dominant assessment instrument and were widely used across General English (GE), English for Academic Purposes (EAP), and English for Specific Purposes (ESP) courses, whether taught in the first or second semester, in a quarterly-based or semester-based system, or at new or established universities. This homogeneity might be due to the nature of some of the foundation year courses, which are unified across all the streams. Therefore, unified practices with small alterations were expected.

Six of the 13 utilized assessment instruments were some variety of oral or written examination (Table 1). Examinations accounted for up to 69.95% of the total grade: final exam (Mean = 42.0476%), mid-term exam (Mean = 23.05%), second mid-term exam (Mean = 15.75%), speaking test (Mean = 14.2857%) and writing test (Mean = 7%). Final and mid-term examinations were administrated in two rounds in some universities. Only one institution based the overall grade on the final exam (60%) and a mid-term exam (40%) with no other assessment tools. Some variation occurred between new and established universities, specifically in the number of mid-term exams and quizzes and in the weight given to various assessments.

Table 1: Descriptive statistics of examinations

	N	*Minimum*	*Maximum*	*Sum*	*Mean*
Examination total	21	30%	100%	1469%	69.95%
Final exam	21	15%	60%	883%	42.0476%
Mid-term exam	19	0%	43%	438%	23.05%
Second mid-term	4	8%	20%	63%	15.75%
Oral final exam	1	15%	15%	15%	15%
Speaking test	7	5%	35%	100%	14.2857%
Writing test	3	1%	15%	21%	7%
Valid N (listwise)	0				

The data were analysed by looking at the universities according to its states: new or established or the ELCs/ELIs system: semester-based and quarterly-based. At established universities, the examination total was Mean = 67.89% and the coursework total was 21%; at new universities, the examination total was Mean = 73.80% and the coursework total was 24.6%. Variations also occurred between semester-based and quarterly-based institutions. Quarter-based gave less weight to examinations (Mean = 49.60%) and more to coursework (Mean = 29.40%) due to assessment in that system being more frequent across different language levels. In semester-based institutions, the examination was Mean = 78.64% and the coursework was Mean = 20.57% of the total grade.

Seven coursework instruments were given a smaller weight on average 22.85% (Table 2). The weight of quizzes, progress tests, and continuous assessment was between 5%–52%, participation (5%), online practice (2%–20%), assignments (5%–14%), and attendance (5%). Attendance was included in two of the examined universities even though, according to Saudi Ministry of Education regulations, attendance and absence should not be included as part of the course grade. 15 out of the 21 universities utilized quizzes or progress tests and continuous assessment; periodical written exams were frequently given (e.g., weekly) and were classified as part of the coursework or classwork. Five institutions instead gave speaking and listening tasks (N = 1; 20%), online practice (N = 2; 5–20%), or an assignment (N = 1; 14%). Correspondent Variance Analysis was used to observe how the variables interact. The findings suggest that all instruments were used frequently with different weights, and that final and mid-term examinations were cornerstone variables.

Table 2: Descriptive statistics of coursework

	N	*Minimum*	*Maximum*	*Sum*	*Mean*
Coursework total	21	0%	57%	480%	22.8571%
Quizzes/Progress tests/ Continuous assessment	15	5%	52%	343%	22.8667%
Participation	2	5%	5%	10%	5%
Speaking/listening tasks	1	20%	20%	20%	20%
Online practice	6	2%	20%	47%	7.83%
Assignments	3	5%	14%	24%	8%
Attendance	2	5%	5%	10%	5%
Valid N (listwise)	0				

Assessment of productive skills (speaking and writing) was extant in some institutions and absent in others; similarly, some institutions assessed one skill but not the other. One institution had an oral mid-term exam (10%) and an oral final exam (15%) in parallel with written examinations. Speaking was assessed through a test in seven institutions and through a speaking and listening activity in one university. The weight given to speaking tests was between 5%–35%. In the five institutions that assessed writing through a writing test (N=3), the test was weighted between 5% and 15%, and was sometimes paired with a writing task (N=3) that was weighted between 1% and 10%. Two universities assessed writing through both tasks and tests that were weighted between 1% and 5%. Speaking and writing tests, which were separately graded, were administrated in a one-on-one environment. Clearly, only a small percentage of student grades was based on speaking and writing tests and other coursework projects. Institutions who innovated in assessment gave speaking and writing tasks a small percentage of the overall grade in order to avoid basing a larger proportion of the grade on an objective assessment, which is still an important step toward shifting the focus from written-based examinations that focus on grammar, vocabulary, listening and reading only.

The weight of the written examinations sent a clear message to teachers and students that written performance was the most important to passing the course, while hands-on activities, such as projects and presentations, were overlooked. If debates and a space for creativity in writing, speaking, narrating stories, and expanding on imagination are limited, then these skills, which are regarded as important to life in the 21st century, are hardly developed for university studies or for later entrance into the job market.

Discussion and conclusion

The data show that English language courses in Saudi universities are assessed mainly with written exams; though a great number of other assessment instruments are utilized, the percentage of non-paper-based examinations is relatively small. The weight given to written examinations is explained as a practical way to produce results that are fair and accurate for all students. However, traditional assessment methods contradict innovations in teaching, learning, and assessment that are needed as part of a 21st-century skillset. Traditional assessment does not allow students to exhibit their language production and development. The argument for written examinations seems to centre on the difficulty of finding a balanced and unified system that can be used to assess, fairly mark, and rank students. However, with training and enhancement of teachers' and students' assessment literacy, some improvements can be realized. These research findings agree with those of Panadero el al. (2019), who conducted a large-scale study across Spanish universities in all disciplines. One reason for this is that teachers tend to teach the way they were taught. Panadero et al. (2019) suggest that higher education regulation should be supportive of any modern assessment methods, including self-assessment and peer-assessment, that enhance the abilities students need to succeed in university and in the workplace.

References

Alaudan, R. (2014). *Saudi student teachers' perceptions of formative assessment* [Doctoral dissertation]. University of York.

Almalki, M. S. (2014). *A preliminary design framework for formative blended assessments in tertiary English as a foreign language (EFL) programs: an exploratory study in Saudi Arabia* [PhD thesis] University of Melbourne.

Almansory, M. (2016). *EFL teachers' beliefs and attitudes towards English language assessment in a Saudi university's English Language Institute* [Doctoral thesis]. University of Exeter.

Almossa, S. (2018). *Developing pedagogy and assessment in EFL: A case study of a Saudi university* [PhD thesis]. King's College London.

Bers, T. H., B. D. Davis, & Taylor, B. (2000). The use of syllabi in assessments: Unobtrusive indicators and tools for faculty development. *Assessment Update*, *12*(3), 4–7.

Brown, G. T. L. (2004). Teachers' conceptions of assessment: Implications for policy and professional development. *Assessment in Education: Principles, Policy & Practice*, *11*(3), 301–318.

Gaffas, Z. (2019). Students' perceptions of the impact of EGP and ESP courses on their English language development: Voices from Saudi Arabia. *Journal of English for Academic Purposes*, *42*(3). Retrieved from: www.sciencedirect.com/science/article/abs/pii/S1475158518306179

Gebril, A., & Brown, G. T. (2014). The effect of high-stakes examination systems on teacher beliefs: Egyptian teachers' conceptions of assessment. *Assessment in Education: Principles, Policy & Practice, 21*(1), 16–33.

Gebril, A., & Hozayin, R. (2014). Assessing English in the Middle East and North Africa. In A. J. Kunnan (Ed.), *The Companion to Language Assessment*. Chichester: John Wiley & Sons. Retrieved from: onlinelibrary.wiley.com/doi/abs/10.1002/9781118411360.wbcla077

Gebril, A., & Taha-Tomure, H. (2014). Assessing Arabic. In A. J. Kunnan (Ed.), *The Companion to Language Assessment*. Chichester: John Wiley & Sons. Retrieved from: hanadataha.com/wp-content/uploads/Published-Wiley-Hanada-Article-Assessing-Arabic.pdf

Panadero, E., Fraile, J., Fernández Ruiz, J., Castilla-Estévez, D., & Ruiz, M. (2019). Spanish university assessment practices: examination tradition with diversity by faculty. *Assessment & Evaluation in Higher Education*, *44*(3), 379–397.

Troudi, S., Coombe, C., & Al-Hamly, M. (2009). EFL teachers' views of English language assessment in higher education in the United Arab Emirates and Kuwait. *Tesol Quarterly, 43*(3), 546–555.

Wiliam, D. (2011). *Embedded Formative Assessment*. Bloomington: Solution Tree Press.

Designing a communicative foreign language assessment program for Dutch secondary schools: A design-based research project within a professional learning community (PLC)

Charline Rouffet
HU University of Applied Sciences Utrecht, the Netherlands

Catherine van Beuningen
HU University of Applied Sciences Utrecht, the Netherlands

Rick de Graaff
HU University of Applied Sciences Utrecht, the Netherlands

Abstract

While Communicative Language Teaching (CLT) is nowadays widely recognized, the implementation of CLT in foreign language (FL) classrooms remains difficult. In the Netherlands, communicative learning goals have been formulated at the national level, but in daily practice assessments and learning activities tend to focus on knowledge of grammar rules and vocabulary out of context.

Under the principle of constructive alignment, assessment and learning activities should be in line with learning objectives in order to enable effective teaching. Evaluation in particular plays a key role, because it has a direct impact on teachers' pedagogical choices and students' learning behavior (i.e., washback effect).

To enhance alignment in the Dutch FL curricula of lower form education, the first author and 21 FL teachers collaborated to create a realistic, theoretically grounded communicative assessment program that would enable positive washback. This paper takes inventory of the challenges faced by teachers during the co-design process and the decisions that have been made to overcome them.

Introduction

Communicative language teaching (CLT) aims to promote the development of communicative language skills in realistic situations. The approach emphasizes the use of language through meaningful spoken and written interaction rather than through memorization and the learning of grammatical rules out of context (Larsen-Freeman & Anderson, 2011). CLT is nowadays widely accepted and recognized as a productive pedagogical approach in foreign language (FL) education. However, the implementation of CLT in daily teaching practices remains difficult (Kissau, Rodgers, & Haudeck, 2015). This situation also seems to occur in the Netherlands. In Dutch secondary education, learning objectives at the national level are formulated from a communicative perspective, but in teaching and assessment practices teachers tend to focus on knowledge of grammar rules, vocabulary and chunks taught out of context (Fasoglio, de Jong, Pennewaard, Trimbos, & Tuin, 2015; West & Verspoor, 2016). As a result, students who perform well during their school career and up to their final exams still have difficulties or are even unable to communicate in the foreign languages they have learned (Schnabel et al., 2016). One of the explanations for this situation could be a lack of constructive alignment in Dutch foreign language curricula.

According to Biggs (1996), the effectiveness of any curriculum depends on its 'constructive alignment': the coherence, or degree of alignment, between learning objectives, learning activities and assessment practices. In the development of an aligned curriculum, Biggs emphasizes the importance of identifying the goals that students need to achieve in assessments before organizing teaching and learning activities, as tests are known to influence both teaching and learning; this influence is known as 'washback' (Alderson & Wall, 1993; Green, 2007). This washback can be positive when tests are aligned with learning goals, but can be negative when alignment is lacking. In the context of FL teaching, where students have to learn to communicate in a FL (learning objective), this means that assessment and learning activities are relevant (i.e., aligned) when they are communicative in nature. To ensure positive washback (or to prevent its negative counterpart), FL teachers should provide communicative tests in which learners are asked to perform authentic tasks similar to those they encounter in real life (Morrow, 2018).

Brown (2005) identifies five requirements that communicative assessment activities should meet:

1. Meaningful communication: Tasks should be connected to students' experiences and perceptions.
2. Authentic situation: Tasks should be linked to real-life situations.
3. Unpredictable language input: Students should be able to respond to unprepared questions or comments.
4. Free production: Students should be able to show that they can give their own content to a conversation.
5. Integration of language skills: Tasks should encourage students to use different language skills in an integrated way, as it is often the case in real-life communication.

In lower–form secondary education in the Netherlands (students aged 12 to 15), test formats and contents are chosen by departments of individual schools. Foreign language teachers are expected to develop their curriculum and tests in line with the communicative objectives formulated at the national level: students should be able to use different language skills in authentic situations in order to communicate effectively in a FL (College voor Toetsen en Examens, 2017). However, teachers often use tests that come with widely-used textbooks, which tend to focus on assessing vocabulary, grammar, and receptive skills (reading and listening). Productive skills (speaking and writing) are tested less often, and if they are, this is usually done by asking students to translate sentences or reading texts out loud (Fasoglio et al., 2015). Such tests do not adequately introduce real-life unpredictability (Brown, 2005) in the form of unknown or unprepared questions and fail to test communicative competence, i.e. the ability to adapt FL knowledge and skills to new situations.

Considering the washback effect of tests, this could explain the difficulty of implementing communicative teaching in Dutch FL classrooms; if tests are not sufficiently communicative, learning activities will not be sufficiently communicative either. In (partly) decentralized educational systems such as the Dutch one, teachers have a lot of control over teaching materials and assessments which would, in principle, grant them with ample opportunity to generate positive washback (Hakim, 2018). Harding (2014), however, reports that individual teachers face challenges in designing communicative language tests, due to practical concerns and different interpretations of CLT. As a result, assessment tasks and rating scales designed by teachers, if any, are often developed or adapted intuitively and are not always able to validly measure students' communicative competence (Fulcher, 2003). It is therefore important to better understand which practical and conceptual challenges teachers face when designing tests and to identify which decisions they can make to overcome them.

The practical objective of the current design study (part of a larger designed based-research project) was to, collectively with FL teachers, develop a valid communicative assessment program for a full school year (a compilation of formative and summative communicative tests) that could easily be implemented in Dutch low-form education and that would incite positive washback. The scientific objective was to further identify and specify the practical challenges (regarding feasibility) and the conceptual challenges (regarding validity) teachers face while designing this communicative assessment program, and to formulate suggestions to overcome them. As such, the research question guiding the current study is: which challenges regarding feasibility and validity do FL teachers face in the design of communicative test materials, and what decisions can be taken to overcome them?

In this paper we present the first results of this co-design study.

Method

Participants

A group of 21 FL teachers of the most commonly taught foreign languages in Dutch secondary schools (English, French, German, and Spanish) from 15 different schools took part in a Professional Learning Community (PLC) with the aim of learning by design, taking both theoretical and practical aspects into consideration in the selection and (re)design process of a communicative assessment program. The participants were all graduated teachers with varied years of experience ranging from 1 to 25 years. They all taught at least one class in lower form.

Participants of the PLC took part in eight interactive working-sessions of 3 hours each. The sessions were organized and supervised by the researcher (i.e., the first author). During the eight sessions, participants shared their vision on communicative teaching and testing practices, compared their practice with theory on communicative testing, shared their testing materials (as used in their current practices), (re)designed tests, and compiled tests to create a feasible communicative assessment program.

Data collection and analysis

Notes on the proceedings of each PLC session as well as the exchanges, remarks, and questioning of the participants were reported by the researcher in a logbook. The report of each session was submitted to each of the participants individually for a member check to ensure transparency. They could complete, modify and/or confirm the report.

The data were analyzed inductively, according to the steps specific to qualitative research analysis (Corbin & Strauss, 2008). In this case, the codes were short sequences of words that described the challenges participants faced during the design process and the decisions they made to overcome them. Codes were then synthesized and categorized into two themes: practical challenges and decisions (regarding feasibility) and conceptual challenges and decisions (regarding validity).

First results

Challenges and design decisions regarding feasibility

Three challenges regarding feasibility were identified:

1. Lack of materials available for learning activities.
2. Lack of time allocated for test administration and scoring.
3. Students' lack of experience with communicative tests.

In response to each of these concerns, decisions were made within the PLC to enhance feasibility in the design of the test materials. The different suggestions to overcome these challenges are reported below.

Lack of materials available for learning activities

Results from the logbook show that the majority of the participants used a textbook and did not have much time to develop their own learning activities in addition to it. Their first concern in the design of communicative tests was to be able to keep using the learning activities from their textbooks while preparing their students for the tests. Most of the learning activities and tests from the textbooks used are organized thematically (e.g., sport, school, holydays) and focus mainly on knowledge of vocabulary, chunks and grammar rules out of context. Communicative learning activities are included, but do not constitute the main focus.

To overcome this problem, we decided to develop communicative tests that can be adapted to the different themes covered by the textbooks. Test tasks focus on language acts to be performed (such as persuading, describing, giving an opinion, etc.) in accordance to the corresponding CEFR level, and can be adapted to any theme. In this way, all activities about a specific theme in the textbook can still be used in preparation of the test, and teachers only need to put more emphasis on the communicative activities in a specific chapter. This implies that teachers use their textbooks as a tool and no longer as the curriculum as such. It is still crucial to select a textbook that provides enough opportunities for communicative activities.

Lack of time allocated for test administration and scoring

The second concern teachers expressed in the PLC meetings is the time allocated to teachers to administrate and score tests. Communicative testing requires more time. Writing, reporting, and speaking are complex operations, and the assessment of these skills requires a lot of attention and expertise from the teachers. One solution to compensate for the time-consuming nature of communicative testing is simply to test less often and introduce more formative activities in between less frequent summative assessments. Another challenge resulting from this decision was to be able to test the four language skills at the intended level, without creating pressure at the end of the year. To resolve this problem, we chose to test the different skills in an integrated way, in accordance with the communicative approach.

The time spent on administering a communicative test in classes of about 30 students was another concern, particularly regarding the assessment of speaking and conversational skills. We suggested using teaching time to assess speaking skills by organizing and scoring presentations during the lessons. This can be justified by the fact that, while the presentation is a test for some students, it can function simultaneously as a learning activity for the others. Besides, presenting in front of an audience as in real life enhances authenticity. To assess conversational skills, we designed test tasks that could be performed by three students at the same time. Per group students receive the same realistic situation in which each of them has to perform a different task.

Finally, teachers were also concerned about the time needed to score communicative tests. We decided therefore to develop rating scales with a holistic part in addition to an analytical part. The holistic part is short and to the point, and can be scored very quickly to give a first indication of the global CEFR level at which a student performed a communicative act. The analytical part zooms in on the quality of communicative competence. The criteria in the analytical part of the scale are descriptive, based on CEFR Can Do statements. The details of the descriptions allow teachers to score quickly, without having to give extra feedback to justify the score.

Students' lack of experience with communicative tests

The last practical challenge mentioned by the teachers was to keep students motivated with fewer grades and tests that do not focus on reproduction only. Tests assessing FL skills are often perceived by students as being less focused and therefore more difficult to prepare for.

To prevent students from 'hiding' and not working regularly or effectively between the formal assessment moments, we suggested the systematic introduction of formative communicative activities during the lessons.

Challenges and design decisions regarding validity

Three challenges regarding the validity of communicative tests were identified.

1. How to introduce unpredicatibility.
2. How to integrate grammar and vocabulary.
3. How to integrate language skills.

To ensure a higher degree of 'communicative validity' in the design of test materials (in order to stimulate positive washback), we made decisions within the PLC to overcome these challenges.

Introduction of unpredictability

Teachers were concerned by the introduction of unpredictability to enhance validity, because they did not know how to prepare their students to react spontaneously and adequately to unprepared questions and situations. To overcome this challenge, we decided in the design of the test tasks to focus on students' ability to perform speech acts that are present in all kinds of realistic situations (reporting, corresponding, sharing experience and information, expressing opinions, etc.). Themes coming from their textbook (sport, holidays, school, etc.) will be known in advance, but students will have to perform one of the acts within a new situation. This increases unpredictability while still providing sufficient guidance to prepare students.

Integration of grammar and vocabulary

A large number of the participating teachers were still testing knowledge of grammar and vocabulary and chunks out of context. It was challenging for them to discard this type of testing and replace it with the assessment of language skills alone. They were afraid that students would no longer learn enough vocabulary, chunks, and grammar rules if these were not tested directly. We decided to create two categories in the rating scales to assess the use of vocabulary and grammar in the context of a communicative task. In addition, we insisted on the importance of using formative activities addressing the development of knowledge of vocabulary and grammar rules next to communicative learning activities.

Integration of language skills

The last challenge faced by the teachers was the integration of language skills within a test to enhance authenticity. Different combinations were possible for this integration. We chose to integrate reading and writing skills on the one hand and the listening and speaking skills on the other, with a distinction between conversational skill (listening and speaking in the context of a conversation) and listening and speaking skills (listening to gather information and then presenting something about it). This combination appears to be the most common in real life.

Teachers were concerned by the fact that one of the skills could be underrepresented in a test or that students could compensate one skill with another, which would not give a clear idea of the level of each skill independently. In order to enhance validity on that point, we introduced Can Do statements in the analytic part of the rating scales belonging equally to each language skill.

Conclusion

This paper presented the first results regarding practical and conceptual challenges faced by FL teachers in the design of communicative test materials and the decisions made to overcome them.

Practical challenges often related to the limited time allocated to teachers to prepare lessons, and to develop or select communicative learning/assessment activities. This time issue appeared to be even more pressing when the textbook used did not include sufficient communicative learning activities. Other practical challenges were related to tests administration and scoring within the allocated time. Teachers were also concerned by students' motivation, with fewer tests focusing only on language skills.

In addition to these practical challenges, teachers also faced conceptual challenges regarding the operationalization of unpredictability, the integration of grammar and vocabulary within a communicative task, and the integration of different language skills within one assessment.

Decisions made to address these challenges are summarized below in the form of suggestions to be taken into account when designing test materials for communicative learning goals:

1. Language skills should be tested in an integrative way to reduce the amount of tests and enhance authenticity.
2. Test tasks should include new situations to ensure unpredictability, but should be formulated with precise instructions to provide enough guidance to prepare students.
3. Rating scales should assess language skills in an integrated way. They should be descriptive, based on Can Do statements belonging equally to each language skill. They should include a holistic part and an analytic part to save time in scoring and to enable efficient feedback.
4. Formative activities should be systematically introduced to keep students motivated and to guide their learning process.

References

Alderson, C., & Wall, D. (1993). Does washback exist? *Applied Linguistics*, *14*(2), 115–129.

Biggs, J. (1996). Enhancing teaching through constructive alignment. *Higher Education*, *32*(3), 347–364.

Brown, J. D. (2005). *Testing in Language Programs*. New York: McGraw-Hill.

College voor Toetsen en Examens. (2017). *Examenprogramma taal vanaf CE 2017*. Retrieved from : www.examenblad.nl/

Corbin, J., & Strauss, A. (Ed.). (2008). *Basics of Qualitative Research* (3rd edn). Thousand Oaks: Sage.

Fasoglio, D., de Jong, K., Pennewaard, L., Trimbos, B., & Tuin, D. (2015). *Moderne vreemde talen: Vakspecifieke trendanalyse 2015.* Enschede: SLO.

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

Green, A. (2007). Washback to learning outcomes: a comparative study of IELTS preparation and university professional language courses. *Assessment in Education*, *14*(1), 75–97.

Hakim, L. N. (2018). Washback effect in language testing: What do we know and what is its effect?. *Jurnal Forum Didaktik*, *2*(1), 59–68.

Harding, L. (2014). Communicative language testing: Current issues and future research. *Language Assessment Quarterly*, *11*(2), 186–197.

Kissau, S., Rodgers, M., & Haudeck, H. (2015). Practicing what they preach ? A comparison of teacher candidate beliefs and practices. *International Journal of Language Studies*, *9*(4), 29–54.

Larsen-Freeman, D., & Anderson, M. (2011). *Techniques and Principles in Language Teaching*. Oxford: Oxford University Press.

Morrow, C. K. (2018). Communicative language testing. In J. I. Liontas (Ed.), *The TESOL Encyclopedia of English Language Teaching* (pp. 342–350). New Jersey: John Wiley & Sons Inc.

Schnabel, P., Ten Dam, G., Douma, T., van Eijk, R., Tabarki, F., Van der Touw, A., & Visser, M. (2016). *Ons onderwijs 2032: Eindadvies.* Retrieved from: www.rijksoverheid.nl/documenten/rapporten/2016/01/23/eindadvies-platform-onderwijs2032-ons-onderwijs2032

West, L., & Verspoor, M. (2016). An impression of foreign language teaching approaches in the Netherlands. *Levende Talen Tijdschrift*, *17*(4), 26–36.

Can we assess the native speaker level? Interview test: a new exam in Russia

Anzhela V. Dolzhikova
Peoples' Friendship University of Russia

Victoria B. Kurilenko
Peoples' Friendship University of Russia

Abstract

The restrictive measures caused by the Covid-19 pandemic have had an uneven impact on the economic systems of different countries; in these circumstances, it would be logical to assume an intensification of global migration processes in the post-pandemic period. These factors significantly pressure the actualisation of the tasks of improving the mechanisms and instruments of the migration policy of countries traditionally receiving large migration flows, including the Russian Federation. An important factor is to increase the efficiency of organising and conducting interview tests for migrants, which are intended to confirm the status of a native Russian speaker, and provide additional opportunities for integration into Russian society for foreigners who share the mental and spiritual values of the Russian world. Interviewing to confirm the status of a native Russian speaker has been held in the Russian Federation since 2014. Since that time, the first strategies for organising and conducting this integration exam have been developed and implemented. One of these strategies was proposed and substantiated by the authors in 2015. However, the changed conditions for the interviews as well as the methodological and testing experience accumulated over the years make it necessary to revise and correct the structure and content of the interview test and develop a new model that meets the fresh requirements and challenges. With this in mind, the authors of this article set themselves the task of identifying the main factors that influence the organisation and administration of an interview for migrants with a high level of Russian language proficiency, a methodological analysis of the effectiveness of existing test strategies, and the development of a new interview test model.

Introduction

The intensification of migration flows is one of the key characteristics of our time. The restrictions caused by the Covid-19 pandemic have slowed down these processes for a while. However, there is every reason to believe that after the situation normalises, their intensity will not only become the same as before but possibly increase. In this regard, the improvement of migration policy and the development of its mechanisms and instruments are urgent tasks facing countries that have traditionally received large migration flows.

As emphasized in *Conception of State Migration Policy of Russian Federation for the period of 2019–2025*[1], the migration policy of the Russian Federation is aimed at providing and developing 'legal, organisational and other mechanisms . . . ensuring voluntary resettlement in the Russian Federation for permanent residence of compatriots living abroad and other persons who are able to successfully integrate into Russian society; . . . foreign citizens wishing to develop economic, business, professional, scientific, cultural and other ties, to study the language, history and culture of our country'. An effective measure for the solution of these aims is to simplify the examination procedure for persons who speak Russian at a high level and apply for the status of a native Russian speaker. 'Native Russian speaker' is as a matter of fact a government programme addressed to those who live in foreign countries but speak Russian and want to move to Russia for permanent residence. These people take the Russian language exam in a lightweight format, i.e., as an interview.

The practice of conducting an interview test in the Russian Federation was introduced in 2014. The first strategies for organising and conducting such tests appeared in Russia during this period. One of these strategies was proposed and substantiated by the authors of this article in 2015 (Dolzhikova, Kurilenko, Ivanova, Pomortseva, & Kulikova, 2015). Due to the changed conditions for the interviews, it is required to correct its structure and content as well as to develop new models of its organisation. It is also

[1] cis-legislation.com/document.fwx?rgn=52502

necessary to take into account the great methodological and testing experience accumulated over the years by the Russian school of testology.

Literature review

The problem of testing through interviews is relatively new for Russian researchers but it has long attracted the attention of European and North American scientists. To date, researchers in these countries have accumulated a large empirical and theoretical knowledge base, proposed various options for the structure of interviews, models for developing their content, analysed the communicative behaviour strategies for testers, and considered the issues of their methodological training (Douglas, 1994; Fulcher, 1996; Hiple, 1987; Lazaraton, 1996; Malone, 2008; Stansfield & Kenyon, 1992; Surface & Dierdorf, 2003; Wigglesworth, 1997; Young & He, 1998). We used the results of these studies in developing the model proposed in this article.

The topic stated in this article also required the study of academic publications describing the level of a native speaker — the Russian language interview test is specifically aimed at confirming it. As our analysis showed, the term 'native speaker' is widely used in testing practice. Unfortunately, with such a long history of use up to the present time, there is no unified approach to its definition (Chalhoub-Deville & Fulcher, 2003; Coppieters, 1987; Lantolf & Frawley, 1985; Lowe, 1986). The differential characteristics of a native speaker most often include language acquisition in childhood (Birdsong & Molis, 2001; Bley-Vroman, 1990; DeKeyser, 2000; Hyltenstam & Abrahamsson, 2000; Karakaç, 2015). As noted in Davies (2003), sometimes the language is learned by a foreigner as native, but such cases are extremely rare, and this is accomplished with great difficulty. Our experience shows that this factor is particularly important in acquiring pronunciation skills. One can perfectly master the grammar of a foreign language even outside the authentic language environment. A person who lives long enough in the country of the target language can develop an 'actual individual vocabulary', i.e. to learn to use foreign words correctly and appropriately. But the 'native speaker's pronunciation', as a general rule, can be formed only in childhood, when the child's cerebral cortex has plasticity, when the imitative psychological mechanisms necessary for mastering the 'mother' language are active. We would like to mention that similar data can be found in American National Corpus, 2015; Cambridge Advanced Learners Dictionary & Thesaurus, 2015; Karakaç, 2015. In the Russian scientific literature, we found only a few articles devoted to this problem (Ilyicheva, Dubinina, Leifland-Berntsson, & Kulikovskaya, 2019; Kharchenko, 2015). Their analysis also indicates the absence of a unified approach to determining the communicative speech indicators of a native Russian speaker, which could serve as a guideline in determining the content of the interview tests. The questions about the essence of the notion 'native Russian speaker' and the characteristics of this level of proficiency in Russian currently remain open, since the state standard for IV Certification Level of Proficiency in Russian as a foreign language (RFL), corresponding to C2 in the European system, has not yet been developed until now.

The model proposed in this article was developed on the basis of the criteria for a native Russian speaker defined and substantiated by the authors of this article (Dolzhikova, Kurilenko, Pomortseva, Ivanova, & Kulikova, 2015).

The main factors determining the structure and content of the RFL interview test on confirming native Russian speaker status

Let us now consider the factors that determine the content and structure of the interview test for migrants. The first important factor is the legally defined format. The federal law (FZ № 62) stipulates that the interview test for migrants should be conducted as an interview. The Dictionary of the Russian Language defines this term as 'a special conversation on a specific topic (topics), aimed at clarifying certain questions' (1999).

The next important factor is related to the requirements of the Russian state testing system: 1) *authenticity*: the degree of closeness of test situations to the real discursive practice of the contingent of testees; 2) *representativeness*: the principles of selection, the degree of the communicative value of linguistic, speech and communicative materials; 3) *content validity*: the full coverage of linguistic, speech and communicative materials corresponding to the C2 Level of Proficiency in Russian; 4) *construct validity*: the degree of conformity of the testees and the tasks with the modern discursive-cognitive model of the communicative process; 5) *external validity of materials*: the level of ethno- and socio-cultural correctness, the degree of correspondence of the content plan to the Russian realities, in which the testees will have to live and work; 6) *reliability:* the amount of tested skills and abilities in operating means of language, the reasonable number of communication tasks as well as the time allotted for the test, the methodologically justified degree of difficulty in the tasks; 7) *feasibility:* the degree of accessibility of instructions and test content for a given contingent of testees; 8) *effectiveness*: the ability to get the maximum amount of reliable linguistic, speech and communicative information about the testees' competence level over a reasonable period of time, etc.; 9) *consistency and structural integrity*: logical coordination and coherence of the components of the test, as well as the assessment principles and techniques.

Methodological analysis of strategies for conducting interview tests implemented in the Russian Federation

In the process of developing the model proposed in this article, we analysed the strategies for conducting interview tests that have become most widespread in Russia in recent years.

The first strategy includes two stages: the so-called admission stage and a thematic conversation. The admission stage is aimed at controlling the level of lexical and grammar skills of candidates. The second stage is conducted as an interview. The admission stage significantly saves the time and effort of the examination commission members. However, this format contradicts the legislative requirements, according to which the status a native Russian speaker should be confirmed in a 'specially organised conversation'.

We have also observed a different interviewing strategy. The candidates write an essay on a socio-cultural topic, then they participate in a conversation with the members of the examination committee on the content of the writing. Like the first one, this strategy cannot be recommended for widespread use, because it does not meet the requirements of Russian legislation.

Another strategy that has become widespread in recent years has the same drawbacks: reading a work of fiction, followed by a conversation on the content of what has been read. As practice has shown, native speakers with a low level of general education show very poor results only because they simply do not know how to conduct conversations on socio-cultural topics, which, moreover, are not always interesting and understandable to them.

Of the foreign testing interview models, the closest to our goals and conditions is the Oral Proficiency Interview proposed by American testers. While positively evaluating the methodological and technological components of this test system, we, nevertheless, revealed some characteristics that do not allow this model to be fully used in testing to determine the level of a native Russian speaker. The most significant is that it is mostly intended to assess the level of the testees's communicative speech competence, while the Russian interview test is intended to comprehensively control and measure the testees' integration (real or potential) into Russian culture.

The innovative model of the RFL interview test for migrants

The proposed model includes two parts. Part 1 is conducted as a thematic conversation, including problematic questions on culturally and socially significant topics, one of which involves a detailed message. The testees act in certain communicative roles, demonstrating their mastery of communication strategies in the 'dialogue', 'monologue in dialogue' and 'polylogue' formats. In the course of performing this task, the testees must show their level of mastery of skills for informing, persuading, protecting their own viewpoints in dialogue-unison, dialogue-dissonance, etc.

Part 2 includes two communicative situational tasks. The testees receive cards with descriptions of authentic communication situations, in accordance with which they solve the communication problems specified in the task.

Test tasks are developed using authentic language, speech and communicative materials; test situations are as close as possible to the conditions of natural communication among Russian speakers. When the testees conduct the conversation of Part 1, films, fragments of news programs, etc. may be demonstrated. Candidates' answers are assessed using rating scales developed in accordance with the requirements of the Russian state testing system.

Much attention is paid to the organisation of methodological support for testers participating in the interviews. For this purpose, special toolkits were developed for examiners. Information support was organised for the candidates; instructions on preparation and participation in the examination test as well as reference and educational materials were developed. Specialised test materials were developed for persons with disabilities, providing conditions for their passing the test, as specified by the legislation of the Russian Federation. The suggested model was then tested in a number of Russian testing centres. We received positive results and feedback from testers and RFL teachers, which allows us to recommend this model for wider implementation.

Concluding remarks

A number of economic, political, and other factors have provoked intensification of world migrant flows and, therefore have given rise to the necessity of improving the migration policy and the development of its mechanisms and instruments in countries that have traditionally received large migration flows. That's why the renovation of the content and structure of language exams for migrants is one of the urgent tasks of Russian testology and methodology of teaching RFL. The interview test for confirming the status of a native Russian speaker is an important part of the language examination system for migrants.

Our analysis indicated that the main testing strategies for confirming the native Russian speaker status currently implemented in Russia do not completely meet modern requirements and, therefore, need content and structural improvement. The model suggested in this article was successfully tested in a number of Russian testing centres. This allows us to recommend it for wider implementation.

References

American National Corpus. (2015). *Who is a native speaker of American English?*. Retrieved from: www.anc.org

Birdsong, D., & Molis, M. (2001). On the evidence for maturational constraints in second language acquisition. *Journal of Memory and Language, 44,* 235–249.

Bley-Vroman, R. (1990). The logical problem of foreign language learning. *Journal of Linguistic Analysis, 20,* 3–49.

Cambridge Advanced Learners Dictionary & Thesaurus. (2015). www.dictionary.cambridge.org

Chalhoub-Deville, M., & Fulcher, G. (2003). The Oral Proficiency Interview: A research agenda. *Foreign Language Annals, 36*(4), 498–506.

Coppieters, R. (1987). Competence differences between native and near-native speakers. *Language, 63,* 544–573.

Davies, A. (2003). *The native speaker: myth and reality*. Retrieved from: www.multilingual-matters.com

DeKeyser, R. (2000). The robustness of critical period effects in second language acquisition. *Studies in Second Language Acquisition, 22,* 499–534.

Dolzhikova A., Kurilenko V., Ivanova A., Pomortseva N., & Kulikova E. (2015). Russian as a foreign language interview test for Russian Federation citizenship applicants: Structure and content. *Mediterranial Journal of Social Sciences, 6*(4), 93–103.

Dolzhikova A., Kurilenko V., Pomortseva N., Ivanova A., & Kulikova E. (2015). The criteria for determining 'the native Russian speaker's level' in the language testing system for migrants. *Indian Journal of Science and Technology, 8*(27), 1–13.

Douglas, D. (1994). Quantity and quality in speaking test performance. *Language Testing, 11*(2), 125–144.

Fulcher, G. (1996). Testing tasks: issues in task design and the group oral. *Language Testing, 13*(1), 23–51.

Hiple, D. V. (1987). The extension of language proficiency guidelines to the less commonly taught languages. In A. Valdman (Ed.), *Proceedings of the Symposium on the Evaluation of Foreign Language Proficiency* (pp. 59–66). Bloomington: Indiana University Press.

Hyltenstam, K., & Abrahamsson, N. (2000). Who can become native-like in second language? All, some, or none? On the maturational constraints controversy in second language acquisition. *Studia Linguistica, 54*, 150–166.

Ilyicheva, I. Y., Dubinina, N. A., Leifland-Berntsson, L. B., & Kulikovskaya, L. Y. (2019). Priznanie statusa nositelya russkogo yazy`ka: format e`kzamena*. Nauchno-prakticheskij zhurnal*, *1*(13), 45–52.

Karakaç, A. (2015). *Foreign accent problem of non-native teachers of English.* Retrieved from: www.academia.edu

Kharchenko, E.V. (2015). Nositel` russkogo yazy`ka kak ob``ekt filologicheskogo issledovaniya/ E.V. Xarchenko//Vestnik Chelyabinskogo gosudarstvennogo universiteta. *Filologiya. Iskusstvovedenie*, *96*, 104–110.

Lantolf, J. P., & Frawley, W. (1985). Oral proficiency testing: A critical analysis*. Modern Language Journal, 69,* 337–345.

Lazaraton, A. (1996). Interlocutor support in oral proficiency interviews: the case of CASE. *Language Testing, 13*(2), 151–172.

Lowe, P. (1986). Proficiency: Panacea, framework, process? A reply to Kramsch, Schulz, and particularly Bachman and Savignon. *Modern Language Journal, 70*, 391–397.

Malone, E. M. (2008). Research on the Oral Proficiency Interview: Analysis, synthesis, and future directions. *Foreign Language Annals, 36*(4), 491–497.

Stansfield, C. W., & Kenyon, D. M. (1992). The development and validation of a simulated oral proficiency interview. *Modern Language Journal, 76*(2), 129–141.

Surface E. A., & Dierdorf E. (2003). Reliability and the ACTFL Oral Proficiency Interview: Reporting indices of interrated consistency and agreement for 19 languages. *Foreign Language Annals, 36*(4), 507–519.

Wigglesworth, G. (1997). An investigation of planning time and proficiency level on oral test discourse. *Language Teaching, 14*(1), 85–106.

Young, R., & He, A. W. (1998). *Talking and Testing: Discourse Approaches to the Assessment of Oral Proficiency*. Amsterdam: John Benjamins Publishing Company.

With responsibility comes power? Teachers insights from language assessment when given responsibility in the assessment

Goedele Vandommele
KU Leuven, Belgium

Inge Reinders
KU Leuven, Belgium

Abstract

For the last five years, the Certificate of Dutch as a Foreign Language (CNaVT) has been assessing the language skills of young learners of Dutch as a foreign language. As assessment can be a powerful tool to further instruction, school policy and learning, this particular assessment offers teachers insights into their students' language proficiency and provides them with much-needed examples of CEFR-based integrated reading, writing, listening and speaking tasks. In order to further empower the teachers and help them gain assessment literacy, the assessment procedure recently has changed, going from a very centralized examination to an exam in which local teachers are involved in the administration of the exam and in assessing language proficiency. Following this recent change, we conducted a survey looking into the teachers' evaluation of the language assessment and into their perceptions of its impact on their teaching practices and school policies. Survey results show that teachers involved (N=36) gained different insights by the new assessment: in general, they value the contribution of assessment experts in their local assessment practices; they also report new insights with regard to student proficiency and aspects of language instruction and testing. What they exactly take away from their involvement in the assessment procedure, however, appears teacher- and context-dependent. A smaller proportion also finds assessment informative for their school policy.

Introduction

Assessment can be a powerful tool to further instruction, school policy and learning. Through assessment, schools and classes (1) collect relevant evidence of language proficiency; (2) interpret proficiency results; and (3) use this evidence in an efficient and adequate manner (Hill & Mcnamara, 2011).

Classroom-based language assessment research typically looks into the first dimension of Hill and Mcnamara (2011) and focuses on validity issues in the collecting of evidence; Bachman's (2000) overview of test research concludes that even though more and more attention is being paid to the dimensions of interpretation and use, the main focus of test research is on the increasing range of complex research techniques to scrutinize test scores.

The available research looking into the second dimension on the interpretation or use of assessment results typically gives evidence of the impact – often negative – of high-stakes examinations such as teaching to the test (e.g., Menken & Solorza, 2014; Palmer,Henderson, Wall, Zuniga, & Berthelsen, 2016). Regarding the third dimension, recent studies emphasize the importance of assessment and test literacy of teachers and school teams in order for them to use assessment results in an effective manner.

However, competences regarding data and test literacy in regular education are often lacking. Also, by taking away teachers' active role and autonomy through centralized tests, little agency, nor the chance to construct their own understanding of procedures remains, resulting in the limited effect on classroom teaching and learning activities (Riazi & Razavipour, 2011). Research in the Flemish context, for example, regarding the use of the obligatory low-stakes screening of language proficiency, confirms the limited use that teachers and schools make of the screening results, which is often only used to identify the least proficient students in order for them to be remedied outside of regular classes. Few teachers take part in the test administration, are aware of its results, or are involved in actions following the results. Moreover, results are hardly taken into account to plan instruction or scaffolding, nor are they regularly used for those students who are in need of an extra challenge (Vanbuel, Vandommele, & Van den Branden, 2020).

Agency and autonomy for stakeholders in low-stakes centralized assessment

One way to achieve a more qualitative use of central exam results is by giving teachers more agency and autonomy – more 'voice' – in the development, delivery and rating of the exam. As teacher autonomy is closely linked with teacher development and teacher professionalism (Benson, 2007), returning the autonomy to the teacher might lead to a better use of the assessment. In this paper, therefore, we will investigate how the involvement of teachers concerning the first dimension i.e., in the collecting of relevant evidence, may affect the other dimensions – teachers' interpretation and use of the evidence.

To do so, we changed the assessment procedure of a low-stakes, centralized examination of Dutch language proficiency. Rather than having all aspects of the examination procedure managed by a central body, teachers and some other stakeholders such as pedagogical advisors and local policy makers were involved in the administration of the exam and in assessing learners' language proficiency. To achieve this, local stakeholders could provide feedback during the test development process and teachers were trained to administer the exams in a standardized way; in addition, we developed an assessment form and provided an assessor training course in which learner products and standardized assessment of these products were discussed. During the administration and assessment of exams, the CNaVT was always available for assistance. In this way, by giving important stakeholders more autonomy and agency in the assessment procedure, we endeavored to better take into account different stakeholders' local contexts and enlarge the effects of the assessment on the specific instructional context.

Research rationale and research questions

With the potential influence of more agency on the interpretation and the use of test results in mind, we were interested in the information teachers and other stakeholders (pedagogical advisors, local policy makers) gained from a renewed procedure allowing more involvement in the assessment. We formulated the following research questions:

- How do stakeholders perceive being involved in the different aspects of a low-stakes centralized test of Dutch language proficiency?
- What do stakeholders learn from being involved in a low-stakes centralized test of Dutch language proficiency?

More specifically, we are interested in:

- What do stakeholders learn from administering the test?
- What do stakeholders learn from rating learners' performances?

Methodology

Context: Centralized examination of Dutch language proficiency

Stakeholders involved were teachers and teacher trainers that participated in the Dutch as a Foreign Language Youth Evaluation. This low-stakes exam is administered in several countries where Dutch is part of the curriculum for young learners (11 to 18 years): Wallonia (French-speaking part of Belgium), Northern France, Germany, Caribbean Islands. 1,362 young learners participated, evenly distributed among schools in Western Europe and the Caribbean Islands.

Increasing autonomy and agency

To reconcile the demands of reliable assessment with more agency and autonomy on the teachers' part, we followed a strict procedure. More autonomy and agency in this situation meant involvement in three phases of assessment: (1) local representatives familiar with the context (pedagogical supervisors, teacher trainers, local policy makers) were given the opportunity to provide feedback in the initial development phase of the test; (2) teachers received training on the test taking, specifically regarding the oral part of the examination, and afterwards took part in the test administration; (3) finally, teachers themselves were involved in the rating (preferably not of their own students). They were given the necessary instructions and model answers and rating scales, combined with assessor training.

Survey

To probe into teachers' evaluation of involvement in the language assessment and their perception of its impact on teaching practices and school policies, we administered a survey among all participating schools and teachers. We probed into their motivation to participate in the centralized examination, using multiple-choice questions. Besides those questions, we included open-ended questions regarding the insights participants take away from the administration, results and rating of the exam. A translation of the exact questions can be requested from the authors.

Participants

The participants were all teachers in schools that organized the low-stakes exam. 38 schools participated in examination in Western Europe; 22 of their teachers completed the survey (58%). For the Caribbean Islands, 24 schools participated in the examination and 14 teachers completed the survey (58%).

Coding

Participants' multiple-choice answers are visualized in figures in the results section. Participants' short open answers were coded starting from the categories that could be distinguished in the literature review, and then supplemented with categories that emerged from the data itself (Grounded Theory, see Strauss & Corbin, 1994).

Results

1. Reliability of the new procedure in which teachers rated their own students

To ensure reliable assessment, a random sample of 10% of student performances in each school was also assessed centrally. A comparison between teacher judgements and central judgements found a high degree of agreement, which points in the direction of reliable rating by teachers after training.

2. Overview of survey answers on reasons for participation

When participants' (N=35) are asked for their reason to participate in this low-stakes language proficiency examination, a clear picture emerges. Figure 1 shows that two main reasons concern insights into students' language proficiency, and insights into the individual teacher's instruction of Dutch. Fewer participants report reasons that have little effect on teaching practices, such as the certificate or being obliged to participate.

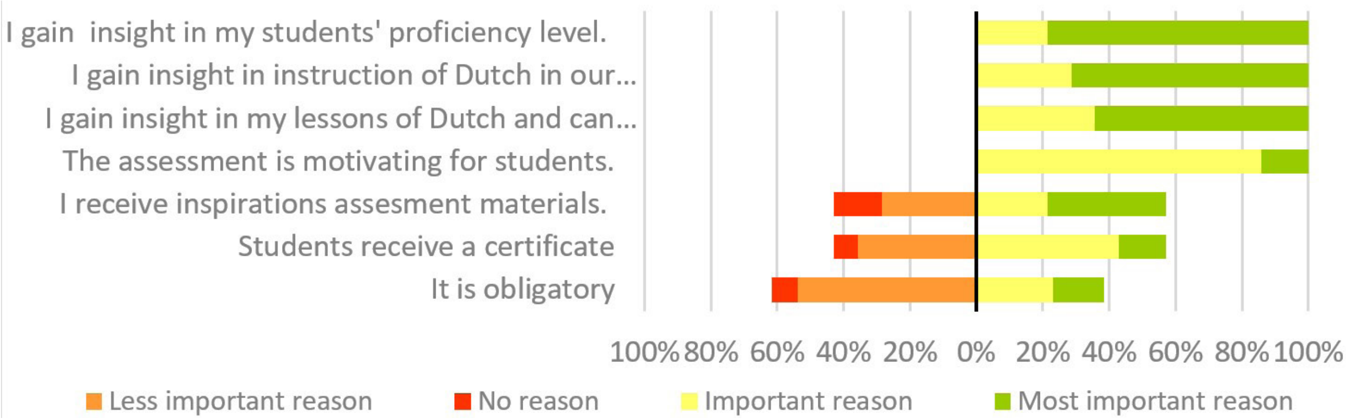


Figure 1 Bar graph of overview of answers on the quantitative part of the survey

3. Reported insights through involvement in the assessment procedure

The coding of the answers on the survey revealed that involvement in different aspects of a central examination led to certain insights into the participants.

All teachers reported gaining insights. On the whole, more insights were reported regarding the effect of involvement in test administration than of involvement in test rating. As few teacher trainers were also involved in the feedback process during test development, we did not specifically probe for effects of this involvement.

The exact insights that teachers gained differed greatly. Teachers reported benefits, insights and questions that came from involvement in test administration and participating in the examination that related to (1) their learners and their competences, (2) qualitative assessment, (3) instruction practices, and (4) school policy measures. Within these categories, participants reported the lowest impact on school policy level.

More specifically, teachers reported the following insights within each category:

- Insights into learners and their proficiency:
  - Insight into instruction and question formulations that can be required for learners to understand
  - Insight into language proficiency levels (high and low)
  - Insight into differential performance on different language proficiency skills
  - Insight into the diversity in language proficiency
- Insights into instructional practices in class:
  - Evidence for small instructional changes: more interaction during classes
  - Evidence for larger instructional changes: balanced instruction of all language skills, lots to learn and to improve
  - Good and concrete examples for teaching, such as the use of audio and images
  - Insights regarding the importance of assessing/teaching all four skills in language assessment
- Insights into qualitative assessment practices:
  - Insights through qualitative or new examples of language proficiency assessment
  - Inspiration/insights regarding question formulation during assessment
  - Insights into (the testing of) receptive skills (listening, reading)
  - Insights into rating practices: e.g., the idea of analytic rating
- Insights into school policy measures:
  - Questions regarding general choices on the school level: requirements regarding teacher competences, learning goals
  - Specific choices on the school level: e.g., changes regarding the school language portfolio
  - Insights regarding the importance of communication to the entire team

The final category, regarding insights gained regarding school policy measures, generated fewer answers than the other categories. However, even though participants did not report many insights or potential changes on the policy level by involvement in the assessment, they do report that impact on school policy level is an important motivation to take part in the examination.

Discussion

In the literature we find evidence that qualitative use of low-stakes language proficiency examinations remains limited. In this small-scale study we highlighted teachers' autonomy and agency in a language assessment procedure and involved them in the process of developing, administering and rating a central examination of Dutch as a Foreign Language. In this way, we attempted to create a basis for the use of such a test in all of the three dimensions of Hill and Mcnamara's (2011) model of assessment in education.

The new procedure was a success and did not negatively impact the reliability of the test. Results of a survey with multiple-choice and open-ended questions probing into teachers' reception of the new procedure for a low-stakes centralized examination revealed that, indeed, teachers that were involved reported insights and 'intentions' that exceed the expectations arising from previous research. Teachers also reported that the insights they gained from their participation in the testing were an important reason for them to participate in the low-stakes exam.

Most importantly, teachers reported insights into students' language proficiency through their students' test performance regarding four skills through the new procedure for centralized language proficiency assessment. Besides use of test results at the student level, teachers also reported insights into their own teaching and instructional practices and their intention to change smaller or larger aspects of this practice. Similarly, but to a smaller extent, teachers reported insights into the assessment of and rating of language proficiency of students that they intended to transfer to their own class and instructional contexts. Finally, teachers reported some changes that they intended to make on a school-policy level, regarding the alignment or implementation of practices. Also on the school-policy level, teachers mentioned insights that needed to be communicated and/or considered with colleagues and/or the larger school team.

All in all, we can conclude from this small-scale study that the involvement of local teachers in a centralized exam may have a positive impact on the insights they take away from such an exam and their subsequent use on the level of learner, class and school, without impacting the reliability of the results. More research and different ways to give agency and autonomy to teachers in this respect provide insights into the exact mechanisms, levers and barriers that can positively influence teachers' use of tests.

Limitations

This small-scale study did not allow for the comparison between test use before and after the procedure was changed. Possibly, therefore, the higher-than-expected test use is not a consequence of the changes in procedure, but inherent to this specific case of a low-stakes exam.

References

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

Benson, P. (2007). Autonomy in language teaching and learning. *Language Teaching, 40*(1), 21–40.

Hill, K., & Mcnamara, T. (2011). Developing a comprehensive, empirically based research framework for classroom-based assessment. *Language Testing, 29*(3), 395–420.

Menken, K., & Solorza, C. (2014). No Child Left Bilingual: Accountability and the elimination of bilingual education programs in New York City schools. *Educational Policy, 28*(1), 96–125.

Palmer, D., Henderson, K., Wall, D., Zúñiga, C. E., Berthelsen, S. (2016). Team teaching among mixed messages: Implementing two-way dual language bilingual education at third grade in Texas. *Language Policy, 15*, 393–413.

Riazi, A. M., & Razavipour, K. (2011). (In) Agency of EFL teachers under the negative backwash effect of centralized tests. *International Journal of Language Studies*, *5*(2), 123–142.

Strauss, A., & Corbin, J. (1994). Grounded theory methodology: An overview. In N. Denzin & Y. Lincoln (Eds.), *Handbook of Qualitative Research* (pp. 273–285). California: Sage Publications.

Vanbuel, M., Vandommele, G., & Van den Branden, K. (2020). *Taalvaardigheid screenen aan de start. De implementatie en implicaties van een verplichte low-stakes taalscreening in lagere en secundaire scholen*. Gent: Steunpunt Onderwijsonderzoek.

Moving in the right direction: Synchronising the assessment components of a high-stakes university presessional EAP programme

Stephen Issitt
University of Birmingham, United Kingdom

Abstract

The integration of task and test design is an important issue (Carless, 2017), especially pertinent to an English for Academic Purposes (EAP) programme for international postgraduates whose entry to UK higher education depends on its successful negotiation. This paper presents an approach to the assessment of readiness for English-medium postgraduate study and describes a reading-into-writing examination which requires candidates to read research texts and explicitly report on them. I outline the assessment programme which serves as a gatekeeping mechanism for postgraduate study, calibrated to individual course entry requirements. I then describe the main features of the exam and present the performances achieved by a 1,000-plus student cohort on two of the components, which indicate a positive correlation. I subsequently discuss the integration of assessment design, showing how students' work was integral to the process. Finally, I consider the test as a conduit for good practice and a barrier to plagiarism.

Introduction: the assessment programme

Our international students who intend to progress onto postgraduate courses of study must pass our programme at the relevant grade level of English proficiency, their competences in subject disciplines being already assessed as satisfactory by their departments. To this end students are accepted for entry according to their level of English as specified by external tests such as IELTS and placed on a course of appropriate length depending on the linguistic demands of their target department. Law students, for example, require a higher score than those studying mechanical engineering and these target scores are calibrated to both IELTS and Common European Framework of Reference for Languages (CEFR, Council of Europe, 2001, 2020) equivalences by matching descriptors across the four skill areas on a best fit basis. The University of Birmingham offers a huge range of postgraduate subjects and the multidisciplinary environment is very much a characteristic of the EAP programme, which explicitly teaches the features of academic English in both general and subject-specific contexts. These features, once taught, form the basis of our teaching syllabus whilst the assessment programme includes them explicitly in the marking descriptors. Other factors which have influenced our choice of assessment methods include the size of the international student cohort, the range of first languages represented (20 different first languages with 85% Mandarin Chinese) and the time available in providing accurate evaluation of over 1,000 students engaging in a four-date entry-level EAP course (April, June, July and August), all finishing on the same date in late September for an October departmental entry. This time compression is compounded by the strict UK Visas and Integration (UKVI) requirements imposing time limits for the completion of documents predicated upon course success. Students are given a conditional acceptance of study (CAS) which is time-limited to the end of the presessional programme and converted to formal registration once the course is successfully completed. The gap between EAP course completion and commencement of the departmental programme is very small, often no more than a week.

Assessment components

There are four assessment components comprising 25% each of the total mark and the first element is writing, which consists of a 2,500-word academic research paper related to the target course of study. This self-selected piece of work is completed towards the end of the programme with tutors offering advice in a similar fashion to that of an MA/MSc supervisor. The key difference being that comment is provided on academic language and not subject content. The features of academic vocabulary, grammar, organisation and register also form the basis of assessment descriptors for this component. The reading-into-writing test is a composite of understanding and elaborating key information used in the academic research paper and a series of unseen academic text-related questions (see the next section). The speaking component is an oral presentation of the academic research paper (ARP), and the listening test consists of a lecture-based series of deconstruction questions designed to assess students' understanding of key points within and the main ideas of an academic lecture. All components are thematically linked and integrated in terms of content. The rationale for this series is at least partly predicated upon the assessment requirements of the students' intended higher degrees, all of which require the production of a 15,000-word dissertation together with related reading activities, and seminar and lecture participation. While the emphasis is very much upon teaching the identified features of academic English required at Master's level, the assessment frame can be said to be congruent with those of UK Masters' programmes as a whole.

The reading-into-writing exam

The reading-into-writing exam has two sections: the first question asks candidates to read two previously unseen subject-neutral academic texts which present broadly opposing or complementary views and invite expression of the extent of agreement. A text is judged to be subject-neutral if its content pertains to the experiences of all the international students in the cohort, irrespective of the target course of study. For example, a suitable text might be one related to the international student experience or the advantages and disadvantages of studying abroad. A text on natural science or literary criticism would be deemed too specialised, not subject-neutral and therefore inappropriate. The second question requires candidates to evaluate an entire academic article to which they have already extensively referred in their academic research paper. Here is an example of the first:

> Some people think that students do not improve their writing over a short time period. Others believe that it is in fact, possible. Which position do you agree with? Support your answer by reference to the texts as appropriate and by consideration of your own experience.

Here is an example of the second:

> Please evaluate your article. Your answer should include:
>
> The contribution it made to your ARP. The strengths and weaknesses of your article. Any unique features it has (e.g. methodology, findings, applications) and any contrast with your other sources.

There are two marking dimensions for each of the two questions designed to reflect an integrative approach to both production and assessment (Chan, 2018) and the following are examples of the highest awarded mark bands of the marking descriptors.

Question one: transformative quality

Full comprehension of source texts. Full reference to source content. Excellent paraphrase of main ideas.

Question one: language quality

A wide range of vocabulary is evident and skilfully used. Complex sentences are frequent and accurately framed. Simple sentences exhibit consistent accuracy. Spelling, punctuation and word class use are very accurate.

Question two: transformative quality

A very good range of reasons for selection of source texts. A very good assessment of the source's contribution to the ARP. A very good assessment of the source's strengths and weaknesses.

Question two: language quality

A wide range of vocabulary is evident and skilfully used. Complex sentences are frequent and accurately framed. Simple sentences exhibit consistent accuracy. Spelling, punctuation and word class use are very accurate.

Integration of assessment design and student performance issues

The ability to read academic texts and report on them I have termed *transformative competence* and this can be considered a key concept underlying the rationale for the assessment format (Asencion-Delaney, 2008). It is also a valuable skill throughout a student's academic career and could be regarded as a marker or index of achievement in the academic field (Shaw & Pecorari, 2013). It implicitly underlies much of written activity in a higher education context and can also be viewed as a 'dynamic ability which interacts with the task demands' (Scardamalia & Bereiter, 1987, p. 142). It is certainly a skill that requires regular honing as progression takes place within the academy. At undergraduate level for example, assessment tends to be exam-based, entailing a lower transformational need, whilst at postgraduate level it is nearly always dissertation-oriented with heavy text interpretation.

At PhD level, with an 80,000-word requirement, transformative competence may be considered even more important, not least in a literature review chapter. Whilst the academic research paper requires students to demonstrate this transformative ability in a dissertation-style context, the reading-into-writing exam requires an explicit demonstration of this competence in a formal examination setting. A key issue therefore is the extent to which the scores obtained on the ARP and the reading-into-writing complement and reinforce each other. This could be termed an issue of synchronisation and there are several factors to consider. Firstly, the composition of the marking teams, which were drawn from the teaching cohort of temporary presessional staff; they were all TEFL-qualified but from different language backgrounds (marking teams were given standardisation sessions and live monitoring to ensure consistency). Secondly, the time compression for the marking programme, which took place over five days at the end of the course; and thirdly, the sheer number of students who had to be graded, processed and placed into various target departments. While these are organisational issues, they certainly created an additional pressure to ensure all marks were accurate. Scatterplot analysis of the student results indicate a positive correlation between the ARP and reading-to-writing scores of *r 0.18*, a *p value* of *0.031* (a relatively low probability of a chance relationship) and *v of 0.35* (a relatively high spread of values). From a statistical perspective, there appears to be no real correlation between the ARP and reading-into-writing scores with the *r* falling well below the 0.7% which usually serves as a marker of significance (Loerts, Lowie, & Seton, 2013). This may indicate that whilst there is a thematic and discoursal link between the two components, the tests may be measuring qualities which are not exactly the same, and this is certainly a topic for further investigation. However, I would suggest the data so far tentatively indicates the ARP and the reading-into-writing exam are reflecting a shared capability, transformative competence, and are, to a certain extent at least, moving in the right direction.

Conclusion

As a practical technique for assessing students' readiness for postgraduate study, the reading-into-writing exam may have a number of strengths and more than partly addresses the areas of concern outlined by Brown and Abeywickrama (2010). It is *practical* (not expensive, stays within time constraints, easy to administer with a time efficient scoring system). It is *reliable* (consistent across administrations, offers clear directions, uniform in terms of rubrics and with unambiguous items). It has *construct validity* (measures what it purports to measure, is empirical, offers useful information about test-takers' ability, and is supported by a theoretical rationale). It is *authentic* (contains natural language, contextualised items, meaningful topics, thematic organisation, and replicates real-world tasks). It provides *washback* (influences teaching, offers chances to prepare, encourages language development and provides conditions for peak performance). By most of these criteria, the test is effective and is easy to apply to other academic contexts. The reading-into-writing test requires the use of texts from an appropriate subject field, which in a multidisciplinary EAP context would be, as discussed, subject-neutral but in other contexts could be more subject-specific. For example, a text from a science or economics journal would be appropriate for a reading-into-writing exam for students studying those disciplines. As such, the test implicitly recognises the contextual variety of academic assignments in both dissertation and examination settings, and can be used effectively in both EAP and subject-specific contexts. It can also be used to check the veracity of assignment submissions in addition to the usual software such as Turnitin (www.turnitin.com). To summarise, this integrated assessment frame may serve to enhance the students' academic English competencies, more or less eliminate the possibilities of obtaining marks by direct plagiarism, and provide a powerful and accurate assessment of the students' competencies in this field, while offering an effective and hopefully enjoyable route into academic life and future success.

References

Asencion-Delaney, Y. (2008). Investigating the reading to write construct. *Journal of English for Academic Purposes*, *7*, 140-156.

Brown, H., & Abeywickrama, P. (2010). *Language Assessment: Principles and Classroom Practices*. London: Pearson.

Carless, D. (2017). Scaling up assessment for learning: Progress and prospects. In D. Carless, S. Bridges, C. Chan, & R. Glofcheski (Eds.), *Scaling up Assessment for Learning in Higher Education* (pp. 3–17). Singapore: Springer.

Chan, S. (2018). Some evidence of the development of L2 reading-into-writing skills at three levels. *Language Education and Assessment*, *1*(1), 9–27.

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 with New Descriptors.* Strasbourg: Council of Europe.

Loerts, H., Lowie, W., & Seton, B. (2013). *Essential Statistics for Applied Linguistics*. Basingstoke: MacMillan.

Scardamalia, M., & Bereiter, C. (1987). Knowledge telling and knowledge transforming in written composition. In S. Rosenberg (Ed.), *Cambridge monographs and texts in applied psycholinguistics. Advances in applied psycholinguistics, Vol. 1. Disorders of first-language development; Vol. 2. Reading, writing, and language learning* (pp. 142–175). Cambridge: Cambridge University Press.

Shaw, P., & Pecorari, D. (2013). Source use in academic writing: An introduction to the special issue. *Journal of English for Academic Purposes*, *12*(2), A1–A3.

Using acoustic features to predict oral reading fluency of students with diverse language backgrounds

Hyunah Kim
Department of Applied Psychology and Human Development, Ontario Institute for Studies in Education, University of Toronto, Canada

Liam Hannah
Department of Applied Psychology and Human Development, Ontario Institute for Studies in Education, University of Toronto, Canada

Eunice Eunhee Jang
Department of Applied Psychology and Human Development, Ontario Institute for Studies in Education, University of Toronto, Canada

Abstract

This study aimed to explore the potential of using acoustic features to predict students' oral reading fluency using machine learning techniques, and compare error patterns of predictions between students with North American accents and those with other regional accents. For 158 Grades 4–6 students' oral reading recordings, 1,581 acoustic features were extracted and used to predict oral reading fluency measured by words correct per minute. The machine-predicted scores were strongly correlated with the human-rated scores (r = .91). The magnitude of prediction error in both absolute and squared terms was not statistically significantly different between the North American and other-regional-accent groups. The findings have promising implications for further use of acoustic features that can substantially reduce time spent on administering oral reading fluency assessment in classrooms while minimizing the bias inherent in existing automated scoring models that rely on yet-to-be reliable automated speech recognition technologies.

Introduction

With rapid development in natural language processing techniques, the role of technological innovations has become particularly instrumental in advancing language assessment (Chapelle & Chung, 2010). Some seminal research contributions have been made to automate the assessment of children's oral reading fluency (ORF) (Bolaños et al., 2013; Duong, Mostow, & Sitaram, 2011), a well-documented, strong predictor of future reading development (Hasbrouck & Tindal, 2006). For example, using speech recognition and machine learning techniques, Bolaños et al. (2013) documented an engine that can automate ORF assessment with words correct per minute (WCPM) scores within three to four words of human scores.

Despite the strong human-machine correlations reported, these automated ORF assessment tools rely on, and are therefore limited by, the accompanying automated speech recognition (ASR) systems. These tools typically have machines transcribe speech data using ASR technologies and then compare the transcription against the original reading text to identify any omissions, insertions, or substitutions, and finally calculate WCPM. However, reports on ASR with children reading from a text have been documented to have word error rates from 10–12% (Cheng, Chen, & Metallinou, 2015; Serizel & Giuliani, 2017). The word error rate appears to further increase for children with diverse language backgrounds, as Hannah, Kim and Jang (Forthcoming) reported it to be 14–18%. Unless adequately trained using a large amount of speech data from diverse individuals, ASR systems are prone to generate errors and even more so for young speakers or speakers with accents (Bolaños, Walsh, Ward, & Cole, 2012; Vergyri, Lamel, & Gauvain, 2010). As such, low ASR accuracy can pose a high risk to valid score interpretation.

To overcome this challenge of relying on yet-to-be sufficiently accurate ASR systems, the current study aimed to explore the potential of using acoustic features to predict students' ORF using machine learning techniques. Specifically, we addressed two research questions:

1. To what extent can acoustic features of oral reading speech from upper elementary students predict their oral reading fluency measured in WCPM?
2. Are the error patterns of WCPM predictions different between students with North American accents and those with other regional accents?

Method

Participants

As part of a larger research project conducted in a linguistically and culturally diverse city in Canada, 158 students in Grades 4–6 (9–12 years old) from three public schools participated in the current study. In all three schools, English was the primary language of instruction. Both gender and grade were evenly distributed among the participants (49.7% female; 37.3% grade 4; 31.0% grade 5; 31.7% grade 6). Fifty-six students (38.1%) were born outside of Canada. Among these non-Canadian-born students, almost half (46.4%) reported that they arrived in Canada at the age of eight or older (grade 2 or later). One native North American English speaker and one non-native speaker coded speech samples based on accent type (North American English vs. non-North American English). Having coded individually (Cohen's kappa = .96), the two coders met to discuss and resolve coding discrepancies. As a result, 28 students (17.8%) were identified as having a non-North American English accent (4.4% of Canadian-born students; 41.1% of non-Canadian-born students).

Speech samples

Students' oral reading recordings were collected through Talk2Me, Jr., an online-based oral language assessment tool (Jang et al., 2018; McCormick et al., 2019). The assessment tool was administered either individually or in a small group of two or three students under a research assistant's supervision. Students were asked to complete (at their own pace) six oral language or cognition tasks, the first of which was the oral reading task. The task asked each student to read aloud a 60-word- long narrative text clearly at a comfortable speed. Although the instructions were provided both in text and orally by the computer, the research assistants ensured that the students comprehended the given instructions before starting the task.

Analysis

Each of the 158 speech samples was transcribed verbatim and double-checked by a group of research assistants. Then, the human-rated WCPM was calculated using the formula below:

WCPM = (the total number of words read correctly) / { (time taken in seconds) / 60 }.

Next, the acoustic features of the speech samples were extracted using openSMILE, an open-source platform for extracting audio data (Eyben, Weninger, Wollmer, & Schller, 2016). Specifically, we used the openSMILE INTERSPEECH 2010 Paralinguistic Challenge feature set that contains 1,582 features. The broad dimensions of the extracted features, which are also called low-level descriptors, include loudness, mel-frequency cepstral coefficients (MFCCs), log mel-frequency bands (log MFBs), line spectral pairs (LSPs), F0 envelope, probability of voicing, jitter, and shimmer (see Alim & Rashid, 2018, for descriptions of some commonly used speech features). For each of these low-level descriptors, certain functionals including the arithmetic mean, standard deviation, skewness, kurtosis, and relative position of maximum and minimum value, are extracted as features (Eyben et al., 2016). These features were extracted using Python (3.7.1) via the openSMILE Python Application Programming Interface. Among the 1,582 features extracted, the total duration of the input was excluded from the predictor set as this factor was used to calculate the human-rated WCPM, which is the outcome variable the machine was asked to predict. As a result, the total number of features used to predict the human-rated WCPM was 1,581.

All machine-learning applications were conducted through WEKA (Frank, Hall, & Witten, 2016). Before training the machine and evaluation algorithms, three transformed datasets were prepared in addition to the raw dataset by applying the following filters: Normalize, Standardize, and AttributeSelection. The Normalize filter normalizes all numeric features to the range of 0 to 1. The Standardize filter rescales all numeric features so that each feature has a mean of 0 and a standard deviation of 1. The AttributeSelection filter selects only the most relevant features in the dataset. When applied to the dataset of this study, this filter created a separate dataset that only includes the selected 72 features.

The machine was then trained to predict the human-rated WCPM as a regression problem using the extracted acoustic features. Twelve commonly used algorithms were applied to the four versions of the dataset (i.e., raw, standardized, normalized, selected features): (1) baseline (ZeroR), (2-4) k-nearest neighbours (IBk) (k = 1, 3, 7), (5-7) decision tree (REPTree) (min = 2, 5, 10), (8) random forest, (9) simple linear regression, (10-11) support vector machine for regression (SVM Reg) (polynomial kernel, exponent = 1, 2), (12) SVM Reg (RBF kernel). Model performance was evaluated using 10-fold cross-validation given the relatively limited sample size (Witten, Frank, Hall, & Pal, 2016). Cross-validation has been reported to result in a less biased estimate of model performance than the simple train/test data split method (Brownlee, 2018). The prediction errors were examined using root mean square error (RMSE) and mean absolute error.

Results

Human-rated WCPM

The mean human-rated WCPM of all 158 students was 134.9 with a standard deviation of 34.8, a minimum of 13.6, and a maximum of 208.2. When examined by accent group, the WCPM of the North American accent group (*NorthAmerica*) was 140.1 on average with a standard deviation of 27.3, while the other regional accent group (*OtherRegion*) had a mean of 111.0 with a standard deviation of 13.6. The difference in mean human-rated WCPM between the two groups was statistically significant, *t*(156) = 4.2, *p* ← .000, suggesting that students with other regional accents had lower oral reading fluency.

Prediction model performance

Table 1 compares the model performance of 48 combinations of four variations of the dataset and 12 algorithms and configurations, using RMSE and correlation coefficient between human-rated and machine-predicted WCPMs. Overall, the machine best predicted the human-rated WCPM when the dataset with the 72 selected features was used. In terms of the algorithm, SVM Reg with the RBF kernel configuration and random forest performed best with the lowest RMSE (13.16 and 15.08, respectively) and highest correlation coefficient (*r* = .91 for both models). The superior performances of these two models were statistically significant compared to other models for the dataset with the selected features at a .05 significance level.

Table 1: Model performance comparison using RMSE and correlation coefficient

Algorithm and configuration	*1,581 features (raw value)*	*1,581 features (normalized)*	*1,581 features (standardized)*	*72 selected features*
ZeroR	*34.34 (.00)	*34.34 (.00)	*34.34 (.00)	*34.34 (.00)
IBk (k=1)	*32.04 (.46)	*31.99 (.47)	*32.04 (.46)	*22.83 (.73)
IBk (k=3)	*28.81 (.53)	*28.82 (.53)	*28.81 (.53)	*19.67 (.82)
IBk (k=7)	*26.82 (.65)	*27.13 (.63)	*26.82 (.65)	*20.04 (.85)
REPTree (min=2)	22.30 (.76)	22.10 (.76)	22.29 (.76)	*21.64 (.78)
REPTree (min=5)	22.70 (.75)	22.19 (.76)	22.66 (.75)	*22.76 (.75)
REPTree (min=10)	*23.78 (.73)	*23.43 (.73)	*23.78 (.73)	*23.59 (.73)
Random forest	19.88 (.89)	19.84 (.89)	20.13 (.89)	15.08 (.91)
Simple linear regression	*25.74 (.68)	*23.78 (.74)	*25.74 (.74)	*25.74 (.68)
SVM Reg (polynomial kernel, exp=1)	17.44 (.85)	17.33 (.85)	17.44 (.85)	*15.88 (.88)
SVM Reg (polynomial kernel, exp=2)	17.20 (.86)	17.15 (.86)	17.19 (.86)	*18.91 (.84)
SVM Reg (RBF kernel)	18.90 (.84)	18.84 (.84)	18.90 (.84)	13.16 (.91)

Note: *Correlation coefficients in parentheses. Asterisks (*) indicate models that performed significantly worse than SVM Reg (RBF kernel) on each dataset variation at a .05 significance level. The best performing models are indicated by shade.*

Error pattern comparison between groups

Table 2 displays the mean and standard deviation of absolute and squared errors for the two best performing models (i.e., SVM Reg with RBF kernel, random forest). When compared between the *NorthAmerica* and *OtherRegion* groups, both absolute and squared errors tended to be smaller for the *OtherRegion* group. Yet, neither of the error values showed a statistically significant difference at a .05 significance level. For both groups, prediction errors were negatively correlated with human-rated WCPM, implying that, regardless of accent, the machine overestimates low performers while underestimating high performers. This negative association between oral reading fluency and prediction error was stronger when using random forest (r = -.72) than SVM Reg with RBF kernel (r = -.48).

Table 2: Two-sample t-test results on absolute and squared prediction errors

	Overall		*North America*		*OtherRegion*		*t-test*		
	M	SD	M	SD	M	SD	*df*	*t*	*p*
SVM absolute error	7.49	8.48	7.93	9.05	5.43	4.63	156	1.42	.157
SVM squared error	127.49	294.58	144.15	321.15	50.11	65.40	156	1.54	.126
RF absolute error	4.19	4.13	4.34	4.13	3.49	4.12	156	0.99	.323
RF squared error	34.54	68.06	35.82	68.26	28.55	68.02	156	0.51	.610

Note: *SVM: support vector machine. RF: random forest*

Discussion

The results suggest that the acoustic features used in this study can predict the human-rated ORF for upper elementary students well. The models using support vector machine or random forest algorithms showed a higher correlation coefficient between the human-machine ORF estimates (r =.91) than those documented in the previous studies that involved ASR-dependent ORF assessment: .78 for Grades 1–4 in Beck, Peng and Mostow (2004) and .75 for Grades 4–5 in Klebanov et al. (2020). More importantly, we found no evidence that supports different model performance for so-to-speak accented speech. Considering the well-documented low ASR accuracy for second language learners (Hannah et al., Forthcoming; Mirzaei, Meshgi, Akita, & Kawahara, 2015), leveraging acoustic features in ORF assessment can be a useful way to minimize the bias inherent in existing ORF assessment tools that fully rely on yet-to-be accurate ASR technologies. The usefulness of acoustic features extracted from oral reading has been supported by recent studies, in which they were employed to predict students' confidence levels (Sabu & Rao, 2020) or detect cognitive impairment in older adults (Nagumo et al., 2020).

Our findings have promising implications for further use of acoustic features as a less biased assessment approach, especially with the ever-increasing linguistic diversity in classrooms. We acknowledge several methodological limitations of this study, including small sample size, unbalanced sample size between two groups, and a single reading passage. Further research is needed to generalize the current study's findings with larger samples, diverse reading passages, and multiple linguistic groups. It is also suggested that future studies focus on how to integrate acoustic features into the existing ASR-based ORF assessment tools so that young speakers and speakers with accented speech are not disadvantaged due to ASR's lower performance towards these populations.

References

Alim, S. A., & Rashid, N. K. A. (2018). Some commonly used speech feature extraction algorithms. In R. Lopez-Ruiz (Ed.), *From Natural to Artificial Intelligence: Algorithms and Applications* (pp. 2–19). London: IntechOpen.

Beck, J. E., Peng, J., & Mostow, J. (2004). Automatically assessing oral reading fluency in a computer tutor that listens. *Technology Instruction Cognition and Learning, 2*, 61–82.

Bolaños, D., Walsh, P. E., Ward, W. H., & Cole, R. A. (2012). *Automatic assessment of oral reading fluency for Spanish speaking ELs*. Retrieved from: www.isca-speech.org/archive/wocci_2012/papers/wc12_040.pdf

Bolaños, D., Cole, R. A., Ward, W. H., Tindal, G. A., Hasbrouck, J., & Schwanenflugel, P. J. (2013). Human and automated assessment of oral reading fluency. *Journal of Educational Psychology, 105*(4), 1,142–1,151.

Brownlee, J. (2018). *Statistical Methods for Machine Learning: Discover How to Transform Data into Knowledge with Python*. San Juan: Machine Learning Mastery.

Chapelle, C. A., & Chung, Y. R. (2010). The promise of NLP and speech processing technologies in language assessment. *Language Testing, 27*(3), 301–315.

Cheng, J., Chen, X., & Metallinou, A. (2015). Deep neural network acoustic models for spoken assessment applications. *Speech Communication, 73*, 14–27.

Duong, M., Mostow, J., & Sitaram, S. (2011). Two methods for assessing oral reading prosody. *ACM Transactions on Speech and Language Processing, 7*(4), Article number 14.

Eyben, F., Weninger, F., Wollmer, M., & Schller, B. (2016). *openSMILE by audEERING, open-source media interpretation by large feature-space extraction*. Retrieved from: www.audeering.com/opensmile/

Frank, E., Hall, M. A., & Witten, I. H. (2016). The WEKA workbench. Online appendix for *Data Mining: Practical Machine Learning Tools and Techniques* (4th ed.). Retrieved from: www.cs.waikato.ac.nz/ml/weka/Witten_et_al_2016_appendix.pdf

Hannah, L., & Kim, H., & Jang, E. E. (Forthcoming). *Investigating linguistic and contextual sensitivity in automated speech recognition software: Implications for use in language assessment.*

Hasbrouck, J., & Tindal, G. A. (2006). Oral reading fluency norms: A valuable assessment tool for reading teachers. *The Reading Teacher, 59*(7), 636–644.

Jang, E. E., Sinclair, J., Lau, C., Vincett, M., Kim, H., Larson, E. J., Barron, C., & McCormick, S. (2018, April). *Promoting young readers' self-regulated literacy competence development through scenario-based literacy assessment design* [Conference presentation]. Paper presented as part of Scenario and simulation-based assessments: Interplay between cognition and assessment symposium at the 2018 Annual Meeting of American Educational Research Association, New York, USA.

Klebanov, B. B., Loukina, A., Lockwood, J., Liceralde, V. R. T., Sabatini, J., Madnani, N., Gyawali, B., Wang, Z., & Lentini, J. (2020, March). Detecting learning in noisy data: The case of oral reading fluency. In Association for Computing Machinery (Ed.), *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge* (pp. 490–495). Retrieved from: dl.acm.org/action/showFmPdf?doi=10.1145%2F3375462

McCormick, S., Kim, H., Sinclair, J., Lau, C., Vincett, M., Barron, C., & Jang, E. E. (2019, March). *Talk2Me Jr: A pre-diagnostic digital language and literacy assessment tool* [Conference presentation]. Assessment demonstration presented at the 41st Language Testing Research Colloquium, Atlanta, USA.

Mirzaei, M. S., Meshgi, K., Akita, Y., & Kawahara, T. (2015). *Errors in automatic speech recognition versus difficulties in second language listening*. Retrieved from: files.eric.ed.gov/fulltext/ED564209.pdf

Nagumo, R., Zhang, Y., Ogawa, Y., Hosokawa, M., Abe, K., Ukeda, T., Sumi, S, Kurita, S., Nakakubo, S., Lee, S., Toi, D., & Shimada, H. (2020). Automatic detection of cognitive impairments through acoustic analysis of speech. *Current Alzheimer Research, 17*(1), 60–68.

Sabu, K., & Rao, P. (2020). *Automatic prediction of confidence level from children's oral reading recordings*. Retrieved from: www.ee.iitb.ac.in/course/~daplab/publications/2020/ks-pr-is2020.pdf

Serizel, R., & Giuliani, D. (2017). Deep-neural network approaches for speech recognition with heterogeneous groups of speakers including children. *Natural Language Engineering, 23*(3), 325–350.

Vergyri, D., Lamel, L., & Gauvain, J. L. (2010). *Automatic speech recognition of multiple accented English data*. Retrieved from: www.isca-speech.org/archive/archive_papers/interspeech_2010/i10_1652.pdf

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data Mining: Practical Machine Learning Tools and Techniques* (4th ed.). Burlington: Morgan Kaufmann.

.

How to think of a multilingual resource for evaluation of language interactions with deafblind children

Alina Khokhlova
Yaseneva Polyana Resource Center Supporting Deafblind and Their Families, Moscow State University of Psychology and Education, Russia

Cedric Moreau
Grhapes EA 7287 – INSHEA – UPL, France

Abstract

The article describes the methodology of developing meaningful communication of parents with deafblind children and children with severe multiple disabilities (MD) through the approach of storytelling. The results of the survey of parents raising children with hearing and visual impairments allow conclusions about the practical need to create a method of storytelling with children to be drawn. The developed method is described through the example of a step-by-step guide for working on the first story: www.deafblindacademy.ru/resource-center/proekt-storitelling (Russia) and ocelles.inshea.fr/ (France)

Introduction

More than anyone else, parents of children with MD understand the importance of natural communication with their child; they dream of being able to tell their child about the world, about people, interesting events and phenomena. They want to be sure that their stories are accessible to their children, and they want to understand what their non-speaking child wants and can say.

Nowadays, many educators and parents know about alternative and augmentative communication and use them in education and everyday communication. But natural parent-child communication includes not only instructions, questions and answers, but also reading fairy tales and telling stories. This article describes technology designed for teaching children and self-learning parents how to tell stories using linguistic and alternative means of communication, and with psychological and pedagogical support.

The process of personal development is conditioned by appropriation of the content of culture in which a more experienced person serves as an intermediary for the less experienced; in particular, the adult acts as a link between the child and cultural experience, helping the child to figure out new knowledge. Meanings and symbols systems play a special role in the process of appropriation of new knowledge (Vygotski, 1997). The development of a child with conditions of deafblindness and MD requires a lot of active support of caregivers, understanding what cultural content is accessible to the child at each stage of their development, and what symbols systems can help convey it.

Development of daily living skills

According to I. A. Sokoliansky's approach to teaching deafblind children, education begins with an adult engaging the child in active interaction. In the process of joint activity, a deafblind child learns about the objects around them, and the teacher tries to develop the everyday skills necessary to meet the needs of life. The child learns how to use a spoon, a plate, how to sit on a chair, at the table, how to lie on the bed, put his head on a pillow, and cover himself with a blanket. Then the child generalizes means of action with objects of a certain kind and tries to repeat them independently (Basilova (Басилова), 2015; Sokolyansky (Соколянский), 1989).

Development of communication

Teaching means of communication should begin with gestures indicating objects and associated actions already familiar to the child. The next stage concerns the development of a child's verbal speech, when the gestures that have been learned are replaced by fingerspelling and Braille (Sokolyansky (Соколянский), 1989).

Development of reading activity

I. A. Sokoliansky developed a system of educational texts. The first texts are composed according to the following rules: they are devoted to situations from a child's everyday life, consist of three to five simple sentences, all objects present must be named (Goncharova (Гончарова), 2018).

Children with sensory impairments and MD experience certain difficulties in mastering oral and written speech, but stories of similar content and built on the same principles can be accessible to children who cannot master reading skills, using gestures as alternative communication.

Cuxac distinguishes two types of narration: narration without an iconic component (intonation, imitation of sounds, movements, facial expressions, and other characteristics) and narration that includes '"illustrating" the situation' (Cuxac, 2007).

According to the theory of iconicity, some structures of sign language, called 'transfers', seem to be very similar to natural gestures during verbal narration and sometimes resemble 'drawing' in three-dimensional space. Thus, sign language makes it possible to convey the shape, size, spatial and temporal characteristics of a situation.

These structures of sign language can be considered as a 'bridge' between non-linguistic (objects, representative objects, tactile pictures) and linguistic means of information transmission to children with sensory impairments and MD.

Evaluation of the practical need for storytelling technology with deafblind children

Thirty-five parents (33 mothers and two fathers of children with sensory impairments and MD) completed a questionnaire containing 14 questions addressing three main themes:

1. Features of child-parent communication in the family (means of communication, topics for communication).
2. Practice of reading or telling stories in a family.
3. Parents' ideas and wishes about new tools, topics, and resources that could help information-sharing with the child.

Briefly, the results of the survey are summarized as follows:

- Most parents read or tell stories to their children every day, but they don't know what the child understands.
- They understand storytelling mainly as verbal speech addressed to the child, and they 'tell' the child, even though they are sure that the child does not understand the content of the story.
- All parents recognize the importance of this activity for different aspects of a child's development (intellectual, emotional, communicative).
- Most parents need themes for new stories.
- All parents emphasize the importance of stories about their children's daily lives.
- All parents, with the exception of those who have children with severe motor disabilities, talk about the possibility of using sign language (including tactile ones) in communication with their children.
- Most parents mention the tactile book as a tool to explain story content to their child.
- Most parents would like to work together to create a story book for their child.

Goals and target groups

Based on the theoretical background, the experience of teaching children with deafblindness, and the results of a parental survey, the goals of developing a methodology of storytelling were figured out. The overall goal is to create the conditions for natural and meaningful communication. In addition, separate goals for children and parents can also be named.

For children:

- Organization of meaningful communication with close adults
- Development of the symbolic function
- Development of the wish and ability to understand the content of stories
- Development of the 'storytelling' skill

For parents:

- Understanding the process of development of communication and language skills in children
- Understanding ways of sharing information with children
- Focus on positive and meaningful interaction with the child

According to these goals, parents and educators are offered an online learning resource that allows them to tell stories and discuss everyday situations with children. Because of the difficulties of verbal communication of children with MD, it uses speech, signs and alternative communication, and describes their step-by-step introduction in discussing each story with the child.

Storytelling methodology

The resource includes a video with the story in sign language; a draft of the tactile page, which parents design on their own with help of instructions; a detailed step-by-step guide for teaching a child to understand the content of the story; and videos with examples.

The basic principle in preparing the story materials is the accessibility for the child's perception and understanding. Accessibility for perception is provided by the simplicity of the image, contrasts, and selection of colors. The accessible level of symbolization is provided by the stages of work with the child from a real situation, through playing with a doll and toys, to the understanding of a flat tactile image. The same verbal and sign explanations are used at each stage. Accessible content is provided by several rules:

- The first stories deal with everyday life situations and a child himself is the hero of the story
- All objects present must be named
- The text of the story contains no more than five phrases consisting of two to three words or signs

Example

The first story is 'Meal'. It is studied with the child in three stages: living the real situation accompanied by a short story in sign language, playing the story with a doll and toys in order to symbolize the situation at the level of representative objects, and studying the same story in a tactile book.

To assess a child's progress in understanding the content of a story, we can distinguish four levels of skill development:

1. The child understands the content presented by the tactile page. This means that they start to manipulate it themselves.
2. The child begins to realize that your story is relevant to the page, then tries to repeat some of your signs, and vocalizes after your words.
3. The child begins to realize that they can also tell a story on a tactile picture, and then reproduces parts of the story.
4. The child reproduces the story independently.

It is important to realize that if a child only demonstrates a Level 1 skill and appropriately manipulates the details of

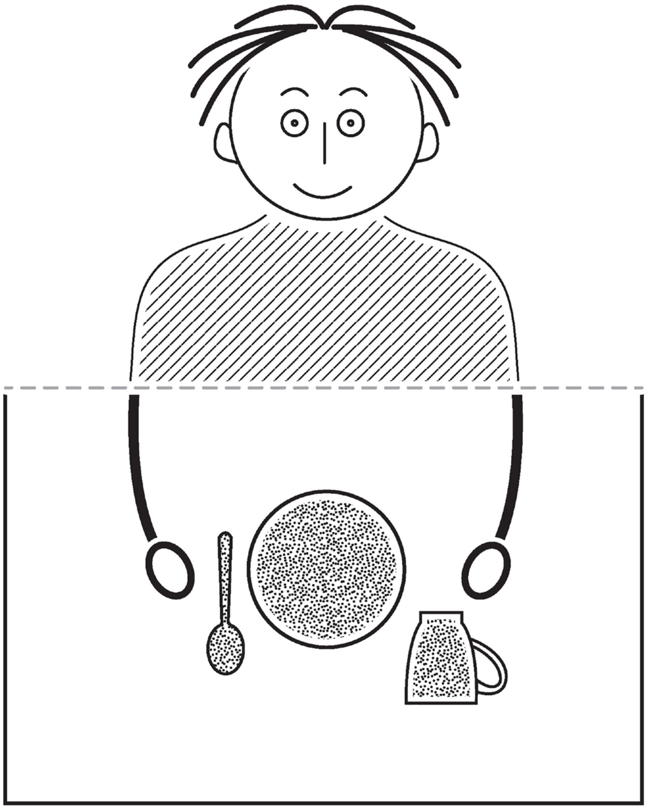


Figure 1 Layout of tactile page; a simple black and white drawing of a child at a dinner table

the tactile picture, we can move on to the next stories, and the connections between the tactile symbols and language tools will develop later, as the child's experience with the stories grows.

Four children with varying degrees of hearing and vision impairment and additional developmental disorders at the age of 8 to 13 years participated in the testing of the methodology. Work with each of the children on storytelling was conducted for five weeks with the frequency of once a week. In all cases all three stages were passed. As a result, one of the participants had a Level 1 result, two had Level 2 results, and one child had a Level 3 result.

Videos of work with these children are included with their parents' permission in the online tutorials and are available on deafblindacademy.ru and ocelles.fr in Russian English and French. All interested parents and teachers can also find the developers' contacts on the websites.

Conclusions

1. The proposed methodology is aimed at meeting the families' request for development of meaningful and natural communication with their children who have deafblindness and multiple disabilities.
2. The methodology is based on classical approaches to training children with deafblindness, but it presents how to use multimodal tools of communication for successful information exchange with a child with deafblindness and multiple disabilities.
3. The testing of methodology with four children with deafblindness and multiple disabilities had shown that all children learned to understand the content of the story but demonstrated different levels of story reproduction skills.
4. The guidelines for parents and teachers on how to use the methodology are available online.

References

Basilova (Басилова), T. A. (2015). *The history of teaching of deafblind children in Russia. (*История обучения слепоглухих детей в *России.).* Moscow (Москва): Eksmo (Эксмо).

Cuxac, C. (2007). Une manière de correuler en langue des signes française. *La linguistique*, *43*(1), 117–128.

Goncharova (Гончарова), E. A. (2018). *Teaching reading and reader's development: a scientific publication.* Moscow: National Education.

Sokolyansky (Соколянский), I. (1989). Training for deafblind children (Обучение слепоглухонемых детей). *Дефектология*, 75–84.

Vygotski, L. S. (1997). *Pensée et langage*. Paris: La Dispute.

Mediation as described in the CEFR Companion Volume (CEFR CV) and reflected in an existing oral examination

Elaine Boyd
UCL Institute of Education, University College London, United Kingdom

Neus Figueras Casanovas
Universitat de Barcelona, Spain

Abstract

The paper reports on the results of a small-scale study which explored how far mediation competences, as described in the Common European Framework of Reference for Languages Companion Volume (CEFR CV) (2020), could be identified at Levels B2 and C1 in Trinity College London's Graded Exams in Spoken English Suite (GESE). The study used an alignment between a selection of the mediation descriptors and the competences assessed in the relevant GESE grades as a starting point and then explored the Trinity Lancaster Corpus of Spoken English for data which might exemplify the identified descriptors.

The results of the study indicate that the tasks within the GESE exams do allow test takers some opportunity to demonstrate aspects of mediation as described in the CV but that there are perhaps insufficient instances to draw significant conclusions. However, the research revealed some interesting aspects about mediation which could be investigated further, principally some uncertainties about the use of mediation by L2 speakers as suggested in the CEFR CV, which may be overestimated, especially at lower levels.

Introduction

Following the impact of the CEFR in the field of language testing and assessment after its publication in 2001, it is to be expected that the CEFR CV will have a considerable impact on assessments offered by reputable exam organisations, especially in relation to the new descriptor scales for mediation and for pluricultural and plurilingual competence. The purpose of the study was to investigate how far mediation was present in the GESE (www.trinitycollege.com/qualifications/english-language/GESE), a test that prizes successful communication above all and focuses on real communication and, specifically, spontaneous interaction. The specific focus of this work has been on the specifications for GESE Grades 7 and 10 (formally linked to CEFR Levels B2 and C1), and on test taker performances at these levels in the Trinity Lancaster corpus (www.trinitycollege.com/about-us/research/Trinity-corpus).

In order to accomplish this purpose, the researchers engaged in the following activities:

- to identify the CEFR CV mediation scales and descriptors relevant to the construct of speaking
- to analyse GESE specifications to find evidence of where mediation might be expected
- to identify the language exponents which illustrate some relevant mediation descriptors
- to analyse corpus data to find evidence of mediation exponents following the coding scheme based on the results from the above.

Challenges

The challenges that the study faced when addressing the tasks outlined above were already suggested in the reports to the EALTA CEFR SIG (Little, 2018) and the EALTA-UKALTA Seminar (Dunlea et al., 2020) held in London on 7–8 February 2020.

On the one hand, and unlike the case with the original CEFR, no guidance or notes for users are available in the CEFR CV and there is still very little research into mediation as understood in the CV, and very few publications on how language mediation is used and can actually be learned, taught and assessed.

On the other hand, whereas the many new scales and descriptors are a welcome tool for designing curricula and teaching programmes, this is not the case in standardized tests even if they thus may become more locally relevant. The introduction of mediation as described in the CV in standardized tests can only come after revisiting existing test constructs and stakeholders' needs, with careful consideration of possible measurement issues.

In fact, the CEFR CV itself – as was the case with the CEFR, first published in 2001– highlights the need to 'interpret' and 'adapt' the scales in relation to context: 'one cannot in practice completely separate types of mediation from each other. In adapting descriptors for their context, therefore, users should feel free to mix and match categories to suit their own perspective' (2020, p. 91).

This is only logical, as mediation cannot be considered a 'stand alone' feature but rather an additional property of human communication, and not all speaking exams have tasks which require group or paired interaction.

In the context of GESE, and in any other standardized exam context, the 24 mediation scales needed careful analysis in terms of context relevance and usefulness, not only in relation to cross-linguistic mediation vs. intra-linguistic mediation but also in relation to scales in the 2001 publication to identify possible conceptual duplicities in the scales featuring interaction (e.g. *Informal discussion* (with friends) in the CEFR vs. *Facilitating collaborative interaction* in the CEFR CV).

Given the limited research on mediation, the researchers decided to document in detail the process followed in order to facilitate future replication(s) of the present study, i.e. the steps taken to identify the relevant descriptors and, having done that, to explore and describe how such descriptors would likely manifest linguistically before tagging any evidence found in the corpus.

The methodology in this project combined qualitative and quantitative approaches. A sequential, exploratory design (Creswell & Plano Clark, 2011) was employed, using qualitative data collection and analysis as a starting point for a quantitative study. Qualitative methods were used when analyzing documentation to identify relevant features, when examining transcripts to identify salient language elements or when searching corpus data to ascertain test-taker's communication purpose(s). Quantitative methods were used to calculate the presence of selected language exponents or exchanges in the corpus and their potential significance.

The Centre for Corpus Approaches to Social Sciences (CASS) at Lancaster University, UK, facilitates access to the corpus through a search platform called Sketch Engine. This enables a variety of search methods and the one utilised for this study was by concordance. All the exponents or sequences identified were searched using the Key Word in Context (KWIC) facility where the query probe is highlighted in the centre of the context. The search was refined through filtering so that the query resource were exemplars that were produced by test-takers at the selected grade, creating a sub-corpus (see Table 1) each for Grades 7 and 10. The exponents were then further searched by task type and by score bands awarded in the test (A, B and C) in order to explore any potential differences generated by the variance in exam tasks and/or test taker performances. The raw frequencies which resulted formed the query outcomes, which were then noted by hand and checked by both researchers before analysis.

Table 1. Number of candidates and tokens in the sub-corpus searched

	Grade 7	*Grade 10*
Total candidates	471	211
Total tokens	619,685	475,813

Results

The results reported in this section are based on the completion of the tasks outlined in the Introduction which required two main phases. Firstly, the careful study and analysis of the GESE specifications at Grades 7 and 9 and of the CEFR CV mediation descriptors, followed by the identification of possible linguistic exponents (identified in a deductive approach) which were summarized in a GESE Mediation Framework. Then, the outcomes of such a Framework were used to search the corpus for evidence(s) of the identified linguistic exponents and to study their implications.

Relevance of the CEFR CV descriptors for the assessment of speaking

The researchers found the explanatory text and the key concepts introducing the scales in the CEFR CV very useful to unpack the content in the descriptors and to relate them to the GESE specifications. The following scales were found relevant at both levels:

- Overall mediation
- Mediating concepts (facilitating collaborative interaction with peers, collaborating to construct meaning, managing plenary and group interaction)
- Mediating communication (facilitating pluricultural space)
- Strategies to explain a new concept (linking to previous knowledge, adapting language and breaking down complicated information)

The descriptors relevant for GESE Grade 7 are at the B2.1. level, and for GESE Grade 10 are at C1. The scale *Processing text in speech* was also found relevant for Grade 10 in GESE's formal topic presentation phase.

Linguistic identification of the selected CEFR CV descriptors

The explanatory text which outlined the key concepts introducing the scales was used as a resource as this listed possible linguistic exponents illustrating the descriptors. The researchers also drew on the literature specifically in the field of corpus analysis (for example, Gablasova & Brezina, 2015) and on their professional expertise in the field to identify exponents, but no doubt other exponents, and even additional ones, might be selected.

Two main types of language mediation exponents were identified: (a) those helping to mediate concepts by using various exponents to explain and elaborate meaning, labelled as 'exemplification', and (b) those contributing to mediate communication and build meaning, labelled as 'collaboration'.

(a) For exemplification, specific language exponents were selected for investigation, namely:

for example, like (preposition), *such as, for instance*. *Similar to* was also explored but the instances were too few to warrant inclusion.

(b) For collaboration, the set of sequences selected for investigation were: *Yes but . . .., Yes and . . ., Yes, also . . ., Yes, in fact . . .*

In addition, collaboration was further investigated by looking at linguistic repetition, specifically where test takers repeated a word uttered by the examiner. Repetition can be a communication strategy used to create a 'tie' and build a rapport with the interlocutor to assist communication and help unfold discourse.

Evidence(s) of mediation in the GESE exam at Grades 7 and 10

Results – albeit not statistically significant – show that there is, in fact, evidence of mediation in test takers' performances at both grades. Table 2 shows the data from the query outcomes for exponents of exemplification in the sub-corpus.

For illustration purposes, some of the instances of exemplification by candidates identified in the study are included below.

Table 2: Total tokens for exemplification (for example, like, such as, for instance)

All tokens	*G7*	*G10*
Total	6.220	7.137

G7

*L17: another type of drugs are erm the **for example** heroine morphine and others*

*G18: it causes erm very bad erm problems to our health **for example** they have hallucinogenic symptoms*

G10

*L02: in a house they have to follow some rules and **for example** if you give one opinion you have to raise your hand*

*L60: we buy things in packages which are not necessary **for example** when we go to the supermarket*

However, no evidence was found of the test takers' supposed marked evolution across levels as expected in the CEFR CV: 'Progression up the scale is characterised as follows: at B1 and B2 the emphasis is on providing repetition and further examples, whereas at the C levels the focus is more on elaboration and explanation, adding helpful detail' (2020, p. 127).

There are some – statistically non-significant – differences across levels but no confirmation of progression in the use of mediation language exponents, and it does appear that instances of mediation might be less prevalent than the exam construct would anticipate. In the next section, reasons for the results are explored and suggested.

Conclusion

This exploratory study successfully identified some aspects of mediation that occur in the spoken exam setting of GESE (Grades 7 and 10). The two aspects which were investigated – exemplification and collaboration to build meaning – were the two key components identified in the alignment process between GESE and the CEFR CV mediation descriptors.

The initial part of the study illustrated that there was a compelling alignment between the GESE performance descriptors, the mediation competences which would allow those to be demonstrated by the test taker and the potential exponents of those competences. This was followed by the evidences of mediation in the corpus analysis manifested by common exponents such as *for example, for instance, like*, and so on, and by numerous instances of repetition (both within and between turns).

The increased facility with the language expected of higher level learners probably explains the higher relative frequency of the exemplification exponents identified in Grade 10, and this may also be due to some task effect as more complex topics and concepts, which are the focus of the higher grade, may need better elaboration. However, it may also be that test takers feel more comfortable with spontaneous interaction at this level, less focused on simply the 'message' and thus better able to elaborate. Interestingly, the raw data suggests that Grade 10 test takers seem to use collaboration dimensions less and this may be because they are relying on more sophisticated pragmatic features of interaction and mediation. The Grade 7 results are more tentative and, while it certainly appears that the GESE gives test takers the opportunity to demonstrate their mediation competences, this is not necessarily happening with even the most overt and frequently taught exemplification exponents (e.g. *for example*). The relatively low occurrence of the selected mediation features may indicate that, even at B2 level, learners struggle with utilising these features of the language and are perhaps still focused on delivering their message rather than managing interaction and/or the understanding or interpretation of their message.

In summary, the study, via the GESE Mediation Framework, did conclude that the GESE exams allow test takers the opportunity to demonstrate their mediation competences and that this facility is expressed in the performance descriptors. However, what the results suggest is that the descriptors in the mediation scales in the CEFR CV might expect too much at too low a level and also that the micro-progression it defines is not evident in L2 speakers.

As ever with exploratory studies, this has potentially raised more questions than it has answered in terms of the GESE suite. For example:

- What is the balance of mediation 'language' between examiner and candidate? It is challenging to untangle the variables of role and proficiency, but a wider study could offer insights here.
- Following on from the above, do examiners use mediation more, e.g. in supporting test-takers?
- How does this impact the test-taker's performance?
- What prompts the test taker to use mediated language, i.e. what comes before the use of mediation. Is it e.g. examiner query, silence, body language?
- What becomes evident if we look at longer chunks of discourse? Would a qualitative study of this yield interesting insights?
- How do age, L1 and culture affect the linguistic manifestations of mediation? And what are the implications for any differences?

Despite the relative newness of the mediation descriptors, and the consequent lack of an established methodology to identify mediation in learner language and the lack of a tradition of explicitly including mediation in language assessments, this study has revealed both the intention (in the GESE specifications) to elicit mediation and a slim vein of mediation in GESE exam transcripts. There is also an indication that the descriptors are perhaps not as well aligned to the reality of learner language as they might be.

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 with New Descriptors*. Strasbourg: Council of Europe.

Creswell, J.W., & Plano Clark, V. L. (2011). *Designing and Conducting Mixed Methods Research*. California: Sage.

Dunlea, J., Figueras, N., Folny, V., Little, D., O'Sullivan, B., & Patel, M. (2020). *The CEFR: towards a road map for future research and development*. Report on a seminar sponsored by EALTA and UKALTA and organized by the British Council. Retrieved from: www.ealta.eu.org/resources.htm

Gablasova, D., & Brezina, V. (2015). Does speaker role affect the choice of epistemic adverbials in L2 speech? Evidence from the Trinity Lancaster Corpus. In J. Romero-Trillo (Eds.), *Yearbook of Corpus Linguistics and Pragmatics Volume 3* (pp. 117–136). New York: Springer.

Little, D. (2018, January). *The CEFR Companion Volume with New Descriptors: Uses and Implications for Language Testing and Assessment* [Conference presentation]. VIth EALTA CEFR SIG, Trinity College Dublin.

Further references

These references were consulted and used in the study reported, and were included in the full report commissioned by Trinity College London. Readers interested in the topic may find them useful.

Boyd, E., & Taylor, C. (2016). Presenting validity evidence: the case of the GESE. In J. V. Banerjee & D. Tsagari (Eds.), *Contemporary Second Language Assessment Volume 4* (pp. 37–60). London: Bloomsbury Academic.

Brezina, V. (2018). *Statistics in Corpus Linguistics.* Cambridge: Cambridge University Press.

Coste, D., & Cavalli, M. (2016). *Education, mobility, otherness The mediation functions of schools.* Retrieved from: rm.coe.int/education-mobility-otherness-the-mediation-functions-of-schools/16807367ee

Council of Europe. (2009). *Manual for Relating Examinations to the CEFR*. Strasbourg: Council of Europe.

De Jong, N. H. (2018). Fluency in second language testing: Insights from different disciplines. *Language Assessment Quarterly*, *15*(3), 237–254.

Gablasova, D., & Brezina, V. (2014). *How to communicate successfully in English? An exploration of the Trinity Lancaster Corpus*. Retrieved from: cass.lancs.ac.uk/wp-content/uploads/2015/02/08-CASS-Trinity.pdf

Gablasova, D., Brezina, V., & McEnery, A. M. (2017). Exploring learner language through corpora: comparing and interpreting corpus frequency information. *Language Learning*, *67*(1), 130–154.

Gablasova, D., & Brezina, V., McEnery, T., & Boyd, E. (2017). Epistemic stance in spoken L2 English: the effect of task and speaker style. *Applied Linguistics, 38*(5), 613–637.

Kulesza, K., Dolinski, D., Huismand, A., & Majewski, R. (2014). The echo effect: The power of verbal mimicry to influence pro-social behavior. *Journal of Language and Social Psychology*, *33*(2), 183–201.

McEnery, A., Brezina, V., Gablasova, D., & Banerjee, J. V. (2019). Corpus linguistics, learner corpora and SLA: employing technology to analyze language use. *Annual Review of Applied Linguistics*, *39*, 74–92.

The student as text producer, mediator and co-assessor: Using student-generated written or spoken texts for practicing and assessing mediation

Daniela Kohn
Victor Babeș University of Medicine and Pharmacy Timişoara, Romania

Gabriel Kohn
West University Timişoara, Romania

Abstract

Medical students and their plurilingual and pluricultural abilities can be involved not only in solving language learning tasks while studying a foreign language, but also in the process of assessing the correspondingly acquired competences. Students can assume diverse roles in language training as text producers by generating authentic texts (oral and written), which can be used in foreign language learning, in mediation tasks, and in the assessment of competences. At the same time, in order to solve mediation tasks, students intensively engage both with their own texts as with those produced by fellow students. Setting out from a specifically designed multi-stage-task situation, this paper focuses on text processing and text mediation tasks and the related (peer-)assessment forms associated with the Romanian language learning process of medical students at the Victor Babeș University of Medicine and Pharmacy Timişoara (Romania).

Introduction: Targets and context

We approach, in this paper, the development of communication skills and assessment of foreign language proficiency from the perspective of a string of multiple tasks, which take, altogether, the shape of a single fluid exercise, a multi-stage-task (MST). Its objective is to involve the student more and more in learning and assessing phases of the process of acquiring general and communicative language competences. The essential feature of the exercise is recursivness. In other words, advancing into the tasks engages a systematic return to a previous phase, in order to assess, set, and correct the student's performance before moving to a subsequent phase. The addition of mediation skills in the learning and assessing process combined with a new mixed assessment (summative and formative) leads towards reorganizing the evaluation of acquired competences in Romanian language of the students at the Victor Babeș University of Medicine and Pharmacy Timişoara (VBUMPT). According to such a model, the teacher waives their absolute authority within the assessment in a targeted manner, sharing it with, or distributing it partially, to the group of students.

To place this assessment in a broader context, it must be noted that at the VBUMPT, knowledge acquisition within its English language programs occurs almost exclusively in English (language of schooling). These students study Romanian as a foreign language (RFL) during their first four semesters at the university. Professionally relevant knowledge of Romanian becomes a pragmatic necessity as soon as the students start their medical practice, and have to read medical documents drafted in Romanian and communicate with local Romanian-speaking patients and medical staff in local hospitals. Therefore, in this phase, knowledge transfer from one language (language of schooling) to the other (foreign language) occurs continuously.

Departing from that basic context, we are describing an example of a multi-stage-task that aims to develop and assess mediation competences during the B2 level RFL course, involving the students at both the learning and the assessing phases. In our scenario, the students are the ones who create starting texts for their fellow students and actively participate in evaluating their peers' output.

Development and assessment of mediation skills

The multi-stage-task (MST): a recursive mediation exercise

The task's aim is to enable the students to understand the way transfer/mediation works or is done, either intralingual or interlingual. The method for pursuing this aim is to actively and constantly involve the students in the process as mediators and evaluators of their own texts or their fellow students' texts. More specifically, we suggest the following exercise roadmap: medical specialty course (English language) → summary (English language) → structured medical history interview (Romanian language).

The starting point is an already existing theoretical course on professional communication with the topic *Structured Medical Interview*, attended by some of the students (group S1), specifically a concrete medical history interview, presented to and analyzed by them within the same course (English language). In our exercise, S1 receives the task to write a summary in English for their fellow students (group S2), including information on the structure of the dialogue (regarding the main structural and stylistic elements), and some key illustrative dialogue sequences. S1 is also required to draft an interview on medical history in Romanian. The subsequent task for S2 is to draft a medical history interview based on the data delivered by the summary. The patient's data, identical to those in the dialogue in English, are available to S1 and to S2 in Romanian, in the patient record (PR). The structure of such a MST may be described as follows:

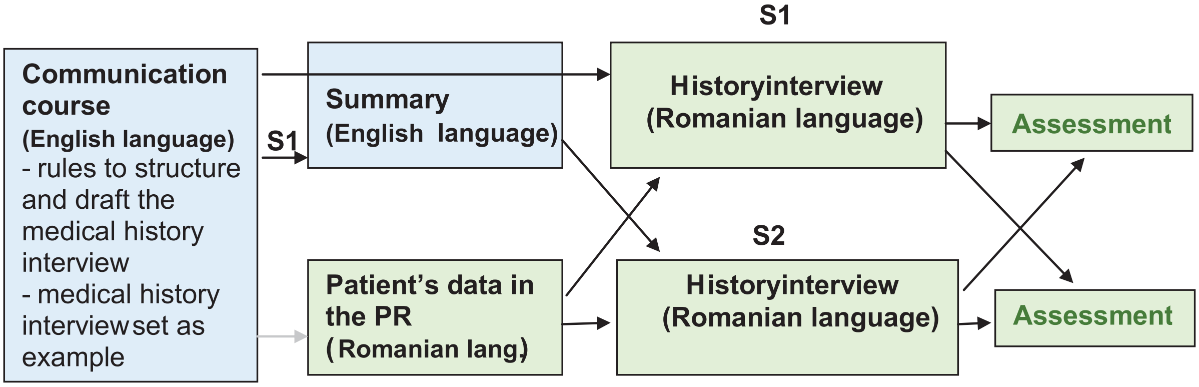


Figure 1 Flowchart of the MST structure

In the first stage we deal with a case of *relaying specific information (mediation)* described in the Common European Framework of Reference for Languages Companion Volume (CEFR CV) (2020): the student is in the position to analytically identify the information in Language A and to properly transfer it to Language B. Moving from the summary to interview, we face an *expansion by operationalization,* and *exemplification*. As well, using the patient's data in the PR, the student also faces a situation of *explaining data*, as described in CEFR CV (2020).

S1 will have the task to transfer/recreate in Romanian the medical history interview which they have studied and analyzed in English. The summary S1 drafts for their fellow students in English will turn into a translation project for themselves.

The test run indicates that groups S1 and S2 rely on two different transfer methods in order to create the medical interview in Romanian. S1 recourses to translating, an indication for the fact that the group is unable to detach from the dialogue template available in English and would rather translate it. S2 recourses to text production + translation. This group builds the interview on the foundation of the briefing received, and tends to translate the illustrating questions included as part of the medical history interview it delivers in Romanian.

After the conclusion of the whole process, the medical history interviews produced by S1 and S2 are to be assessed by the other students. The assesment needs to adress/reflect the different strategies of S1 and S2 and the subsequent different processes deriving from those choices.

Peer review in assessing mediation

The purpose of assessing the interviews created by S1 and S2 focuses on a single fundamental objective: success or failure of mediation, i.e. if the information from the English communication course was transferred properly into the doctor–patient dialogue in Romanian. The following question arises: is the holistic or analytical type the most suitable grid for the assessment made by the students? We consider the analytical grid the most suitable one, especially for less experienced examiners, as Barkaoui (2010) concludes. We have chosen for the students' peer assessment a dichotomous +/- grid, extended by an internal differentiation on the '+' side.

Two assessment grids are built, and they depend on the mediation process by which the target text was achieved. In order for the students to be able to deliver an assessment, they need to understand the evaluation criteria. Their involvement in creating a part of the evaluation grid facilitates the understanding process. S1, the English language summary producer, builds an assessment grid for the medical interview in Romanian drafted by S2, based on the ideas selected to be entered in the summary. S1 can define – with some guidance by the teacher – the content part of the assesment grid, focussing on the adequacy to the typological features of the medical interview (components of the interview, exploration of patient's problems, question style, attentive listening signals, expression of empathy, clarification etc.).

The dialogues created by the member of the group S1, mainly drafted by translation, are subject to a different assessment grid, even if the latter has a common core to that applied to the texts drafted by S2. It is a combined grid, which includes the performance of translation. Our option regarding the complex issue of inserting the assessment of the translation performance within the foreign language learning process is the following: defined as a 'fifth skill'[1], translation returns to the foreign language teaching and learning process at least in 2020 (CEFR CV) under the umbrella of mediation. Hence, the evaluation of translation performance serves mainly to appreciate the quality of the mediation performance.

Translation and mediation are pragmatic processes fundamentally determined by their success or failure in real communication. In our case, in order to make a useful assessment, we are compelled to make a distinction between two stages of the translation and foreign language teaching relationship: (1) the current stage, when the teacher and the student have intuitive translational skills and pre-theoretical knowledge on translation, and (2) a desirable stage when such skills and knowledge ('fifth skill') are not a privilege of translation professionals and find their natural place in foreign language training. Starting from the realistic scenario (1), we suggest the use of the following two assessment criteria: the observance of the principle of relevance[2] and the functional adequacy of the translation. The first indicator may be applied to the peer review process when the S1 and S2 groups discuss if mediation has been successful or failed, i.e. if the information from the English language communication course was properly transferred to the medical interview in Romanian[3]. The evaluation of the functional adequacy of the translation (*Skopostheorie4*) belongs rather to the assessment area of the teacher, as the one able to assess in macro-textual matters, but it can also be worked out by the student under the guidance of the teacher. It is not the mimetic loyalty ('fidelity') to the original that prevails, but the functional viability of the text in the target culture/language. Therefore, the assessment grid dedicated to the translation performance will focus on the difference between the so-called *binary errors*, i.e. unacceptable, categorical ('It's wrong!') and *non-binary errors*, i.e. acceptable and perfectible, ambiguous ('It's correct, but . . .'). As a basic principle, the grid will reward the reduction of binary (categorical) errors and will avoid penalizing non-binary errors by: (1) rewarding the 'ability to generate a TT (target text) series of more than one viable term (TT1, TT2, . . . TTn) for a ST (source text)' and (2) rewarding the ability to select from such series functionally adequate units (Pym, 1992, p. 282).

Conclusions

The MST presented generates a series of positive effects transforming mediation into a reflected practice:

- Students' focus on content while producing a text in a familiar language (mother tongue or language of schooling) is beneficial for mediation processes.
- The transition to focusing on formal linguistic aspects in a foreign language is gradual.
- Active involvement of the students in self- and peer-assessing processes in Language B.
- Involvement in the construction of the assessment grid leads to a higher awareness of students regarding expectations they are confronted with in foreign language learning. For instance, identifying and correcting possible design errors of their assessment grid can become part of the learning process. Errors can be fruitful.
- Mediation is practically assessed in its natural environment, where it occurs.

[1] Alongside speaking, listening, writing, and reading, according to Pym, Malmkjær and del Mar Gutiérrez-Colón Plana (2013).

[2] 'In relevance theoretic terms, an input is relevant to an individual when its processing in a context of available assumptions yields a *positive cognitive effect*. A positive cognitive effect is a worthwhile difference to the individual's representation of the world – a true conclusion, for example. . . . an input is relevant to an individual when, and only when, its processing yields such positive cognitive effects. . . . relevance is not just an all-or-none matter but a matter of degree' (Wilson & Sperber, 2006, p. 608).

[3] In other words, if the proposed translation yields the intended interpretation in a proper unit, i.e. 'without putting the audience to unnecessary processing effort' (Gutt, 1991/2000, p. 107).

[4] See Vermeer and Reiss (1984).

This formative assessment system also comprises some disbenefits: the longer length of the process, and the need for endurance and sustained motivation of the students in order to participate in a comprehensive and demanding learning and assessment process. And last but not least, the effort needed to aknowledge the fellow student as a valid actor of one's assessment.

References

Barkaoui, K. (2010). Variability in ESL essay rating processes: The role of the rating scale and rater experience. *Language Assessment Quaterly*, *7*, 54–74.

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

Gutt, E. A. (1991/2000). *Translation and Relevance. Cognition and Context*. Oxford: Blackwell Publishing/Manchester: St. Jerome Publishing.

Pym, A. (1992). Translation error analysis and the interface with language teaching. In C. Dollerup & A. Loddegaard (Eds.), *The Teaching of Translation* (pp. 279–288). Amsterdam: John Benjamins.

Pym, A., Malmkjær, K., & del Mar Gutiérrez-Colón Plana, M. (2013). *Translation and Language Learning: The Role of Translation in the Teaching of Languages in the European Union*. Strasbourg: Publications Office of the European Union.

Vermeer, H.J., & Reiss, K. (1984). *Grundlegung einer allgemeinen Translationstheorie*. Tübingen: Niemeyer.

Wilson, D., & Sperber, D. (2006). Relevance theory. In L. R. Horn & G. Ward (Eds.), *The Handbook of Pragmatics* (pp. 607–632). Oxford: Blackwell Publishing Ltd.

Using CEFR 'Mediation' scales to inform CLIL pedagogy: constructing a history lesson plan based on academic language proficiency descriptors

Stuart D. Shaw
Cambridge International, United Kingdom

Abstract

The use of academic language descriptors based on Common European Framework of Reference for Languages (CEFR, Council of Europe, 2001) mediation scales affords a practical means of implementing content-based language learning. A scaffold of pedagogic and linguistic support allows learners to access curriculum content. Given that the primary role of assessment is to support learning, a key issue central to successful content and language integrated learning (CLIL) practice is the achievement of intended content and language learning outcomes. The CLIL lesson plan suggested here focusses on CEFR mediation activities.

Proposing an academic language proficiency scale

This study is framed within the context of an international awarding body offering programmes of learning and assessment through the medium of English to schools in a variety of multilingual and educational contexts. A key function of Cambridge Assessment International Education ('Cambridge') is to prepare students whose first language is not necessarily English as candidates for international high-stakes assessments in a variety of bilingual education settings.

The Cambridge International General Certificate of Secondary Education (IGCSE) History, a general education qualification for 14- to 16-year-olds, constitutes the focus of study.[1] The study seeks to determine some of the generic features of an academic language proficiency scale that could be used in the student learning process in a variety of CLIL contexts, and how (and to what effect) such a scale can be used in assessments in CLIL classrooms.

One purpose of developing an academic language scale is to use it to a positive effect in specific contexts of learning of interest to Cambridge, where typically students will be non-native speakers of the language of schooling. This is critical: what is proposed is at once a description of language use in school classrooms, and a tool supporting intervention in how language is used in school classrooms.

CLIL and the CEFR 'mediation' scales

The CEFR is a socio-cognitive model of language use – that is, it is about cognition as it is manifested in, and developed through, social interaction. The CEFR places '*the co-construction of meaning* (through interaction) at the centre of the learning and teaching process. This has clear implications for the classroom. At times, this interaction will be between teacher and learner(s), but at times, it will take a collaborative nature between learners themselves' (Council of Europe, 2020, p. 27; emphasis by author).

[1] Cambridge IGCSE History focuses on both historical knowledge and on the skills required for historical research. It encourages learners to raise questions and to develop and deploy historical skills, knowledge and understanding in order to provide historical explanations. Two of the syllabus aims encourage the development of arguments and communication skills.

The most recent published CEFR Companion Volume (Council of Europe, 2020), intended as a complement to the CEFR, provides illustrative descriptors for new areas not in the original text. Mediation is one such area and includes reactions to creative text/literature; mediating communication, texts or concepts. The concept of mediation in the CEFR, which has assumed great importance with the increasing linguistic and cultural diversity of global societies, takes in a range of communicative tasks and strategies relating to collaborative team work, integrated skills, relaying and synthesising text and meanings, and fostering better understanding among others. The CEFR approach to mediation is broader than cross-linguistic mediation but also covers 'mediation related to communication and learning as well as social and cultural mediation' (Council of Europe, 2020, p. 36). CEFR scales pertinent to this study relate to text mediation and in particular 'Processing text' scales. (Detailed descriptions of these scales can be found on pp. 99–101 of the Companion Volume.) Text processing entails comprehending information and/or arguments encountered in the source text, then condensing and/or reformulating these to another text, in a manner that is appropriate to the context (whilst focusing on the main points and ideas in the source text).

This more expansive approach has particular relevance to CLIL because 'mediation is increasingly seen as a part of all learning, but especially of all language learning' (Council of Europe, 2020, p. 36).

Using CEFR academic language proficiency descriptors in the CLIL classroom

The construction of an academic language scale whose model of reference is the CEFR has clear implications for CLIL pedagogy: it is believed that CLIL learning can be improved through a CEFR task-oriented approach. Such a scale could constitute the basis of CLIL history lesson planning by facilitating, for example, the creation of conditions for communication and cooperation through mediation activities.

Meaningful content resides at the very heart of CLIL and it is crucial that content teachers identify and teach core content concepts in a rigorous manner. Language constitutes an opening to the knowledge and skills of any given content subject. Therefore, it is important that content teachers and their students are language-aware. An academic language scale would support CLIL teachers and schools in implementing bilingual education by making content teachers teaching through the L2 aware of students' L2 language knowledge and needs. CLIL's multi-faceted approach serves to motivate students through a range of diversified teaching methods. An academic language scale could be used to make a foreign language programme more motivating by teaching real content through the target language.

CLIL introduces a cognitive dimension not explicitly treated in the CEFR, adding a new competence – *using language to learn* (Barbero, Damascelli, & Vittoz, 2014, p. 1). One specific context not best considered by the CEFR relates to CLIL – highlighting a clear differentiation between conversational fluency and academic language proficiency. Given that student communication through the L2 may be limited to the CLIL classroom, content teaching through the L2 needs to be as effective as possible. The use of academic language descriptors can both build on, and encourage, students' linguistic repertoires (academic proficiency in L1 and L2) and knowledge (epistemological) repertoires in both spoken and written contexts. Such a scale, when used purposefully, can enrich bilingual programmes including CLIL.[2]

An illustration of how an academic language scale may be employed in the CLIL classroom is in the application and use of *learning outcomes*. Both the content subject and the language used as the medium of instruction are similarly involved in defining the learning outcomes. Achievement of intended content and language outcomes is a key point central to successful CLIL practice (Mehisto & Ting, 2017, p. 214).

A CLIL lesson plan should include, amongst other things, details of how the lesson is intended to proceed and should take account of what is to be taught (learning objectives) and what is to be achieved by the learners (learner outcomes, content and language). Clear intended content and language learning outcomes are fundamental to building and maintaining learner motivation (Gardner, 1985; MacIntyre, 2002), and afford opportunities for students to establish their own learning targets and create openings for teachers to plan their lessons, facilitate course development and create learning resources, as well as providing a mechanism for assessing student learning.

Learning outcomes define what a learner 'can do' by the end of a course of study. Pedagogic exploitation of academic, communicative can do statements has the potential to inform planning and delivery of lessons, negotiation of syllabus content with learners and, more generally, build an effective learning environment. The clarity of content and academic learning outcomes can be enhanced with references to academic CEFR descriptors. Academic can do descriptors, if clear and specific, not only guide students more effectively in their learning but also provide measurable outputs for teachers (thereby performing

[2] See Shaw, Imam and Hughes (2015) for practical insight into how languages are actually understood and used for both teaching and learning 'on the ground' in bilingual settings within international schools from the perspective of an international awarding body.

an *assessment for learning* function). Students would need to be presented with exemplars of the types of language use in order to achieve such outcomes. (For instance, exemplars could be based, in part, on authentic student responses.) Teachers will also need to provide students with evaluative criteria and appropriate scaffolding. Scaffolding affords a very practical means of implementing authentic language and content-area learning, and provides a framework of pedagogic and linguistic support which permits the learner access to the curriculum content through the teacher provision of templates, guides, frames and verbal supports. The efficacy of scaffolding is contingent upon learners building on what they already know in order to make sense of, and take on board, new knowledge. Thus the scaffold acts as a bridge between prior knowledge and new material.

Effective lesson planning entailing the use of academic language proficiency descriptors enables:

- the teacher to set clear targets for content-area learning
- explicit teaching of the language needed to participate in content-area learning
- acknowledgement of the needs of CLIL learners
- learner participation in classroom activity based on an understanding of their language development
- the use of cognitively challenging tasks that require learners to engage with cognitive academic language
- the provision of models of authentic language in use and opportunities to practise it.

When introducing students to specific language skills associated with the topic of the learning outcome, the teacher may draw upon cognitive academic language proficiency (CALP)[3] can do statements as a means for signposting when: students need to learn how to use language in a new way; using authentic material to illustrate the kind of language that is used for an activity; and leading learners step by step through the different stages of practising the skills.

The teacher must also consider the abilities of students within a CLIL class. CLIL classrooms tend to consist of mixed ability language students. Instructional *differentiation* provides an effective framework for classrooms that includes diverse students. As a discipline, history is ultimately expressed through advanced language structures and cognitive discourse functions which necessarily demand cognitive maturity and language competence amongst learners. Effective differentiation (by CEFR level, and mark scheme level, say) is dependent upon an awareness of each learner's current level of understanding and achievement and their individual learning needs. Armed with this knowledge, teachers would be able to provide appropriately differentiated learning tasks and activities to mediate learners based on a range of CEFR level cognitive descriptors. Notwithstanding the challenges of implementing the CEFR in the CLIL classroom, the inherent value of the CEFR in assessment for learning is not in determining learning outcomes but in how knowing what learners can do can direct further teaching and learning. (See Leontjev & deBoer, 2020.) Ultimately, assessment *for* learning and feedback improve teaching and learning practices, which is why CLIL needs to assume a formative orientation.

An example CLIL history lesson plan

By way of illustration, consider the proposed lesson plan (Table 1). The lesson focusses on CEFR descriptors relating to 'mediating a text' (in particular, the 'Text processing' scales provided by the Companion Volume). Given that the proposed lesson entails small group and collaborative tasks, interactions between learners that occur will have mediating functions (such as organising collective work and the relationships between participants, and facilitating access to, and the construction of, knowledge).

[3] In order to address the problem of Second Language (L2) learners' underperformance in mainstream schooling arising from their insufficient command of L2, Cummins (1981) proposed a minimal level of linguistic competence – a threshold of cognitive academic language proficiency (CALP) – that a student must attain to function effectively in cognitively demanding academic tasks. Cummins distinguished CALP from basic interpersonal communication skills (BICS). The distinction was intended to highlight the longer time period needed by students to acquire academic proficiency in their L2 compared to conversational fluency in that language.

Table 1: CLIL History lesson plan (based on Ellis, 2003, p. 217)

Focus points:

Why was there opposition to Soviet control in Hungary in 1956 and Czechoslovakia in 1968?

How did the USSR react to this opposition?

Learning outcome(s):

Learners understand the issues underpinning opposition to Soviet control.

Learners are aware of why and how the USSR reacted the way they did.

CALP descriptors:

Can understand in detail a wide range of lengthy, complex authentic historic texts.

Can summarise in writing and speech long and complex historical source texts, respecting the style and register of the original, interpreting the content appropriately through the meanings of content-compatible language.

Can use high-level phrases, idiomatic and colloquial language in response to historic stimulus material.

Can use appropriate content-obligatory terminology which could include phrases relating to specific historic periods/events, topics and concepts in the curriculum (mainly nouns and proper nouns).

Can facilitate understanding of a complex historical issue by highlighting and categorising the main points, presenting them in a logically connected pattern and reinforcing the message by repeating the key aspects in different ways.

Can recognise a complex historical source text in order to focus on the points of most historic relevance to target audience.

Final task	**Type of input (Scaffolding)**	**Instructional differentiation**	**Processes**	**Micro-tasks** **(focus on one aspect of language)**	**Assessment**
Present a version of historical facts	L1 and L2 textbooks, authentic documents	CEFR Level C1: *Reading for information and argument* History Mark Scheme Levels 4 & 5 C1: Mediation – Conveying clearly and fluently in well-structured language the significant ideas in long, complex historical texts.	Individual work Group work Oral and written production	*Vocabulary*: according to topic *Lexico-grammar*: structures that present an interaction of time and causes and the expression of temporal markers.	Students providing sticky notes with reasons to a whole-class diagram for discussion. A small group exercise involving Information and Communications Technology (ICT) and asking the groups to produce a short script for a radio news bulletin to be broadcast to the West immediately after the Soviet response. Where possible, details could be based on authentic material from the time. Following presentations, the different approaches could be discussed.

Example use of cognitive academic language descriptors based on the CEFR 'Mediating a text' descriptors for *Relaying specific information in speech and writing* (from long, complex historical text) and *Processing text in speech and writing* (from long, complex historical text):

- *Can summarise in writing and speech long and complex historical source texts, respecting the style and register of the original, interpreting the content appropriately through the meanings of content-compatible language*:
  - *Understanding content-compatible language from co-text*
    - 'From very early in 1968, other Communist leaders in Eastern Europe were alarmed by developments in Czechoslovakia. It was clear to them that the growing freedom could be highly infectious.'
  - *Identifying non-essential language to know in order to understand the text*
    - 'Indeed, it was not long before demonstrating Polish students shouted, 'We want a Polish Dubcek!' The first sustained pressure put on the Czechoslovak leadership came at a meeting with five member states of the Warsaw Pact in March 1968.'
  - *Identifying language that needs to be translated*
    - 'The meeting in early August between the Czechoslovak leaders and the Soviet and East European leaders produced a compromise document. At the very time when this agreement was being reached, the Soviet leadership were sent a letter they had been asking for to justify an invasion.'
  - *Identifying essential to know yet difficult to translate language*
    - 'It was a request from the hard-line members of the Czechoslovak leadership calling for intervention. The final decision to launch an invasion was taken between 15 and 17 August.'

This kind of activity gives learners the opportunity to develop their own ideas and understanding with direct input from the CLIL teacher. It also allows learners to employ mediation strategies to explain new concepts as well as strategies to simplify a text.

Through lesson tasks (such as the kind described here) learners are able to engage in ways which require complex language derived from curricular complex relations (instantiated in historical source texts, for example). Language is the mediating tool through which content and language are co-constructed in the learning environment: language is used to mediate content knowledge and content is used to mediate language. Mediation takes place through teacher and learner talk in interaction. Academic content and language descriptors scales offer a direct means for assessing learning – how much (or how little) takes place.

Learner mediation here helps to develop historical concepts and ideas by talking ideas through and articulating the thoughts, thereby facilitating understanding and communication. As such, the use of CEFR cognitive mediation scales is specifically relevant for the CLIL context where small group, collaborative tasks constitute the focus of lesson activities.

This CLIL activity provides opportunities for learners to experiment with the disciplinary language of the given subject and in so doing promote 'a level of talking and interaction that is different to that of the traditional language classroom' (Coyle, 2006, p. 11). The activity also affords learners the chance to use language for various purposes in order to meet curriculum expectations; to employ a balanced and proficient use of the four macro skills of language; and to use authentic literary-specific materials.

The use of a scale of academic language proficiency as a learning tool has additional implications for assessment. Assessment serves as a tool for the learning of content and language. A fundamental role of assessment (whether formative or summative assessment) is to support student learning. Assessment for learning (including feedback) is a core driver of learning. Assessment outcomes can be used to identify teaching and learning needs and subsequent actions (Black, Harrison, Lee, Marshall, & Wiliam, 2002). A distinction needs to be made, however, between the assessment of content learning and the learning of academic language. Mehisto and Ting (2017) contend that assessment literacy is inextricably linked to making explicit intended learning outcomes for both content and language, and for students to work with and practice assessing exemplars of poor, satisfactory and excellent work.

The primary aim of assessment for learning in CLIL 'is to support the learning of content and language, as well as to foster critical thinking about both, and *ultimately to improve teaching and learning practices*' (Mehisto & Ting, 2017, p. 213; emphasis in original). The lesson plan described here provides opportunities for learners in the CLIL classroom to reflect on their own work as well as on the work of others. For reflection to be efficacious, however, learners need to be aware of the learning outcomes (in terms of content and language) as well as having an appreciation of how their learning will be measured. At the same time, teachers must have clear evidence of learning in the CLIL classroom in order to make informed decisions about future teaching and learning of both content and language, whilst providing learners with appropriate feedback on how to progress in terms of achieving the intended learning outcomes.

Summary

The construction of a scale for academic language proficiency based on the CEFR is a complex endeavour. The multidimensional nature of the subject is clear, and it would take a drastic degree of abstraction to entirely reduce it to a single dimension describing something called 'academic language'. By accepting this, the interesting challenge becomes to identify the minimal set of constructs and parameters that would address the complexity of the task. If successful, what would emerge would be a more complex, composite picture of an individual's language profile in relation to dealing with academic subject matter. It is believed that the approach described here will provide insights which will prioritise professional development of CLIL skills and heighten language awareness in the assessment process.

References

Barbero, T., Damascelli, A. T., & Vittoz, M-B. (2014). *Integrating the Common European Framework of Reference (CEFR) with CLIL. CLIL Practice: Perspectives from the field*. Retrieved from: https://www.unifg.it/sites/default/files/allegatiparagrafo/20-01-2014/barbero_damascelli_vittoz_integrating_cefr_with_clil.pdf

Black, P. J., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2002). *Working Inside the Black Box: Assessment for Learning in the Classroom*. London: King's College London School of Education.

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 with New Descriptors*. Strasbourg: Council of Europe.

Coyle, D. (2006). Content and Language Integrated Learning. Motivating learners and teachers. *The Scottish Language Review*, *13*, 1–18.

Cummins, J. (1981). Age on arrival and immigrant second language learning in Canada. A reassessment. *Applied Linguistics*, *2*, 132–149.

Ellis, R. (2003). *Task-based Language Learning and Teaching.* Oxford: Oxford University Press.

Gardner, R. C. (1985). *Social Psychology and Second Language Learning: The Role of Attitudes and Motivation*. London: Edward Arnold.

Leontjev, D., & deBoer, M. (2020). Conceptualising assessment and learning in the CLIL context. An Introduction. In D. Leontjev & M. deBoer (Eds.), *Assessment and Learning in Content and Language Integrated Learning (CLIL) Classrooms: Approaches and Conceptualisations* (pp. 1–27). Switzerland: Springer.

MacIntyre, P. D. (2002). Motivation, anxiety and emotion in second language acquisition. In P. Robinson (Ed.), *Individual Differences and Instructed Language Learning* (pp. 45–68). Philadelphia: John Benjamins.

Mehisto, P., & Ting, T. Y. L. (2017). *CLIL Essentials for Secondary School Teachers*. Cambridge: Cambridge University Press.

Shaw, S. D., Imam, H. C., & Hughes, S. K. (2015). *Language Rich: Insights from Multilingual Schools*. Cambridge: Cambridge University Press.

La evolución de las pruebas de dominio de idiomas: ¿Cómo se evalúa la mediación en los exámenes de expresión oral?

Victoria Peña Jaenes
Cambridge Assessment English, United Kingdom

Abstract

Este artículo comienza ofreciendo una visión general de la relación entre enseñanza y evaluación. A continuación, se reflexiona sobre el papel reforzado que ha adquirido el concepto de mediación desde la publicación del Companion Volume (Consejo de Europa, 2018), y se analiza una muestra de descriptores de mediación y la relación que guardan con algunos descriptores de otros modos de comunicación. Por último, se intenta determinar si la mediación intralingüística se evalúa en la actualidad considerando una muestra de tareas y criterios de evaluación de los exámenes de Cambridge Assessment English, y se estudian las oportunidades y limitaciones que evaluar la mediación puede presentar.

Introducción

El concepto de lengua, su enseñanza y aprendizaje han evolucionado debido a fenómenos socio-políticos y a la investigación en la materia. En la actualidad, vivimos en un mundo interconectado en el que la movilidad, así como la diversidad lingüística y cultural enriquecen nuestras sociedades, favoreciendo que las lenguas y las culturas trasciendan las fronteras nacionales (Piccardo, 2020). En este contexto, hablar un idioma extranjero ha pasado de ser un lujo a ser una necesidad pues es un requisito para migrar o para progresar académica o profesionalmente (Jaimechango, 2009; citado por Chávez-Zambano, Saltos-Vivas, & Saltos-Dueñas, 2017, p. 761).

La investigación en lingüística y pedagogía ha ido de la mano del cambio social y ha resultado en la percepción de la lengua como un vehículo de comunicación y como una actividad conjunta que nos permite coordinar comportamientos, compartir experiencias y mantener relaciones sociales (Fusaroli, Rączaszek-Leonardi, & Tylén, 2014). La lengua permite al ser humano –entendido como individuo capaz de crear sentido y pensar (Piaget citado por Rüschoff, 2020) – articular nuestros pensamientos. Este uso de la lengua para crear significado pone de manifiesto la importancia de la mediación como elemento clave del entendimiento, el pensamiento y la colaboración (Piccardo, 2020).

Esta percepción de la lengua se ve reflejada en la enseñanza, aprendizaje y evaluación de los idiomas. Hasta comienzos del siglo XX, el método de enseñanza más arraigado estaba basado en la gramática y la traducción debido, en parte, a la influencia del griego y el latín clásicos. Sin embargo, el fin de las guerras mundiales y el aumento de la movilidad favorecieron el interés por la expresión oral, lo que llevó al modelo comunicativo de Canale y Swain (1980), posteriormente desarrollado por Canale (1983), Bachman (1990), y Bachman y Palmer (1996), todos citados por Galaczi y Taylor (2018, p. 220), y que hoy en día es el método más extendido en la enseñanza de idiomas. La publicación del *Marco Común Europeo de Referencia para las lenguas* (MCER) por el Consejo de Europa en 2001 marcó un hito en la internacionalización de la enseñanza y de la evaluación de lenguas extranjeras y supuso un cambio significativo en la formación en idiomas. El Companion Volume (Consejo de Europa, 2020) vino a reforzar y desarrollar aspectos ya presentes en el MCER como la mediación.

Mediación

Aunque la mediación haya recibido una mayor atención a raíz de la publicación del Companion Volume, no es un concepto nuevo, pues forma parte de nuestra esencia como individuos capaces de pensar. De hecho, ya estaba presente en el MCER y en el trabajo de reconocidos autores como Swain (2006; citado por Piccardo, 2020) o Green (2014, p. 5).

La mediación tiene lugar siempre que hace falta un intermediario para construir o transmitir significado. Podemos mediar entre lenguas diferentes pero también entre variedades y registros de una misma lengua. Asimismo, mediamos cuando expresamos de forma oral lo que aparece de forma escrita y viceversa. Por último, la mediación se hace fundamental a la hora de superar las diferencias culturales para facilitar la comunicación (Engquist, 2019).

La conexión entre la vida real, la enseñanza y la evaluación de idiomas se pone de manifiesto en el enfoque orientado a la acción que se sigue en la enseñanza y en el hecho de que el objetivo final de la evaluación sea la comunicación en la vida real. La mediación es parte de nuestro día a día, por lo que resulta lógico que se incorpore al aula y a la evaluación, aunque presente un desafío para los docentes y los evaluadores.

El estudiante como agente social

Si bien en el Companion Volume se reconoce la utilidad de representar el uso del idioma en cuatro destrezas, también se desarrolla de forma más extensa la idea de los modos de comunicación como una forma más apropiada de capturar la complejidad de la comunicación. En la producción y la recepción, el estudiante actúa como productor o receptor de información y los aspectos que se consideran son, entre otros, la complejidad del mensaje, la corrección con la que se produce o se comprende y la fluidez. En la interacción, participa en la actividad comunicativa y negocia el significado. En la mediación, construye puentes para facilitar la comunicación tanto entre distintos idiomas como en el mismo. Independientemente del modo de comunicación, el estudiante asume el papel de *agente social*, es decir, el de un individuo que emplea sus habilidades lingüísticas para construir significado de manera conjunta (Margonis-Pasinetti, 2019). Como agente social, en la mediación, el alumno participa activamente en la comunicación empleando sus recursos cognitivos, emocionales, lingüísticos y culturales, a la vez que reflexiona y pone en práctica estrategias para planificar, controlar y adaptar la comunicación con otros individuos (North, 2020). Sin duda, se trata de una actividad compleja que puede parecer poco adecuada para usuarios que dan sus primeros pasos en el aprendizaje de un idioma. Sin embargo, Engquist (2019) pone de manifiesto que cualquier estudiante de cualquier nivel puede mediar y pone como ejemplo una actividad en la que dos amigos se encuentran en un restaurante en el que la carta está en inglés, uno de ellos habla este idioma y el otro no. La persona que habla inglés debe ver el menú y seleccionar los platos que su amigo puede comer teniendo en cuenta que es vegetariano. Esta actividad de mediación entre lenguas no requiere que se empleen las destrezas de un traductor profesional, pero permite que el hablante seleccione información relevante para su amigo y la comunique.

Análisis de los descriptores de mediación y su aplicación

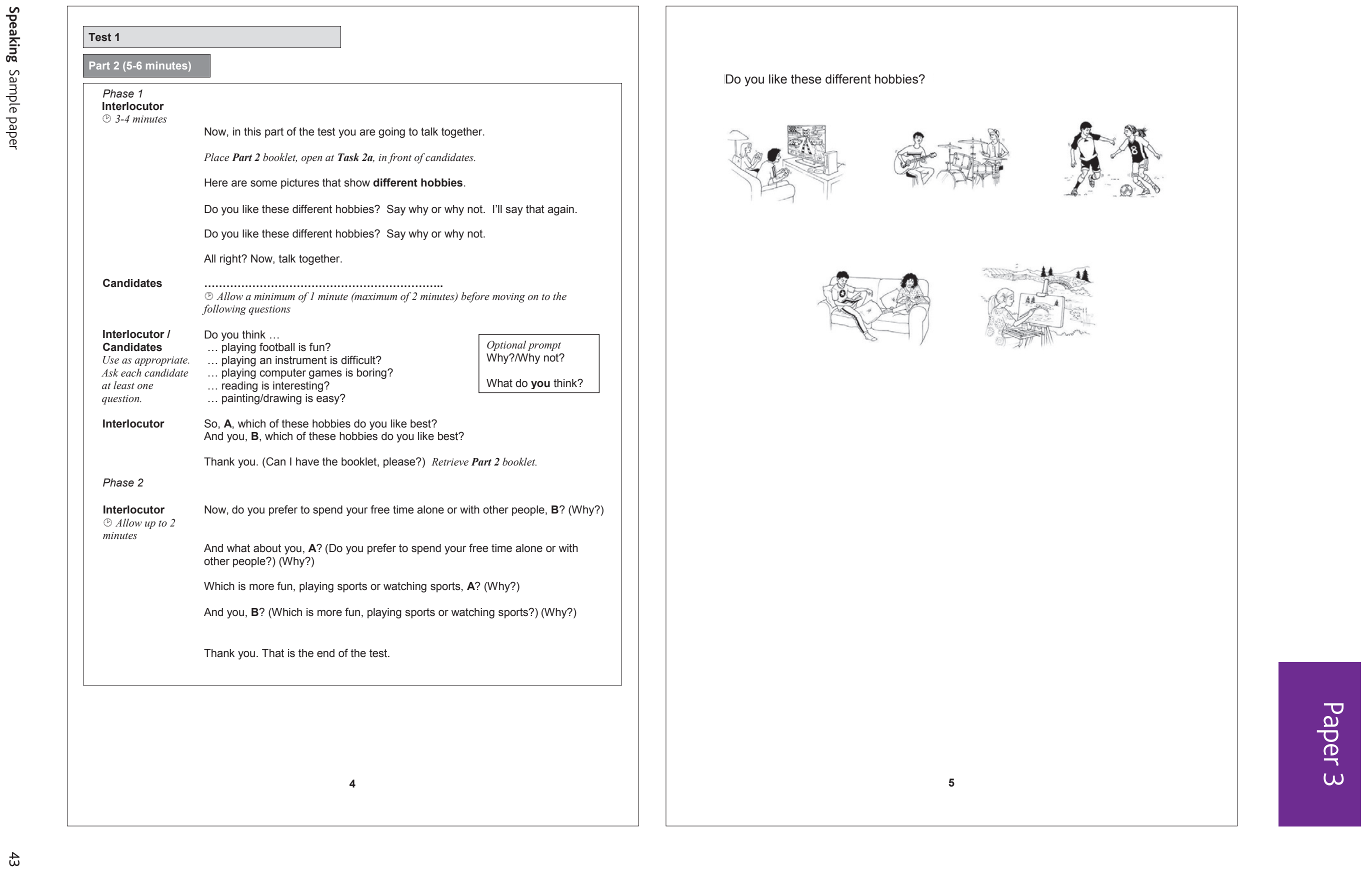
Probablemente la mejor forma de entender la mediación y cómo los usuarios pueden mediar es consultar sus descriptores, lo que, a su vez, nos permitirá comprender mejor la relación entre la mediación y otros modos de comunicación como la interacción. Así, proponemos tres descriptores de mediación para tres niveles del MCER.

En el nivel C1 del MCER, en la categoría de *Mediación → Colaboración en grupo → Facilitar la interacción colaborativa con iguales* (Consejo de Europa, 2020, p. 119) se espera que los usuarios "sean capaces de mostrar sensibilidad hacia distintas perspectivas expresadas dentro de un grupo de hablantes, reconociendo las contribuciones de otros y expresando su reserva, desacuerdo o crítica, de forma que evite o minimice cualquier tipo de ofensa"[1]. Este descriptor refleja un concepto similar al que vemos en la categoría de *Interacción oral → Conversaciones formales en reuniones* (Consejo de Europa, 2020, p. 87) en el que se espera que los usuarios de nivel C1 del MCER "sean capaces de reiterar, evaluar y cuestionar lo expresado por otros participantes sobre un tema de su especialidad académica o profesional".

Seleccionamos ahora otro descriptor, en este caso en la categoría de *Mediación > Mediación de la comunicación > Facilitar el espacio pluricultural,* que indica que un usuario de B2 del MCER "es capaz de apreciar visiones del mundo distintas a la suya y expresarse de forma adecuada al contexto" (Consejo de Europa, 2020, p. 123). Esta habilidad se podría poner en práctica si tenemos alumnos de distintos países y les pedimos que mantengan una conversación sobre un aspecto cultural como puede ser la edad a la que se suelen independizar los jóvenes.

Por último, analizamos un descriptor que muestra como la mediación es posible en los niveles de dominio más básicos. De acuerdo con el descriptor de la categoría de *Mediación > Mediación de conceptos > Liderar trabajo en grupo facilitando las conversaciones sobre conceptos*, un usuario de nivel A2 "es capaz de preguntar lo que otra persona piensa sobre una idea concreta" (Consejo de Europa, 2020, p. 121). Este descriptor podría ponerse en práctica dando a los alumnos / candidatos un esquema con actividades para el tiempo libre del tipo que podemos encontrar en el examen A2 Key for Schools (University of Cambridge Local Examinations Syndicate, 2019, p. 43) (Figura 1) y pidiéndoles que mantengan una conversación sobre lo que piensan de las distintas opciones.

[1] Las traducciones de los descriptores no son oficiales sino que las ha realizado la autora para este artículo.



Speaking Sample paper

Test 1

Part 2 (5-6 minutes)

Phase 1
Interlocutor
3-4 minutes

Now, in this part of the test you are going to talk together.

*Place **Part 2** booklet, open at **Task 2a**, in front of candidates.*

Here are some pictures that show **different hobbies**.

Do you like these different hobbies? Say why or why not. I'll say that again.

Do you like these different hobbies? Say why or why not.

All right? Now, talk together.

Candidates ………………………………………………………………

Allow a minimum of 1 minute (maximum of 2 minutes) before moving on to the following questions

Interlocutor / Candidates
Use as appropriate. Ask each candidate at least one question.

Do you think …
… playing football is fun?
… playing an instrument is difficult?
… playing computer games is boring?
… reading is interesting?
… painting/drawing is easy?

Optional prompt
Why?/Why not?

What do **you** think?

Interlocutor So, **A**, which of these hobbies do you like best?
And you, **B**, which of these hobbies do you like best?

Thank you. (Can I have the booklet, please?) *Retrieve **Part 2** booklet.*

Phase 2

Interlocutor
Allow up to 2 minutes

Now, do you prefer to spend your free time alone or with other people, **B**? (Why?)

And what about you, **A**? (Do you prefer to spend your free time alone or with other people?) (Why?)

Which is more fun, playing sports or watching sports, **A**? (Why?)

And you, **B**? (Which is more fun, playing sports or watching sports?) (Why?)

Thank you. That is the end of the test.

4

Do you like these different hobbies?

5

43

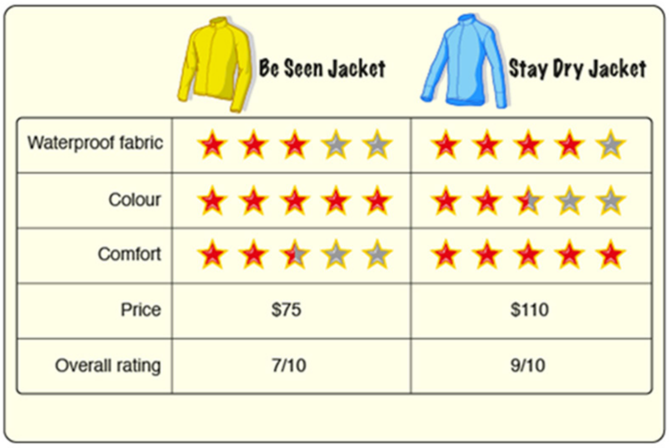
Paper 3

Figura 1: Modelo de la prueba de Expresión oral de A2 Key for Schools (University of Cambridge Local Examinations Syndicate, 2019, p. 43)

Evaluar la mediación: oportunidades y desafiós

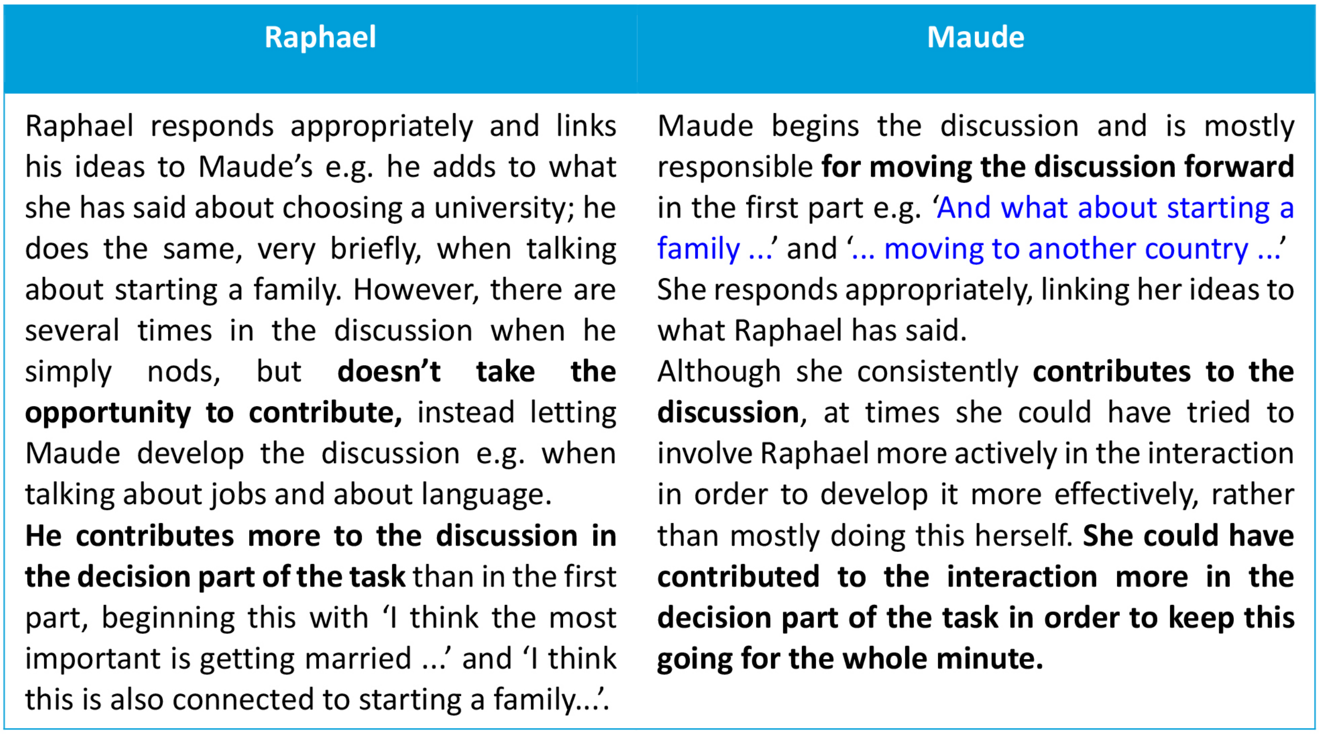
Para mediar, los alumnos deben emplear estrategias que faciliten el entendimiento y, por este motivo, la mediación encaja especialmente bien con actividades de colaboración y con la evaluación formativa. Desde el punto de vista de la participación y de la motivación, las tareas de mediación requieren un papel más activo y una actitud más reflexiva, que las hacen enriquecedoras y motivadoras. Desde el punto de vista lingüístico, permiten a los alumnos / candidatos poner en práctica habilidades lingüísticas para entender, resumir o parafrasear un mensaje. Asimismo, son tareas que favorecen la integración de destrezas e incluso el uso de la tecnología, y como ejemplo, citamos esta tarea de la prueba de Expresión oral de Linguaskill (University of Cambridge Local Examinations Syndicate, 2021) (Figura 2), en el que los candidatos deben leer información sobre prendas de ropa y recomendar una a su amigo en función de sus necesidades, que deben haber sido claramente especificadas.

Por lo anteriormente mencionado, las tareas de mediación suponen una oportunidad. No obstante, también presentan desafíos. Para que los alumnos/candidatos puedan mediar satisfactoriamente se deben tener en cuenta una serie de aspectos. En primer lugar, la tarea ha de estar diseñada de acuerdo con alguno de los descriptores de mediación del Companion Volume y con el nivel que se desea evaluar. Este aspecto constituye la pregunta de investigación número 2 de Jing y Pinnington (2020) que se resolvió con resultados positivos en su estudio sobre mediación. El siguiente desafío es diseñar la tarea con unos objetivos claros, con un tiempo adecuado y con suficientes detalles sobre la situación comunicativa. Además, debe permitirnos discriminar entre candidatos con un rendimiento más o menos óptimo. Este último aspecto fue investigado por Jing y Pinnington (2020)



	Be Seen Jacket	Stay Dry Jacket
Waterproof fabric	★★★☆☆	★★★★☆
Colour	★★★★★	★★⯪☆☆
Comfort	★★⯪☆☆	★★★★★
Price	$75	$110
Overall rating	7/10	9/10

Figura 2: Modelo de la prueba de Expresión oral de Linguaskill (University of Cambridge Local Examinations Syndicate, 2021)



Raphael	Maude
Raphael responds appropriately and links his ideas to Maude's e.g. he adds to what she has said about choosing a university; he does the same, very briefly, when talking about starting a family. However, there are several times in the discussion when he simply nods, but **doesn't take the opportunity to contribute,** instead letting Maude develop the discussion e.g. when talking about jobs and about language. **He contributes more to the discussion in the decision part of the task** than in the first part, beginning this with 'I think the most important is getting married ...' and 'I think this is also connected to starting a family...'.	Maude begins the discussion and is mostly responsible **for moving the discussion forward** in the first part e.g. 'And what about starting a family ...' and '... moving to another country ...' She responds appropriately, linking her ideas to what Raphael has said. Although she consistently **contributes to the discussion**, at times she could have tried to involve Raphael more actively in the interaction in order to develop it more effectively, rather than mostly doing this herself. **She could have contributed to the interaction more in the decision part of the task in order to keep this going for the whole minute.**

Figura 3: Comentarios de examinadores. Prueba de Expresión oral de C1 Advanced, Raphael y Maude (University of Cambridge Local Examinations Syndicate, 2014)

en su pregunta de investigación número 3 y sus resultados pusieron de manifiesto la dificultad a la hora de discriminar entre los niveles más altos – C1 y C2.

Asimismo, es necesario desarrollar criterios que evalúen la mediación y que sean lo suficientemente claros y precisos para que distintos examinadores puedan aplicarlos con fiabilidad con la formación necesaria. Como ejemplo de criterios de evaluación relacionados con tareas de mediación en el nivel C1, mostramos los comentarios de los examinadores de Cambridge Assessment English (University of Cambridge Local Examinations Syndicate, 2014) (Figura 3), que citan los criterios empleados y los relacionan con el rendimiento de los candidatos en la prueba de Expresión oral de C1 Advanced. Las partes en negrita reflejan la mediación de conceptos y la colaboración en grupo, concretamente se ve cómo el candidato formula preguntas al compañero con el objetivo de que elabore sus ideas e intenta apoyarse en ellas para avanzar y llegar a la solución de la tarea.

Conclusión

La mediación como práctica da respuesta a las necesidades de la sociedad actual, caracterizada por su dinamismo y variedad lingüística y cultural. Sin embargo, la incorporación de la mediación en la enseñanza y la evaluación ha supuesto un cambio al que profesores y evaluadores nos estamos adaptando. La investigación en la materia y su divulgación aportan luz y permiten ser conscientes de las oportunidades que la mediación ofrece, además de contribuir a superar los principales desafíos que presenta.

Referencias

Bachman, L. (1990). *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press.

Bachman, L., & Palmer, A. (1996). *Language Testing in Practice*. Oxford: Oxford University Press.

Canale, M. (1983). From communicative competence to communicative language pedagogy. In J. C. Richards & R. W. Schmidt (Eds.), *Language and Communication* (pp. 2–27). London: Longman.

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics, 1*, 1–47.

Chávez-Zambrano, 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.

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

Consejo de Europa. (2020). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Companion Volume with New Descriptors*. Strasbourg: Council of Europe.

Engquist, B. (2019). *The challenge of developing mediation tasks for our student*s. Retrieved from: eltlearningjourneys.com/2019/06/05/the-challenge-of-developing-mediation-tasks-for-our-students/

Fusaroli, R., Rączaszek-Leonardi, J., & Tylén, K. (2014). Dialog as interpersonal synergy. *New Ideas in Psychology*, *32*, 147–157.

Galaczi E., & Taylor, L. (2018) Interactional competence: Conceptualisations, operationalisations, and outstanding questions. *Language Assessment Quarterly, 15*(3), 219–236.

Green, A. (2014). *Exploring Language Assessment and Testing*. Abingdon: Routledge.

Jaimechango. (2009). *Importancia del inglés en la educación*. Retrieved from: es.slideshare.net/jaimechango/importancia-del-ingles-en-la-educacon

Jing, X., & Pinnington, A. (2020). *Developing computer-based speaking tasks to assess the 'mediation' construct* [Vídeo]. Retrieved from: www.youtube.com/watch?v=iQo4EdLwnHs&feature=youtu.be

Margonis-Pasinetti, R. (2019, febrero, 22–23). *The CEFR/CV: Reinforcing the CEFR Message – Learning, Teaching, Assessment* [Conference presentation]. Congreso En Escuelas Oficiales de Idiomas de Canarias, Santa Cruz de Tenerife (España).

North, B. (2020, diciembre). *CEFR Companion Volume for Learning, Teaching and Assessment* [Online seminar]. MiLLaT online seminar.

Piccardo, E. (2020, diciembre). *The new CEFR descriptors for mediation and the dynamic nature of language learning: from the conceptualization to the class* [Online seminar]. MiLLaT online seminar.

Rüschoff, B. (2020, diciembre). *The CEFR Companion Volume Descriptors for online interaction and transaction* [Online seminar]. MiLLaT online seminar.

Swain, M. (2006). Languaging, agency and collaboration in advanced second language learning. In H. Byrnes (Ed.), *Advanced Language Learning: The Contributions of Halliday and Vygotsky* (pp. 95–108). London: Continuum.

University of Cambridge Local Examinations Syndicate. (2019). *A2 Key for Schools Handbook for Teachers*. Cambridge: Cambridge Assessment English.

University of Cambridge Local Examinations Syndicate (2014) *Cambridge English: Advanced Speaking (from 2015). Sample test with examiner's comments*. Cambridge: Cambridge Assessment English.

University of Cambridge Local Examinations Syndicate. (2021). *Materiales de práctica.* Retrieved from: www.cambridgeenglish.org/exams-and-tests/linguaskill/information-about-the-test/practice-materials/

Considerations of Ethics and Justice in the Multilingual Context of Assessment

Assessing language needs of adult refugees and migrants in the Greek context

Anna Mouti
Aristotle University of Thessaloniki, Greece

Christina Maligkoudi
Aristotle University of Thessaloniki, Greece

Gogonas Nikos
National and Kapodistrian University of Athens, Greece

Abstract

The Council of Europe has developed a toolkit to support member states in their efforts to respond to the challenges posed by unprecedented migration flows, as part of the project Linguistic Integration of Adult Migrants (LIAM). The main purpose of this study is to report from an informal implementation of aspects of the Toolkit for Language Support for Adult Refugees, for the language needs analysis of adult refugees and migrants in Greece. The findings show high diversity in terms of language competence, literacy, and linguistic backgrounds among the students and points to the necessity for new tools for course planning and materials design based on refugees' and migrants' needs analysis.

Introduction

Beacco, Krumm and Little (2017, p. 4) highlight that 'there is no such thing as a typical migrant'. Language classes organized for adult refugees and migrants are heterogeneous and students in these educational settings differ across various aspects, including language competences, educational background and levels of literacy. Regarding linguistic heterogeneity, it is much higher among immigrant learners because they have extremely different linguistic biographies, depending on the status of their first language(s) in the country of origin, the other languages they have used during their migration and the language contacts in the host society (Krumm & Plutzar, 2008). Assessing adult refugees and migrants' subjective language needs should be a prerequisite for the design and development of tailor-made courses and learning material as 'it makes no sense to provide just one type of language course for adult migrants or to impose the same language requirement on everyone' (Beacco et al., 2017, p. 4) .

As indicated in the LAMI Guide (p. 18, alte.wildapricot.org/resources/Documents/LAMI%20Booklet%20EN.pdf) 'a needs analysis, for immigration policy decisions, can provide an overview of information about the different types of migrants coming to the host country', collecting information on topics such as countries of origin, language and skills backgrounds, likely linguistic demands of the contexts that migrants will find themselves in, different migrant groups in terms of purpose for entry, and different migrant profiles in terms of age, gender, linguistic and cultural repertoire, and literacy skills.

Beacco , Little and Hedges (2014) suggest that a structure of parameters and categories, presented in the Common European Framework of Reference for Languages (CEFR) (Council of Europe, 2001, p. 44), may provide a starting point for needs analysis. More specifically, parameters like the context of language use (including domains and language use situations), communication themes, communicative tasks and purposes, communicative language activities and strategies, and communicative language processes and texts are relevant to the design of language learning programmes, and thus relevant to the design and analysis of communicative language needs. In a similar way, van Avermaet and Gysen (2008) present the way the needs analysis could lead to the design of tailor-made courses by establishing relevant domains and language use situations and then deriving tasks from lists of language use situations.

As part of the project Linguistic Integration of Adult Migrants (LIAM) of the Council of Europe's major programme on language policy, the Council of Europe has developed the Toolkit for Language Support for Adult Refugees, to help volunteers, teachers and refugees in member states face the linguistic challenges caused by large-scale movements of refugees. The toolkit comprises 57 tools and other resources contained in the various sections of a website. Among these 57 tools, there are seven for focusing on needs analysis to help teachers and volunteers prepare language support activities[1]: Tool 24 - Identifying refugees' most urgent needs; Tool 25 - Finding out what refugees can already do in the target language and what they need to be able to do; Tool 26 - First steps in the host country language; Tool 27 - Refugees' linguistic profiles; Tool 28 - Finding out more about refugee's own linguistic resources and capacities; Tool 29 - What are the most important things to learn? The refugees' point of view; and Tool 30 - Observing situations in which refugees need to use the target language.

The main purpose of this study is to assess language needs of adult refugees and migrants through the implementation of the aforementioned toolkit's needs analysis tools, focusing mainly on what refugees and migrants can already do in the target language and what they need to be able to do.

Method

The aim of our research is to explore the implementation of various empirical needs analysis case studies through the Council of Europe's needs analysis toolkit, and to provide a detailed needs analysis study of specific groups of refugees and migrants to form the basis of tailor-made ad hoc language learning material. Case studies were conducted by postgraduate students and teachers attending a Master's programme entitled 'Language Education for Refugees and Migrants' (LRM)[2] offered by the Hellenic Open University. It is designed as an initial/primary research programme to form the basis of innovative research. The tookit has not been extensively used in Greece and to our knowledge there is no relevant study published in the Greek context[3]. Regarding research design, multiple case studies were chosen and, in particular, 20 case studies were conducted in various educational settings throughout Greece.

Participants: Research assistants and adult learners

The study was conducted through an exploratory research procedure, without attempting to offer final and conclusive answers to specific research questions, and completed by 20 students (as research assistants) attending the Module LRM53[4] during three semesters (Fall 2018, Spring 2019, and Fall 2019)[5].

Approximately 80 learners participated in the study and their ages ranged from 17 to 55 years. The learners spoke a variety of languages, with Arabic, Farsi, Kurdish and Urdu being the most common ones. Other languages included Bangla, Bengali, Banjabi, Turkish, Bulgarian, Albanian, Russian, Armenian, Igbo and Tigynya. English was reported to be spoken by several participants thus retaining its lingua franca function. The participants' countries of origin were: Afghanistan, Syria, Iraq, Bangladesh, Iran, Pakistan, Albania, Georgia, Bulgaria, Armenia, Egypt, Senegal, Turkey Cameroon, Sudan, Eritrea, Italy, Spain, and Nigeria. The most frequent nationalities identified were: Afghani, Syrian, Iraqi, Bangladeshi and Kurdish, and less frequent ones included Italian Spanish, Egyptian, Georgian and Armenian.

In the majority of the cases, learners used Greek at an A1/A2 level but they did not seem confident using it. English was also used in many cases along with elementary/basic Greek in a translanguaging/mediating role. As far as the participants' educational experience is concerned, most of them were literate as they had attended formal education in their country of origin. There were varying levels of educational and vocational backgrounds with various professions and university degrees. However, there were also students who were illiterate in their L1, or who spoke languages with a writing system typologically distinct from Latin or Greek (e.g. Arabic, Georgian)

Educational settings

A variety of case studies were spotted: 20 educational settings, largely in South Greece (Crete, Patras, Andravida, Nafpaktos, Arcadia, Kos, Komotini) and 10 of them in Athens. The students were attending courses in non-formal educational programmes run by NGOs (e.g. Greek Council for Refugees, Pyxis, Steki Metanaston, Kato ta thrania, Piso Thrania). Modern Greek was the target language in almost all the educational settings, and German was taught in one case. Various levels of Greek competence were noted (A1:8, A2:4, A2-B1:5, B1:1, C1:1) and A1 level in German.

[1] Tools 24–30: www.coe.int/en/web/language-support-for-adult-refugees/needs-analysis

[2] LRM is a postgraduate programme designed for teachers and graduates who wish to complete or deepen their knowledge and skills of language teaching in various languages. Find more details in Makri et al. (2017).

[3] See the toolkit implementation in the Italian context coordinated by Lorenzo Rocca: www.coe.int/en/web/language-support-for-adult-refugees/piloting

[4] Language teaching for adult refugees and migrants.

[5] The research assistants are a special group of teachers/students as also presented and pointed out in Gkaintartzi, Mouti, Skourtou and Tsokalidou (2019).

Instruments

It was suggested to the research assistants to approach the migrant students' needs based on the relevant seven tools provided by the toolkit[6]. The research instruments were mainly qualitative, for example using the toolkit's needs analysis tools for questionnaires completed by the students or the structured interview guides for interviews with the students.

Results

The majority of the research assistants opted for needs analysis Tool 25 ('Finding out what refugees can already do in the target language and what they need to be able to do'), Tool 28 ('Finding out more about refugee's own linguistic resources and capacities') and Tool 29 ('What are the most important things to learn? The refugees' point of view').

The needs were analysed according to van Avermaet and Gysen's (2008) suggestions but their analysis was also based on Beacco et al.'s guide (2014). Thus, the first step was to explore the needs by considering the context(s) of language use, establishing relevant domains, and specifying the communication themes and language use situations the target group should be able to cope with. Deriving tasks from lists of language use situations was the last step. The domains were ranked in the following order based on their occurrence in the 20 case studies conducted: public domain (mentioned in 17 cases), occupational (mentioned in 11 cases), personal (mentioned in six cases) and educational (mentioned in five cases).

It is obvious that communicative language needs in the public and occupational sector are more urgent and present in most of the cases studied. Personal and educational domains, on the other hand are only discussed in one quarter of the cases. These findings seem to agree with van Avermaet and Gysen (2008) and Androulakis, Gkaintartzi, Kitsiou and Tsioli (2017) in Greece, indicating that the urgency of needs of the learners is diverse. More precisely, participants' needs in specific sub-domains and communication themes were ranked as follows: work/business (11 cases), formal social contacts/ public services (9), formal social contacts/communicating with doctors (5), education/training/children's education (5), informal social contacts/need for socializing (3), formal social contacts/obtaining permanent residence (3), informal social contacts/shopping in the market/shops (2), informal social contacts/finding accommodation (1).

During the needs analysis, more refined language use situations and functions emerged. Although the main domains and sub-domains already presented were all catered for in the toolkit, the more refined language situations and functions, appearing throughout our study, are not all catered for in the toolkit. Of course, not all the questions in the tools have to be used, and others can be added as the tools are simply intended as a guide for this kind of exchange with refugees. In any case, some of these more detailed language use situations could be included, for example in Tool 29 (What are the most important things to learn? The refugees' point of view). To demonstrate, we present some of the most typical language use situations and functions from the 20 case studies, in two of the two sub-domains mentioned above:

Workplace/Business: Complete an online CV and submit a cover letter; Search for a job via the Internet and interact with employers in a more effective way during interviews; Introduce themselves regarding working experience in formal situations; Use the Internet in order to look for a job and apply for employment; Clearly describe their professional profile; Define their intended work environment.

Formal social contacts/children's education: Understand about enrolment of a new pupil in a school; Be able to communicate in informative meetings for parents organised by the school; Inform school about sickness of the child; Become familiar with the procedure of parents getting school reports; Communicate with children's teachers.

Discussion and further research

The need for studying second language acquisition and language education issues for refugees and migrants in the Greek context has recently increased, mainly after and due to the 2015 refugee crisis. In our paper we have made an attempt to implement needs analysis in a new dimension of language education for adult refugees and migrants. This procedure was supported and positively commented on by the MA students as it made clear to them the need for tailor-made courses and materials designed for adult refugees and migrants based on needs analysis. The use of these seven tools made this procedure less challenging and more 'feasible' to them, and it helped them in the design of adequate ad hoc learning material. The needs analyses tools of the toolkit seemed to work appropriately and sufficiently in the multilingual classrooms, enhancing plurilingual and multilingual awareness among all the participants and fulfilling focal students' language needs. It is a series of tools that could be used not only with volunteers and teachers actively involved in language education for refugees and migrants, but also in the training of both pre- and in- service language teachers.

[6] Tools 24-30: www.coe.int/en/web/language-support-for-adult-refugees/needs-analysis

The need to identify descriptors of language proficiency for first levels in migration contexts, specifically, for levels that are lower than those of the CEFR, has become apparent, supporting the ongoing Council of Europe project of developing a reference tool for teachers of low-literate learners, describing language skills at four levels below CEFR Level A1 (LASLLIAM). It seems that the language education provided to adult refugees and migrants in Greece is limited and fragmented (Kantzou, Manoli, Mouti, & Papadopoulou, 2017) and generally designed to let students reach only a basic level of language competence (A1–A2). In Italy a similar issue was addressed through a syllabus and descriptors for Italian L2 for migration contexts (Borri, Minuz, Rocca, & Sola, 2014) as a tool to plan courses, to create teaching materials and to prepare diagnostic and achievement tests at levels preceding A1, and for addressing needs of non-literate and low-literate migrants.

Greece and Italy share a double role both as host and transition countries, as the main EU entry points for refugees and migrants. This role complicates linguistic integration for the refugees who are either residing temporarily or permanently settled, and use either lingua francas or their mother tongue to communicate with other non-Greeks. These exact needs in the Italian context have been identified by Bianco and Ortiz Cobo (2019, p. 12). In conclusion, we can remark that, according to the real needs of the refugees, other languages should also be taken into account, especially regarding the learning of languages for refugees who intend to relocate because, as Bianco and Ortiz Cobo (2019, p. 12) mention: 'Multilingual needs are typical of refugees due to their migration path and their migration plans' .

Acknowledgments

We would like to express our gratitude to the students of the MA programme 'Language Education for Refugees and Migrants' of the Hellenic Open University attending the module 'LRM53-Language Teaching to adult refugees and migrants' during the academic year 2018–2019 and 2019–2020, who helped collect the data.

References

Androulakis, G., Gkaintartzi, A., Kitsiou, R., & 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: De Gruyter.

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.

Beacco, J. C., Little, D., & Hedges, C. (2014). *Linguistic Integration of Adult Migrants: Guide to Policy Development and Implementation.* Strasbourg: Council of Europe.

Bianco, R., & Ortiz Cobo, M. (2019). The linguistic integration of refugees in Italy. *Social Sciences*, *8*(10), 1–15.

Borri, A., Minuz, F., Rocca, L., & Sola, C. (2014). *Italiano L2 in contesti migratori. Sillabo e descrittori dall'alfabetizzazione all'A1.* Turin: Loescher.

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

Gkaintartzi, A., Mouti, A., Skourtou, E., & Tsokalidou, R. (2019). Language teachers' perceptions of multilingualism and language teaching: The case of the postgraduate programme 'LRM'. *Language Learning in Higher Education, 9*(1), 33–54.

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.

Krumm, H. J., & Plutzar, V. (2008). *Tailoring Language Provision and Requirements to the Needs and Capacities of Adult Migrants*. Strasbourg: Council of Europe.

Makri, A., Papadopoulou, M., Tsokalidou, R., Skourtou, E., Arvaniti, E., Gkaintartzi, A., Kantzou, V., Manoli, P., Markou, E., Mouti, A., Palaiologou, N., & Kitsiou, R. (2017). Tutor practices in new HOU programmes. Stories from the trenches: the case of LRM (Language Education for Refugees and Migrants). In *Proceedings of ICODL 2017 (International Conference in Open and Distance Learning)* (pp. 70–80). Thessaloniki: Aristotle University of Thessaloniki.

van Avermaet, P., & Gysen, S. (2008). *Language learning, teaching and assessment and the integration of adult immigrants. The importance of needs analysis*. Retrieved from: rm.coe.int/CoERMPublicCommonSearchServices/DisplayDCTMContent?documentId=09000016802fc1d5