Educational Linguistics

Xun Yan
Slobodanka Dimova
April Ginther *Editors*

Local Language Testing

Practice across Contexts

Springer

Educational Linguistics

Volume 61

Educational Linguistics is dedicated to innovative studies of language use and language learning. The series is based on the idea that there is a need for studies that break barriers. Accordingly, it provides a space for research that crosses traditional disciplinary, theoretical, and/or methodological boundaries in ways that advance knowledge about language (in) education. The series focuses on critical and contextualized work that offers alternatives to current approaches as well as practical, substantive ways forward. Contributions explore the dynamic and multi-layered nature of theory-practice relationships, creative applications of linguistic and symbolic resources, individual and societal considerations, and diverse social spaces related to language learning.

The series publishes in-depth studies of educational innovation in contexts throughout the world: issues of linguistic equity and diversity; educational language policy; revalorization of indigenous languages; socially responsible (additional) language teaching; language assessment; first- and additional language literacy; language teacher education; language development and socialization in non-traditional settings; the integration of language across academic subjects; language and technology; and other relevant topics.

The *Educational Linguistics* series invites authors to contact the general editor with suggestions and/or proposals for new monographs or edited volumes. For more information, please contact the Editor: Marianna Georgouli, Van Godewijckstraat 30, 3300 AA Dordrecht, The Netherlands.

All proposals and manuscripts submitted to the Series will undergo at least two rounds of external peer review.

This series is indexed in Scopus and the Norwegian Register for Scientific Journals, Series and Publishers (NSD).

Xun Yan • Slobodanka Dimova • April Ginther
Editors

Local Language Testing

Practice across Contexts

Springer

Editors
Xun Yan
Department of Linguistics
University of Illinois at Urbana Champaign
Urbana, IL, USA

April Ginther
Department of English
Purdue University
West Lafayette, IN, USA

Slobodanka Dimova
Centre for Internationalisation and Parallel Language Use, Faculty of Humanities
University of Copenhagen
Copenhagen, Denmark

ISSN 1572-0292 ISSN 2215-1656 (electronic)
Educational Linguistics
ISBN 978-3-031-33540-2 ISBN 978-3-031-33541-9 (eBook)
https://doi.org/10.1007/978-3-031-33541-9

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Contents

Contributors

Ivy Chen Language Testing Research Centre, University of Melbourne, Melbourne, Australia

Lixia Cheng Purdue Language and Cultural Exchange, Purdue University, West Lafayette, IN, USA

Sara T. Cushing Department of Applied Linguistics and English as a Second Language, Georgia State University, Atlanta, GA, USA

Slobodanka Dimova Centre for Internationalisation and Parallel Language Use, Faculty of Humanities, University of Copenhagen, Copenhagen, Denmark

April Ginther Department of English, Purdue University, West Lafayette, IN, USA

Marita Härmälä Finnish Education Evaluation Centre, Helsinki, Finland

Raili Hilden Department of Education, University of Helsinki, Helsinki, Finland

Annemiek Huisman Language Testing Research Centre, University of Melbourne, Melbourne, Australia

Martina Hulešová Institute for Language and Preparatory Studies, Charles University, Prague, Czechia

Yasin Karatay Department of English, Iowa State University, Ames, IA, USA

Ute Knoch Language Testing Research Centre, University of Melbourne, Melbourne, Australia

Martyna Kozlowska The Université du Québec à Montréal, Montreal, Canada

Luigi Lerose British Sign Language and Deaf Studies, University of Central Lancashire (UCLan), Preston, UK

Xiaorui Li Oral English Proficiency Program, Purdue University, West Lafayette, IN, USA

Sasithorn Limgomolvilas Chulalongkorn University Language Institute, Chulalongkorn University, Bangkok, Thailand

Xiaohua Liu The Chinese University of Hong Kong, Shenzhen, China

Stephen Daniel Looney Department of Applied Linguistics, Pennsylvania State University, University Park, PA, USA

Sabrina Machetti Certification of Italian as Second/Foreign Language Centre, University for Foreigners of Siena, Siena, Italy

Paola Masillo Certification of Italian as Second/Foreign Language Centre, University for Foreigners of Siena, Siena, Italy

John Read The University of Auckland, Auckland, New Zealand

Haoshan Ren Department of Applied Linguistics and English as a Second Language, Georgia State University, Atlanta, GA, USA

Suzanne Springer The Université du Québec à Montréal, Montreal, Canada

Kateřina Vodičková Institute for Language and Preparatory Studies, Charles University, Prague, Czechia

Jirada Wudthayagorn Chulalongkorn University Language Institute, Chulalongkorn University, Bangkok, Thailand

Xun Yan Department of Linguistics, University of Illinois at Urbana Champaign, Urbana, IL, USA

Local Language Testing Practice across Contexts

Xun Yan, Slobodanka Dimova, and April Ginther

Abstract Local language testing, though the most common form of language assessment practice, has not been formally recognize until very recently. In this chapter, we provide a working definition of local language tests and an overview of local assessment issues addressed in each chapter. We argue that local language tests present unique challenges that are shaped by test purpose, test policy, and other contextual factors; however, local contexts can also provide rewarding opportunities for testing research and practice when stakeholders work collaboratively to address the assessment issues. This collaborative practice is conducive to the development of language assessment literacy within the local community.

Keywords Local language test · Test purpose and use · Language policy · Language assessment literacy

Despite the ubiquity of local language test practices, local tests rarely stand in the spotlight. Nonetheless, interest in local language testing research among scholars is stunning. Our call for the 2022 special issue of *Language Testing*, Local Tests, Local Contexts, attracted 104 submissions, not only a record number in the 40 year history of the journal but also a window to the rich diversity of empirical research that underlies many local assessment practices. Local testing represents a major research trend in language testing and assessment (Ockey & Green, 2020).

X. Yan (✉)
Department of Linguistics, University of Illinois at Urbana Champaign, Urbana, IL, USA
e-mail: xunyan@illinois.edu

S. Dimova
Centre for Internationalisation and Parallel Language Use, Faculty of Humanities, University of Copenhagen, Copenhagen, Denmark
e-mail: slobodanka.dimova@hum.ku.dk

A. Ginther
Department of English, Purdue University, West Lafayette, IN, USA
e-mail: aginther@purdue.edu

X. Yan et al. (eds.), *Local Language Testing*, Educational Linguistics 61,
https://doi.org/10.1007/978-3-031-33541-9_1

With this edited volume, we continue the discussion of local language testing practice. The chapters in this volume discuss testing needs that emerge within local contexts, the instructional practices that develop in association with those needs, and the investigations enabled by analyses of test scores for program evaluation and research. In these discussions, local testing practices are grounded in the theoretical principles of language testing (Bachman & Palmer, 1996; Chapelle, 2012; Fulcher, 2012; Kane, 2006; Weir, 2005); in addition, based on their experiences with local testing, the authors supplement the theoretical content with practical examples of local language tests and how they can be designed, revised, and maintained, and how the process of validation works within and across different institutional contexts.

Given the importance of the local context, each of the chapters offers description of the local instructional and/or testing context, identifies the local need for test design or revision, and provides insights gained from test development practices and data analyses. Each chapter presents local assessment solutions to sets of problems, which can lead to one or a sequence of revisions to the local tests – some optional, and attractive, others required. The complementary processes of construct examination and definition, operationalization, and item development and analyses, lay strong foundations for subsequent score use and interpretation. At the same time, local language testing practice requires a considerable investment of time and labor, and all of the development efforts described in this volume were conducted across multiple years.

1 Test Purpose(s)

Local language tests and assessments are developed for use within a specific context and serve a variety of purposes, including entry-level proficiency testing, admissions testing, placement testing, international teaching assistant testing, testing for immigration purposes, and program evaluation (Dimova et al., 2020, 2022). These purposes cannot be addressed effectively through the use of either broadly-purposed, standardized, commercial instruments or specialized, focused classroom assessments. Read (2015, 2016) and Elder (2017) articulate features [GA3] of their importance, functionality, and widespread use, paving the way for a broader representation of local language testing practitioners to contribute to the conversation.

This volume presents research related to local language tests designed for various purposes. For instance, in the chapter "Using Multi-faceted Rasch Analysis to Examine Raters, Prompts, and Rubrics in an Office-Hour Role-Play Task," Ren et al. discuss the use of a local language test for screening prospective international teaching assistants for English language proficiency, which is standard practice at most large public institutions in the United States with large populations of international graduate students (ITAs). One advantage of local ITA tests is the ability to select tasks and items that represent features and needs of the local student populations and academic environment. ITA tests at some institutions, as is the case in this

chapter, include tasks intended to replicate aspects of the ITA context, in this case, an interactive task where the test takers are required to participate in role-plays relevant to instructor/student interactions. Where multiple versions of these tasks exist, e.g., making a request to change a grade or make up an exam, item difficulty becomes a critical quality control issue. Differences across difficulty estimates pose a threat to technical quality and validity if differences in prompts systematically affect scoring, which in this case, could not be ruled out.

Ren et al.'s examination revealed that raters, who assigned one of four prompts as part of the test admin process, displayed preferences for particular prompts when assigning prompts. While assumed to be comparable in terms of difficulty, the authors' observation that raters prefer particular items leads them to examine whether difficulty might underlie raters' preferences.

Then, Hulešová and Vodičková address three related problems in their examination of the Czech Language Certificate Exam (CCE), a high-stakes examination ranging from the CEFR A1 to C1 levels, developed at Charles University, but used for university admissions and immigration purposes across Czechoslovakia. Their study, conducted over 7 years, was undertaken (1) to address improvements suggested by an ALTE certification project involving realignment to the CEFR, and (2) to provide a link/explanation for how CCE level cut scores to align with the local, university-wide requirement that all passing cut scores be set at 60%, and (3) to examine/establish stable item difficulty estimates, allowing their inclusion in a CCE item bank.

The researchers explain that the CCE is not administered under conditions that allow the use of IRT-based approaches to test equating and task banking, i.e., a large and continuous supply of test takers for pre-testing, so they present a popular solution to the problem of establishing test form comparability: standard setting. The Direct Comparison method was chosen for the listening and reading subsections and a modified Body of Work approach for the writing and speaking subtests. Ultimately, their CCE revision ended up addressing not only their original concerns, but also additional issues. The authors caution that test developers should anticipate spending a considerable amount of time and effort to complete similar undertakings; however, they conclude that the outcomes, a revised CCE exam with a solid validity argument and a development/research team that has grown in terms of knowledge, experience, professionalism, self-motivation and self-confidence, was well worth the effort.

When developing a test in a local context, a possibility arises to take into consideration multiple test uses. The local test could be designed with multiple purposes in mind or revised and adapted for the additional uses as they arise. One way to enhance the applicability of the local test for multiple purposes, especially if institution-external uses are anticipated, is to align the test with national or international standards. In their chapter "Speaking "CEFR" About Local Tests: What Mapping a Placement Test to the CEFR Can and Can't Do," Springer and Kozlovska present the alignment procedures undertaken in order to (1) report English proficiency scores that can be used for placement in ESL courses that are linked to CEFR proficiency levels (a low-stakes use), (2) provide evidence of the proficiency levels

required for admission and placement, and (3) provide evidence of level needed for graduation (another high-stakes use), as well as to help instructors to develop CEFR benchmarked materials (a programmatic use).

In their chapter "Designing a New Diagnostic Reading Assessment for a Local Post-admission Assessment Program: A Needs-Driven Approach," Liu and Read present a problem associated with the extension of their local test's purpose from screening to diagnosis. The problem involves a mismatch between the representation of academic reading, indicated on the post-entry test DELNA by a single score, and the more expansive set of academic reading skills required of entry-level students but not represented on the test. The test revision they propose is based on examination of the literature that provides the foundation for a needs analysis, first as a pilot with a small group of students, interviewed in regards to their performance of particular reading tasks, followed by the development of a questionnaire, subsequently administered to a larger group of incoming students (N = 220).

Cheng and Li present another case where an extended test purpose is required, and their chapter "Pre-post Elicited Imitation: Documenting Student Gains in the Purdue Language and Cultural Exchange" highlights the importance of the relation between assessment and instruction for a local program. In fact, the continued existence of the program in which the test was embedded was predicated on students' demonstration of gains in proficiency after a two-semester, two-course sequence of instruction. For many good reasons, including time and expense, students' proficiency gains at the end of a language course are seldom investigated. The institution's requirement that the program demonstrate accountability in terms of proficiency gains presented challenges: would the available instruments lend themselves to the extended purpose of measuring gains and if so, which subsection(s) might be sensitive enough to demonstrate gains? The authors decide to use the elicited imitation (EI) subsection of their local proficiency test, the ACE-In as the potential arbiter of gain. Like Liu and Read, they begin with research literature in support of the reliability of EI as a measure of implicit knowledge/general language proficiency. In addition, EI offers some advantages in terms of ease of item development and rater training when ongoing use is required. The performance of the EI subsection of the Ace-IN is examined in terms of standard, expected technical qualities using classical test theory (CTT) for both the pre- and post-test EI administrations. Their analyses also include hedges-g, a measure of instructional sensitivity, an addition to standard technical analyses, allowing the authors to address item quality in relation to instruction.

2 Local Language Tests as Policy Instruments

The need for development of local language tests often arises as a result of the implementation of an implicit or an explicit language or educational policy in a particular context. Although each local test presented in the chapters of the volume respond to language policies (e.g., policies regarding ITA employment, university

admission, placement in language courses, immigration, national foreign language assessments), two chapters specifically focus on issues associated with the role of local language tests in the implementation of fair and just language policies.

In their chapter, "Placement Test Overriding Rules: Tension Between Stakeholders' Conceptualizations of Fairness," Chen et al. discuss the role of local language testing in the establishment of fair placement policies of students at appropriate proficiency level courses in Chinese and French at the University of Melbourne. This chapter highlights the issues that arise from stakeholders' discrepant perceptions of the roles of equity and equality, which are the major considerations for the establishment of fair placement policies. Test fairness is often evaluated through analyses of the psychometric qualities of the local test. In those examinations, test reliability is considered to be the foundation for equal and fair treatment of test takers. Equal treatment of students under the testing procedures and for instructional decision-making based on test scores is essential. However, equality may not suffice when there is student group diversity in terms of their learning needs. In order to establish equity in learning outcomes across the groups, subsequent instruction may require different kinds of support and accommodations. [GA7] As such, when a local test caters to different test taker groups, it sometimes becomes difficult to determine the most appropriate and equitable testing policy. The ideal outcomes of fair and equitable placement in university language courses would include students' satisfaction and confidence with placement, accurate placement in terms of proficiency levels, and homogenous classes so that the classes are not too difficult for some students and too easy for others. However, due to the variation in students' first-languages and language learning experiences, the attainment of the ideal outcomes is complicated.

Machetti and Masillo, on the other hand, bring up an important discussion regarding the role of local language testing as a policy tool in the European context, where an increased number of countries require evidence of a certain language proficiency level in the national language for migration and integration purposes. They raise concerns about implementation of language tests as part of an immigration policy as the misuse of scores may lead to discrimination, exclusion, and injustice especially if they lack validity. However, the authors argue that despite their sociopolitical background, local language tests can gain value if they are integrated in local instructional programs and contribute to development of migrants' language skills based on their individual needs. In the Italian context, where instead of one large standardized language test administered at national level, the Ministry of Interior requested that language tests be developed and administered locally at the public adult educational centers throughout the country, i.e. the Provincial Adult Education Centres (CPIA). The problem that emerged with this policy decision was that CPIAs lacked teachers trained in language testing and assessment who could develop and administer reliable and valid language tests based on the ministerial guidelines that include information about test structure and proficiency level of A2 on the CEFR scale. Therefore, language teachers engaged in ad hoc testing practices with limited quality control, which introduced a wide range of discrepancies in the testing procedures across the different CIPAs.

3 Assessment of Language for Specific Purposes

Language for specific purposes tests (LSP) can be either local or large-scale, depending on their scope and stakes (Dimova et al., 2022). LSP tests undertake local test characteristics if they are designed to assess domain-specific language in a particular target language domain (TLU), with the intention to have an impact within the local context, especially by being embedded in an educational program. Local LSP tests are developed to elicit performances in specialized TLU domains, support instruction, and/or provide information about whether the learning goals and objectives have been reached.

Three chapters in the volume discuss local LSP testing practices with the purpose to assess to what degree the course goals and objectives were reached in different disciplinary courses. In the chapter "Identifying the Phonological Errors of Second-Modality, Second-Language (M2-L2) Novice Signers Through Video-Based Mock Tests," Larose discusses a constant challenge that local language testers, teachers, and program administrators face, i.e. how to make scores on language tests useful for teaching and learning. Regardless of the wide array of quality control procedures performed on a test, if the test is not embedded in a language program or does not make fine-grained diagnostic information directly useful for instructors and students, it is unlikely to be highly useful in a local context. In his study, Lerose tackles this problem in the assessment of British Sign Language (BSL) in his university context, where he noticed a predominance of phonological errors among novice BSL learners, and the learners appeared to make slow progress on improving their phonological skills. Through a small-scale investigation, Lerose examined how the mock tests can be used to identify the common types of phonological errors and help learners improve their phonological accuracy. For the mock test devised to meet this local language need, ESL learners had to send a video clip of themselves signing about an assigned topic. To categorize the errors, the teacher-researcher used the five phonological parameters of BSL: handshape, location, orientation, movement, and non-manual features.

Wudthayagorn and Limgomolvilas focus on assessing English for pharmaceutical sciences, where medication distribution is considered one of the most important skills mandated for all pharmacy students (see "An Indigenous Rubric for Assessing Dispensing Drugs in English of Thai Pharmacy Students: Reliability and Perception Issues"). Their paper presents efforts they made at Chulalongkorn University in Thailand to create and validate an indigenous rubric for assessing dispensing medication in English of Thai pharmacy students. The authors first drew on the indigenous criteria developed by domain experts and insights gained from teaching English for Pharmaceutical sciences courses to create a rubric for assessing Thai pharmacy students' medication distribution skills in English. The rubric consists of three major criteria, including pharmaceutical knowledge, language use, and strategic competence. To further validate the effectiveness of this indigenous rubric, the authors performed a Many-Facet Rasch Measurement (MFRM) on scores raters assigned based on the rubric and qualitatively analyzed interview data with the raters about their perceptions of the rubric and students' performance.

In the chapter, "Using Spoken Dialog Systems to Assess L2 Learners' Oral Skills in a Local Language Testing Context," Karatay discusses the lack of adequate oral proficiency assessment, which can negatively impact language instruction in ESP contexts in Turkey. In the focus of the chapter is a task-based Tourism English oral performance assessment, designed using a specialized spoken dialogue system (SDS) in which a computer program was designed to take the role of a hotel guest and test takers act as a hotel employee. To investigate whether task administration conditions and the rubric for scoring are appropriate for providing evidence of targeted language ability, Karatay used a mixed-methods research design that included data from student oral performances (n = 30), students' post-test questionnaire responses, and semi-structured individual interviews. Results suggest that a dialogue-based oral assessment in ESP programs may lead to positive washback on teaching and learning by introducing more communicatively-oriented language instruction than had been the case prior to the introduction of the assessment.

4 Languages, Contexts, Scope

Issues related to local testing and assessment practices are applicable across different languages, contexts, and scope. Given the spread of English, it is inevitable that the majority of chapters featured in this volume focus on the assessment of English. These chapters deal with needs for assessment that arise in the local context and cannot be addressed with the existing large scale, international English tests. While only one chapter deals with English assessment in obligatory education (chapter "Exploring the Alignment of Test Tasks with Curriculum Goals Among Finnish 6th Graders"), most of the chapters deal with ESL at university level, where the authors are concerned with various issues, such as the relationship between test analysis, test quality, and test administration in oral assessment of international teaching assistants (see, for example, chapter "Using Multi-faceted Rasch Analysis to Examine Raters, Prompts, and Rubrics in an Office-Hour Role-Play Task"), establishment of diagnostic assessment of university students (see, for example, chapter "Designing a New Diagnostic Reading Assessment for a Local Post-admission Assessment Program: A Needs-Driven Approach"), provision of adequate data for instructional program evaluation (see, for example, chapter "Pre-post Elicited Imitation: Documenting Student Gains in the Purdue Language and Cultural Exchange"), alignment of local tests with CEFR (see, for example, chapter "Speaking "CEFR" About Local Tests: What Mapping a Placement Test to the CEFR Can and Can't Do") as well as relevant language assessment associated with specific disciplinary courses (see, for example, chapters "An Indigenous Rubric for Assessing Dispensing Drugs in English of Thai Pharmacy Students: Reliability and Perception Issues" and "Using Spoken Dialog Systems to Assess L2 Learners' Oral Skills in a Local Language Testing Context").

Nonetheless, the field of applied linguistics as a whole is experiencing a multilingual turn (Gorter & Cenoz, 2017; Schissel et al., 2018; Shohamy, 2011). Along with the trend, discussions of proficiency (especially in the assessment of English) have gradually shifted from the native vs. nonnative (or first vs. second language) dichotomy to a model of diversity, where the ability to communicate in more than one language is conceptualized with a more positive light rather than viewed as a form of deficiency. Moreover, the multilingual turn has also prompted more attention on the assessment of languages other than English. In this volume, one third of the contributing chapters focus on assessment of languages other than English, which reflects the recent years' trend of a rising interest in local tests of less commonly taught languages is evident (Yan et al., 2020). These include a chapter on sign language (chapter "Identifying the Phonological Errors of Second-Modality, Second-Language (M2-L2) Novice Signers Through Video-Based Mock Tests"), Chinese and French (chapter "When Student Background Overrides Score-Based Placement: Tension Between Stakeholders' Conceptualizations of Fairness"), Czech (chapter "Validation of the Czech Language Certificate Exam with Respect to the Local Context"), and Italian (chapter "Local Tests, Local Contexts: The Italian Language Testing Scenario"). Local language testing is especially relevant for other languages due to the lack, or limited selection, of international, or off-the-shelf standardized tests.

In terms of scope, while it may seem reasonable to assume that local language tests are designed and used within a particular institution, we argue that the test scope is not a defining characteristic of local language tests. Regardless of whether the test is used to inform decisions in a particular course or an instructional program or used for decision-making across institutions and at national level, what characterizes the test as local is not its scope but its curricular and instructional embeddedness, often achieved through active involvement of local actors in test design and administration.

Härmälä and Hilden discuss a broader scope of local language testing by analyzing the alignment of the written, official Finnish curriculum and a national test of English-language learning outcomes at the end of 6th grade (see chapter "Exploring the Alignment of Test Tasks with Curriculum Goals Among Finnish 6th Graders"). They examine alignment of the items through the lens of the revised Bloom's Taxonomy and by determining the CEFR levels of the classified items through a standard-setting procedure. While not the focus of the study, the researchers do mention that the test is now administered online. Bloom's Taxonomy was chosen as an interpretative instrument because of its popularity for evaluation of skills in other content areas in Finland where analyses revealed concerns with representation of lower- and higher-order Bloomian skills, e.g., items requiring the identification of factual information versus those requiring evaluation/analysis/creation of information. The authors classify complexity of knowledge according to a system where verbs and nouns comfortably combine to represent levels, where entry level interaction tasks are represented as exchange (v) thoughts/information (n); entry-level interpretation as understanding texts/speech/messages; and entry-level production tasks as describing topics. The verb/noun pairs are intended to capture language

functions and forms in association with task complexity and item difficulty. The authors report that task difficulty in all of the four skills corresponded well to the targeted CEFR levels, and that the representation of skills across Bloomian levels was appropriate, i.e., without an over-representation of lower-level skills.

5 Language Assessment Literacy

As we argued in Dimova et al. (2020, 2022), one of the advantages of local language testing is the empowerment of testers, teachers, and other stakeholders who are involved in the assessment practice. In the case of large-scale proficiency tests, most stakeholders (except item writers) are consumers of the test who rely on test developers' input to interpret the test scores. In contrast, in most local assessment contexts, stakeholders tend to enact a more active role in developing assessment solutions. This collaborative practice gives them agency and confidence in the assessment practice, which helps enhance their LAL. More importantly, the development of LAL involves extensive apprenticeship and reflection (Yan & Fan, 2021), and the involvement in an actual test development and revision project creates direct opportunities for reflections upon assessment practices. Although not all chapters are framed directly under language assessment literacy (LAL), the impact of local language testing on LAL development is evident from the insights the researchers gained from the local assessment practices.

For example, Springer and Kozlowska reflected in their chapter that although their efforts on mapping the local tests to CEFR still need ongoing validation, this alignment project enabled them to become familiar with the procedures involved linking and standard setting. Therefore, the outcome of the assessment practice is not only a mapping between the local test and CEFR but also a clearer idea of the remaining questions and approaches to undertake to address those questions. Similarly, in Hulešová and Vodičková's chapter, the validation of the Czech Language Certificate Exam (CCE) facilitated growth of shared assessment knowledge, and the LAL development further changed the attitudes among the team members, especially in terms of their perceived responsibilities on the test and the ability to follow a principled approach to gather and examine validity evidence. Involvement in the project also made the process of gathering validity evidence part of the day-to-day quality control procedures for the CCE exams.

6 Insights Gained

Although each chapter in this edited volume represents a specific context with unique language testing needs, the pervasive feature among most chapters is the continuous development of language tests based on immediate feedback from various local actors or based on the changing needs in the local context. Given that the

selection of an appropriate testing method depends on the testing purpose and the local educational values and norms, what may be considered an appropriate testing practice varies across contexts.

Research in local language testing has great potential to contribute to our conceptualizations of the assessed constructs and our theoretical discussions about validity in the field. However, the primary purpose of local language testers' engagement in validation research tends to be improvement of the local testing practices and provision of validity arguments in discussions with local stakeholders. Although the diversity of language testing practices may pose challenges in making generalizations about particular methodological approaches in language testing, if the publication trend of local language test research increases over time, certain contextual, practical, and methodological patterns are likely to emerge.

Acknowledgements We would like to thank our external reviewers for their contributions to the volume. In addition, our many thanks to Weijian Yan and April Fidler at Purdue University for their support with management and editing.

References

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford University Press.

Chapelle, C. A. (2012). Validity argument for language assessment: The framework is simple…. *Language Testing, 29*(1), 19–27. https://doi.org/10.1177/0265532211417211

Dimova, S., Yan, X., & Ginther, A. (2020). *Local language testing: Design, implementation, and development*. Routledge. https://doi.org/10.4324/9780429492242

Dimova, S., Yan, X., & Ginther, A. (2022). Local tests, local contexts. *Language Testing, 39*(3), 341–354. https://doi.org/10.1177/02655322221092392

Elder, C. (2017). Assessment in higher education. In E. Shohamy, I. G. Or, & S. May (Eds.), *Language testing and assessment* (pp. 271–287). Springer. https://doi.org/10.1007/978-3-319-02261-1_35

Fulcher, G. (2012). *Practical language testing*. Routledge. https://doi.org/10.4324/980203767399

Gorter, D., & Cenoz, J. (2017). Language education policy and multilingual assessment. *Language and Education, 31*(3), 231–248.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (pp. 17–64). American Council on Education/Praeger.

Ockey, G. J., & Green, B. A. (2020). *Another generation of fundamental considerations in language assessment: A festschrift in honor of Lyle F. Bachman*. Springer. https://doi.org/10.1007/978-981-15-8952-2

Read, J. (2015). *Assessing English proficiency for university study*. Palgrave Macmillan. https://doi.org/10.1057/9781137315694

Read, J. (2016). *Post-admission language assessment of university students*. Springer. https://doi.org/10.1007/978-3-319-39192-2

Schissel, J. L., Leung, C., López-Gopar, M., & Davis, J. R. (2018). Multilingual learners in language assessment: Assessment design for linguistically diverse communities. *Language and Education, 32*(2), 167–118.

Shohamy, E. (2011). Assessing multilingual competencies: Adopting construct valid assessment policies. *Modern Language Journal, 95*(3), 418–429.

Weir, C. (2005). *Language testing and validation: An evidence-based approach*. Palgrave Macmillan. https://doi.org/10.1057/9780230514577

Yan, X., & Fan, J. (2021). "Am I qualified to be a language tester?": Understanding the development of language assessment literacy across three stakeholder groups. *Language Testing, 38*(2), 219–246. https://doi.org/10.1177/0265532220929924

Yan, X., Lei, Y., & Shih, C. (2020). A corpus-driven, curriculum-based Chinese elicited imitation test in US universities. *Foreign Language Annals, 53*(4), 704–732. https://doi.org/10.1111/flan.12492

Using Multi-faceted Rasch Analysis to Examine Raters, Prompts, and Rubrics in an Office-Hour Role-Play Task

Haoshan Ren, Stephen Daniel Looney, and Sara T. Cushing

Abstract Locally-administered placement tests for International Teaching Assistants (ITAs) impact the quality of ITAs' graduate studies and the quality of undergraduate education. However, in-house testing programs face challenges in terms of funding, time, and expertise to ensure the reliability and address the validity of their ITA tests. This chapter examines comparability of prompts used in an office-hour roleplay task on an in-house ITA test. The role-play task requires test-takers to respond to one of four office-hour scenarios where the test-taker plays the role of the teacher and one of the two raters plays the role of the undergraduate student. We used Many-Faceted Rasch Measurement (MFRM) to investigate prompt difficulty, rubric consistency, and selected interactions. All raters were self-consistent, but we found significant differences in raters' relative severities. Further examination reveals that, although raters were instructed to select one of the four prompts randomly, some raters selected certain prompts to the exclusion of others. The rater x prompt interaction shows that raters' internal consistency is related to selection, while variation in rater severity is an independent effect. This chapter provides insights into practices that connect test analysis, test quality, and test administration in a local context.

Keywords Local language testing · International Teaching Assistants (ITAs) · Roleplay tasks · Many-faceted Rasch Measurement (MFRM)

H. Ren (✉) · S. T. Cushing
Department of Applied Linguistics and English as a Second Language, Georgia State University, Atlanta, GA, USA
e-mail: hren2@gsu.edu; stcushing@gsu.edu

S. D. Looney
Department of Applied Linguistics, Pennsylvania State University, University Park, PA, USA
e-mail: sdl16@psu.edu

X. Yan et al. (eds.), *Local Language Testing*, Educational Linguistics 61,
https://doi.org/10.1007/978-3-031-33541-9_2

1 Introduction

International Teaching Assistants (ITAs) compose a sizable portion of U.S. universities' teaching labor force. They are typically not first language English speakers, and rarely have teaching experience prior to entering U.S. universities. Since the 1980s, a discourse has emerged which frames ITAs as lacking the English proficiency needed to teach (Bailey, 1983, 1984). Over the years, American universities have taken steps to ensure the English proficiency of graduate student instructors in an effort to improve the educational experiences for both ITAs and the undergraduate students they teach. A manifestation of these efforts has been the establishment of screening procedures to assess ITAs' oral proficiency levels upon their entrance to the programs. The decisions based on ITA screening test scores are critical to the length and quality of ITAs' graduate studies and the quality of undergraduate students' educational experiences. If unqualified ITAs are assigned independent teaching positions, both ITAs and departments may encounter negative learning outcomes from undergraduates, as well as complaints from their parents (Bailey, 1984).

A commonly used approach for proficiency screening is to set a cut-off score for the speaking section of a standardized language test such as the Test of English as Foreign Language (TOEFL) or the International English Language Testing System (IELTS), which usually varies by individual schools or specific departments. In addition to this score, many universities choose to use locally administered in-house tests with the specific purpose of measuring prospective ITA's oral proficiency and teaching ability. Typically, these locally administered ITA tests measure prospective students' readiness for teaching responsibilities in addition to their ability to study in an English dominant academic environment (Elder, 2017). Accordingly, such tests assess a more targeted and restricted range of language abilities (Dimova et al., 2020). These tests are also used for placement purposes in institutions that provide ITA programs where prospective ITAs may enroll in coursework to prepare for their teaching roles.

In terms of task design, in-house ITA tests take into consideration the features and needs of their local student population and academic environment. Many ITA tests follow the format of a typical three-part oral proficiency test (Sandlund et al., 2016), in which items start with testing lower-level abilities and gradually increase in complexity, moving from monologue via dialogue to interactions. However, ITA tests at some institutions include tasks intended to replicate aspects of the ITA context, such as interactive tasks where the test takers are asked to participate in role-plays relevant to university teaching. Where multiple versions of these tasks exist, an important validity issue is the extent to which differences in prompts create construct-irrelevant variance by eliciting differences in performance or differences in scoring.

Another essential aspect of validity for an in-house test is the rating scale used to score performances. In theory and practice, scales developed for second language assessment always go through recursive stages of validation and revisions (Knoch et al., 2021, pp. 83–84). In rater-mediated, performance-based assessments, the

quality of rubrics is one of the critical aspects that affect rater behaviors (Dimova et al., 2020). To examine rater behaviors on a rating scale, language testers often use Rasch modeling which provides detailed analysis on rater variability influenced by various factors (McNamara et al., 2019).

Motivated by local needs for addressing these challenges, this chapter describes the process of providing validity evidence for two critical aspects of an in-house ITA test, namely, the rating scale used and the prompts involved in one of the test tasks.

1.1 The Oral English Test (OET)

The Oral English Test (OET) is a locally administered in-house ITA test used in a large public university in the northeast U.S. Like many ITA screening tests, the OET is administered prior to the start of each semester, and the scores are used for placing prospective ITAs into appropriate English courses or recommending them as ready instructors in their home departments. The test is administered to over 400 international graduate students each year.

There are four tasks in total in the OET. This study focuses on the fourth task: an office-hour role-play. In this task, the examinee is given a card chosen from four possible alternatives with printed descriptions of a hypothetical office-hour scenario commonly faced by teaching assistants. The four prompts (Appendix A), approximately 70 words each, respectively present the following scenarios: negotiating attendance policy, discussing chronic lateness, requesting an extension, and requesting a makeup test. The test-takers are given 1 min to read the scenario and prepare. During the task, one of the two raters plays the role of a student, and the test-taker plays the role of the teaching assistant.

These prompts were chosen to simulate situations in which ITAs may need to negotiate sensitive course-related issues that are common in interactions with students. This task, together with the other tasks in the OET, has gone through several rounds of revisions in the history of this ITA test. The role-play format was introduced in 2005 to replace a picture description task (Gevara, 2016). This change was made to better reflect the real situations ITAs will face in their academic context in the university. The four prompts developed for these task also went through revisions based on the quality and quantity of their elicited language from the test-takers.

All four tasks in the OET are rated on a 4-point scale using the same analytic rubric (Appendix B). The rubric was informed by Bachman's and Palmer's (2010) framework of language ability and was developed iteratively over several semesters based on two latent trait statistical models (see Gevara, 2016) and feedback provided by test raters. The current version of the analytic rubric includes the following eight criteria: Grammatical Constructions; Thought Groups; Tone Choice, Prominence, Task Response, Transition, Multimodality, and Question Response.

All raters went through training sessions where they were shown video clips of performances at different levels to calibrate their rating using the OET rubric. The

training sessions take place 2 days before the OET tests begin, and each rater, on average, receives 12 h of training before the actual rating. During each test, two raters are randomly paired with one test-taker, and each rater rates independently the test-taker's performance on all four tasks. If the summed scores given by the two raters are 10 points apart (out of a total of 128; i.e., 4 tasks * 8 criteria * 4 scale points) or cross a placement threshold, a third rater will watch the recording of the test and give a third score. All three scores are averaged to determine the test-taker's final score.

2 Testing Problem Encountered

The OET raters' rating decisions are influenced by a few task-specific factors pertaining to the OET office-hour role-play task. Although all raters for the OET test are experts in the field of Applied Linguistics, and they go through training sessions, they differ in their linguistic backgrounds and rating experience. These rater-related factors have been studied extensively in previous literature and are shown to be significantly influential to rating quality (i.e., Attali, 2016; Barkaoui, 2010; Isaacs & Thomson, 2013; Zhang & Elder, 2011). These factors are influential not only by themselves, but also manifested in their interaction with the use of rating criteria and task features. For the office-hour role-play task, one of the raters plays the role of an undergraduate student. Thus, the discourse during the task is co-constructed between the rater and the test-taker, and it inevitably imposes a more influential role on the rater to the test-taker's performance (Brown & Yule, 1989; Fulcher & Davidson, 2007; McNamara, 1996). In addition, the use of an analytic rubric in the OET test also likely requires a higher level of experience and expertise due to its high cognitive demand (Seedhouse et al., 2014). These factors could potentially lead to inconsistent ratings of the task on the eight criteria in the rubric. Therefore, a primary goal of this project is to gather information about how raters use the eight criteria in the rubric to rate test-takers' performances at different levels, hoping to inform rater training and improve the validity of the OET test.

Another main concern has to do with the validity and fairness of using different versions of prompts in the office-hour role-play task. As described above, the task involves four different prompts, and each test-taker only has one chance to respond to one of them. Although all prompts are designed to represent common scenarios of ITA office hours, the different content may elicit different amounts of ratable language, and thus unfairly advantaging or disadvantaging students given different prompts. In addition, although the raters were instructed to choose one of the four prompts randomly, a few raters tend to use one prompt over the others due to various reasons. This raises the concern of whether rater selection threatens the fairness of the test, especially if a rater prefers a prompt that is of different difficulty than the others. Therefore, another goal of this project is to test whether the four prompts are comparable in difficulty and despite rater preferences can still be assigned fairly. If item difficulty and preference interact to advantage particular groups, test designers and administrators can adjust prompt usage and rater training accordingly.

3 Literature Review

Previous validation studies on local ITA tests mainly focused on rubrics constructs features of ITA language production during the tests compared to their actual teaching contexts. Several studies contributed to the validation of commonly used constructs in ITA rubrics such as grammar (Thirakunkovit et al., 2019; Jameel & Porter-Szucs, 2014), pronunciation, and listening comprehension (Plough et al., 2010). Using both qualitative and quantitative methods, Plough et al. (2010) examined the criteria used in GSI OET (The Graduate Student Instructor Oral English Test, University of Michigan). They found that pronunciation and listening are the most influential categories for determining ITA's eligibility by the test. In a more focused effort to investigate the construct of lexico-grammatical features, Jameel and Porter-Szucs (2014) examined the office-hours task in the Graduate Student Instructor Oral English Test (GSI-OET) at University of Michigan. They found that the ITAs actually did not differ much from their native-speaking peers in their use of *native-like formulaic sequence*, but the ITAs do use significantly less contractions. Taking a different approach to the construct of grammar, Thirakunkovit et al. (2019) focused on grammatical complexity of ITA test-takers' language production at different proficiency levels. They compared features found in the spoken production to the language used in academic and non-academic speaking and writing genres. They found that higher proficiency ITAs use more grammatical features typically appearing in academic written genres, even though all ITAs generally use grammatical features found in academic spoken genres, such as modal verbs and finite adverbial clauses.

In an effort to depict the target language use domain for the ITAs, Cotos and Chung (2019) composed an ITA corpus including three instructional genres across 11 disciplines. They explored the use of functional language as categorized by the Knowledge Structure (KS) framework, and analyzed the linguistic instantiation of these functions across all types of texts in the forms of n-grams. This study reveals distinguishable patterns of functional language use in classes taught by the ITAs across three disciplines and genres, thus providing a detailed map for the ITA's target language use domain.

Although we see a growing number of validation studies on local ITA tests focusing on rubric constructs, there has not been any assessment literature that contextualizes prompt effects in local ITA tests. This may have to do with the fact that not many local ITA tests include tasks involving different versions of task prompts. Nevertheless, the effects of prompt variability on rating quality have been studied from various perspectives for both written and spoken tests. In literature focusing on written tasks, prompt comparability has been discussed in relation to prompt subjects (e.g., Hamp-Lyons, 1990; Weigle, 2002), test-taker backgrounds (e.g., Lim, 2009), language elicited by the prompts (Spaan, 1993), and interactions between these factors and rater behaviors (e.g., Reid, 1990; Weigle, 1999). In spoken assessments, prompts are found to significantly influence the interactional pattern of test-takers' performances (Leaper & Riazi, 2014). In addition, it was shown that prompt formation (i.e., question type, number of questions) may also play an important role

in shaping test-takers' performance during the task (Gan, 2010; Van Moere, 2006). In these studies, Rasch Modeling is commonly used as the quantitative method to detect prompt effects, whereas Conversation Analysis is used to illuminate qualitative differences in test performance elicited by different prompts.

4 Methods

Building on previous literature, we use Many-Faceted Rasch Measurement (MFRM, Linacre, 1989) to investigate the interactive role-play tasks specifically, especially regarding the difficulty of different versions of roleplay prompts. If differences in difficulty are found, we then examine whether rater selection preferences are related to difficulty. Specifically, we examine the following research questions that directly address the issues concerning the OET office-hour role-play task:

1. Do the eight criteria in the rubric function effectively and consistently when used to rate the office-hour role-play task?
2. What are the relative difficulties of the four prompts?
3. Is rater selection of an item related to item difficulty?
4. How does raters' use of prompts interact with their scoring of examinees' performance in the office-hour role-play task?

4.1 Participants and Role-Play Prompts

The OET scores used in the current analysis came from 225 international graduate students who took the OET in Fall 2019. Among a total of 28 raters, 12 were from the U.S., eight were from China, two from South Korea, two from Russia, one from Peru, one from Finland, one from Sweden, and one from Spain. Raters' ages range from 25 to 37. Sixteen of them were female, and 12 were male. All raters were faculty and graduate students in the Department of Applied Linguistics.

4.2 Data Analysis

The method of analysis used in this study is Many-Faceted Rasch Measurement (MFRM, Linacre, 1989), which is a general psychometric modeling approach to process many-faceted data of rater mediated assessment. In MFRM, a facet is a component of a test that may systematically influence test scores. For example, in rater-mediated tests such as the OET, examinees' performances are rated subjectively by human raters using rubrics that specify certain target criteria of the performances; therefore, the scores are likely to be influenced by many different facets

involved in this process. MFRM provides detailed information indicative of construct-relevant and construct irrelevant factors that influence human raters' rating. Results from a MFRM analysis can be used for monitoring fair use of tasks and rubrics, examining rater-bias and provide insights for rater training, and most importantly, making more informed decisions about the ITA's abilities.

The MFRM analysis was performed using the computer program *Facets* (Version 3.83.0; Linacre, 2019). *Facets* is a Windows-based software which assists in many applications of the Rasch model, particularly in the areas of performance assessment and paired comparisons. To perform MFRM, *Facets* uses joint maximum likelihood estimation (JMLE), a mathematical procedure that estimates the measurements of each of the facets (e.g. ability measurements for test-takers; severity measurements for raters, etc.) in terms of probabilities, and compares the estimates with the actual observed data to generate the best prediction model through multiple rounds of iteration. In this study, *Facets* uses scores of the examinees on each of the tasks given by each rater on each rubric aspect to estimate the proficiency levels of examinees, severity levels of raters, task difficulties, criteria difficulties, and scale category difficulties. All of these facets are calibrated onto the same scale using logit transformation. Logit transformation converts raw scores onto a log-odds equal-interval scale so that an equal distance between two data points represents an equal difference in the represented measurement (Bond & Fox, 2013). This process allows us to analyze and interpret each facet within a single frame of reference.

Prior to running *Facets*, data were handled to ensure appropriate modeling. First, *Facets* analysis requires each score level to have more than 10 ratings, and no levels should be disordered or too close to each other to avoid accidents in the data biasing (Linacre, 2019). Therefore, prior to running *Facets*, score level 1 and 2 were collapsed into one level because of the extremely low frequency of score 1. In addition, because each examinee encountered only one of the four prompts in the office-hour role-play, the existence of subsets was detected by *Facets*. The presence of subsets may render difficulties in deciding whether the score difference is caused by examinees' ability (construct relevant), or by the difficulties of the prompts (construct-irrelevant). To eliminate ambiguities caused by subsets in the data, all examinees were group-anchored at 0, which allows the analysis to assume that test-takers using all prompts are equivalently distributed, so that the prompts is the only influential variable and they will be positioned according to their estimated difficulty level (Schumacker, 1999).

Finally, a partial-credit MFRM model (Masters, 1982) was applied to reflect how the structure of a rating scale was used for each rubric criterion. The final four-facets partial-credit MFRM model is defined as follows:

$$\ln\left[P_{nijk} / P_{nijk-1} \right] = \theta_n - \beta_i - \alpha_j - \gamma_g - \tau_{ik}$$

where:

P_{nijk} = probability of examinee *n* receiving a rating of k on criterion *i* from rater *j*.

$P_{nijk\text{-}1}$ = probability of examinee *n* receiving a rating of *k−1* on criterion *i* from rater *j*.
θ_n = proficiency of examinee *n*.
β_i = difficulty of prompt *i*.
α_j = severity of rater *j*.
γ_g = difficulty of rubric criterion *g*.
τ_{ik} = difficulty of receiving a rating of *k* relative to a rating of *k−1* for criterion *i*.

To review, this study analyzes raters' use of an analytic rubric with eight criteria for a human-mediated ITA test and the raters' use of different prompts in the office-hour role-play task. The Facets program outputs a series of statistics for different criteria of the model. Before analyzing the statistics for the target research questions, the core assumptions, unidimensionality and local independence, for MFRM were checked and confirmed following the methods discussed in McNamara et al. (2019, pp. 153–155). In the following section, a model summary is provided using a Wright map, followed by a focused discussion of the rubric, the four prompts, and their interaction with raters through interpretations of the Facets output.

5 Results and Discussion

We begin with a presentation of candidate ability, rater severity, and item difficulty.

5.1 The Wright Map

Figure 1 presents the Wright map that visually summarizes the measures for candidate ability, rater severity, and the difficulty of the prompts and the rubric categories. The scale in the left-most column shows the logit measures (*Measr*) used as a ruler for all facets on the map. The second to the left column (*Examinees*) represents the positively oriented examinees' ability variation. Within this column, each asterisk (*) represents 4 examinees, and each dot (.) represents one. "Positively oriented" means that the higher an examinee ranks on the logit scale, the higher their estimated ability level is in the model. The third column provides information on the *Raters*. This ranking is negatively oriented to represent rater severity. Raters ranked higher on the rulers are more severe in their ratings. In this column, each asterisk represents one individual rater. The fourth column represents the negatively oriented measures for *Prompts* ranked by difficulty level. The higher a prompt is located, the more difficult the prompt is, and vice versa. As we can see in this column, Prompt 2 is the lowest on the scale. This means that it is easier for a given examinee to be rated at a certain score if they were assigned to answer Prompt 2 compared to the other prompts. The widest column labeled *Constructs* provides difficulty rankings of the eight constructs used in the OET rubric. For the eight criteria

in this study, Grammatical Constructions is ranked the highest, which means it has the highest difficulty level. The final four columns to the right present the rating scale structure for each of the four prompts. The rating scales were collapsed from a 4-level scale into a 3-level scale by regrouping all level 1 scores and level 2 score as one level; therefore, we see number 4 as the highest number on the top for each scale, and number 2 at the bottom. For the rating scales, each horizontal dashed line within the scale columns represents the threshold between two score levels. For example, Prompt 2, among all, has the widest range for score level 3. Specifically, the slightly higher threshold between level 3 and level 4 indicates that it is more difficult for an examinee to be awarded a score 4 if they used Prompt 2; while the lower threshold between level 2 and level 3 on Prompt 2 indicates that it is easier for examinees to be awarded a score of 3 compared to other prompts.

5.2 Rubric Criteria

Research question 1 investigates the usage of the 8 criteria in the rubric. The Wright map above in Fig. 1 provides a simple visual representation of the difficulty levels of each of the criterion. A more detailed report is presented in Table 1, which shows the measurement (in logits) of each criterion and their fit statistics. *Facets* also

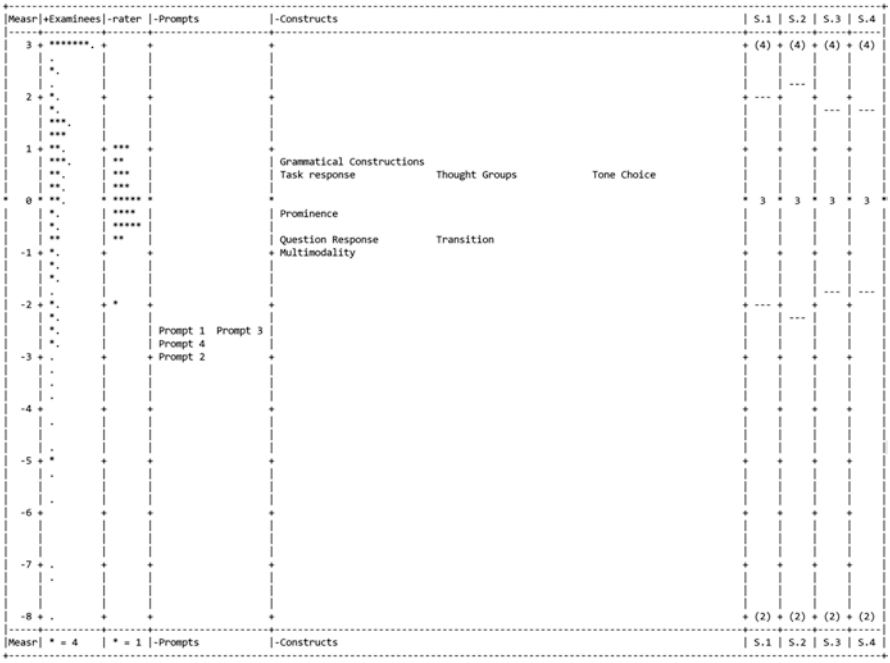


Fig. 1 Wright map for summary of all facets

Table 1 Output criteria table arranged by logit measurement

Criteria	Measure	S.E.	Outfit	Infit
Multimodality	−0.94	0.13	1.16	1.15
Question response	−0.68	0.12	0.80	0.91
Transition	−0.66	0.12	0.50	0.69
Prominence	−0.18	0.11	1.02	0.93
Task response	0.50	0.11	1.43	1.31
Thought groups	0.61	0.11	1.05	1.04
Tone choice	0.62	0.10	1.03	1.02
Grammatical constructions	0.72	0.10	0.90	0.92

RMSE .11, Adj. S.D. = .63, Separation = 5.61, Reliability = .97, Fixed (all same) chi-square = 251.5, d.f. = 7, significance (probability) = .00

provide statistics that report the reliability measures at the bottom of the table. These statistics help us to understand (1) if the criteria are functioning to measure the same underlying construct; (2) if the criteria are interpreted and used by the raters in the same way.

In Table 1, the *Measure* column provides the logit value for each criterion, ranked from the lowest to the highest. Because the criteria facet is negatively oriented, the higher the measure, the lower the score a criterion is expected to receive, which indicates a higher difficulty level. In this case, Multimodality has the highest logit score (lowest difficulty), and Grammatical Construction has the lowest logit score (highest difficulty). The *standard error (S.E.)* helps to estimate a confidence interval of the measurements. If we compare Multimodality and Grammatical Constructions, we can be very confident (99.5%) that the true difficulty level for Multimodality is within 0.42–1.02, and the Grammatical Constructions between −1.33 and −0.55. This result indicates that, in general, Multimodality is rated significantly higher than Grammatical Constructions. In other words, raters are likely to have interpreted these two criteria separately when rating test-takers' performances, which is preferred in this case.

The *Facets* MFRM analysis also reports individual-level statistics for rubric consistency via outfit and infit mean square residuals (i.e., *Outfit* and *Infit*). These mean square values show the size of randomness. Outfit mean square residual is the unweighted mean squared standardized residual, which is particularly sensitive to outliers. Infit mean square residual is the weighed mean-squared residual, which is less sensitive to outliers but more sensitive to unexpected rating patterns (Myford & Wolfe, 2003). The ideal value for both infit and outfit is 1, indicating a balanced amount of variability and consistency. While outlying ratings can be statistically adjusted, unexpected rating patterns are normally hard to remedy (Linacre, 2002). Therefore, we take a better look at the infit values for the criteria. Infit value less than 0.7 indicates an overfit, which means the criterion is not rated with much variability. An infit value of more than 1.3 indicates a misfit, which means the criterion has many unexpected ratings that does not fit the model (McNamara et al., 2019,

p. 45). In Table 1, Transition has an infit value of 0.69, which indicates slight overfit, meaning that raters tend to give invariant ratings on this criterion. On the other hand, Task Response has an infit value of 1.31, indicating a slight misfit. The other criteria were used consistently with good fit statistics.

Facets also provides statistics to interpret the degree of overall variability among the elements within each facet, which, in this case, refers to the criteria in the rubric. These statistics are provided at the bottom of Table 1. The *separation index* specifies how many distinct levels into which the model separated the criteria. For rubric criteria, a larger separation index is preferred because it indicates that the different criteria are measuring different levels of traits calibrated along the target construct (Myford & Wolfe, 2003). The separation index is calculated by the ratio of the corrected standard deviation (Adj. S.D.) to the root mean-square standard error (RMSE). A separation of 5.61 shows that the OET rubric for rating office-hour role-plays provides a fairly good separation of the different criteria of the target traits. The reliability index shows the degree to which the analysis reliably distinguishes the different levels of difficulty. A value of .97 indicates that the current analysis is highly reliable in separating the different levels of criteria in the rubric. Lastly, the Chi-square statistics tests the null hypothesis that no difference exists in the use of the eight criteria. The chi-square value of 251.5 with a degree of freedom 7 is significant at p = .00, indicating that the null hypothesis must be rejected.

5.3 Prompts

Table 2 shows the prompt measurement report. For the office-hour role-play task to be fair to all participants tested with different prompts, each prompt should have similar difficulty levels, and the scale rating levels should be used the same way by all the raters for each of the four prompts. In other words, the null hypothesis tested by the chi-square test (all four prompts are rated the same) is one that we do not wish to reject in this situation. However, the chi-square value of 35.2 at degree of freedom 3 is significant at p = .00, which means the four prompts are not at the same difficulty level. In addition, as indicated by the separation index and reliability index, the prompts are shown to be reliably separated into 2 levels, meaning not all four prompts are rated on the same difficulty level.

In terms of the functioning of rating scale categories, the MFRM analysis supports that the OET rubric used for each of the 4 prompts functions consistently. Table 3 shows the MFRM rating scale statistics that provide evidence for the functioning of the score levels used for each prompt. The rating scales for each prompt were examined following the *seven guidelines* summarized by McNamara et al. (2019) from Linacre's earlier work. First, the *Ctg. Total* column shows how many observations each category has. As mentioned in the methods, to ensure meaningful

Table 2 Prompt measurement report

Prompts	Measure	S.E.	Infit MnSq
Prompt 2	−3.04	0.08	1.02
Prompt 4	−2.63	0.13	1.05
Prompt 1	−2.50	0.08	0.97
Prompt 3	−2.49	0.06	1.00

RMSE = .09, Adj (True) S.D. = .20, Separation = 2.20, Reliability = .83, Fixed (all same) chi-square = 35.2, d.f. = 3, Significance (probability) = .00

Table 3 MFRM rating scale statistics

Score	Ctg. Total	Avge. Meas	Exp. Meas	Rasch-Andrich Thresholds	S.E.
Prompt 1					
2	138	−2.21	−2.31		
3	345	0.49	0.59	−2.00	0.14
4	621	3.65	3.60	2.00	0.10
Prompt 2					
2	45	−1.84	−1.98		
3	373	1.34	1.35	−2.17	0.20
4	654	3.64	3.65	2.17	0.09
Prompt 3					
2	113	−1.56	−1.56		
3	434	1.38	1.38	−1.76	0.15
4	989	3.3	3.300	1.76	0.07
Prompt 4					
2	19	−0.45	−0.78		
3	112	0.84	0.93	−1.74	0.28
4	174	3.08	3.05	1.74	0.16

analysis, level 1 and 2 were collapsed so that each level in each prompt has more than ten observations (*Guideline 1*). The average measures as indicated in the fourth column *Avge. Meas* showed that each score point increases at approximately the same value for each of the prompts (*Guideline 2*). The *Avge. Meas* also showed that measures for levels are not disordered (*Guideline 3*). *Guideline 4* dictates that the frequency of data points in each category should result in a smooth distribution. The cumulative distribution shows that our data meets this requirement as well. Next, examining the discrepancy between the average measure (*Avge. Meas*) and the expected measure (*Exp. Meas*) confirms that all measures are near their expected values. The data also meets the rest of the guidelines which require the mean-square value to be less than 2 and the Rasch-Andrich thresholds to increase at each level by a logit value larger than 1.4 but smaller than 5.

5.4 *Prompt and Rater Interaction*

To better understand the different usage of prompts by raters, a further analysis was carried out to investigate the interaction between raters and prompts. *Facets* reports details of which prompt is rated differently by which rater. Table 4 presents the items that show a significant bias in the bias/interaction report of the two facets: rater and prompt, arranged in the order of bias size. The second column indicates the specific prompt that the rater in the first column shows bias for. For example, the first row indicates rater 10 gave 56 scores under prompt 1, which means they rated seven examinees considering for each examinee a rater gives eight scores (the eight criteria in the rubric).

The column *Avg. diff* represents the difference calculated by the observed value minus the expected value, so a positive value indicates the rater's observed score is higher than the expected score. In this case, rater 10 rates Prompt 1 significantly more leniently than expected. The other rows can be interpreted using the same formula described for rater 10. Overall, we see that rater 10 tends to rate Prompt 1 more leniently, and Prompt 3 more strictly. Rater 3 rates prompt 4 more strictly, and rater 8 rates Prompt 2 more strictly.

Once the bias is detected, the more important task is to interpret the bias (Xi, 2010). In this section, raters' harshness, inter-rater reliability, and the number of prompts used by each rater are examined. Using *Facets* raters' report, Table 5 presents the logit measurement for rater severity and the infit value for rater consistency for the three raters that showed bias in rating the four prompts (rater 10 appears twice). Raters' total ratings given for each prompt were collected from the original data. The data show that raters did not balance the number of picks of each prompt. Nevertheless, they all exhibit good inter-rater consistency.

As shown in Table 5, all three raters' infit measures fall within the range .7 to 1.3. The total number of ratings they gave is relatively low, and Rater 28 exhibits a noteworthy imbalance in their choice of prompts – 18 uses of Prompt 3 and zero of Prompt 4. Similarly, rater 10 also did not use Prompt 4 at all. In terms of severity, rater 10 is the most severe, at the logic measure of 1.04. Rater 28 is the most lenient, at the logit value of −0.47. But both of them are within a reasonable range. We notice that the most lenient rater rated Prompt 2 significantly more leniently than expected, and this rater also used Prompt 2 much less frequently than Prompt 3.

Table 4 Interaction/Bias report between rater and prompt (significant items)

Rater	Prompt	Count	Avg. diff	*t*-value	prob.
10	1	56	−.13	−2.18	.03
03	4	32	.17	2.41	.02
10	3	24	.20	2.12	<.05
28	2	16	.28	2.14	<.05

Table 5 Rater information

Rater	Measure	Infit	Prompt 1	Prompt 2	Prompt 3	Prompt 4
03	−.07	.96	6	8	4	5
10	1.04	1.06	7	9	3	0
28	−.47	1.25	2	2	18	0

However, rater 28 used Prompt 1 equally scarcely, but the interaction analysis did not detect a bias. This may lead to a tentative conclusion that when a rater is faced with an unfamiliar prompt, their ratings are more likely to be influenced by prompt difficulty and their own harshness compared to the rest of the raters.

6 Insights Gained

In rater-mediated performance tests like the OET, systematic variance caused by rater effects poses a great threat to the validity of scores and score interpretations. This variance, ideally, could be tempered, although not eliminated, by rater training (Eckes, 2019). Previous literature has pointed out that rater training generally helps with enhancing intra-rater reliability while lacking in effects on inter-rater reliability (Lim, 2011; Weigle, 1994, 1998, 1999). As pointed out by McNamara (1996), the best approach to address variability between raters is to train raters to be internally consistent, then use statistical measures to adjust for the differences in rater severity.

6.1 OET Raters

This study confirms that OET raters generally exhibit good intra-rater reliability, which means that they provide internally consistent ratings of test-takers performances. In addition, the MFRM analysis provides a measure for each candidate called the "fair average," which indicates the score the candidate would have received from a rater of average severity on tasks of average difficulty. These scores may be useful for score reporting, as they are adjusted to remove the influence of the severity of the specific raters or difficulty of the specific tasks encountered.

The interaction analysis reveals which specific rater may have differently perceived the quality of a performance for a specific prompt, and therefore giving scores more leniently or more severely compared to the other raters. The MFRM results untangled the complex interaction between raters and prompts, which showed that even though raters' severity may vary, their internal consistency seems to be achieved by sticking to one or two prompts. These results provide valuable insights in future OET rater training in terms of choosing prompts used for the office-hour role-play task.

Although during the rater training sessions, raters were instructed to choose one of the four prompts randomly, some raters tended to stick to one or two prompts that they felt comfortable acting as a student. The behavior of repeatedly using the same prompts may also arise from raters' needs to reduce their cognitive load given that one of the two raters will need to act as a student while also providing ratings of the test-taker. From a rater training perspective, actions will be taken to thoroughly familiarize raters with all prompts during the training session, then reinforce the instruction for raters to randomly choose from the different prompts. In addition, for the current or any future iterations of the OET test, once this issue is detected by a MFRM analysis, it can be addressed by communicating to the specific rater about their rating patterns. Certainly, this solution is based on the presumption that all prompts are comparable in difficulty. Our results show that prompt 2 tends to elicit unexpected ratings, especially when not used often. This aspect will be addressed in the following section on task prompts.

6.2 The Prompts

The results show that the four prompts are not at the same difficulty level, with Prompt 2 (chronic lateness, with a logit measure of −3.04) rated significantly more severe than the others. The rater-prompt interaction analysis also indicates that Prompt 2 elicited the most unpredictable ratings: as the most difficult prompt, a lenient rater rated it significantly more lenient than expected. A closer look at the four prompts reveals that Prompt 2 elicits a different communication pattern than the others. Specifically, the situation described in Prompt 2 specifies that the ITAs should initiate the conversation during the office-hour interaction because they notice an issue about a student's attendance. All other three prompts involve a more passive and reactive role from the ITAs, where the student (played by one of the evaluators) initiates the conversation and leads into the task.

Evidently, test-takers who responded to Prompt 2 experience a different communication pattern, for which they have to initiate the conversation with potentially less time to prepare and less language context to work with. In other words, this prompt potentially introduces construct-irrelevant variance that threatens the validity of the task. As shown in the analysis, there's a higher risk for test-takers using Prompt 2 to receive lower or unpredictable scores. To address this issue, the OET test administration could take actions to revise or remove Prompt 2 from the currently circulating prompts in the OET test. If replacing or revising, test designers should take into consideration the types of interactions elicited by the prompts.

6.3 The Rubric

The results from the MFRM analysis show that the rubrics are used fairly and consistently by the OET raters, and the raters generally distinguish the different constructs in the rubrics. It is worth noting, however, that some constructs in the rubrics display slight misfit or overfit for the infit statistics, which shed light on the relationship between the nature of the office-hour role-play task and the usefulness of each criterion in the rubric.

As shown in Table 1, the criterion Transition shows a slight overfit (0.69), which means that raters tend to give invariant ratings on this criterion. This rating behavior can be explained by the nature of office-hour interactions: in a realistic problem-solving-oriented office-hour situation, the student-ITA communication is typically filled with short turns. There is inherently less chance for using explicit transitional language as one would use in registers that allow a longer stretch of speech. In contrast, scores for Transition may show more variance in tasks that require the ITAs to structure a larger amount of information in a monologue, such as in a lecture task. The invariant ratings on this task confirm that raters did not interpret this construct to be one that distinguishes different levels of test-takers' performance in the office-hour role-play task; in addition, it provides information for test designers regarding the usefulness of this construct to be included in the rubric.

Meanwhile, the criterion Task Response has a slight overfit (1.31), indicating that raters tend to give invariant ratings on this criterion. This result has important implications for rater training. In the rubric, a score 4 in this criterion indicates that the test-taker "completes the task in a pragmatically appropriate manner" (see Appendix B). Compared to the other categories, this criterion places more focus on pragmatics and completion, which involves more subjective judgment than those focusing more explicitly on linguistic features, such as grammar constructions or pronunciation. Thus, during the rater training sessions, more attention will be put into calibrating and negotiating raters' understanding of this specific construct when it comes to the office-hour role-play task.

7 Conclusion

Motivated by practical needs to improve rater consistency and rubric utility, our study provides insights into practices that connect test analysis, test quality, and test administration in a local context. Specifically, this study focuses on a standardized, locally developed placement test for ITAs at a large public university – OET. We investigated the eight criteria in the OET rubric and the difficulty levels of the four prompts used in the office-hour role-play task.

Results from the Facets analysis show that the eight-aspect rubric was used appropriately for rating the office-hour role-play task, although certain aspects display a less than expected variance (i.e., transition). Our results also show that the four prompts vary in their difficulty levels, and each of them interacts with certain raters. Further examination reveals that some raters selected certain prompts to the exclusion of others. The rater-prompt interaction shows that raters' internal consistency seems to be achieved by sticking to one or two prompts, but the variation in rater severity is not directly related to their choices of prompts. These results provide useful insight for rater training and rubric development for the OET test. They inform the OET test designers of a potential prompt effect in the office-hour role-play task, thus possibly threatening the accuracy of decision making regarding ITA test-takers' ability and placement.

In terms of raters, generally, all raters were self-consistent in rating the office-hour role-play task, but we found significant differences in raters' relative severities. These differences can be addressed by focusing on aforementioned aspects of the rubrics or the prompts during the rater training session, or using the adjusted scores provided by the MFRM analysis.

Appendices

Appendix A: Role-Play Prompts

Task Four: Office Hours Role-Play

Role-Play Scenarios

 # 1: (Attendance Policy)

At the end of class, an undergraduate student in your class comes to you and tells you that he/she will be absent from your class at least 10 times because of a serious illness in the family. Imagine that [Name of Evaluator] is your "student," and explain to him/her why you will, or will not, allow the student to be absent.

 # 2: (Chronic Lateness)

One of the undergraduate students in your class comes to your office because you have asked him/her to come to see you. He/she is always 20 min late to your class and is in danger of failing the course. Imagine that [Name of Evaluator] is your student and discuss the situation with him/her.

 # 3: (Extension Request)

An undergraduate student in your class comes to your office and asks you for more time to complete an assignment. He/she usually does a good job on assignments, but has missed several due dates during the semester. Imagine that [Name of Evaluator] is your student, and explain to him/her why you will, or will not, accept the late assignment.

 # 4: (Makeup Test Request)

An undergraduate student in your class comes to your office and asks you to let him/her take an exam that he/she missed last week without notice. The student has done well on his first two exams but has not attended class for 2 weeks. Your department has a policy that no late or makeup work is accepted. There is one exam remaining and it is 50% of the final grade. Imagine that [Name of Evaluator] is your student, and discuss the situation with him/her.

Appendix B: OET Rubric

	1 (3 semesters ESL)	2 (2 semesters ESL)	3 (1 semester ESL)	4 (Exempt)
Grammatical constructions	Generally incorrect and labored constructions	Consistently incorrect and labored constructions	Noticeable incorrect constructions	Infrequent incorrect constructions
Thought groups	Unintelligible thought groups and volume	Limited effectiveness of thought groups and inappropriate volume	Difficulties with thought groups, or inappropriate volume	Thought groups adequate with appropriate volume
Tone choice	Monotone	Mostly irregular falling, rising, and/or level tones	Uses rising, falling, and level tones but occasionally misleading	Uses rising, falling, and level tones in appropriate manner
Multimodality	Minimal eye contact, consistent body orientation away from audience; poor use of non-verbal resources	Frequent retreat to body position away from audience and insufficient use of non-verbal resources	Infrequent periods of orientation away from audience supplemented by effective use of non-verbal resources	Negligible orientation away from audience and effective use of nonverbal resources

(continued)

	1 (3 semesters ESL)	2 (2 semesters ESL)	3 (1 semester ESL)	4 (Exempt)
Transition	Minimal transitions used	Task answer organized with restricted use of transitions	Task answer organized but varied transitions are lacking	Task answer organized with appropriate and varied transitions
Prominence	Key words not identified through stress	Minimally effective identification of key words through stress	Consistent identification of key words through stress with noticeable omission	Effective identification of key words via stress throughout task
Task response	Initial task response does not address the task	Task response pragmatically inappropriate and/or several gaps of information	Completes task with pragmatic appropriateness; any gap of information requires minimal repair	Completes task in pragmatically appropriate manner
Question response	Answer does not address raters' concerns; other-initiated repair unsuccessful	Partial answer even after multiple repair sequences	Answer negotiation after other-initiated repair; successful completion	Resolved with minimal repair

References

Attali, Y. (2016). A comparison of newly-trained and experienced raters on a standardized writing assessment. *Language Testing, 33*(1), 99–115. https://doi.org/10.1177/0265532215582283

Bachmann, L., & Palmer, A. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford University Press.

Bailey, K. M. (1983). Foreign teaching assistants at U.S. universities: Problems in interaction and communication. *TESOL Quarterly, 17*(2), 308–310. https://doi.org/10.2307/3586658

Bailey, K. M. (1984). The "foreign T.A. problem". In K. M. Bailey, F. Pialorsi, & J. Zukowski-Faust (Eds.), *Foreign teaching assistants in U.S. universities* (pp. 3–15). National Association for Foreign Student Affairs.

Barkaoui, K. (2010). Variability in ESL essay rating processes: The role of the rating scale and rater experience. *Language Assessment Quarterly, 7*(1), 54–74. https://doi.org/10.1080/15434300903464418

Bond, T. G., & Fox, C. M. (2013). *Applying the Rasch model: Fundamental measurement in the human sciences*. Psychology Press.

Brown, G., & Yule, G. (1989). *Discourse analysis*. Cambridge University Press.

Cotos, E., & Chung, Y. R. (2019). Functional language in curriculum genres: Implications for testing international teaching assistants. *Journal of English for Academic Purposes, 41*, 100766. https://doi.org/10.1016/j.jeap.2019.06.009

Dimova, S., Yan, X., & Ginther, A. (2020). *Local language testing: Design, implementation, and development*. Routledge. https://doi.org/10.4324/9780429492242

Eckes, T. (2019). Many-facet Rasch measurement: Implications for rater-mediated language assessment. In *Quantitative data analysis for language assessment volume I* (pp. 153–175). Routledge.

Elder, C. (2017). Language assessment in higher education. *Language Testing and Assessment, 271–286*. https://doi.org/10.1007/978-3-319-02261-1_35

Fulcher, G., & Davidson, F. (2007). *Language testing and assessment*. Routledge.

Gan, Z. (2010). Interaction in group oral assessment: A case study of higher-and lower-scoring students. *Language Testing, 27*(4), 585–602. https://doi.org/10.1177/0265532210364049

Gevara, J. (2016). *Confirming the impact of performance tasks on latent class membership and placement decisions*. Unpublished doctoral dissertation. The Pennsylvania State University.

Hamp-Lyons, L. (1990). Second language writing: Assessment issues. In B. Kroll (Ed.), *Second language writing: Research insights for the classroom*. Cambridge University Press.

Isaacs, T., & Thomson, R. I. (2013). Rater experience, rating scale length, and judgments of L2 pronunciation: Revisiting research conventions. *Language Assessment Quarterly, 10*(2), 135–159. https://doi.org/10.1080/15434303.2013.769545

Jameel, U., & Porter-Szucs, I. (2014). Nativelike formulaic sequences in office hours: Validating a speaking test for international teaching assistants. *Research Notes, 100*(55), 28–34.

Knoch, U., Fairbairn, J., & Jin, Y. (2021). *Scoring second language spoken and written performance: Issues, options and directions*. Equinox Publishing.

Leaper, D. A., & Riazi, M. (2014). The influence of prompt on group oral tests. *Language Testing, 31*(2), 177–204. https://doi.org/10.1177/0265532213498237

Lim, G. S. (2009). *Prompt and rater effects in second language writing performance assessment*. University of Michigan.

Lim, G. S. (2011). The development and maintenance of rating quality in performance writing assessment: A longitudinal study of new and experienced raters. *Language Testing, 28*(4), 543–560.

Linacre, J. M. (1989). *Many-facet Rasch measurement*. MESA Press.

Linacre, J. M. (2002). Optimizing rating scale category effectiveness. *Journal of Applied Measurement, 3*(1), 85–106.

Linacre, J. M. (2019). *A user's guide to Facets Rasch-model computer programs*. Retrieved March, 09, 2020 from www.winsteps.com/facetman/webpage.htm

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*(2), 149–174. https://doi.org/10.1007/BF02296272

McNamara, T. F. (1996). *Measuring second language performance*. Addison Wesley Longman.

McNamara, T., Knoch, U., & Fan, J. (2019). *Fairness, justice & language assessment*. Oxford University Press.

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement, 4*(4), 386–422.

Plough, I. C., Briggs, S. L., & Van Bonn, S. (2010). A multi-method analysis of evaluation criteria used to assess the speaking proficiency of graduate student instructors. *Language Testing, 27*(2), 235–260. https://doi.org/10.1177/0265532209349469

Reid, J. (1990). Responding to different topic types: A quantitative analysis from a contrastive rhetoric perspective. In B. Kroll (Ed.), *Second language writing: Research insights for the classroom* (pp. 191–209). Cambridge University Press. https://doi.org/10.1017/CBO9781139524551.017

Sandlund, E., Sundqvist, P., & Nyroos, L. (2016). Testing L2 talk: A review of empirical studies on second-language oral proficiency testing. *Language and Linguistics Compass, 10*(1), 14–29. https://doi.org/10.1111/lnc3.12174

Schumacker, R. E. (1999). Many-facet Rasch analysis with crossed, nested, and mixed designs. *Journal of Outcome Measurement, 3*, 323–338.

Seedhouse, P., Harris, A., Naeb, R., & Üstünel, E. (2014). *The relationship between speaking features and band descriptors: A mixed methods study* (IELTS research reports online series) (p. 30). IELTS Partners.

Spaan, M. (1993). The effect of prompt in essay examinations. In D. Douglas & C. Chapelle (Eds.), *A new decade of language testing research* (pp. 98–122). TESOL.

Thirakunkovit, S., Rodríguez-Fuentes, R. A., Park, K., & Staples, S. (2019). A corpus-based analysis of grammatical complexity as a measure of international teaching assistants' oral English proficiency. *English for Specific Purposes, 53*, 74–89. https://doi.org/10.1016/j.esp.2018.09.002

Van Moere, A. (2006). Validity evidence in a university group oral test. *Language Testing, 23*(4), 411–440. https://doi.org/10.1191/0265532206lt336oa

Weigle, S. C. (1994). Effects of training on raters of ESL compositions. *Language Testing, 11*(2), 197–223.

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language testing, 15*(2), 263–287.

Weigle, S. C. (1999). Investigating rater/prompt interactions in writing assessment: Quantitative and qualitative approaches. *Assessing Writing, 6*(2), 145–178. https://doi.org/10.1016/S1075-2935(00)00010-6

Weigle, S. C. (2002). *Assessing writing*. Cambridge University Press. https://doi.org/10.1017/CBO9780511732997

Xi, X. (2010). How do we go about investigating test fairness? *Language Testing, 27*(2), 147–170. https://doi.org/10.1177/0265532209349465

Zhang, Y., & Elder, C. (2011). Judgments of oral proficiency by non-native and native English speaking teacher raters: Competing or complementary constructs? *Language Testing, 28*(1), 31–50. https://doi.org/10.1177/0265532209360671

Validation of the Czech Language Certificate Exam with Respect to the Local Context

Martina Hulešová and Kateřina Vodičková

Abstract In 2015 a quality management system was introduced into the Czech Language Certificate Exam (CCE), a high-stakes examination ranging from A1 to C1 level. The research goal focused on addressing the following broad questions: (1) how to meet standards for quality and fair assessment in a specific situation when the cut-off score is set in advance and (2) when local resources are limited. The specific research questions were: (1) Are test forms comparable in terms of difficulty and (2) how should the 60% cut-off score across the suite of exams be interpreted? A mixed-methods approach was applied. In the first phase, the constructs, specifications and content were revised. In the second phase, two standard setting methods to address the difficulty were used: The Direct Consensus method for the Listening, Reading and Grammar-lexical subtests, and a locally modified Body of Work method for the Writing and Speaking subtests. To achieve score comparability of different test forms, linear equating was applied, and subscale scores were reported on a single, shared reporting scale. Both methods solved the problem with a pre-set cut-off score, helped to verify comparability of the assembled test forms, and prepared the ground for further research.

Keywords Local language tests · High-stakes exams · Standard setting · Comparable test forms · Difficulty · Validation · CEFR alignment

1 Introduction

Test results may have serious consequences for the test takers, such as denial of permanent residency, citizenship, or admission to a university. Exam providers are accountable to their stakeholders and should adhere to the principles of good practice and professional standards at all stages of the test development process

M. Hulešová (✉) · K. Vodičková
Institute for Language and Preparatory Studies, Charles University, Prague, Czechia
e-mail: martina.hulesova@ujop.cuni.cz; katerina.vodickova@ujop.cuni.cz

X. Yan et al. (eds.), *Local Language Testing*, Educational Linguistics 61,
https://doi.org/10.1007/978-3-031-33541-9_3

(item writer training and item writing, test administration, statistical analyses and the interpretation of the results, among others) (AERA et al., 1999/2014) and demonstrate the validity and appropriateness of test results interpretation. Validity as we understand it refers to the degree to which evidence and theory support the interpretations of test scores for the proposed uses of tests (AERA et al., 1999/2014, p. 11). Validity evidence is derived from various sources (test construct, test content, response processes, marking processes, training, administration processes, security issues, use of the results, etc.) that need to be evaluated together as a whole.

The Czech Language Certificate Exam (CCE) is a suite of high-stakes exams. The consequences for test takers who do not pass the exams might be the denial of admission to Czech Universities, where students can study for free in Czech-medium instruction programs, or denial of access to certain job opportunities. In 2015, a quality management system was implemented into the CCE development cycle. Consequently, revision began with the aim to revalidate the processes related to the exam development and use. Revision and revalidation are intertwined by nature, as the decisions taken during the revision process influence the validity evidence interpretation. In addition, we wanted to overcome some of the issues which were considered weak points of the exam and less well researched areas from the validity point of view. The long-term research goal focused on addressing the question of how to meet standards for quality and fair assessment for Czech (1) as a less widely spoken and taught language, (2) in a context where the cut-off score is set in advance and cannot be changed, and (3) where local resources are limited.

Both qualitative and quantitative methods were used. In the first phase, the construct, specifications and test content of each level were revised by the internal team (individually and through team discussion) by comparing and contrasting the original test specifications and results with the CEFR Companion Volume (CEFR CV) using the intuitive approach and the ALTE framework of minimum standards; the construct definitions for all five CCE levels were defined, a new set of descriptors were developed for the CCE test specifications, a new rating approach and rating criteria were derived from the CEFR CV. In addition, new sets of tasks for live exams were developed by internal and external teams, and the products went through pilot testing and an external review procedure.

In the second phase, before the exams went live, the research focused on providing an interpretation for the pre-set cut-off score, which had had no clear interpretation up to that point, it also focused on the alignment of the exams to the CEFR, and on the issue of test form comparability. As the cut-off score was already set by the university's assessment policy at 60%, the test provider team knew that a) it could not be changed easily, b) we were not able to create test forms of the same difficulty. Amongst the reasons for this was the fact that the numbers of candidates do not allow for the use of IRT-based task banking system, c) we cannot set and use different cut-off scores for every new test form as this would have been very difficult to explain to the stakeholders.

As a result, two standard setting methods were used during the second phase. For receptive tasks (listening, reading) and linguistic competences (grammar/lexical tasks), the Direct Consensus (DC) method (Cizek & Bunch, 2007, pp. 97–102) was

used as it was found suitable for the local context and it also circumvents certain disadvantages of other standard setting methods (i.e. Angoff or Basket method). For writing and speaking, the Body of Work (BoW) method (Kingston et al., 2001) was used, with modifications described in the article by Verhelst et al. (2019). In the standard setting events, external and internal judges took part and an EALTA Expert member provided consultation and initial supervision, especially for the statistical analyses and the interpretation of the results.

Both approaches dealt successfully with the issue of a pre-set cut-off score, and provided test developers with useful information for assembling comparable test forms. In the local context, one of the most important outcomes, although unplanned, was the deeper involvement of the internal and external team which led to the increase in their knowledge. Stakeholders within the broader local context (e.g., test takers themselves, Czech universities, employers, the Ministry of Education, youth and sports organizations, etc.), were also positively impacted thanks to face-to-face meetings, presentations, and the dissemination of information online.

2 Test Purpose and Testing Context

The CCE is provided by the Research and Test Centre (RTC) of the Institute for Language and Preparatory Studies (ILPS) of Charles University.[1] The CCE is a suite of general language proficiency exams ranging from A1 to C1, linked to the Common European Framework of Reference (CEFR). Any of the exams may be taken by any L2 speakers of Czech aged 16 and over who want to or who need to certify their level of language proficiency in Czech regardless of their L1, country of origin, or level of education, etc.

Typical CCE test takers[2] are students interested in studying in Czech university programs or people interested in working in the Czech Republic. This is the reason why the B2 and B1 exams are the most popular whereas the A1, A2 and C1 levels have fewer candidates, and the needs of the candidates are different (see Tables 1 and 2). Most of the candidates indicate they want to take the exam for study purposes (usually B2 level) whilst more than 25% state that it is for employment

[1] Founded in 1348, Charles University (https://cuni.cz/UKEN-1.html) is the one of the oldest universities in the world and the largest as well as the best-rated university according to international rankings in the Czech Republic. There are currently 17 faculties at the University plus 3 institutes, one of which is the scientific-educational ILPS. The ILPS offers a large selection of Czech as a FL/L2 courses at varying levels of difficulty and intensity. See https://ujop.cuni.cz/en/ The intensive year-long preparatory courses are offered both in Czech and English and specialist subjects (specialisations: Economics, Engineering, Medicine and Humanities). The ILPS provides support to Czech language teachers via a number of methodological courses. It is also a highly respected test provider, for example, for its internationally recognised CCE exam and its Czech Language for Citizenship exam.

[2] The information is provided by test takers in the pre-test questionnaire, which is part of the registration form.

Table 1 Predominant reasons for taking the CCE exam according to level

Level	Declared test results use
A1	Progress test; Permanent residency
A2	Permanent residency; Study purposes
B1	Study purposes; Citizenship; Job purposes (educators)
B2	Job purposes (state employees, educators); Study purposes
C1	Job purposes (mainly state employees); Study purposes (faculties of medicine)

Table 2 Number of candidates per year and level 2015–2021

Level	2015	2016	2017	2018	2019	2020	2021	Total 2015–2021
A1	28	38	35	38	22	3	27	191
A2	66	70	77	70	95	28	98	504
B1	112	154	274	321	370	171	283	1685
B2	319	328	465	434	458	336	326	2666
C1	49	94	75	75	85	54	82	514
Total	574	684	926	938	1030	592	816	5560

purposes (B2, B1, C1 levels). As a result, the CCE is a high-stakes exam for most candidates.

The CCE exams consist of four subtests (Reading, Listening, Writing, Speaking) at each level. For B2 and C1 level exams, a fifth subtest, a Grammar-lexical subtest, is included. Candidates have to reach at least 60% in each subtest to pass the whole exam and obtain the certificate required by their prospective university or employer. The certificate for those who pass includes the total percentage score, a *Pass* statement, additional information about the percentage scores in all the subtests, and a description of the performance standards.[3]

The CCE, first introduced in 2007, has developed into an international proficiency exam of Czech as a foreign language. The test frameworks, the theoretical construct, test specifications, rating criteria and training materials were developed in a very short time after the CEFR appeared in 2001, and no significant content or procedural changes were made until 2014. Shortly after the publication of the CEFR, expertise and its use as a descriptive tool were still at a very early stage in the Czech Republic as well as the rest of Europe.

Until 2020, the CCE was paper-based and centralized in many aspects i.e. printing, completing and checking the test materials in the central office, delivering them to the examination center and back to the central office by the exam supervisors. The pandemic in 2020 accelerated planned changes and some steps towards decentralization and computer-assisted administration have been taken. As a result, the exam centers have been given more responsibility for the exam administration and organization. In addition, exam sessions abroad have been organized in a hybrid

[3] The additional information is not required by the stakeholders.

form combining administration of the written part on site with online examining of speaking and asynchronous marking after the session is closed.

3 Testing Problem Encountered

Charles University, specifically the ILPS, has been a full member of the Association of Language Testers in Europe (ALTE) since 2009, which audits the CCE at all levels for quality every 5 years. In 2014, the RTC compiled a validity argument for the ALTE re-audit.[4] During the first audit in 2009, it was recommended to revise the way the standards (passing boundaries) had been set.

The passing score for the CCE had been pre-set by the internal examination regulations, which are applied across Charles University. However, after the internal and external reviews and first standard setting at ILPS using the Direct Consensus method (described in detail in Sect. 4), conducted in 2014 and discussed in depth, identified several problematic areas that needed to be addressed. They were concerned the missing validity-related documentation, the reliability and interpretability of the pre-set standard, the weak relation between the standard and the CEFR, and the validity of score interpretations. This was due not only to the construct coverage and the weak alignment to the CEFR, but also due to the lack of evidence regarding test form comparability. The nature of the findings clearly showed that both the way of setting the standard as well as other aspects of the exams had to be revised. The main areas of concern are summarized as follows:

- The CCE claim that the exams are targeting a particular CEFR level was not perceived as such by the group of experts involved in the standard setting procedure in 2014.
- The original test specifications did not seem to be derived clearly from the CEFR, they seemed to be over-localised and lacking connection with the CEFR; the increase of difficulty across levels was not perceived as clear.
- The test construct was not clearly reflected in the exams or tasks and some of the testing techniques used in the exams did not measure the intended subconstructs well.
- The cut-off score for the CCE was pre-set according to the Charles University regulations and its alignment to the CEFR was not established by any procedure.
- There were no mechanisms in place that allowed for the assembling of comparable test forms for different exam sessions. This means there was no system within the test development processes or pilot testing that would allow for comparisons of results to be made (i.e. a reporting scale, equating procedure, linked design in pretesting) and for the interpretation of scores from different test forms in the same way.

[4] The exams must be re-audited every 5 years to maintain the quality Q mark – https://www.alte.org/Setting-Standards

The following research areas were identified:

– Re-defining the construct and the alignment of the CCE (test specifications, rating schemes, cut-off scores) to the CEFR;
– Finding a way to deal with the pre-set cut-off scores that cannot be changed and to set a reliable and defendable passing standard;
– Establishing a procedure to ensure comparability of alternative test forms for different sessions.[5]

4 Literature Review

Validation can be understood as a complex process through which validity is described, evaluated, interpreted, and validity evidence gathered, justified, and documented. In this sense, validation does not concern the test or assessment tool only, but also the processes related to its development, use, interpretation, and consequences. Most of the currently used validation models such as Kane's argument-based approach (1990, 2011), Bachman and Palmer's Assessment Use Argument Model (2010), and Weir's Socio-cognitive Model (2005) emphasize the usefulness of the testing tool, the meaningfulness of test score interpretation and the use of the results, and the justifiability of the decisions taken on their basis. Most of the validation frameworks build directly on Samuel Messick's view of validity (1987, 1989), which called into question validity as an exclusive characteristic of a test. Messick (1989) characterizes validity as a level of agreement between the theoretical rationale and its operationalisation, score interpretation and use, and the consequences of the score use.

The Standards for Educational and Psychological Testing (AERA et al., 1999/2014) incorporate the different views of validity and define validation as a long-term, evidence-based building argumentation. However, as Bachman (2012, p. 1) states, this validity argument needs to be localized, since it is closely connected to the local conditions. This was also the approach adopted by the RTC team: to scrutinize all the processes, to discuss their appropriateness, to carry out necessary changes, justify them, and gather evidence about the revision processes and outcomes.

The question as to whether test forms are comparable and whether scores from these alternative test forms can be interpreted as being equivalent in terms of their interpretation and use is one of the topical issues of language testing (von Davier, 2011; Holland, 2007). Despite its importance, test form comparability seems to be rather a marginal, under-researched topic in the Central European context of

[5] We use the term *comparable test forms* when referring to the state where the construct is operationalised and measured in the same way across the test forms based on the same test specifications, which is a necessary precondition for the comparison of results (van der Vijver & Poortinga, 2005) and also for the application of equating procedures (Livingston, 2004).

standardized testing. There are virtually no studies or documents describing how test providers in the Czech Republic have been dealing with this area of research (Anýžová, 2013). It might be deduced that the comparability of test forms is not mentioned at all, that it is taken for granted or, at the very least, not questioned. Users may assume that test forms are comparable because forms are based on test specifications associated with proficiency levels, and the assumption is generally not investigated. This has been the case with many existing exams in the Czech Republic, and it was also the case with the CCE at the beginning of its existence.

Given the current state of the art within the language testing field, we claim that test forms comparability is one of the key prerequisites for a meaningful interpretation of the scores and for fair and justifiable use of the scores. As test providers and test developers, we are also convinced of our responsibility for test qualities, especially for those related to fairness, validity and justice in the score use.

Although different test forms of the same exam, but containing different tasks, cannot be considered to be of the same difficulty even if they are based on the same test specifications, they might be considered comparable test forms if several conditions are fulfilled. By test forms comparability, we understand the notion whereby the scores or the results of different test takers who took different test forms of the same exam reflect the same construct and measure it in the same way, and thus, the scores may be meaningfully compared.

Test forms comparability might be usually achieved by equating procedures. There are different methods of equating (Livingston, 2004) but most of them usually require the equated test forms to have something in common: common-persons taking different test forms where the test-takers ability is considered as being stable; or common-item design where the item difficulty is stable across the two forms. Equating procedures allow for scores from the two forms to be put on the same scale and compare them or transform them into the equated score that is then reported.

The limitations in live testing or pretesting, especially due to the small sample size and not being able to control the sample characteristics on the available samples during pretesting, do not allow for the use of the above-mentioned methods. However, the method of standard setting, namely the DC method, shares some aspects of common person design: every task is judged by a panel of judges who are selected from a large pool. This is analogous to the situation where different test forms are taken by the same group of test takers.

5 Methods, Approaches, and Solutions

After the first standard setting study in 2014, which pointed out the areas in the test development process that were not covered well (listed in Sect. 2), it was decided to start re-validation in three main stages: first, to revise the theoretical framework, the construct and test specifications, which included, among other changes, those concerning the item writing process, item writer and rater training); second, to align the revised exams to the CEFR and document the local adaptations; third, to introduce

the so called *continuous standard setting* for reading, listening and grammar/lexical tasks. Continuous standard setting means that every new task has to undergo the process of estimating its cut-off score i.e. the score that a minimally competent candidate at one of the CEFR levels would obtain in this task. Only then is the task stored in a task bank. In this task bank, all tasks are stored with their respective characteristics (topic, length, testing technique, subskills measured, etc.) and psychometric characteristics from pretesting, standard setting, and live sessions. This banking system allows for the assembly of comparable test forms.

For productive skills, the standard setting approach differed not only in the method, but also in terms of the timeframe. It was divided into two phases although only the first phase has been implemented so far. In this first phase, due to the small number of samples of written and spoken performances for each task, we had to assume the tasks were equally challenging for the test-takers.[6] In the second phase, standard setting will be repeated with more samples of performances from live testing and task forms will be treated as being different tasks, following the procedure described in Verhelst et al. (2019). The standard will be set separately for each task form. Samples from the first stage will be included in the second stage as anchor tasks.

5.1 Analysis of the CEFR and Construct Revision

The initial step entailed the re-definition of the theoretical framework, construct, test specifications, and rating schemes. First, detailed analysis of both the existing CCE test materials and the CEFR CV was carried out. Relevant scales and descriptors were identified, their relevance for the CCE exams and their match with the existing test specifications were discussed with the internal team. The final set of agreed CEFR scales and descriptors was compiled, local modifications were documented and the outcomes were translated into Czech. Language testing literature and a set of new descriptors aligned to the CEFR were used to define the constructs for each level of the exam. Models of language ability by Bachman (1990), Bachman and Palmer (1996), Weir's socio-cognitive framework (2005) were considered when defining the constructs. For the subtests, we worked with models for each skill as described for example by Luoma (2004) for speaking; Fulcher (2010), in Purpura (2004), for grammar; Buck (2001) for listening; Weigle (2002) for writing and Alderson (2004) for reading. The theoretical construct definitions were elaborated in more detail and operationalised in the first drafts of test specifications and in the rating schemes for productive skills. All the descriptors from test specifications and from the rating schemes were derived from the CEFR, and as it was described above, the local modifications or additions were documented.

[6] In fact, the task forms differ very little, mainly in terms of the topic, but still, we cannot be sure whether there might be a difference in their difficulty.

After the cycle of internal discussions and redrafts that took into consideration the local context and needs, new test specifications and rating schemes were finalised, and internal trial task writing rounds and discussions about the test specifications were carried out to ensure the test specification enabled item writers to replicate tasks according to test specifications. Rating criteria for the productive skills, derived from the revised and localised CEFR and CV descriptors, also underwent several internal try-outs to see whether they reflected the construct and the intended CEFR level well, and whether they were understood well by the raters. All the newly produced materials (theoretical framework, construct definition, test specifications, rating criteria and mock test forms for all five levels) were scrutinised by external reviewers. The new test specifications and rating schemes and also the local descriptors describing typical features of the tasks, which are aligned to the CEFR levels.

When the new test specifications were finalised for each level, item writers were trained and wrote new tasks. Every new task was coded i.e. item writers indicated the specific goal (from test specifications) targeted by every item. In this way, items and tasks were aligned with the test specifications as well as with the CEFR. Item writers also revised the tasks used before the revision process started. Those tasks that corresponded to the new, revised test specifications, were also coded and described by the specific goals, and then included in the task bank. All the new tasks and the rating schemes were pretested. Post-test analyses and interviews with the pretestees were carried out. Only pretested and well-functioning tasks were stored in the task bank and used later in the standard setting.

5.2 Standard Setting

The next stage was intended to solve the problem of the lack of mechanisms for setting the passing standards. After discussing and analyzing the local resources and the local context, the following logistical challenges were recognised:

- CCE has relatively low numbers of candidates. Therefore, the standard setting method could not rely on large data sets,
- Pilot testing and pretesting should be carried out in the intensive year-long courses[7] at the ILPS; that means it depends on students' and teachers´ willingness to participate and on their time schedules, so there is no space for a specific pre-testing design,

[7] Preparatory courses are organized by the ILPS. These courses are usually one-year in length and are for students who plan to take an entrance exam to Czech universities and so need to be prepared in the Czech language and in the content subjects. During the academic year, students proceed from A1 to B2 or C1 levels. Therefore, it is practical and logistically feasible to organize pretesting for these courses.

– The exam development team is a relatively small one, and without the regular support of a psychometrician, so the standard setting methods need to be relatively simple and easy to apply.

What emerged from the internal review was that the CCE does not follow the principles for the use of sophisticated, mainly IRT-based approaches to test equating or task banking and, therefore, another solution was needed to assemble comparable test forms. A decision based on thorough literature research was taken to implement two standard setting procedures that do not require students' answers i.e. data from live testing: the DC method was chosen for receptive skills and a modified BoW approach was implemented for subtests in productive skills.

Carrying out the standard setting followed a scheme described in the literature (Coe, 2009; Cizek and Bunch, 2007; Cizek, 2012;): the sessions began with familiarization of the CEFR levels and, in the case of the DC method, also with the concept of the minimally competent candidate (MCC). The standardization phase followed whereby model tasks were discussed, and the selected method was applied. Once a technical check had been completed, there followed one (BoW method) or two rounds of individual judgment with a discussion in between (DC method). The sessions ended with a presentation of the results and a closing discussion.

The definition of the MCC for each level and skill was established between 2016 and 2018 in a small-scale project in which internal and external teams collaborated. It consisted of a detailed analysis of the CEFR descriptors and the local test specifications with the minimally acceptable expected performance in each subskill as described by the descriptors identified. Several drafts of the MCC definitions were prepared for each level which included illustrations of the MCC performance in particular tasks. The tasks and performances were analyzed and commented on and the decision as to whether a task was difficult, easy or adequate for an MCC were justified and described. The outcomes were five sets of descriptors describing a hypothetical MCC at each of the five levels contrasted with the original descriptors in the CEFR.

5.3 Familiarization

During the familiarization phase, judges undergo thorough familiarization with the CEFR reference levels, the descriptors and task specifications as well as the description of the MCC. Familiarization involves a wide range of activities, varying from sorting descriptors into levels and analyzing key words, through to answering multiple choice questions, to discussing and summarizing salient features of the respective levels. An important aspect is that all familiarization activities and materials are in Czech. This was an important decision, as we came to realize during the CEFR analysis, because translating had led to a more precise understanding and explanation of the meaning of the original descriptors, and it shed light on some

inconsistencies within and across the CEFR. The continuous and systematic work on creating a bank of localized descriptors[8] with a clearly defined relationship to the original CEFR and CV descriptors serves as the evidence of the alignment of the CCE to the CEFR.

5.4 Standard Setting Method for Receptive Skills: Direct Consensus Method

For receptive skills, a performance-centered method in the form of the DC method was chosen. The key feature of the DC method is that judges estimate the score that an MCC would receive at a certain level in a particular task. The definition of a minimally competent candidate for each level had to be developed prior to its implementation. This DC method was chosen in preference to other methods as it is focused on the perceived relative difficulty of the tasks or items in terms of the expected test performance of candidates who are minimally competent. It is also considered to be relatively simple; it does not require large amounts of data (unlike, for example, the Bookmark method) and it does not work with the complex concept of probability of the correct answer (as the Angoff method, for example). It focuses on the task characteristics and the concept of the MCC. The key question for the judges in the standard setting was: What *score would a student who is exactly on the borderline (a minimally competent student) get in this test?* Since its successful application on CCE in 2014–2015, DC method has since been used twice or three times a year as a regular part of the test development process. All new pretested tasks go through the standard setting process and only on completion can tasks be stored in the task bank and used for assembling test forms. In the local context, it means that standard setting is carried out for every new task at the end of the task development process.

For each session, 10–14 judges are usually invited. Most of them are L1 speakers of Czech, teachers, language testers, item writers, but we also include several L2 participants, proficient speakers of Czech and experienced teachers of other languages. All judges are given booklets with tasks in the same form as they are to be presented to the candidates as well as printed answer sheets. There is also a list of links to a spreadsheet where judges record their estimated score. This is done after each round.

For each task, the judges, who are aware of the maximum score for the task, have to estimate the score that a minimally competent candidate would obtain. They cannot give fewer points than the score likely achieved by random guessing. They can also use decimal numbers in a limited way. For example, for a task with 5 MCQ

[8] such as the ones dealing with the written literary Czech vs. the interdialect referred to as "common Czech" spreading even into mass-media

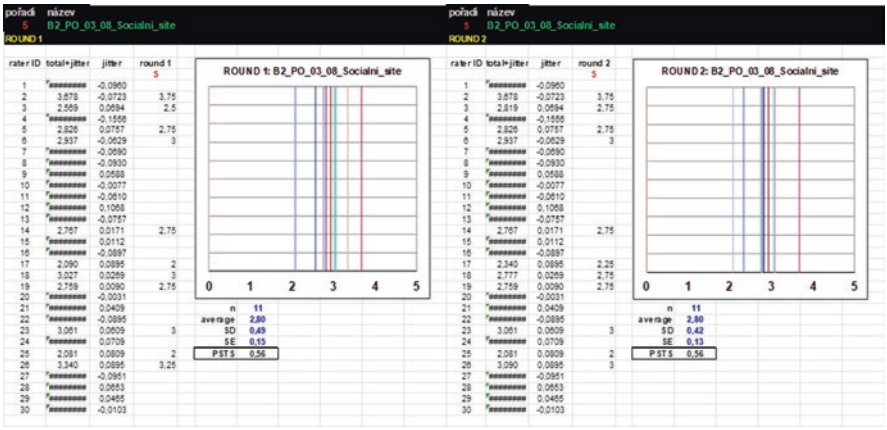


Fig. 1 Outcomes for Round 1 and Round 2: Direct consensus method

items with four alternatives, where each item is scored 1 point, the maximum is 5 points so the minimum that can be given by a judge is 1.25. Judges can work with quarter points too (0.25, 0.5, 0.75 points, etc.). After reading and judging all the tasks in the booklet, the judges send their estimation for Round 1, and they are shown the results (Fig. 1) – the partial cut-off scores for each task and other information. This is the group mean of their judgment. They also see the standard error of the cut-off score and how much they may differ within the group (Fig. 1). In the moderated discussion after Round 1, judges are asked to explain and justify their judgment for each task and use the CEFR descriptors, the MCC descriptors, and task features to support their argumentation. Round 2 is the final round of individual judgements. Judges can consider the arguments used in the discussion and change their initial estimation. The value given to each task in Round 2 by the group of judges is taken as the cut-off score for the task.

Validity Checks

The results of the standard setting for the receptive skills are checked immediately at the end of each session by discussing the values agreed by the panel (the standard and its standard error) of the judges, as well as retrospectively[9]: judges received cumulative information on their consistency and their tendency towards leniency or harshness across the skills and levels for all standard setting events they took part in. Judges are also sent an evaluation questionnaire where they can comment on different aspects of the standard setting event and express their opinions about the results, procedure, familiarization and other aspects.

[9] This has been done regularly for the DC method and it is planned for the BoW method.

5.5 Standard Setting Method for Productive Skills: Body of Work Method

For the subtests of productive skills,[10] the BoW method was chosen (Kingston et al., 2001; Cizek and Bunch, 2007; Cizek, 2012; CoE, 2009) and modified according to the way it was used in a similar standard setting event by the Saint Petersburg University team (Verhelst et al., 2019). The common feature of this method, regardless of its modification, is that each judge, as a member of a larger panel, has a unique folder with a subset of written or spoken performances selected according to a special linked design, and has to answer a question *Is a candidate who writes (or speaks) like this at a certain level or above?* The 'body' in the name of this method refers to one of the characteristics of the BoW: the *body* refers to several performances produced by one student or candidate. Theoretically, as described by Cizek and Bunch (2007), the body might be represented by a mixture of constructed response tasks and selected response tasks. In our case, the body was a written or spoken performance by one student which consisted of responses to two or three different tasks.

Data Collection

Performances were collected during the pretesting of the new, revised productive skills tasks. They were scored previously by at least three senior raters and a consensual score was assigned to each of the performances. This score was not known by the panel members at any point, but it was used for the selection of the performances for the standard setting and later for the analysis in the statistical program.

The performances included in the STS generally came from candidates with scores within the 30–100% range for each level. This was done to avoid very weak performances with multiple problems and a high probability of them being far below the target level. The intention was also to have uniform score distribution along the score range. This turned out to be a condition very difficult to fulfill in the early stage of the new exam development since performances could only be collected during the pre-testing phase, which typically have a lower number of participants. As a consequence, not all score points were equally well represented, in some cases only by one performance with a particular score point. Therefore, some of the performances had to be included twice with different IDs to improve the score scale coverage[11]; they were coded and treated as different performances. In the design

[10] At the time of writing, only the standard setting for Writing has been fully completed.

[11] In fact, including some of the performances twice with different ID helped only in the sense that there were no gaps in the score scales and for every score point there were some performances. It would be better if for each score there existed more different performances. In our case, they were not available at the time of the standard setting as the exam had not gone live yet and only samples from pretesting could be included.

code script	score	judge 1	judge 2	judge 3	judge 4	judge 5	judge 6	judge 7	judge 8	judge 9	judge 10	judge 11	judge 12	judge 13	judge 14
03000000	19			14			7		11			3			
19000001	15		7				9	9					5		
27000002	19	14			6						8			12	
17000004	25				12			13		5		7			
47000005	21						5		5		4			8	
11000007	12		4			1				2				13	
13000008	21	1	10						13	6					
16000009	14		12						6					4	7
15000010	13	7			4						1				4
18000011	15	12						3		9		13			
20000009	14							1		10	13	1			

Fig. 2 Example of the linked design prepared for C1 Writing (part)

(see Fig. 2 below) we had to ensure each judge had a set of unique performances by avoiding the inclusion of these doubled performances in the judges´ folder.

Panel Composition

Between 12 and 16 judges were present in each panel. All of them were Czech L1 speakers, experienced language teachers of foreign languages, mostly of Czech, with a very good knowledge of the CEFR. Approximately half of them were external collaborators (mainly teachers or raters) and half of them were internal team members (test constructors, rater trainers etc.).

Design

For the session, individualized folders for judges following a specific design (Fig. 2) were prepared. There were several conditions essential for reliable calculations: (a) every performance is judged by four judges (indicated by the number of yellow/gray cells in each row), (b) every judge shares at least half of the performances with another judge (for instance, J07 shares script 200,000,009 with J09, J10 and J11), (c) each folder contains a unique combination of randomly selected performances, and these are ordered randomly to avoid the influence of weak or good performances clustered together (the sequence numbers are the numbers in the cells, going from 1 to the maximum numbers of performances per judge); (d) the order of the performances in each folder was random, but it has to be strictly followed by the judge; (e) the workload (number of judged performances) for each judge is the same.

In the modified BoW method, the judge's task is to read a script or listen to a performance and consider whether the person who writes or speaks like this is at a particular level of the CEFR or above. For all levels, each performance (the body of work) consisted of two or three pieces. Judges had to consider all of them, evaluate their good and weak aspects and make only one final judgment, indicating YES, i.e. that the performance is at the intended level or above, or NO, that the performance

is not at the intended level. One of the added features was that judges were asked to express how much doubt was involved in their judgment of each performance (Table 3).

The statistical model in the LOGREG[12] program used for the calculations operates on two basic assumptions: first, students who are at a certain level will score on average higher than students below the intended level (which is related to the validity of the scores), and second, they are also more likely to obtain a YES from the judges (related to the validity of the judgment). The cut-off score is defined as the score for which the probability of obtaining a YES judgment equals 0.50 (Verhelst et al., 2019, p. 284). LOGREG calculates (among other kind of output) the frequency distribution of the scores, the number of positive judgments per score (YES), parameter estimates β1 and β2 (the slope and the intercept), the cut-off score of the set of samples included in the standard setting and its standard error, and the estimation of rater effect: the theta value (leniency of the judges) and the residual sum of squares that indicates how much a judge matches the model prediction, or in other words, how well a judge discriminates between the performances (with a high score and a high probability of YES judgements and those with a lower score and a lower probability of YES judgements). At the end of each session the results are presented to the judges – the value of the standard they set, and the output information provided by LOGREG as well as the accompanying Excel sheet with additional calculations, which shows the cut-off score (Fig. 3) and judges' leniency and consistency. These outcomes are also displayed graphically (Fig. 4).

Table 3 Example of the sheet with data provided by judges. (The score was visible only to the session moderator)

judges	sequence number in the folder	code script	score	**Is this a performance at an X level or above?** YES=1 NO=0	**How much doubt was involved in your decision?** No doubts – 0 Some doubts – 1 A lot of doubts - 2
J01	1	13000008	21	**1**	**0**
J01	2	23000824	12	**0**	**0**
J01	3	30000020	22	**1**	**0**
J01	4	29000019	18	**1**	**0**
J01	5	21000043	15	**1**	**0**
...	...	...	...	...	...
J02	1	37000593	22	**0**	**1**
J02	2	31000042	19	**0**	**2**
J02	3	21000014	13	**0**	**2**

[12] The name LOGREG stands for the method of logistic regression implemented in the program. The regression coefficients are estimated by a maximum likelihood method. More technical details can be found in Verhelst et al., 2019.

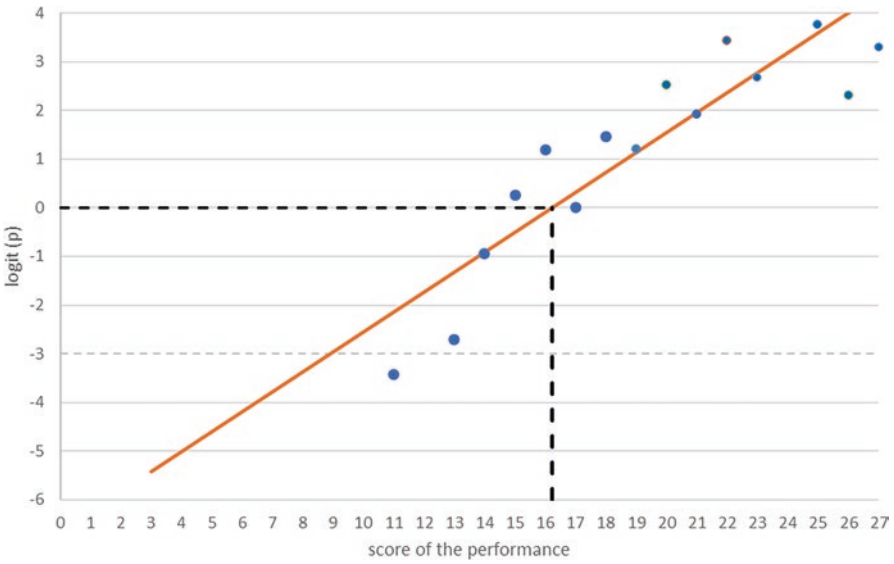


Fig. 3 Standard set at 0.5 probability point

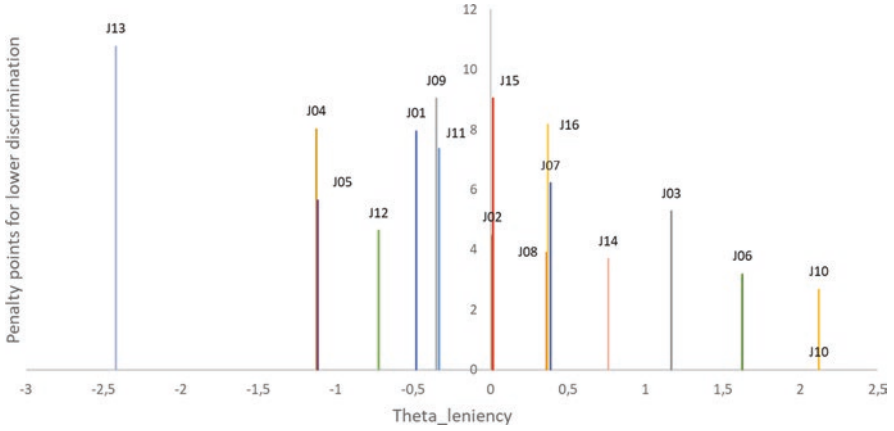


Fig. 4 Leniency of the judges (the more to the right the judge is, the more lenient is the judge) and the amount of unpredictability in the judgement (the higher the vertical line, the more unpredictable the judge's behaviour is)

Validity Checks

For the judges, the session ends with the discussion of the presented results, and they are asked to complete a feedback questionnaire that provides us with evidence of the procedural validity. The results are also validated by checking whether the results correspond to the following expectations: (a) there is a strong positive correlation between the higher scoring performances and the probability of getting a

YES answer; (b) there is no statistically significant difference between cut-off scores set by subgroups (i.e. teachers x testers, internal x external team); meaning the panel was well selected and the familiarization and trial round followed by discussion led to the shared interpretation of the level characteristics across the groups; (c) the level of uncertainty (the additional question we added to the original BoW method – see the far right column in Fig. 2) is accumulated around the point of the set standard and decreases towards both ends of the score scale. We expect the highest uncertainty around the standard. The uncertainty should decrease towards both ends of the scale. Concerning the judges' harshness/leniency, the question is whether the level of uncertainty is a personal characteristic of each judge, and to what extent it is related to the panel composition or influenced by temporal factors (performance features, the way familiarization and discussion are carried out, the quality of materials and information provided, the influence of the moderator, etc.). For this, the final discussion and the evaluation questionnaire is crucial and could help to shed light on this issue.

6 Assembling Comparable Test Forms

The processes described above in this section are a necessary step towards another aim, namely the assembling of comparable test forms.

The reading, speaking and grammar/lexical subtests as well as the tasks produced (all of which were previously pretested and described in terms of content characteristics, such as the length, topic, testing technique, can be now described by another parameter – the task cut-off score, which equals the score that a minimally competent candidate (linked to the 60% score on the score scale) would obtain). Thanks to the continuous standard setting, a task bank containing tasks of receptive skills with estimated cut-off scores and their related standard error is being compiled. When assembling test forms for a particular exam session, tasks from the bank are combined according to both a pre-established content matrix and with regard to their cut-off score values. The aim is to assemble a test form with the total cut-off score[13] as close as possible to 60% (the value pre-established for all exams at Charles University as the pass point). If the total cut-off score deviates from 60% (bearing in mind the confidence intervals calculated automatically for each test form at the time they are assembled), a linear score transformation method is used. Raw scores of the candidates taking this particular test form are transformed and reported on a reporting scale.

For productive skills, the standards set by the panels are very close to 60%. But even so, a reporting scale is always prepared for every test form at each level and candidates' scores are transformed using linear transformation, where the common point on both scales is 60%. The scores below and the score above the 60% point are transformed separately. Candidates' certificates contain the transformed percentage scores.

[13] calculated as the average of tasks cut-off scores

7 Insights Gained

The internal review the RTC team had to undergo brought multiple benefits for the CCE exams, the stakeholders and for the team itself. The aims formulated at the beginning of the revision process have been fulfilled: the content of the exams was revised and related to the newly defined construct and the CCE exams were aligned to the CEFR. The new specifications have been working well, providing good guidelines for item writers and enabling them to create tasks and items that are comparable in content and as well as in terms of subskills measured. As it is not expected item writers would create tasks of equal difficulty, the task difficulties, operationalised by the score the MCC would get, were estimated during the standard setting procedures, where judges considered the interaction between task characteristics and the test-takers' ability i.e. the MCC description. Standard setting for receptive skills has become an ongoing process, happening three times a year. Every new task must pass through the standard setting estimation and only then can it be stored in a task bank. Task banking is also one of the new features of test development and one of its major benefits is that it allows for the assembling of comparable test forms. The results/scores are reported on a reporting percentage scale, where the pre-set cut-score is linked to the cut-off score based on the standard setting and if necessary, linear transformation is applied.

The CCE exams are now not only new, but they are explicitly in line with good practice in the field of language testing, as described by the ALTE Minimum Standards, EALTA Guidelines for Good Practice, or the Standards for Educational and Psychological Testing. The need and will to keep pace with the latest developments in the field, the regular critical checks of the revised exams, and the implementation of the Quality Management System in 2015, led to several key improvements:

1. The growth of shared internal knowledge changed attitudes. The team members feel much more responsible for the exams as the newly developed exams and material were clearly based on evidence and thorough analysis.
2. The processes and products (item writing, item writer training, raters' behavior and their satisfaction with the rating schemes, rater training, test content, pretesting, etc.) improved significantly, as reported in the feedback from those involved in the test development. New rating schemes are coherent across levels and linked to the CEFR. Raters report being more effective and comfortable when using them, the agreement is easier for them to reach, and the wording helps to justify their ratings.
3. Validation, gathering evidence and building the validity argument has become part of the test development process and the CCE exams can prove explicitly what has been done in terms of fairness and validity.

Although the benefits brought by the revalidation prevail, there are still areas that require attention, e.g., an examination of the cumulative data from live testing and to see whether the cut-off score values for the tasks correlate with the operational data.

For productive skills, as was mentioned, tasks for writing and speaking have been treated as comparable although we have no evidence so far for this claim. We plan to continue to investigate the difficulty question and repeat the standard setting when we have more performances from live sessions available. This will allow us to confirm the previously set standards, to investigate to what extent the cut-off scores set for different tasks differ, and adjust them accordingly.

The experience with validation has been very challenging and time consuming. The first impulse for the revision was the need to pass the second ALTE audit and to address the issues revealed by the first audit. The deeper we went into the research, the more aspects that needed to be improved were discovered, and the revision and revalidation have taken more than 7 years for the five exam sets.

However, the endeavor led to many insights and gave the RTC team an opportunity to learn a lot, something that the RTC team has made full use of. It was necessary to realize that (1) the revision and revalidation would take a long time and new issues were likely to arise, (2) the process would need to be segmented into smaller steps, and (3) that the phases needed to be well planned and scheduled to make sure the work proceed forward and it was not necessary to return back to some steps. Much welcome was the fact the wide range of individuals working on the CCE understood the importance of every single step to re-design a valid and reliable examination with positive washback. The outcome of the revalidation has been a new, revised CCE exam with a solid validity argument and an RTC team that has grown in terms of knowledge, experience, professionalism, self-motivation and self-confidence.

References

AERA, APA, & NCME. (1999/2014). *Standards for educational and psychological testing*. AERA.

Alderson, C. J. (2004). *Assessing reading*. Cambridge University Press.

Anýžová, P. (2013). Ekvivalence položek v mezinárodních datech: základní vymezení a možnosti analýzy. *Data a výzkum – SDA Info 2013, 7*(1), 29–56. https://doi.org/10.13060/1802-8152.2013.7.1.2

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

Bachman, L. F. (2012). *Justifying the use of language assessments: Linking interpretations with consequences*. Conference paper. Retrieved January, 10, 2015, from http://www.sti.chula.ac.th/conference. https://doi.org/10.20622/jltajournal.18.0_3

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford University Press.

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice: Developing language tests and justifying their use in the real world*. Oxford University Press.

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

Cizek, G. J. (Ed.). (2012). *Setting performance standards. Foundations, methods, and innovations*. Routledge.

Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. Sage Publications Ltd..

Council of Europe. (2009). *Relating language examinations to the common European framework of reference for language: Learning, teaching, assessment (CEFR): A manual*. Council of Europe.

Fulcher, G. (2010). *Practical language testing*. Hodder Education/Routledge.
Holland, P. W. (2007). A framework and history for score linking. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 5–30). Springer.
Kingston, N. M., Kahl, S. R., Sweeny, K. P., & Bay, L. (2001). In G. J. Cizek (Ed.), *Setting performance standards. Concepts, methods, perspectives* (pp. 219–248). Erlbaum.
Livingston, S. A. (2004). *Test score equating (without IRT)*. Educational Testing Service. Retrieved July, 17, 2016 from www.ets.org
Luoma, S. (2004). *Assessing speaking*. Cambridge University Press.
Messick, S. (1987). *Validity*. ETS Research Report Series. https://doi.org/10.1002/j.2330-8516.1987.tb00244.x.
Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher, 18*(2), 5–11. https://doi.org/10.3102/0013189X018002005
Purpura, J. (2004). *Assessing grammar*. Cambridge University Press.
van de Vijver, F., & Poortinga, Y. (2005). Conceptual and methodological issues in adapting tests. In R. K. Hambleton, P. F. Merenda, & C. D. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 39–63). IEA Lawrence Erlbaum Associates, Publishers.
Verhelst, N., Figueras, N., Prokhorova, E., Takala, S., & Timofeeva, T. (2019). Standard setting for writing and speaking: The Saint Petersburg experience. In A. Huhta, G. Ericson, & N. Figueras (Eds.), *Developments in language education: A memorial volume in honour to Sauli Takala*. University Printing House.
von Davier, A. A. (2011). A statistical perspective on equating test scores. In A. A. von Davier (Ed.), *Statistical models for test equating, scaling, and linking* (pp. 1–17). Springer.
Weigle, S. (2002). *Assessing writing*. Cambridge University Press.
Weir, C. (2005). Language testing and validation. In *An evidence-based approach*. Palgrave/MacMillan.

Speaking "CEFR" about Local Tests: What Mapping a Placement Test to the CEFR Can and Can't Do

Suzanne Springer and Martyna Kozlowska

Abstract Many local tests were developed before the widespread use of the Common European Framework of Reference for Languages (CEFR), and therefore were not designed to reflect the categorizations of CEFR levels. Institutions wishing to express their local tests in terms of CEFR benchmarks are encouraged to carry out a mapping procedure (Council of Europe, Relating language examinations to the common European framework of reference for languages: learning, teaching, assessment (CEFR). Language Policy Division. https://rm.coe.int/1680667a2d, 2009). Although the CoE offers guidelines for such undertakings, the brunt of the challenge remains with the institutions, administrators, and language professionals who must reconcile numerous practical and methodological considerations specific to their local contexts. In this chapter, we report on an exploratory CEFR mapping study involving a local ESL placement and proficiency test at a Canadian university. We highlight the ways in which this specific type of testing research is useful to report linkages, but that the real benefit of the exercise is that it encourages documentation and analysis of test constructs and assumptions and facilitates informed reflections about the implications for language programs and local tests, especially those outside Europe, to create or maintain an association to the CEFR.

Keywords Local language testing · Common European framework (CEFR) · Francophone instructional contexts · English as an additional language

S. Springer (✉) · M. Kozlowska
The Université du Québec à Montréal, Montreal, Canada
e-mail: springer.suzanne@uqam.ca; kozlowska.martyna@uqam.ca

X. Yan et al. (eds.), *Local Language Testing*, Educational Linguistics 61,
https://doi.org/10.1007/978-3-031-33541-9_4

1 Introduction

The Common European Framework of Reference for Languages (CEFR) has become very influential in shaping language proficiency descriptors worldwide (Read, 2019). While North (2020) stresses that the English Language Teaching (ELT) industry has played a disproportionate role in the spread of the CEFR, its currency as a useful tool to speak about learner levels is undeniable and a large part of its appeal for institutions, particularly those who recruit international students. Regardless of country or context, language practitioners can share a common understanding, albeit approximate, of a learner's language competency. Indeed, the precise criteria that lead to a B2 judgment about a student's ability, for example, may not be the same criteria used in a different context; as long as the two classifications do not differ too greatly, this nomenclature "works." It's generally understood that variation in criteria is not only acceptable but to be expected and in fact, desired. As North (2020) stated, "The CEFR is deliberately open-ended. This is because it is intended to be used in a wide variety of different contexts: for different languages, for different age groups, for different types of learning goals, in different pedagogic traditions" (p. 14).

The adaptable nature of the CEFR speaks to its universality, but it brings challenges when operationalizing the descriptors for a specific context. Of particular concern is the lack of set limits, in terms of level descriptor content and scope, on how much an institution can modify the CEFR rubric to fit their specific context (Harsch & Hartig, 2015). Savski (2021) summarizes some of these tensions as follows:

The key issue is that in order to achieve maximum universality, any large-scale global framework like CEFR must be minimally flexible, since significant variation between how it is interpreted and used may endanger its universality. This is at odds with the general need for policy of any kind to be open-ended enough to allow local actors an appropriate amount of leeway to take decisions based on their knowledge of the context in which they are working. Such flexibility is particularly key when it comes to language education policy at the global scale, since there are significant differences between different language ecologies and between the practical conditions individual educational actors have to consider when making decisions (p. 61).

With this caveat acknowledged, the Council of Europe (CoE) issued a Manual entitled *Relating language examinations to the Common European Framework of Reference for languages: Learning, teaching, assessment* (2009) providing a blueprint and several options for how to conduct a test mapping study. Subsequent complementary resources for use with the CEFR such as the *Manual for language test development and examining* (CoE, 2011) and Martyniuk's special volume on *Aligning tests with the CEFR* (2010) offered further insight on the mapping process and objectives. Nonetheless, the sheer scope of alignment methodologies, statistical analyses, and potential outcomes may dissuade the very language professionals the CoE hopes to encourage to validate CEFR claims. However, embarking on a CEFR mapping study constitutes a meticulous and comprehensive learning process about CEFR, the local test, the local learning contexts, objectives, and outcomes for all involved stakeholders.

This chapter reports on an exploratory mapping study that operationalized the CEFR levels to describe test taker performance on an ESL[1] language placement examination at a francophone university in the province of Quebec (Canada), henceforth the ANG test. We will note the ways in which, with limited resources (a hallmark of many local testing contexts), language practitioners can maximize this type of research in order to not only relate tests and the CEFR benchmarks, but to improve skills in Language Assessment Literacy (LAL),[2] begin critically analyzing test constructs and assumptions, and make recommendations for future claims about language tests and their relationships to CEFR levels. This cycle, which the Framework creators argue is in fact the true intention of working with the CEFR in language programs (North, 2020), leads to discoveries about what using the CEFR *can* and *cannot* do while navigating the challenges inherent to its application in standard setting.

2 Local Testing Context

Before going into the details about the ANG test's creation, test candidates, and target language use (TLU) domains, it's important to contextualize the complex linguistic setting in which it was developed and administered.

2.1 The University Context

This university is located in the French-speaking province of Quebec. English is not the primary language of instruction and therefore not an entrance requirement for the vast majority of students. Historically, this university positioned itself as an open and democratic higher learning institution, and part of its initial mission was to welcome francophone students from Quebec families, many of whom were the first to have access to post-secondary educational opportunities, in contrast to a higher percentage of anglophone Quebecers attending university at the time (Venne, 2019). Policy makers continue to position the university as a staunch defender of the French language and Quebecois culture that actively promotes French in the dissemination of academic and scientific knowledge and advancements.[3] Over time, program directors have had to contend with the usefulness and even necessity of being

[1] As this university is in the French-speaking Canadian province of Quebec, *English as an Additional Language* (EAL) may more accurately describe the English programs in this institution's French as a Second Language (FLS) context.

[2] See Coombe et al. (2020) for an overview of the multifaceted modes and constructs involved in LAL.

[3] See the Francophone Association for Knowledge (*Association francophone pour le savoir*) or ACFAS (acfas.ca) for more about efforts to maintain and promote French in the dissemination of academic and scientific information.

proficient in English in a growing number of academic and professional situations both in large multicultural Quebecois cities like Montreal as well as the rest of Canada, the United States, and beyond. English programs and the inclusion of an English proficiency level as a graduation requirement for some programs grew out of this reality, yet linguistic tensions between French and English remain, both at the university and in Quebecois society in general (Bruemmer, 2018). The pertinence of English tests such as the ANG test are acknowledged university wide, albeit begrudgingly by some policy makers and even some students.

2.2 *Local Test Purpose*

The ANG test, designed for placement in the university's credit ESL programs, was developed at the university's School of Languages in 2000 and has been in use since 2004. The test consists of nine sections that are mapped to the ESL courses at five proficiency levels. The sections of the test (*reading, listening, vocabulary, speech perception, grammar, critical reading, speaking, writing,* and *pronunciation*) and how their combinations are mapped to courses are illustrated in Fig. 1 for communicational courses and Fig. 2 for grammatical courses.

The ANG test is a computer-based exam. Test-takers complete multiple-choice questions, a section of true/false/not enough information questions, a written production subtest, pronunciation recordings, and a spoken monologue, as listed in Table 1 below.

The production subtests are typed or recorded through a computer interface. Receptive test sections are automatically scored through the computer interface,

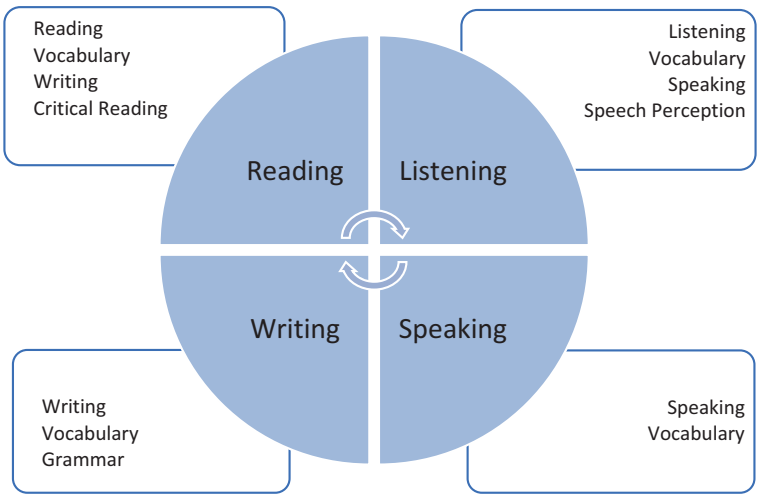


Fig. 1 Test sections used for scoring and placement – communication courses

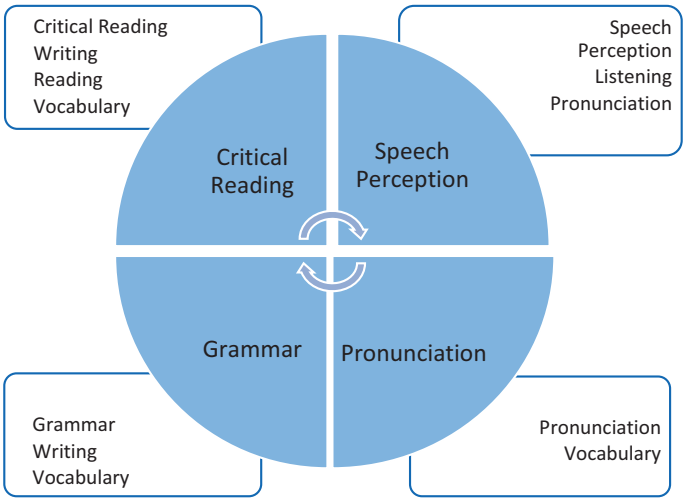


Fig. 2 Test sections used for scoring and placement – grammatical courses

Table 1 ANG test subsections

Test section	Question type
Grammar	Multiple choice
Vocabulary	Multiple choice
Speech perception	Multiple choice
Reading	Multiple choice
Listening	Multiple choice
Critical Reading	True, false, not enough information
Speaking	Constructed response
Writing	Constructed response
Pronunciation	Read aloud

while productive test sections (writing, pronunciation, and speaking) are evaluated concurrently by a pair of raters using an analytic scale, but a single holistic score is submitted.

2.3 The Evaluation Scales

The production sections of the ANG test (writing, speaking, and pronunciation) use a continuous marking scale, but raters select scores at 5-point intervals, ranging from 17 to 97. As Table 2 shows, each band corresponds to a proficiency/course level, starting with beginner, split into low-mid-high, and continuing up to the fluent proficiency/course level, also split into low-mid-high points. Evaluators choose borderline scores according to these intervals, but in reality, the cut-off scores lie between these two integers.

Table 2 Production subtest marking scale

Score		*Level*
17	–	(indicates a technical error with the test, usually with audio recordings)
25	**Cut-off score**	
27	Low	Beginner
32	Mid	
37	High	
40	**Cut-off score**	
42	Low	Intermediate I
47	Mid	
52	High	
55	**Cut-off score**	
57	Low	Intermediate II
62	Mid	
67	High	
70	**Cut-off score**	
72	Low	Advanced
77	Mid	
82	High	
85	**Cut-off score**	
87	Low	Fluent
92	Mid	
97	High	

The multiple-choice and true/false/not enough information subtests also use a continuous scale from 10 to 100, and cut-off scores between levels are at the same points as in the scale used in production subsections (Table 2). Students taking the ANG test for course placement do not receive performance scores. Instead, their test results are issued as a report indicating personalized course choices based their scores on the nine evaluated competencies.

2.4 ESL Program Structure and the CEFR

The ESL programs under consideration were not designed with the action-oriented approach espoused by the CoE at its core; instead, they are structured based on distinctions between discrete grammatical and communicational competencies. This design sought to reflect the diverse language objectives of the university's student population, from recent immigrants to more established Quebecois students, by offering a flexible curriculum and diverse pedagogies.

As part of a recent modification to the university's ESL programs, a team of language instructors had mapped course levels to the CEFR and the Canadian

Language Benchmarks based on a review of course descriptors, objectives, and minimum content. Even though certain grammatical courses (such as Speech Perception) could not be easily mapped to the 2001 CoE "can do" descriptors, the communicational courses could.

This mapping procedure allowed for the association of the program course levels (from beginner to fluent) with their corresponding CEFR levels, from A2 to C2 respectively.

2.5 Other University Stakeholders

In addition to its primary TLU for placement in English courses and programs, the ANG test is used to report on English proficiency levels for certain programs that have added this as a graduation requirement. Some students also take the ANG test in order to attest their English proficiency for study abroad in English institutions. One reason for the ANG test's appeal is that students taking it for attestation purposes receive not only a report of their subsection scores for communicative and grammatical competencies but also scores that serve for ESL course placement. This type of placement information is not available through other proficiency measures such as IELTS or TOEFL and is only possible thanks to a close collaboration between the language practitioners at the university's School of Languages and administrators of programs with language requirements. The English program director is frequently consulted in order to understand and formulate language proficiency objectives, or TLUs, for graduates in programs such as business administration, marketing, or linguistics; those proficiency objectives are then expressed in terms of a score on the ANG test, along with courses that allow the attainment of the desired level if necessary. The language of the CEFR benchmarks is very useful in these discussions as the descriptors are often more accessible for those less familiar with language acquisition and proficiency terminology. These level calibrations are periodically negotiated and revised as students who do not have the prerequisite level must take additional courses and add any associated fees to their tuition, which invariably puts additional pressure on the programs requiring an English level.

3 Testing Problem Encountered

Due to the expanded use of the ANG test and its reporting of scores for both students requiring an English proficiency score (a high-stakes use) and its original purpose for placement in ESL courses (a low-stakes use), we found ourselves with a number of stakeholders to account to: students placing in courses mapped to CEFR levels, students requiring a certain level of proficiency to graduate, international partners who rely on attestations for their required admissions levels in terms of CEFR, and instructors developing courses using CEFR benchmarked materials.

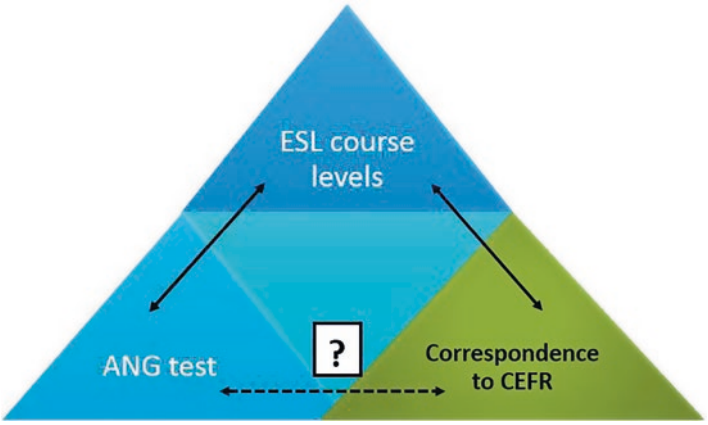


Fig. 3 Pre-mapping study verified and unverified correspondences

We knew that mapping courses to CEFR may have been the first step in establishing CEFR linkages, but that we were indirectly making unverified claims about linkages to the CEFR through the ANG test. This situation is illustrated in Fig. 3.

The time was overdue to validate these assumed ANG test linkages to better inform our score reporting. While we could have simply opted out of using any correspondences between the ANG test and the CEFR, the pressure to make valid CEFR claims is too important for a large university such as ours with established international outreach to ignore. To address this issue, we designed a study to find correspondences between our ANG test and the CEFR benchmarks despite challenges we anticipated as the ANG test was not developed based on the action-oriented framework underpinning the CEFR.

4 Literature Review

A number of internationally recognized high stakes tests such as TOEFL, TOEIC, and IELTS have undergone mapping studies linking their test scores to the CEFR (Tannenbaum & Wyle, 2004; Taylor, 2004; Tannenbaum & Wylie, 2008; Tannenbaum & Baron, 2011; Lim et al., 2013). However, CEFR mapping is challenging for institutional proficiency measures that are used for localized purposes, such as university language entrance exams, placement tests, or exit level tests. Even if a given institution undertakes a study relating their test scores to the CEFR levels, operationalizing the CEFR descriptors is not always straightforward (North, 2014), and the mapping process is resource intensive (O'Sullivan, 2010).

Deygers et al. (2018) highlight many of the issues that language evaluators face when making claims about a local test and the CEFR. While North (2020) refutes some of the claims made with regards to institutional use (or misuse) of the CEFR

as a normative standard setting tool, Deygers et al. (2018) offer the following insights and suggestions:

Decision making would benefit from a realistic assessment of the CEFR's limitations in this respect, recognizing it as a general theoretical framework that needs to be supplemented by language-specific and context-specific descriptors. This does not imply that linking a test to the CEFR is futile. On the contrary: CEFR linking provides a reference point, but without the assumption of exact equivalence between all tests at the same level. Furthermore, linking may provide valuable insights into those characteristics of the test that are language-specific and context-specific. Results of such a linking should not be regarded as the one truth but as one of many valuable aspects of validation (p. 13).

North (2020), the co-author of the CEFR (CoE, 2001), sees the publication of the CEFR Companion Volume (CoE, 2020), which he also co-authored, as an occasion to take stock of some of the criticisms that have been expressed over the years towards CEFR. He refutes complaints such as the lack of a clear theoretical underpinning for CEFR and its apparently weak link to research on learner language as issues that, in his view, are but misconceptions with regard to the methodology employed for descriptor development and calibration. Perhaps the most pertinent for the present study are the criticisms with respect to *the formulation* and *the intended scope* of the descriptors, as these are instrumental in a linking study such as ours.

In response to the criticism of a seemingly random and "subjective" occurrence of constructs at different levels of the scale and the lack of their systematic development at every level (Alderson et al., 2006; Tracy, 2017), North clarifies that the choice made in favor of the "salient features approach" rather than the "systematic" approach in the formulation of the CEFR level descriptors was deliberate. It was to provide a sort of a map for curriculum development (North, 2020), *but not criteria for a rating scale.* Furthermore, the use of relative rather than binary distinctions between levels, he explains, was driven by the inherent relative nature of a common framework, such as the CEFR, which in turn determines the scope of the CEFR itself. Its application is intended for varied educational contexts, in relation to different languages and different learner profiles. Hence, North argues the CEFR must be taken solely as a set of guidelines interpretable in varied contexts and adaptable to varied specific needs. These guidelines may even be used for elaborating new descriptors as some have done (Díez-Belmar, 2018; Dendrinos & Gotsouilia, 2015; Szabo & Goodier, 2018; North & Jaroscz, 2013; Shackleton, 2018 all as cited in North, 2020). Overall, it appears that even after several modifications of the CEFR itself (CoE 2001, 2018, 2020), its original fundamental vocation remains unchanged: "We have NOT set out to tell practitioners what to do or how to do it. We are raising questions, not answering them. It is not the function of [the CEFR] to lay down the objectives that users should pursue or the methods they should employ" (CoE, 2001, p. iv).

It is in the spirit of these debates and recommendations that we undertook our study – to establish a starting point from which to find links and better understand the characteristics of our local ANG test and the implications of our making CEFR claims in score reports.

5 Methods

Our discussion of methods begins with objectives.

5.1 Linkages Between the ANG and the CEFR

The research project reported here centers around a mapping study carried out between the years 2017 and 2020. We were primarily looking for linkages between the ANG test and the CEFR. As this type of study is very resource intensive, we maximized its scope to serve us in collecting information that will be of use in the future. Namely, we added steps to allow us to begin making more informed claims about our students and their language proficiency levels in terms of the CEFR and to better understand the implications of incorporating the CEFR benchmarks in future iterative test evaluation, development, and revision cycles (Milanovic & Weir, 2010, pp. xiii–xiv).

The research questions for this study are:

1. What are the linkages between the ANG test scores and CEFR cut-off scores (communicational sections)?
2. What recommendations can be made, based on ANG test and CEFR linkages, for future iterations of the ANG test?
3. In light of our findings in questions one and two, what role should the CEFR have in future program and ANG test claims?

5.2 Mapping Procedure

The CoE's Volumes (CoE, 2001, 2020)[4] and Manuals (2009, 2011) proved indispensable to understand important concepts in language assessment validity and the CEFR's origins, underpinnings, development, and uses. These resources detailed the mapping stages and their rationales, thus demystifying the procedure and offering helpful first steps.

A panel of judges[5] participated in the first four of five interrelated mapping steps outlined in great detail in the 2009 Manual.

1. Familiarization
2. Specification

[4] It is important to note that the study began before the publication of the CoE 2018 preliminary version of the updated Companion Volume.

[5] Due to financial limitations, the panel included the authors of this study. Although this may raise some concerns, we felt confident in the integrity of the panel judgments, which were not mitigated by inter/intra-rater reliability nor performance rankings as part of the scope of this project.

3. Standardisation training/benchmarking
4. Standard setting
5. Validation

The last step, validation, is ongoing and based in part on recommendations from this exploratory study.

The resources available for the study were: (1) a team of five evaluators, all experienced English language instructors (each with over 10 years' teaching experience), familiar with the test-takers' and the test itself; (2) access to test scores and responses (for both the constructed response and multiple-choice sections of the test) through the university's centralized testing center[6]; (3) a research assistant working in concert with the testing center to both prepare and anonymize existing responses and code panel judgments for analysis; and (4) statistical analysis expertise offered by the university data analysis consulting service.

The study was carried out in phases that grouped communicational competencies according to the format of the test questions:

Phase 1 – examinee centered sections, concerned with analyzing constructed responses (*writing* and *speaking*)

Phase 2 – test-centered sections, concerned with analyzing multiple choice questions (*listening, reading, critical reading* and *vocabulary*)

Familiarisation Sessions

At the onset of the study, the panel coordinator gave a brief presentation of the CEFR (its objectives and applications) and the Common Reference Levels. The panelists were then asked to read relevant sections of the 2001a CEFR document, including a holistic summary table of Common Reference Levels (Table 1, p. 24). What ensued was a series of familiarization activities, such as sorting and ordering the text from Table A1 and highlighting the salient features for each level.Panelist responses were then compared and discussed.

Familiarization activities for written production consisted of activities such as finding salient features from the level descriptors from Table C4 of the Manual (CoE, 2009, p. 187), writing a text in a second language (French) common to all panelists, self-evaluating the French text, receiving an expert judgment on the French CEFR writing level,[7] and lastly, rating benchmarked samples from the Helsinki project training material[8] (supplemented by a small sample of writing performances from the ANG test). Familiarization activities for oral production

[6] Only responses from test-takers who consented to having their results used for research were considered.

[7] A colleague from the university's French as a Second Language program generously provided this feedback for the panelists.

[8] https://blogs.helsinki.fi/ceftrain/2019/11/27/welcome-to-the-ceftrain-project-training-materials-site/

Table 3 ESL program course levels mapped to CEFR levels

ANG course level	CEFR level
Beginner	A2
Intermediate I	B1
Intermediate II	B2
Advanced	C1
Fluent	C2

consisted of identifying salient features of each level descriptor from Table 3 (CoE, 2009, p. 185), a discussion of the latest Companion Volume with updated descriptors for sustained monologues (CoE, 2018, pp. 70–72), and practice rating sample speaking performances from the ANG test. Discussion was encouraged during each activity, and all sessions were recorded.

Lastly, panelists completed a questionnaire regarding their understanding of the CEFR pre and post familiarization and how prepared they felt to apply the CEFR benchmarks to evaluate writing and speaking performances from the ANG test. In general, all panelists felt well prepared to apply the CEFR benchmarks, but there were some ongoing struggles with the validity of inferring what a student *can do* based on a single writing or speaking performance, as summarized below in the feedback from.

Rater A:

> There is the problem of competence and performance. Sometimes the descriptors use CAN and sometimes just DOES… We're probably better to say USES rather than CAN USE. Even better for our purposes – USED-which reflects the one-time nature of the assessment, and doesn't overestimate what a one-time assessment can do.

Through subsequent discussions, it was decided that a hybrid "Can do" rating scale be adopted. As not to stray too far from the CEFR's original descriptors, the overall level descriptor was maintained. However, descriptors related to specific aspects of competency (i.e., coherence and accuracy) were turned into past tense statements in order to describe the actual student performance.

Specification

After the panel had gone through the familiarization process, we proceeded to the next stage of the mapping procedure involving an analysis of the ANG test and its content in relation to the CEFR benchmarks.

A test specification document had to be drafted as part of the project. As is often the case with many locally designed proficiency measures, such a report had not been created by the original test authors. Therefore, information needed to be collected from various sources: the test itself, original test developers, and the university testing center technicians.

The drafting of the specification document offered one of the first positive outcomes of the current mapping procedure – a result anticipated in the CoE mapping Manual (2009, p.27). The necessity of delving into the properties of each of the test

sections and marking rubrics allowed us to analyze the test content's strengths and weaknesses, reflect on the testing and scoring procedures, and then document those observations. As suggested by the CoE, the specification process serves as an awareness raising exercise which, ultimately, increases the transparency of the test content, its quality, and how it positions itself vis à vis the CEFR.

The Manual proposes numerous procedures and specification tools to analyze test content; however, many of them go beyond the scope of what was practically attainable in the context of the current study. Nevertheless, the panel contributed to the documentation of ANG test properties and constructs by means of checklists from the 2009 Manual (pp. 126–132).

Standardization Training

Chapter 5 of the Manual (CoE, 2009) provides detailed information about the purpose and procedure for this step. The examinee-centered method relies on judgments about the performance of the examinees on a test, whereas the test-centered method relies on judgments about test items. For the purpose of the current study, both of the methods were used: the examinee-centred method in standard setting for constructed response performance, and the test-centred method for multiple-choice test items (Kane, 1998).

Standardization – Benchmarking The writing and speaking subtests (direct tests) were marked holistically. Typical samples to illustrate the six CEFR levels were reviewed and analyzed as part of the familiarization sessions. As this study is exploratory in nature, benchmarking was ongoing, and certain student productions were earmarked for both this present project and future validation studies. Concretely, panelists provided their CEFR judgments accompanied by a short rationale (i.e., which parts of the CEFR level descriptor was particularly salient in the sample), collected through Google Forms. As noted in the familiarization sessions section of this chapter, CEFR evaluation grids for writing and speaking were modified slightly to maintain the "Can do" statements while also using descriptive language for discrete sub competencies. Samples for which a number of salient points were noted became part of a "benchmarking bank" for both this and future studies in accordance with group agreement. All benchmarking sessions were recorded.

Setting Standards This study sought to find numerical scores on the ANG test to express CEFR level cut-off points. The mapping looked at similarities and differences between the two scores (those from the ANG test and those arrived at using CEFR). As such, performance standards needed to be set for both the direct ANG test sections (writing and speaking) and indirect ANG test sections (reading and listening). The Body of Work method (Kingston et al., 2001) was used for writing and speaking and the Minimally Acceptable Person (MAP) version of the Tucker-Angoff method (1971) for reading and listening. While performance standards were established using benchmarking for writing and speaking, the panel co-constructed a description of an imagined MAP for each reading and listening competency and CEFR level. These descriptors incorporated the updated "Can do" statements from

the 2018 version of the CEFR Companion Manual (CoE, 2018) and served as a common reference in the standard setting step. All MAP creation sessions were also recorded.

Standard Setting

Modified Body of Work Method Phase one of the study concerned constructed written and oral responses covering a range of numerical scores that correspond to 12 levels on the ANG placement test.[9] These responses had been previously scored using ANG test rubrics. As the test rotates 96 writing prompts and 96 speaking prompts, we also designed the study to include responses from all the possible test prompts. This was done to supplement the ANG test specification document regarding these sections, in particular content analysis. While it is recommended to use this method with a panel of 15 evaluators, we compensated for our smaller panel by collecting more judgments, a modification that was found to also provide technically appropriate results and an option for other local test practitioners who also may not have the resources to include a large panel.

Rangefinding – Writing and Speaking In rangefinding, panelists are presented with work samples in folders. In our study, the evaluators were presented samples in six folders per session, with 12 samples per folder covering possible scores from the low to high ANG scores. The ANG test score was not revealed. We used Google Forms to organize the samples and collect judgments. Raters were also requested to substantiate their responses by indicating which salient features allowed for each classification and then entering this information in Google Forms as well. The information about salient features helped identify which samples to earmark as "expertly" benchmarked. Directly following each folder submission, there was time for discussion about the panel members' reflections and comments on their rating experience. As recommended by the CoE, the panel aimed for consensus but was not required to reach it – which comes with the caveat of not analyzing the raters themselves.[10] All sessions were recorded.

In order to collect data about all the ANG test production prompts (96 for writing and 96 for speaking), every folder contained four questions (Q1, Q2, Q3, Q4 etc.) and three responses (R1, R2, R3) per question. Table 4 illustrates the structure of each session, and there were three sessions total per production competency.

While the project was concerned with all score linkages, we were mostly interested in the cut-off point between A2-B1, B1 being an entry level to the program,

[9] This table indicates the ANG test scores of the majority of the samples. When possible, samples at scores lower than 32 and higher than 87 were also evaluated.

[10] Inter-rater and intra-rater reliability was outside the scope of this exploratory study.

Table 4 Rangefinding session 1

ANG test level	ANG test score	Folder 1 (Google form 1)	Folder 2 (Google form 2)	Folder 3 (Google form 3)	Folder 4 (Google form 4)	Folder 5 (Google form 5)	Folder 6 (Google form 6)	Evaluator task
Beginner – (A2)	32 37 42	Q1, R1 Q1, R2 Q1, R3	Q5, R1 Q5, R2 Q5, R3	Q9, R1 Q9, R2 Q9, R3	Q13, R1 Q13, R2 Q13, R3	Q17, R1 Q17, R2 Q17, R3	Q21, R1 Q21, R2 Q21, R3	Place at CEFR level (A1 – C2)
Intermediate I – (B1)	47 52 57	Q2, R1 Q2, R2 Q2, R3	Q6, R1 Q6, R2 Q6, R3	Q10, R1 Q10, R2 Q10, R3	Q14, R1 Q14, R2 Q14, R3	Q18, R1 Q18, R2 Q18, R3	Q22, R1 Q22, R2 Q22, R3	Place at CEFR level (A1 – C2)
Intermediate II – (B2)	62 67 72	Q3, R1 Q3, R2 Q3, R3	Q7, R1 Q7, R2 Q7, R3	Q11, R1 Q11, R2 Q11, R3	Q15, R1 Q15, R2 Q15, R3	Q19, R1 Q19, R2 Q19, R3	Q23, R1 Q23, R2 Q23, R3	Place at CEFR level (A1 – C2)
Advanced – (C1)	77 82 87	Q4, R1 Q4, R2 Q4, R3	Q8, R1 Q8, R2 Q8, R3	Q12, R1 Q12, R2 Q12, R3	Q16, R1 Q16, R2 Q16, R3	Q20, R1 Q20, R2 Q20, R3	Q24, R1 Q24, R2 Q24, R3	Place at CEFR level (A1 – C2)

B1-B2, B2 being an important score for proficiency attestations, as well as C1-C2 where C2 was judged too high to be admitted in the university's ESL programs. Samples judged at below A1 were coded as A0.

Pinpointing – Writing and Speaking The second round of judgements, *pinpointing*, is similar to rangefinding, but with several important differences. Folders still continued to rotate through the possible writing and speaking prompts on the ANG test, and samples were still presented from low to high ANG test scores, but the panelists judged a more limited range of scores targeting the cut-off scores. Again, the ANG scores were not disclosed. As indicated in Table 5, there were now eight samples per folder, spread out over 6 folders total. Importantly, evaluators submitted their judgments and salient feature information only after the consensus discussion. Finding CEFR judgment consensus was strongly encouraged, but still not obligatory.

In total, approximately 1800 CEFR judgments were collected through both the rangefinding and pinpointing stages.

The Tucker-Angoff Method – The Minimally Acceptable Person (MAP)

The Minimally Acceptable Person (MAP) version of the Tucker-Angoff method (Angoff, 1971) was adopted for the purpose of standard setting for relevant multiple-choice sections of the ANG test (*listening, reading, critical reading,* and *vocabulary*) in phase two of the project. This MAP is judged to have skills and competences at a given level, but only minimally. The panel created a description of the competency of an imagined MAP student for each section of the test under investigation

Table 5 Pinpointing session 1

ANG test level	ANG test score	Folder 1 (Google Form 1)	Folder 2 (Google Form 2)	Folder 3 (Google Form 3)	Folder 4 (Google Form 4)	Folder 5 (Google Form 5)	Folder 6 (Google Form 6)	Evaluator task
Beginner (A2)	37	Q73, R1	Q77, R1	Q81, R1	Q85, R1	Q89, R1	Q93, R1	Place at CEFR level (A1 – C2)
Intermediate I (B1)	42	Q73, R2	Q77, R2	Q81, R2	Q85, R2	Q89, R2	Q93, R2	Place at CEFR level (A1 – C2)
	47	Q74, R1	Q78, R1	Q82, R1	Q86, R1	Q90, R1	Q94, R1	
	52	Q74, R2	Q78, R2	Q82, R2	Q86, R2	Q90, R2	Q94, R2	
Intermediate II (B2)	57	Q75, R1	Q79, R1	Q83, R1	Q87, R1	Q91, R1	Q95, R1	Place at CEFR level (A1 – C2)
	62	Q75, R2	Q79, R2	Q83, R2	Q87, R2	Q91, R2	Q95, R2	
	67	Q76, R1	Q80, R1	Q84, R1	Q88, R1	Q92, R1	Q96, R1	
Advanced (C1)	72	Q76, R2	Q80, R2	Q84, R2	Q88, R2	Q92, R2	Q96, R2	Place at CEFR level (A1 – C2)

and for each CEFR level. The *Yes-No* version of this method was deemed most suitable for the ANG test as a range of difficulty is inherent in the tasks and a range of scores is possible. For each of the multiple-choice sections of the ANG test, the panel decided if such a person, based on the MAP descriptors, could (marked by a score 1), or could not, (marked by a score 0) provide a correct answer to the prompts under investigation. The sum of the scores for each section became a raw result for such a hypothetical MAP. Those raw scores were then combined into composite scores for the constructs of reading and listening (see the Findings section) and compared to a sample (*n = 45 per level per competency*) of actual student ANG test scores of those who were just barely placed in the reading and listening courses in the ESL programs.

6 Results and Discussion

The following section shows the linkages we found through the standard setting procedures suggested in the 2009 Manual (CoE) for the communicational subtests of the ANG test. To answer the first research question with regards to linkages between the ANG test scores and the CEFR cut-off scores, the mappings here suggest strong linkages in certain areas and weaker linkages in others. At times, the same competency may even have both strong and weak links (i.e. the mapping of the reading construct). At this point in this exploratory study, we were concerned with documenting the mapping. Further research is needed to interpret these results.

6.1 Productive Skills – Writing

Tables 6 and 7 indicate frequency judgments at both the rangefinding and pinpointing stages for the writing production section of the test.

From the rangefinding stage, tendencies began to emerge about where the cut-off scores would fall. A category for students who could not perform the task also had to be created (A0), as the panel felt that judging a performance at A1 indicates some competence. In cases where the student did not write more than a few words or wrote in French (the language of the university), the student was categorized as A0 for the performance.

In the pinpointing stage, the panel focused on a more restrictive range of scores, focusing on borderline cut-off scores. The scores for placement in the first full level of the ANG program (between scores 37 and 42, intermediate I – B1 level), and the last full level of the ANG program (between scores 67 and 72, advanced – C1 level) was of particular interest. The cut-off score for the level often required for language proficiency attestation (between scores 52 and 57, intermediate II – B2 level) was also important.

In the pinpointing stage, judgments were focused on a much narrower range of scores. The CoE (2009) Manual suggests logistical regression to establish cut-off points (Table 8 for the rangefinding stage – writing and Table 9 for the pinpointing stage – writing).

Using the logistic regression coefficients and cut-off scores from the pinpointing rounds, it was now possible to map the CEFR score range results by level onto the ANG score ranges for writing (Fig. 4).

The results of this mapping suggest that using the ANG writing rubrics for standard setting leads to different cut-off points than when using the CEFR level descriptors, but that there is also quite a bit of overlap.

6.2 Productive Skills – Speaking

The following tables indicate frequency judgments at both the rangefinding and pinpointing stages for the speaking production section of the test (Table 10).

As we saw with the writing, rangefinding indicates when judgments about a CEFR are split, such as at the ANG score 47, and when CEFR judgments are more unanimous, such as at the ANG score 52. Recall that the use of an A0 category was used for students who did not speak more than a very limited number of words or who spoke in French (Table 11).

Through the pinpointing rounds, cut-off point frequencies over a narrower range of scores are collected. Logical regression analysis establishes what these boundary scores are (Table 12 for the rangefinding stage; Table 13 for the pinpointing stage of the study).

Table 6 Rangefinding judgment frequency – writing

ANG Program Level	ANG Test Score	CEFR "A0"	CEFR A1	CEFR A2	CEFR B1	CEFR B2	CEFR C1	CEFR C2	Total judgments
Beginner (A2)	32	32	52	1					85
	37	**9**	**66**	**15**					90
Intermediate I (B1)	42		**33**	**50**	**7**				90
	47		7	58	24	1			90
	52			**11**	**74**	**5**			90
Intermediate II (B2)	57			**1**	**45**	**41**	**3**		90
	62				7	72	11		90
	67				**1**	**44**	**28**	**7**	80
Advanced (C1)	72					**11**	**49**	**15**	75
	77					1	14	20	35
	82					**4**	**21**	**10**	35
Fluent (C2)	87						**5**	**5**	10

Table 7 Pinpointing judgment frequency – writing

ANG Program Level	ANG test Score	CEFR A1	CEFR A2	CEFR - B1	CEFR B2	CEFR C1	Total judgments
Beginner (A2)	37	**21**	**8**	**1**			30
Intermediate I (B1)	42	**5**	**24**	**1**			30
	47		17	13			30
	52		**3**	**24**	**3**		30
Intermediate II (B2)	57			**15**	**15**		30
	62				25	5	30
	67				**15**	**15**	30
Advanced (C1)	72					**30**	30

Table 8 Logistic regression coefficients and cut-Offs scores, rangefinding – writing

a	b	Cut-off score
A0/A1-10.61	0.35	30.70
A1/A2-15.66	0.38	41.02
A2/B1-20.79	0.43	48.24
B1/B2-26.30	0.46	57.23
B2/C1-17.14	0.25	68.44
C1/C2-7.46	0.09	83.67

Using the logistic regression coefficients and cut-off scores from the pinpointing rounds, it was now possible to map the CEFR score range results by level onto the ANG score ranges for speaking (Fig. 5).

The results for the mapping of the speaking section of the test suggests that using the ANG speaking rubrics for standard setting leads to different cut-off points than when using the CEFR level descriptors, and that the overlap is similar to what we found in writing.

Table 9 Logistic regression coefficients and cut-offs scores, pinpointing – writing

a	b	Cut-off score
A0/A1-	–	–
A1/A2-20.85	0.54	38.89
A2/B1-20.81	0.44	47.50
B1/B2-32.81	0.58	56.46
B2/C1-33.57	0.51	66.21

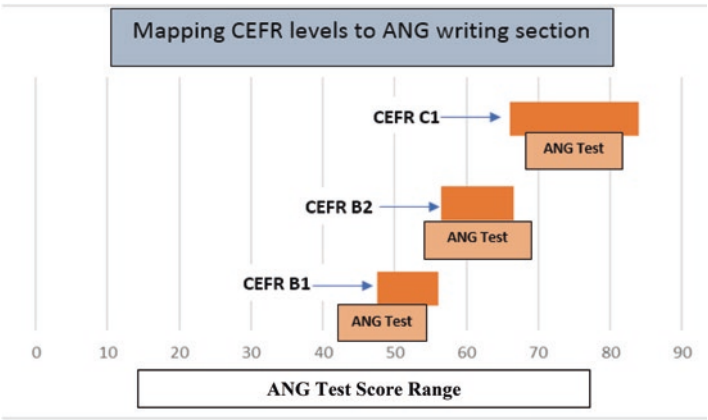


Fig. 4 Mapped cut-off scores and score ranges – writing

Table 10 Rangefinding judgment frequency – speaking

ANG Program Level	ANG Test Score	CEFR "A0"	CEF R A1	CEFR A2	CEF R B1	CEFR B2	CEFR C1	CEFR C2	Total judgments
Beginner (A2)	27	29	6						35
	32	26	22	6	1				55
	37	**15**	**60**	**13**	**2**				90
Intermediate I (B1)	42	**2**	**48**	**34**	**6**				90
	47		2	43	43	2			90
	52			**10**	**69**	**11**			90
Intermediate II (B2)	57			**1**	**27**	**56**	**6**		90
	62				5	48	34	3	90
	67					**32**	**55**	**3**	90
Advanced (C1)	72					**8**	**57**	**25**	90
	77					1	46	28	75
	82						**10**	**20**	30
Fluent (C2)	87						**8**	**32**	40
	92						1	14	15
	97							5	5

Table 11 Pinpointing judgment frequency – speaking

ANG Program Level	ANG test Score	CEFR A1	CEFR A2	CEF R B1	CEF R B2	CEFR C1	Total Judgments
Beginner (A2)	37	**10**	**20**				30
Intermediate I (B1)	42		**12**	**13**	**5**		30
	47		5	15	10		30
	52			**10**	**20**		30
Intermediate II (B2)	57				**11**	**19**	30
	62				30		30
	67				**15**	**15**	30
Advanced (C1)	72					**30**	30

Table 12 Logistic regression coefficients and cut-offs scores, rangefinding – speaking

a	b	cut-off score
A0/A1-10.05	0.31	32.41
A1/A2-12.18	0.29	41.68
A2/B1-14.10	0.30	46.55
B1/B2-23.65	0.43	55.69
B2/C1-17.20	0.28	64.33
C1/C2-13.86	0.18	78.65

Table 13 Logistic regression coefficients and cut-offs scores, pinpointing – speaking

a	b	cut-off score
A0/A1-145.16	3.94	36.82
A1/A2-18.22	0.42	42.97
A2/B1-10.70	0.22	48.00
B1/B2-228.43	4.02	56.87
B2/C1-274.81	4.10	66.99

6.3 Comprehension Skills – Reading

The results of the mapping to the reading construct were reported differently due to the nature of the methodology we adopted. We did not have judgment frequencies, but rather a comparison between composite scores derived from MAP judgments on the four sections of the test that comprise the reading construct (reading, critical reading, vocabulary, and writing) and the composite scores of students those who were minimally placed at each level of the reading courses in the ESL programs. The results are in Table 14.

Cut-off scores by level are similar at the Beginner (or A2) and Intermediate (or B1) levels. However, there is quite a discrepancy between the MAP cut-off scores

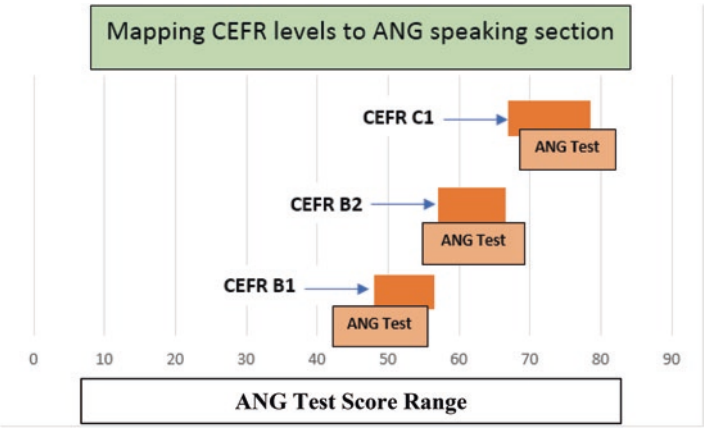


Fig. 5 Mapped cut-off scores and score ranges – speaking

Table 14 Minimally acceptable person CEFR cut-off scores and ANG test cut-off scores – reading

Reading construct (reading – 50%; vocabulary – 20%; writing – 15%; critical reading – 15%)			
Yes-No MAP Level	Reading construct composite score	ANG course level	ANG confidence interval bounds
MAP – A2	24.17	ANG TEST – Beginner (A2)	26.30 25.31
MAP – B1	43.13	ANG TEST – Intermediate I (B1)	41.00 40.11
MAP – B2	70.47	ANG TEST – Intermediate II (B2)	54.23 53.70
MAP – C1	87.52	ANG TEST – Advanced (C1)	67.95 67.95

and the cut-off scores on the ANG test at the Intermediate II (B2) and Advanced (C1) levels. This suggests that the panel felt that the MAP student would begin getting most questions correct already at the B2 level, and almost all the questions correct by the advanced level. This was not observed in the actual scores of students minimally placed in reading courses (Fig. 6).

6.4 Comprehension Skills – Listening

The Tucker Angoff method (Angoff, 1971) was also used in mapping the listening section of the test, determined by a calculation of the listening, vocabulary, speaking, and speech perception sections of the ANG test. As speech perception does not

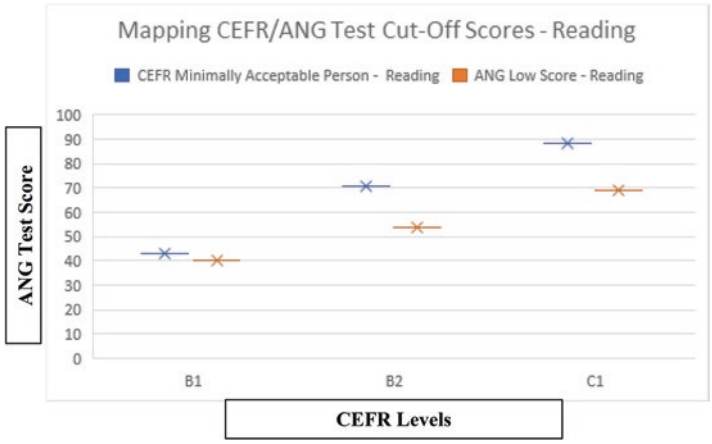


Fig. 6 Mapping of the MAP CEFR cut-off scores and ANG test low scores for the reading construct sections (composite scores) on the ANG test

Table 15 Minimally acceptable person CEFR cut-off scores and ANG test cut-off scores – listening

Listening construct (listening – 40%; vocabulary – 20%; speaking – 25%; speech perception – 15%)			
Yes-No MAP Level	Listening construct composite score	ANG course level	ANG confidence interval bounds
MAP – A2	25.28	ANG TEST – Beginner (A2)	31.86 30.10
MAP – B1	40.2	ANG TEST – Intermediate I (B1)	43.27 41.75
MAP – B2	50.76	ANG TEST – Intermediate II (B2)	54.89 53.59
MAP – C1	60.41	ANG TEST – Advanced (C1)	64.99 64.14

have a clear CEFR descriptor, we scaled the composite scores to exclude this calculation, both for the MAP CEFR scores and the ANG test confidence intervals. These results are in Table 15.

Cut-off scores are within five points for each level. As the MAP CEFR score is consistently lower than the cut-off scores observed by actual test takers who were just barely placed in the ESL program's listening courses, this suggests that the panel felt the MAP student would not have answered as many questions correctly as the actual students taking the ANG test (Fig. 7).

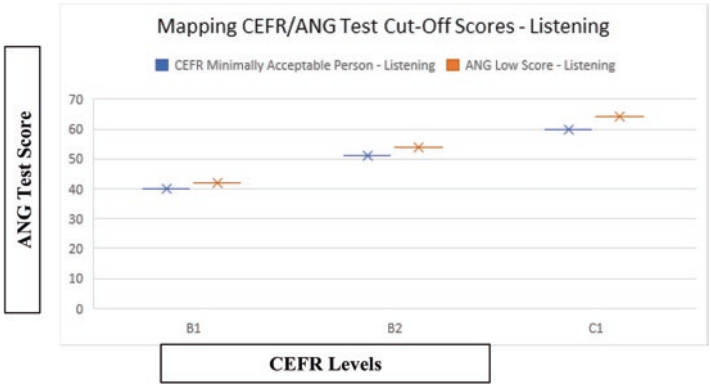


Fig. 7 Mapping of the MAP CEFR cut-off scores and ANG test low scores for the listening construct sections (composite scores) on the ANG test

7 Insights Gained

We began this exploratory study with an urgency to validate our provisional claim that the links we found between the CEFR benchmarks and our course content and objectives, particularly for communicative competencies, also shared similar links to the ANG test. The results of this mapping thus far are just one part of an ongoing validation process. We now have a better idea of how the ANG test maps to the CEFR,[11] but through the mapping process, a number of questions meriting further study emerged in order to strengthen any future claims. While we have more questions than the scope of this chapter permits us to elaborate, the following lines of inquiry were recurrent in our panel discussions:

- The implication of having cut-off scores that vary from a few points to over 15. While the Council of Europe (2009) suggests mapping studies find a test's relationship to the CEFR, we must still establish criteria to determine what constitutes a sufficient, or valid relationship for our context and score reporting needs;
- The impact of prompt type on written and spoken production to elicit an optimal response for a learner's level, particularly as the updated CEFR Manual (2020) suggests that lower-level learners *can* discuss more concrete topics, such as personal experiences, while higher level learners *can* "state a case" or be persuasive;

[11] It goes without saying that if we had adopted different methodologies, we might have different results. For example, we could have used the Extended Tucker-Angoff method instead of the Minimally Acceptable Person for standard setting. ROC statistical analysis may also yield more valid results with the Body of Work method (Kingston et al., 2001).

- Investigate what, if any, changes should be made to writing and speaking rubrics to take into account the learner's motivation for taking the ANG test (placement or proficiency);
- Now that we have mapped the ANG test to the CEFR using CEFR descriptors that were minimally modified, would further modifications to descriptors strengthen correspondence claims between the CEFR and the ANG test? Panelists were particularly concerned with the underspecified nature of the C2 level, which repeats many of the *can do* statements from the C1 level.

That we have so many questions following the study suggests that validating links is important, but that the real goal of the exercise is to gain a more heightened awareness about how the local test works and the complexity of making any valid CEFR claims between a local test and CEFR. To that end, the entire CEFR mapping process actually brought us on a more important journey towards understanding language testing issues in general and specific challenges for local tests. It was a crash course in language assessment literacy and the CEFR rolled into one, both of which have benefits for future ANG test modifications and for all stakeholders connected to the ANG test.

7.1 Benefits for Test Revision and Quality Control

This mapping study provided a pretext to find the much-needed resources to complete a thorough review of the communicative sections of our local ANG test. ANG test specifications needed to be documented and test content, in particular for item bias, confusing distractors, and areas of underspecification in marking rubrics were due for periodic review (CoE, 2011). Rather than being a one-off activity, we sought to investigate the pertinence of integrating the CEFR into current and future procedures for revising the ANG test. Indeed, through this study, we were able to collect data for future investigation on certain speaking and writing prompts as well as test items on the multiple-choice sections of the test. In addition to revising some test items immediately, we also put together a follow-up project to revise the current writing, speaking, and pronunciation rubrics. These rubrics and their integration into the test will inform future decisions regarding the use of the ANG test for placement or proficiency, especially in light of mapping discussions among the panelists that suggest ANG evaluators may benefit by using rubrics differently for placement and for proficiency scoring. This study provided some important data on our way to paving a path forward for future iterations of the ANG test.

7.2 Benefits for Stakeholders

While most of the evaluator panel was familiar with the CEFR, most had never attempted to apply the level descriptors in a meaningful way. The mapping study encouraged a dialog about the challenges of using the CEFR as well as language testing in general. As transparent as the CEFR claims to be, descriptors are often underspecified, leaving the evaluator panel to deduce the CEFR's intended meaning. While the panel understood that some modifications were permitted, many concerns were raised about maintaining the integrity and authenticity of the mapping with modified descriptors. Evaluators are also used to describing a performance in terms of salient features in the production. With the CEFR, the task became to consider inferences that can be made from the performance about how the student would do in an authentic, real-world context.

Beyond the rubrics, the mapping study encourages reflections about the implications of using the CEFR as performance level descriptors (PLDs). In the context of Canada, other PLDs are available, notably the Canadian Language Benchmarks/ *les Niveaux de compétence linguistique canadiens* (CLB/NCLC). Bournot-Trites et al. (2020) argue that if context is such an important feature of the CEFR, educators and evaluators should be more interested in using PDLs specifically developed for Canadian students. They warn that by continuing to use PLDs related to CEFR, language educators may be complicit in "the Europeanization of language in Canada (which) brings a risk of losing language diversity and cultural specificity" (p. 157). While the CLB and CEFR have been linked (North & Piccardo, 2018), the promotion of the CEFR at a moment when the CLB was still undergoing certain validations meant that the CEFR enjoyed an early adopter advantage over other PLDs. However, Bournot-Trites et al. (2020) make a compelling case for the revision of these systemic practices. This debate merits ongoing reflection.

Another unexpected benefit of this study for the language evaluators and instructors was the overall gains in language assessment literacy (LAL) among all members of the panel. Fulcher (2012, 2020) advocates for a comprehensive version of this type of literacy, defined as follows:

> 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 of testing on society, institutions and individuals (2012, p. 125).

Indeed, embarking on a CEFR mapping study requires reflection on not only situating standard-setting in argument-based validity practices (Papageorgiou & Tannenbaum, 2016) but also understanding the underpinnings of the CEFR and its use in a multitude of educational and decisional contexts. While some of these concepts were familiar to some panel members, it was only through applying theory to the mapping project did the members have a concrete reason to add procedural and socio-political depth to this knowledge.

8 Implications for Test Developers and Users

The CoE (2009) states at the outset of the linking process that, "Relating an examination or test to the CEFR is a complex endeavor" (p. 7). For local contexts in which resources may be limited, the complexities are even more acute. After 4 years, the validation of our CEFR testing claims is still ongoing. Yet we have come to appreciate the overall relating process not for the valid claims we will one day be able to make, but for the relating process itself. As this modest experimental study suggests, embarking on a mapping study kickstarts important dialogues about the CEFR and standard setting in general between language practitioners and language assessors. It is through applying the CEFR that even more stakeholders can add to a growing body of research related to the pertinence of adopting the CEFR in non-European contexts (Read, 2019; Normand-Marconnet & Lo Bianco, 2015). This is perhaps the most tangible benefit of what mapping a local test to the CEFR can do as local actors become highly sensitive to the tension between the advantages of comparability (with the CEFR as a common language) and the potential cost to local agency. We have found that relating a test to the CEFR has actually enhanced our agency. Despite the challenges, we hope our experience sets realistic expectations about the mapping process while still encouraging other local testing stakeholders to view their tests through a CEFR lens in order to add their voices to ongoing debates.

References

Alderson, J. C., Figueras, N., Kuijper, H., Nold, G., Takala, S., & Tardieu, C. (2006). Analysing tests of reading and listening in relation to the common European framework of reference: The experience of the Dutch CEFR construct project. *Language Assessment Quarterly, 3*(1), 3–30. https://doi.org/10.1207/s15434311laq0301_2

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508–600). American Council on Education.

Bournot-Trites, M. G., Friesen, L., Ruest, C., & Zumbo, B. D. (2020). A made-in-Canada second language framework for K-12 education: Another case where no prophet is accepted in their own land. *Canadian Journal of Applied Linguistics, 23*(2), 141–167. https://journals.lib.unb.ca/index.php/CJAL/article/view/30434

Bruemmer, R. (2018, June 16). *Anglophones and francophones have distorted views of each other: Survey*. Montreal Gazette.. https://montrealgazette.com/news/local-news/anglophones-and-francophones-have-distorted-views-of-each-other-survey

Coombe, C., Vafadar, H., & Mohebbi, H. (2020). Language assessment literacy: What do we need to learn, unlearn, and relearn? *Lang Test Asia, 10*(3), 1–16. https://doi.org/10.1186/s40468-020-00101-6

Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge University Press. https://rm.coe.int/1680459f97

Council of Europe. (2009). *Relating language examinations to the common European framework of reference for languages: Learning, teaching, assessment (CEFR).* Language Policy Division. https://rm.coe.int/1680667a2d

Council of Europe. (2011). *Manual for language test development and examining*. Language Policy Division. https://rm.coe.int/manual-for-language-test-development-and-examining-for-use-with-the-ce/1680667a2b

Council of Europe. (2018). *Common European framework of reference for languages: Learning, teaching, assessment.* Companion volume with new descriptors. Council of Europe. https://rm.coe.int/cefr-companion-volume-with-new-descriptors-2018/1680787989

Council of Europe. (2020). *Common European framework of reference for languages: Learning, teaching, assessment.* Companion volume. Council of Europe. https://rm.coe.int/common-european-framework-of-reference-for-languages-learning-teaching/16809ea0d4

Dendrinos, B., & Gotsouilia, V. (2015). Setting standards for multilingual frameworks in foreign language education. In B. Spolsky, O. Inbar, & M. Tannenbaum (Eds.), *Challenges for language education and policy. Making space for people* (pp. 23–50). Routledge

Díez-Belmar, M. B. (2018). Fine-tuning descriptors for CEFR B1 level: Insights from learner corpora. *ELT Journal, 17*(2), 199–209

Deygers, B., Zeidler, B., Vilcu, D., & Carlsen, C. H. (2018). One framework to unite them all? Use of the CEFR in European university entrance policies. *Language Assessment Quarterly, 15*(1), 3–15. https://doi.org/10.1080/15434303.2016.1261350

Fulcher, G. (2012). Assessment literacy for the language classroom. *Language Assessment Quarterly, 9*(2), 113–132. https://doi.org/10.1080/15434303.2011.642041

Fulcher, G. (2020). Operationalizing language assessment literacy. In D. Tsagari (Ed.), *Language assessment literacy: From theory to practice* (pp. 8–28). Cambridge Scholars Publishing.

Harsch, C., & Hartig, J. (2015). What are we aligning tests to when we report test alignment to the CEFR? *Language Assessment Quarterly, 12*(4), 333–362. https://doi.org/10.1080/15434303.2015.1092545

Kane, M. (1998). Choosing between examinee-centered and test-centered standard-setting methods. *Educational Assessment, 5*(3), 129–145. https://doi.org/10.1207/s15326977ea0503_1

Kingston, N. M., Kahl, S. R., Sweeny, K. P., & Bay, L. (2001). Setting performance standards using the body of work method. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods and perspectives* (pp. 219–248). Lawrence Erlbaum.

Lim, G. S., Geranpayeh, A., Khalifa, H., & Buckendahl, C. W. (2013). Standard setting to an international reference framework: Implications for theory and practice. *International Journal of Testing, 13*(1), 32–49. https://doi.org/10.1080/15305058.2012.678526

Martyniuk, W. (Ed.). (2010). *Aligning tests with the CEFR: Reflections on using the Council of Europe's draft manual*. Cambridge University Press.

Milanovic, M., & Weir, C. J. (2010). Series editors' note. In W. Martyniuk (Ed.), *Aligning tests with the CEFR: Reflections on using the Council of Europe's draft manual* (pp. viii–xx). Cambridge University Press.

Nomand-Marconnet, N., & Lo Bianco, J. (2015). The common European framework of reference down under: A survey of its use and non-use in Australian universities. *Language Learning in Higher Education, 5*(2), 281–307. https://doi.org/10.1515/cercles-2015-0014

North, B. (2014). Putting the common European framework of reference to good use. *Language Teaching, 47*(2), 228–249. https://doi.org/10.1017/s0261444811000206

North, B. (2020, June). Trolls, unicorns and the CEFR: Precision and professionalism in criticism of the CEFR. *CEFR Journal – Research and Practice, 2*, 8–24. https://cefrjapan.net/images/PDF/Newsletter/CEFRJournal-2-1_BNorth.pdf

North, B., & Piccardo, E. (2018). *Aligning the Canadian language benchmarks (CLB) to the common european framework of reference (CEFR)*. https://www.language.ca/wp-content/uploads/2019/01/Aligning-the-CLB-and-CEFR.pdf

North, B., & Jaroscz, E. (2013). Implementing the CEFR in teacher-based assessment: Approaches and challenges. In E. Galaczi & C. J. Weir (Eds.), *Exploring language frameworks: Proceedings of the ALTE Krakow Conference, July 2011* (Studies in Language Testing Series 36, pp. 118–134). Cambridge University Press.

O'Sullivan, B. (2010). The City & Guilds Communicator examination linking project: A brief overview with reflections on the process. In W. Martyniuk (Ed.), *Aligning tests with the CEFR: Reflections on using the Council of Europe's draft manual* (pp. 8–24). Cambridge University Press.

Papageorgiou, S., & Tannenbaum, R. J. (2016). Situating standard setting within argument-based validity. *Language Assessment Quarterly, 13*(2), 109–123. https://doi.org/10.1080/15434303.2016.1149857

Read, J. (2019). The influence of the common European framework of reference (CEFR) in the Asia-Pacific region. *LEARN Journal: Language Education and Acquisition Research Network Journal, 12*(1), 12–18. http://files.eric.ed.gov/fulltext/EJ1225686.pdf

Savski, K. (2021). CEFR as language policy: Opportunities and challenges for local agency in a global era. *The English Teacher, 50*(2), 60–70. https://doi.org/10.52696/aide2513

Shackleton, C. (2018). Developing CEFR-related language proficiency tests: A focus on the role of piloting. *CercleS 8*(2), 333–352. https://doi.org/10.1515/cercles-2018-0019

Szabo, T., & Goodier, T. (2018). Collated representative samples of descriptors of language competence developed for young learners. *Resource for educators*. https://www.coe.int/en/web/common-european-framework-reference-languages/bank-of-supplementary-descriptors

Tannenbaum, R. J., & Baron, P. A. (2011). *Mapping TOEFL ITP scores onto the common European framework of reference*. Educational Testing Service. https://www.ets.org/s/toefl_itp/pdf/mapping_toefl_itp_scores_onto_the_common_european_framework_of_reference.pdf

Tannenbaum, R. J., & Wylie, E. C. (2004). *Mapping test scores onto the common European framework: Setting standards of language proficiency on the test of English as a foreign language (TOEFL), the test of spoken English (TSE), the test of written English (TWE), and the test of English for international communication (TOEIC)*. Educational Testing Service. http://www.ets.org/Media/Tests/TOEFL/pdf/CEFstudyreport.pdf

Tannenbaum, R. J., & Wylie, E. C. (2008). Linking English-language test scores onto the common European framework of reference: An application of standard-setting methodology. *ETS Research Report Series, 2008*(1), i–75. https://doi.org/10.1002/j.2333-8504.2008.tb02120.x

Taylor, L. (2004). Issues of test comparability. *Research Notes, 15*, 2–5. https://www.cambridgeenglish.org/Images/23131-research-notes-15.pdf

Tracy, R. (2017). Language testing in the context of migration. In J.-C. Beacco, H.-J. Krumm, D. Little, & P. Thalgott (Eds.), *The linguistic integration of adult migrants/L'intégration linguistique des migrants adulte : Some lessons from research/les enseignements de la recherche* (pp. 45–56). De Gruyter. https://doi.org/10.1515/9783110477498-007

Venne, J.-F. (2019, October 5). *UQAM: 50 ans de démocratisation de l'éducation universitaire*. Le Devoir. https://www.ledevoir.com/societe/education/563910/uqamuqam-50-ans-de-democratisation-de-l-education-universitaire

Designing a New Diagnostic Reading Assessment for a Local Post-admission Assessment Program: A Needs-Driven Approach

Xiaohua Liu and John Read

Abstract The Diagnostic English Language Needs Assessment (DELNA) is a post-entry language assessment program designed to identify students' academic language needs at the University of Auckland. Compared with results yielded from its writing section, the single reading band score was not fulfilling its diagnostic purpose. Therefore, an initiative was undertaken to re-design the current DELNA reading assessment so that more fine-grained information could be reported to students and language advisors. We developed the new assessment based on both theoretical and empirical evidence. First, the relevant literature is reviewed, including a discussion of the principles for designing diagnostic language assessments. Following this we present how we established the construct model of the new assessment, taking into account the results of a needs analysis. We then turn to operationalization, and discussion of the number and length of input texts, the types of tasks, and the statistical models used to synthesize information across tasks, and proposed specifications for the new reading assessment. The proposed revision consists of three separately timed modules to measure academic language knowledge, careful reading and expeditious reading. This study demonstrates that it is both important and feasible to design a local diagnostic language assessment based on local needs.

Keywords Local language testing · Reading assessment · Reading instruction · Diagnostic assessment · Post-admission assessment

X. Liu (✉)
The Chinese University of Hong Kong, Shenzhen, China
e-mail: liuxiaohua@cuhk.edu.cn

J. Read
The University of Auckland, Auckland, New Zealand
e-mail: ja.read@auckland.ac.nz

X. Yan et al. (eds.), *Local Language Testing*, Educational Linguistics 61,
https://doi.org/10.1007/978-3-031-33541-9_5

1 Introduction: Test Purpose and Testing Context

At the University of Auckland, the need for a local language test arose in the 1990s from a recognition of the growing linguistic diversity of the student population, resulting from a liberalization of New Zealand's (N.Z.) immigration policy. This meant that large numbers of young people from non-traditional source countries, particularly in East Asia, were entering the university, and faculty members expressed increasing concern about their perceived lack of language proficiency to cope with English-medium academic study. International students had to achieve a minimum score on IELTS (the preferred measure of English proficiency at the time) as a condition of admission but, according to NZ education law, students from similar linguistic and educational backgrounds who had obtained permanent residence could not be required to take a language test if they matriculated through one of the pathways available to NZ citizens. At the same time, there were concerns about the academic literacy of other groups of students who had come through the national education system, notably ethnic minority students being recruited on equity grounds and mature students with no recent experience of formal study (for a fuller account, see Read, 2015, Chap. 2).

In order to cope with this complex situation that could not be fully addressed by international standardized tests such as IELTS, the University introduced in the early 2000s a program of post-admission language assessment called the Diagnostic English Language Needs Assessment (DELNA) (see Fig. 1). In principle, DELNA encompasses all incoming students including doctoral candidates, regardless of their language background, but our focus here is on first-year undergraduates. The initial screening phase is essentially designed as an efficient means to exempt those who are unlikely to face language-related difficulties in their studies. Students who fall below the screening cut score in this assessment are asked to collect their results in person from a language advisor, who will then review with them their assessment profiles and give them specific advice on the resources that could be utilized to improve their language skills. Thus, the assessment is not simply a placement test, in the sense of directing students into one or more of a suite of English language courses. There are in fact a variety of options available on campus for students to enhance their language and literacy skills, depending on the policy of their faculty and their individual initiative to take advantage of the opportunities available (Fig. 1).

2 Testing Problems Encountered

Since advising has a central role in the program, it is desirable for the language advisors to have access to good-quality diagnostic information. From this perspective, there is an obvious mismatch among the three sections of DELNA. The writing section generates fairly rich information: it is scored analytically by two

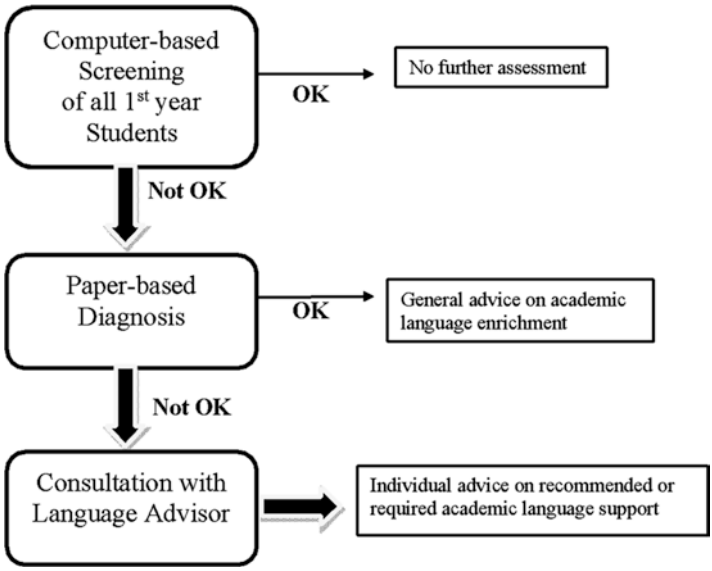


Fig. 1 The DELNA assessing and advising process at the University of Auckland

raters according to three main criteria, each of which includes three sub-scales, and it even yields counts of different types of error. In addition, the advisor has the writing script available to review with the student in the advisory session. By contrast, the listening and reading sections were not designed to generate this kind of diagnostic information, and so the reported test result is simply a six-level band score based on the number of correctly answered items. Consequently, both test takers and post-assessment language advisors have found the coarse-grained nature of the score reporting for the two sections rather unsatisfactory and would welcome more detailed feedback on each student's performance. Therefore, there is a practical need to transform the two sections into more diagnostic instruments.

The current reading and listening sections of DELNA were originally developed at the University of Melbourne for that university's Diagnostic English Language Assessment (DELA) (Elder & Read, 2015) in the early 1990s. Thus, they were influenced at the time by the design of communicative proficiency tests such as IELTS. There was in fact an initiative by Lumley (1993) to derive diagnostic information from the DELA Reading test, but it proved to be too time-consuming and impractical for operational use. Although, over the years, new forms have been developed and modifications made to the DELA/DELNA reading and listening tests, the test specifications still state the target skills only in broad terms. It remains unclear if and to what extent the skills assessed in these two sections are in line with the real-life academic needs of the students studying at this university.

These, then, constitute the motivations for a project that would lay the groundwork for a re-designed assessment of academic reading skills in order to provide informative diagnostic results to both advisors and students.[1]

3 Literature Review

In the following section we will address diagnostic language assessments, principles for defining the constructs for a diagnostic language assessment, and academic reading.

3.1 Diagnostic Language Assessments

Recent years have witnessed a surge of interest in diagnostic assessment in the field of language assessment, due to increasing calls for more detailed assessment results that can offer insights into individual students' strengths and weaknesses (Lee, 2015). However, existing assessments developed with diagnosis in mind have been scarce, especially for those of receptive language skills (i.e., reading and listening). This is probably because such skills, with their implicit nature, are difficult to elicit, observe and analyze in any direct way. Recent advances in psychometrics such as the application of cognitive diagnostic models (CDM) have shed promising light on how to analyze reading and listening test tasks in order to produce fine-grained assessment results (Leighton & Gierl, 2007; Rupp et al., 2010), an approach that is now commonly known as cognitive diagnostic assessment (CDA). In the absence of true diagnostic instruments, such techniques have been increasingly applied to existing proficiency or placement tests to generate diagnostic information, through *post hoc* analyses of task content and statistical modeling of test data (Chen & Chen, 2016; Jang, 2009; Lee & Sawaki, 2009; Li & Suen, 2013; among many others).

Nevertheless, this retrospective approach has been criticized for its problems, of which one is the usefulness of the skills identified and reported (Alderson, 2010; Alderson et al., 2015). After collating and summarizing the skills identified in past CDA studies, Liu (2018) found that some of them were more like test-taking strategies (such as "evaluating the plausibility of distractors" and "lexical matching between options and texts") rather than real-world reading skills. Reporting such skills would be of little help to assessment users, particularly teachers and students who expect meaningful information to guide their subsequent teaching and learning.

[1] A separate project by another researcher to redesign the DELNA listening section is currently underway.

In contrast to the reverse-engineering approach, one of the few assessments that was designed from the outset to be diagnostic is DIALANG, the online diagnostic system sponsored by the Council of Europe that measures knowledge and use of 14 European languages on a self-access basis (Alderson, 2005). As a first step, test takers are encouraged to complete language knowledge tasks in their chosen target language that have a placement function. The results determine the difficulty level of subsequent reading, listening, and writing assessments, each of which measures explicitly several sub-skills. All assessment results are automatically generated and reported by sub-skill to the test-taker, accompanied by some brief suggestions for future learning.

A significant issue associated with the retrofitted CDAs and even DIALANG is that their construct models were derived mainly from general theories, frameworks, or standards of language abilities, which may not adequately represent the learning needs of specific groups of students. This is particularly relevant for a local assessment like DELNA. It is evident that a significant mismatch between a theoretically-based construct and students' actual needs are likely to undermine the meaningfulness, generalizability, and sufficiency of the diagnostic information thus produced due to construct irrelevance and underrepresentation (Bachman & Palmer, 2010). To avoid such risks and to produce useful diagnostic results, we need systematicity in assessment development.

3.2 Principles for Defining the Constructs for a Diagnostic Language Assessment

While a diagnostic assessment is apparently distinct from placement and proficiency assessments in aspects such as score reporting, they share many similarities, especially in the way that they are developed. Therefore, there are two main procedures that interact in designing a diagnostic assessment: construct definition and operationalization. Due to space limitation, this section focuses only on the major principles for construct definition. Issues associated with operationalization will be discussed in detail based on the DELNA context in the Findings section. With regard to construct definition, the defining feature of diagnostic assessment is that the knowledge, skills, and abilities associated with the target language use domain can be decomposed into subcomponents, forming multiple constructs. In the study of reading, there is a long tradition of compiling taxonomies of reading skills, which means that this skill area lends itself particularly well to diagnostic assessment. From a validity perspective, the multiple constructs should represent the typical sub-areas of language use in the target domain (academic reading in this case) so that test performance can be used to predict performance in those sub-areas in real-life study situations. In addition, in order to ensure that the assessment results are effectively utilized, the constructs should be defined at a level of specificity that is appropriate for the needs of the assessment users, in this case the DELNA language advisors and the students.

To enhance the generalizability of test results, there are two possible sources of a construct definition: one is a well-established theory that analyzes the linguistic and cognitive demands of language use in the target domain; the other is an empirical analysis of the needs of language users in the domain (Bachman & Palmer, 2010). Although they are not mutually exclusive, empirical needs analysis is likely to play a more important role in defining the relevant constructs for a diagnostic assessment that produces meaningful results for the users. Theoretical accounts may describe the subcomponents of the target domain at a fine grain size that makes little sense to assessment users. For example, a theory of reading may include the orthographic and phonological processing of words (such as the Lexical Quality Hypothesis by Perfetti, 2007), which is too detailed and technical to be useful for students. However, both approaches to defining the construct model will come up against practical constraints in operationalizing the number of constructs that have been identified. A larger number of constructs means that more tasks are needed to assess them, with the result that the diagnostic test may become unreasonably long. In practice, then, it is necessary to give priority to more central and useful constructs and to assess constructs at a relatively coarser-grained level, rather than in fine detail.

3.3 Academic Reading

A great body of research has been conducted on the cognitive processes of reading comprehension for many years. Findings have generally shown that fluent reading requires the efficient orchestration of multiple skills, at the lower levels of word recognition, sentence parsing, and meaning encoding, as well as those involved in comprehension at the higher-order levels (Grabe & Stoller, 2013). A variety of reading models have also been built, focusing on particular subprocesses or the entire process of reading (McNamara & Magliano, 2009). For instance, the widely-known Construction-Integration Theory distinguishes between two types of comprehension, textbase construction, and situation model building, which respectively refer to the comprehension of literal information (such as explicitly stated details, main points, and relations) and implied information (i.e., information that is inferred from both textual information and world knowledge) (Kintsch, 2013). In addition, discourse analysts have identified another two types of mental representation of text commonly constructed during reading: the author's pragmatic communication (such as their goal, attitude, belief, and intended audience) and rhetorical information (such as genre type and organizational patterns) (Graesser & Forsyth, 2013).

However, most of the research and models paid attention only to comprehension of single texts through careful reading. Academic reading, especially at the tertiary level, often requires skills that go beyond this kind of comprehension. For instance, reading scholars are now paying growing attention to skills such as synthesizing information from multiple documents, which often necessitates the accommodation and integration of sources that may differ significantly in modality, readability, perspective, and other features (Richter & Maier, 2017; Rouet et al., 2019). Such added

complexity in information processing may pose significant challenges to university students. Another closely related skill is source evaluation (Macedo-Rouet et al., 2019). Leu et al. (2011) contend that this is especially pertinent to sourcing through the internet, which can return a massive amount of information of varied quality. Therefore, being able to critically evaluate the value, reliability, relevance, and currency of information found online will largely determine the quality of learning through reading (Goldman et al., 2012; Macedo-Rouet et al., 2019).

Strategic reading is another important feature of academic reading (Grabe & Stoller, 2013). This includes the ability to shift efficiently between different reading speeds according to specific reading purposes initiated by reading tasks. For example, Carver (1992) identified five typical reading speeds, scanning, skimming, rauding (or reading at one's average speed), learning, and memorizing. More broadly, Weir and his colleagues differentiated between careful versus expeditious reading, arguing that it is the latter that often poses challenges to learners, yet often gets neglected by research and instruction (Khalifa & Weir, 2009; Weir et al., 2000). Empirical surveys have indeed found that fast reading is a much-needed skill among university students, who often have to cope with the tension between large amounts of reading and a tight study schedule (Hellekjær, 2009; Liu & Read, 2020; Weir et al., 2012). In short, academic reading in higher education often requires the use of skills that exceed the traditionally conceptualized terrain of reading comprehension. This suggests that we need to go beyond mainstream reading theories and models in order to develop an adequate construct definition.

4 Methods

Considering that the constructs to be assessed in the re-designed reading assessment should represent real-life skills, we decided to adopt a principally needs-driven approach to developing the assessment. Therefore, a needs analysis was conducted at the University of Auckland in 2018 to understand students' needs in academic reading. The first stage consisted of individual semi-structured interviews with undergraduates ($n = 22$) as well as language teachers and advisors ($n = 7$). Questions in the respective interview guides focused on students' reading tasks (e.g., reading material and task type), teachers' course or consultation work, students' reading difficulties, both students' and teachers' perceptions of important reading abilities, and their opinions on aspects of a useful reading assessment (e.g., skills that need to be assessed, task format, and feedback).

These interview findings fed into the development of a follow-up student questionnaire, which was administered to a quota sample of 221 undergraduates from different disciplines and program years. In analyzing the questionnaire data, we conducted factor analyses to explore the subdomains of academic reading, as reported by the students. For the factors or subdomains identified in the analyses, we compared their mean values and calculated the effect size of each mean difference to rank order the subdomains in terms of both *need* and *difficulty*. Details of the

interviews and questionnaire survey have been reported elsewhere (Liu, 2018; Liu & Brown, 2019; Liu & Read, 2020). Our final rank-ordering of those subdomains and their component skills was applied to the construct definition of the DELNA reading assessment.

5 Results and Discussion

The findings will be discussed first in terms of the underlying construct.

5.1 Construct Definition

Based on results of the needs analysis, the construct model for the test, as presented in Table 1, is defined at three levels of specificity, or "granularity," to use the term favored in diagnostic assessment. The model incorporates the component skills as the basic, fine-grained elements, since they will be more meaningful to the users (students and advisors) and thus will facilitate the utilization of assessment results. The skills also provide a better basis for writing test items that assess particular aspects of reading ability. Almost all of the component skills were reported by students in the needs analysis as being needed more than half of the time during their academic studies; this is evidence to support the generalizability of those skills, which relates the test performance to real-life applications (academic reading in this case).

The component skills are grouped under subdomains, based on the results of the factor analyses of the student questionnaire responses. Being less directly measurable, the subdomains represent a higher-level abstraction of the skills, informed also by theoretical accounts of the nature of reading. The subdomains are in turn hierarchically ordered according to the students' judgements about their relative need and difficulty, as classified in the left-hand column of the table.

5.2 Refining the Construct Model

The construct model is obviously elaborate and, from the point of view of practicality, it is unrealistic to expect that all of the components can be included in a single test such as the DELNA reading assessment. Such a comprehensive approach would make the test unacceptably long, as well as putting pressure on administration, scoring, and feedback provision. Thus, on practical grounds a number of skills can be eliminated. First, assessing the skills "summarizing information using one's own words" and "paraphrasing information using one's own words" in any direct way involves a significant amount of writing by the test takers. Although open-ended

summarizing and paraphrasing tasks have long been promoted as more valid and authentic methods of assessing reading comprehension (Cohen, 1993; Riley & Lee, 1996), scoring them is costly and time-consuming, if they are to be assessed reliably (Alderson, 2000; Cohen, 1993). Even though recent advances in natural language processing are making automated scoring of such tasks increasingly feasible (Carr, 2014), restricted access to such tools would still limit their application to an institutional diagnostic assessment such as DELNA. Removing such skills from a construct model of academic reading does not necessarily lead to a substantial construct under-representation as long as: (a) items in other formats can be constructed to target reading comprehension at both the local and global levels, and (b) language production is assessed by other modules in the same diagnostic assessment program (Table 1).

To further simplify the list of skills in the construct model, we may ask ourselves questions such as "Which constructs are less relevant to, or less likely to predict performance in the target domain?" and "Is there any redundancy in the constructs identified?" For example, in the needs analysis the subdomain *understanding pragmatic and rhetorical communication* was regarded as being the least needed one. Besides, it was rated as being only moderately challenging compared with other subdomains. Features such as genre and rhetorical functions belong to the formal aspect of textual knowledge and thus are rarely the target of comprehension, although successfully recognizing them can accelerate comprehension (Grabe & Stoller, 2013). Meanwhile, the skill "understanding author's purpose of writing a text" may seem unnecessary in many university study situations, given that written discourse at the tertiary level tends to be predominantly informational in purpose (Biber et al., 2004). Therefore, this subdomain would be a good candidate for elimination. Nonetheless, the skill "understanding the author's point of view (such as attitudes, beliefs, or opinion) based on a text" may be kept for two reasons. First, this skill is ranked in the middle in terms of need, indicating that it was considered somewhat necessary by students. Second, theoretically this skill is less closely associated with text genre than the previous one, but more related to text content and topic

There are also skills that are potentially redundant. For example, both "understanding the details clearly presented in a text" and "understanding the meaning of complex or long sentences" describe comprehension at the local level – although the latter emphasizes more the use of grammatical knowledge – and thus we may want to assess only one of them. Considering that the latter skill was reported by students to be much more challenging than the former one, we may want to keep the latter and exclude the former in order to focus on diagnosing potential weaknesses rather than strengths (Alderson, 2005; Harding et al., 2015). Regarding expeditious reading, "reading and understanding a text as fast as possible" is more or less a general description of speed reading, and "deciding what information is important and what is not" could be considered as an integral component of information locating or searching. Thereby, they both can be represented by the other two expeditious reading skills: locating specific information and skimming for the general idea.

Table 1 A relative importance scale of the components of academic reading

Importance and indicators		Subdomains and component skills
1	Need: High; difficulty: High	*Academic language knowledge* Understanding the meaning of academic words or jargon. Understanding the meaning of complex or long sentences.
2	Need: High; difficulty: Moderate	/
	Need: Moderate; difficulty: High	*Expeditious reading* Quickly locating useful, important, or needed information in a text. Skimming through a text to have a quick understanding of its general content. Reading and understanding a text as fast as possible. Deciding what information is important and what is not.
3	Need: High; difficulty: Low	*Textbase comprehension* Identifying the main ideas or main points of a text. Understanding the details (such as specific facts, descriptions, or opinions) clearly presented in a text. Understanding the implied meaning of a sentence or sentences (such as figures of speech).
	Need: Moderate; difficulty: Moderate	*Information reconstruction and intertextual model building* Paraphrasing information using one's own words. Summarizing information using one's own words. Integrating (such as comparing and contrasting) information from different texts. Drawing implications or conclusions based on multiple texts. *Global situation model building* Understanding the relationships between ideas or different parts of a text. Understanding the information structure or logical development (such as compare/contrast or cause/effect) of a text. Figuring out the situation (such as environment, event, or relationship) implied in a text.
	Need: Low; difficulty: High	/
4	Need: Moderate; difficulty: Low	/
	Need: Low; difficulty: Moderate	*Understanding pragmatic and rhetorical communication* Understanding the author's point of view (such as attitude, belief, or opinion) based on a text. Understanding the author's purpose of writing a text. Understanding the function (such as introducing, summarizing, or giving examples) of particular parts of a text. Recognizing the genre or type (such as narrative, persuasive, expository, or descriptive) of a text.
5	Need: Low; difficulty: Low	/

Note: 1 = most important, and 5 = least important

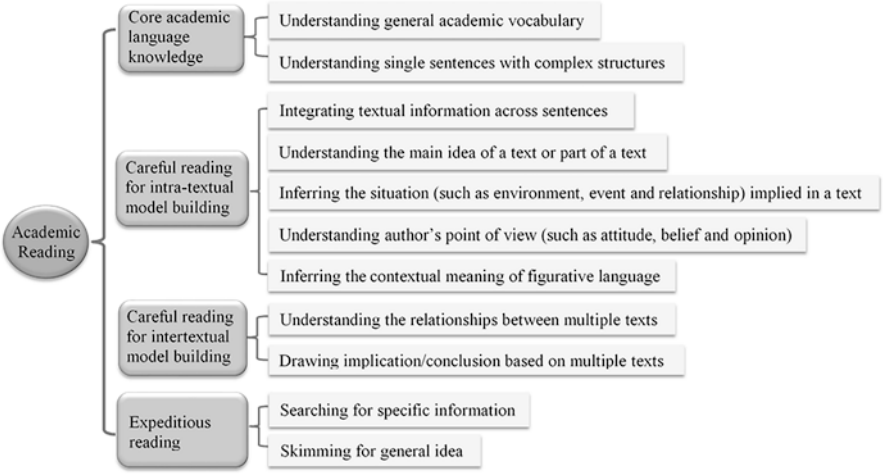


Fig. 2 A recommended construct model for DELNA reading assessment

After going through this process, a construct model as illustrated by Fig. 2 is recommended for the new reading assessment. Skills in this model are also subsumed under four redefined subdomains based on the needs analysis and other commonalities and differences between them (Fig. 2).[2]

5.3 Operationalization

Once a construct model is established, the next step is to operationalize this model by constructing assessment specifications, which include detailed descriptions of prototypical tasks targeting each construct, scoring methods and rules, statistical models for synthesizing information about test-taker abilities across tasks and for producing assessment results, as well as an assembly model assembling all of these into a complete assessment of reasonable length and task distribution (Mislevy et al., 2003). Table 2 summarizes the key features of the specifications for the redesigned reading assessment based on the abovementioned construct model. This new design has three modules. Due to space limitation, we will discuss selectively the major principles that we followed in operationalizing the construct model (Table 2).

Module I aims at assessing core academic language knowledge through two separate tasks targeting general academic vocabulary and grammar respectively. Both tasks adopt a discrete-point multiple-choice format. The first task (20 items)

[2] For example, the skills from the subdomain *careful reading for intra-textual model building* were originally from three different factors in the needs analysis. Given that they all concern careful reading within single texts, they were grouped within the same subdomain.

Table 2 Recommended specifications for DELNA reading assessment

	Module I: Core academic language knowledge		Module II: Careful reading			Module III: Expeditious reading	
Task	Task 1	Task 2	Task 1	Task 2	Task 3	Task 1	Task 2
Construct	Understanding general academic vocabulary	Understanding single sentences with complex structures	(1) identifying the main idea of a text or part of a text; (2) integrating textual information from across sentences; (3) inferring the situation (such as environment, event and relationship) implied in a text; (4) understanding author's point of view (such as attitude, belief, opinion); (5) inferring the contextual meaning of figurative language; (6) understanding the relationship between two texts; (7) drawing implication/conclusion based on two texts			Searching for specific information	Skimming for general idea
Text number	–	–	2	2	2	6	6
Text length	–	–	About 300 words per text	About 300 words per text	About 300 words per text	About 250 words per text	About 200 words per text
Text genre	–	–	Mainly informational	Mainly informational	Mainly informational	Mainly informational	Mainly informational
Item type	MC	MC	MC, sentence-completion, short-answer, true/false, information-transfer	MC, sentence-completion, short-answer, true/false, information-transfer	MC, sentence-completion, short-answer, true/false, information-transfer	MC	MC
Time constraint	5 mins	10 mins	15 mins	15 mins	15 mins	6 mins (1 min per item)	6 mins (1 min per item)
Interaction medium	Computer-based	Computer-based	Computer-based	Computer-based	Computer-based	Computer-based	Computer-based
Scoring	Dichotomous	Dichotomous	Dichotomous	Dichotomous	Dichotomous	Dichotomous	Dichotomous
Statistical model	Rasch	Rasch	Rasch (CDM)	Rasch (CDM)	Rasch (CDM)	Rasch	Rasch
Item number	20	15	14 (2 items per construct)	14 (2 items per construct)	14 (2 items per construct)	6	6

takes the form of a traditional vocabulary size test, and the second one (15 items) asks test takers to choose the correct paraphrase for each target sentence. A major issue considered for these two tasks was whether to have input materials for each item. In reading assessments, it is common to find test items eliciting the meaning of particular terms from reading texts, usually low-frequency or academic words. Nonetheless, analyses of test-taking processes based on verbal reports have revealed that word comprehension embedded in context can evoke compensatory skill use (Jang, 2009; Liu & Read, 2021). Consequently, it would be difficult to discern if a correct response is attributable to accessing vocabulary knowledge or lexical inferencing based on the context, which would affect the accuracy of diagnostic results. This may also apply to assessing the ability to process complex sentence structures, which could be compensated for or assisted by contextual meaning as well, if such contexts are provided. Therefore, both tasks are designed to be without supporting context with the purpose of yielding clear and explainable diagnostic information.

Module II measures careful reading with three tasks, each targeting seven skills (see Table 2, Chap. 3). For each task, two items are used to assess each skill as the primary attribute. Thus, across the three tasks there would be six items targeting each of the seven skills. A major concern in designing the tasks for this module is the length of stimuli. In traditional integrative reading assessments, multiple tasks are commonly constructed based on one single text to achieve administrative efficiency (Alderson, 2000). Texts in such cases are relatively long, and thus are more representative of academic study materials compared with brief excerpts of one paragraph or less; furthermore, both global as well as local reading skills can be assessed. One potential drawback of this format, however, is the likelihood of interdependence among items due to overlapped target information (Liu & Read, 2021; Marais & Andrich, 2008), which causes a "testlet" effect (DeMars, 2012) that creates obstacles to clear inferences about ability state at the skill level. A possible solution is to base each item on a separate short text. However, it has been found that tasks with short stimuli such as those consisting of one or two sentences are mostly only capable of measuring word decoding or local comprehension rather than advanced or global comprehension skills (Keenan, 2012). Therefore, it seems that a balance needs to be found when designing the input materials for a diagnostic reading assessment, by employing, for instance, multiple texts of medium length (or texts of variable lengths, according to the skills being targeted). Also, in order to increase local independence, each text should not be loaded with too many items. Based on these considerations, each task in Module II is designed to contain two 300-word texts, which are related in theme so that items can be constructed to measure the last two skills (i.e., "understanding the relationship between two texts" and "drawing implication/conclusion based on two texts").

Another critical step in designing tasks targeting the seven skills is choosing between different test methods or formats. As discussed earlier, open-ended questions often raise issues of both scoring and cost efficiency. Despite these drawbacks, they are often regarded as more valid measures of reading comprehension by reducing construct-irrelevant factors such as random guessing (Alderson, 2000). As alternatives, task types that can be scored objectively, such as multiple choice, have been

found to be capable of measuring a range of reading skills (Freedle & Kostin, 1993; Jang, 2009). Meanwhile, they may introduce construct-irrelevant factors (Li & Suen, 2013; Rupp et al., 2006), such as test-taking strategies that are not generalizable to real-life academic reading. Liu and Read (2021) found that, nonetheless, some of these factors may not be associated with the relevant task types *per se*, but with the way that items are written. Therefore, in addition to choosing the right task types, good item writing practices are crucial for the construct validity of tasks.

It should also be noted that items embedded in larger tasks (such as multiple-matching and banked cloze) tend to influence each other in one way or another, due to their shared item stem or option pool (Liu & Read, 2021). Therefore, for a diagnostic assessment which aims to provide reliable skill-level information based on individual test items, items nested within larger tasks should be avoided or used with caution, given their potential interrelatedness, a situation that violates the local independence assumption of major statistical models estimating latent abilities (such as Item-Response Theory models or cognitive diagnostic models) (Rupp et al., 2010). After taking all the above issues into account, tasks from Module II are designed to be in a range of formats, including multiple-choice, sentence-completion, short-answer, true/false, and information-transfer.

Module III is devoted to assessing expeditious reading. In a diagnostic reading assessment that treats careful and expeditious reading as separate constructs, the time constraints for tasks measuring expeditious reading in particular should be strictly implemented to ensure that the intended skills are engaged (Weir et al., 2000). According to Carver (1992), while careful reading for learning information is typically done at a speed of around 200 words per minute (w.p.m.), expeditious reading is usually done at 450 w.p.m. or faster. Note that these criteria were set for L1 readers. Since the DELNA reading test population is primarily composed of L2 speakers of English, standards set for similar populations may be more appropriate. A model could be the Advanced English Reading Test (AERT) developed for Chinese university students (Weir et al., 2000), in which the reading speed for careful reading tasks was about 60–90 w.p.m., while that for expeditious reading tasks was about 100–150 w.p.m. In the College English Test (CET) of China, expeditious reading has been measured at similar speeds – 100 w.p.m. for CET Band 4 and 120 w.p.m. for CET Band 6 (Zheng & Cheng, 2008). Therefore, 100–150 w.p.m. could be a starting point for setting the time limits for expeditious reading tasks to be included in the new DELNA reading assessment.

Using the above criteria as references, the first task of Module III, which focuses on measuring information searching, contains six 250-word texts, with each text being presented *below* one multiple-choice question that asks the test taker to search for specific information in that text within 1 min. These six mini-tasks will be presented one after another, and each one is supposed to be finished within 1 min. The second task assesses skimming for the general idea. It also has six slightly shorter texts, each around 200 words in length and followed by one multiple-choice question. Each sub-task will also be presented one by one and finished in 1 min.

Tasks from all three modules will be presented and completed on computer. As shown in Fig. 1, the screening phase of DELNA is already computer-based and it makes sense for the diagnosis to transition to an online platform as well. For diagnostic reading assessment in particular, the computer is more efficient than paper and pencil in manipulating the order in which input materials and tasks are presented as well as constraining the time allotted to each task. Both considerations are critical for eliciting the intended reading behaviors, especially for the expeditious reading tasks. Weir et al. (2000) observed that when the time allotment for a paper-and-pencil reading test was not strictly controlled, test takers spent most of their time carefully reading the first few of a set of tasks that were designed to measure expeditious reading, while leaving insufficient time for answering the rest of the tasks. Computer-delivered tests will also boost the efficiency of scoring and score reporting, while reducing administrative costs over time.

All items will be scored dichotomously. To analyze the results, primary use will be made of the Rasch analysis, along with cognitive diagnostic modeling (CDM, for Module II). Rasch analysis can be used to synthesize information across items targeting the same construct. For Module II, CDM may be additionally employed, given that each item may tap into other skills beside the primary one being targeted. Although CDM has become increasingly popular with diagnostic assessments due to its capability of modeling sophisticated relationships between constructs and item responses as well as producing fine-grained results, it generally requires complex software and specialist expertise, which may not be available to many developers of local assessments. In such cases, simpler methods such as the Rasch model and even Classical Testing Theory would appear more viable, as long as they can produce acceptably reliable and accurate information on the target constructs.

Since with Rasch models test takers and test items measuring the same construct from different administrations can be plotted on a common scale, each test taker can be informed about their position or developmental status of this construct against that scale. Apart from the construct- or skill-specific information, feedback can also be provided on performance at the subdomain level (such as academic language knowledge, careful reading within text, careful reading across texts, and fast reading) as well as the level of overall academic reading. Information at different levels can be reviewed by the language advisor and the test taker together during individual consultation so that they can also identify weak skills and/or subdomains and discuss future learning plans, including resources that can be utilized to improve those skills or areas. At present DELNA assessments are taken only once by individual students, but if the re-designed reading test proves to have value as a diagnostic assessment, it may be possible to allow students to retake the test after a period of study to measure improvement, following the model of the Diagnostic English Language Tracking Assessment at Hong Kong Polytechnic University (Urmston et al., 2016).

6 Insights Gained

This study demonstrates, though in a preliminary fashion, how it is possible to develop a diagnostic reading assessment which incorporates learners' needs in a principled way. As Karakoç et al. (2022) note in their recent survey of academic reading requirements for first-year students at another New Zealand university, there have been few studies of this kind published internationally, and this in itself provided a justification for investigating reading needs at the local level. However, undertaking the survey also gave us more confidence that the design of the test addressed reading needs that had been identified by both students and language tutors at the university as important and challenging for undergraduates in their academic studies. This in turn would lend credibility to the advice given by the DELNA language tutors to students who performed poorly in the diagnostic reading test about how to develop their academic reading skills.

Results from the local needs survey show us that skillful academic reading at the tertiary level requires more than passive comprehension of texts, which most traditional reading models focus on, to include skills that reflect more the strategic, purposeful, evaluative, and productive nature of academic reading (Grabe & Stoller, 2013), such as expeditious and multiple-text reading, evaluation, and information reconstruction skills. Due to practical constraints on administration and scoring, these latter skills are often neglected by conventional proficiency and placement reading tests. However, for a local diagnostic assessment that aims to generate useful feedback for subsequent learning, such constraints may need to be re-considered in the light of concern for the relevance, sufficiency and utility of the diagnostic information that is obtained. Expeditious and intertextual reading skills could readily be incorporated in a new version of DELNA reading assessment based on the design presented in Table 2, Chap. 3, though the inclusion of other skills such as summarizing and paraphrasing may not be so feasible.

Regarding the construct definition of diagnostic assessments, the issue of the grain size or granularity of the constructs to be assessed has been raised and discussed on various occasions in the literature (e.g., Harding et al., 2015; Jang, 2009; Lee, 2015); however, no empirical research has been conducted to explore the possible grain sizes at which reading constructs can be defined. As a result, past assessment practices had the tendency of focusing only on individual skills. The subdomains identified in this study show how a diagnostic reading assessment could also target constructs of a larger grain size. That is, in addition to assessing reading skills in detail, such assessments could also provide information on broader subdomains of academic reading. That said, it remains to be investigated as to which type of information will be more effectively communicated and taken up in developing students' academic reading skills.

7 Implications for Test Development and Use

One implication that can be drawn from the study is that we need to take a cautious attitude towards retrofitting the interpretation of test results for diagnostic uses from assessments designed for non-diagnostic purposes, as alerted by Alderson (2005, 2010). A major risk of doing so is that we may end up with feedback that is of insufficient or little use to students due to construct under-representation, which cannot be remedied through retrofit simply based on *post hoc* data re-analyses. However, in tertiary-level academic situations, these neglected constructs may be more needed than those that are already assessed, as shown by the needs survey in this study. Therefore, a useful diagnostic reading assessment prior to or at the outset of a tertiary program ought to assess them and provide valid information on them to students. To achieve this, an empirical needs analysis would become a useful means to define the constructs that are relevant and critical to the specific needs of the target student population. In other words, a needs-driven approach to designing a diagnostic language assessment, especially a local one, has the advantage of addressing specific needs in a more straightforward and meaningful way, compared with the approach of reverse-engineering tests designed for other purposes and targeting a larger student population.

In the study described in this chapter, results of the needs analysis led to a redesigned reading assessment measuring a series of knowledge areas and skills, which can be theoretically subsumed under different subdomains of academic reading. However, the feasibility and usefulness of this new design is yet to be validated. To do this, sample tasks and items need to be developed and piloted. Different types of evidence also need to be collected and evaluated during the pilot, such as evidence about reliability, content validity, and construct validity. While the ways in which different evidence is collected are similar to those involved in validating proficiency and achievement tests, for a diagnostic assessment it is necessary to gather validity evidence for each component of the assessment based on which results will be interpreted and reported.

Another important type of evidence that needs to be collected, not only in the pilot, but also during on-going use of the assessment, concerns language teachers' and students' utilization of the diagnostic information, including their perceptions of and attitudes towards its usefulness in guiding consultation and learning as well as the impact it brings to students. After all, the ultimate goal of a diagnostic language assessment is to bring about a positive influence on students' language development. To what extent this intended goal is achieved should be the central focus of the validation of such assessments. Such investigations may take the form of interviews, questionnaire surveys, diary research, further testing, or a combination of these.

References

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

Alderson, J. C. (2005). *Diagnosing foreign language proficiency: The interface between learning and assessment*. Continuum.

Alderson, J. C. (2010). "Cognitive diagnosis and Q-matrices in language assessment": A commentary. *Language Assessment Quarterly, 7*(1), 96–103. https://doi.org/10.1080/15434300903426748

Alderson, J. C., Haapakangas, E.-L., Huhta, A., Nieminen, L., & Ullakonoja, R. (2015). *The diagnosis of reading in a second or foreign language*. Routledge.

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford University Press.

Biber, D., Conrad, S., Reppen, R., Byrd, P., Helt, M., Clark, V., et al. (2004). *Representing language use in the university: Analysis of the TOEFFL 2000 spoken and written academic language corpus*. Educational Testing Service.

Carr, N. T. (2014). Computer-automated scoring of written responses. In A. J. Kunnan (Ed.), *The companion to language assessment* (Vol. II, pp. 1063–1078). John Wiley & Sons. https://doi.org/10.1002/9781118411360

Carver, R. P. (1992). Reading rate: Theory, research, and practical implications. *Journal of Reading, 36*(2), 84–95. https://www.jstor.org/stable/pdf/40016440.pdf

Chen, H., & Chen, J. (2016). Exploring reading comprehension skill relationships through the G-DINA model. *Educational Psychology, 36*(6), 1049–1064. https://doi.org/10.1080/01443410.2015.1076764

Cohen, A. D. (1993). The role of instructions in testing summarizing ability. In D. Douglas & C. A. Chapelle (Eds.), *A new decade of language testing research: Selected papers from the 1990 language testing research colloquium* (pp. 132–160). Teachers of English to Speakers of Other Languages.

DeMars, C. E. (2012). Confirming testlet effects. *Applied Psychological Measurement, 36*(2), 104–121. https://doi.org/10.1177/0146621612437403

Elder, C., & Read, J. (2015). Post-entry language assessments in Australia. In J. Read (Ed.), *Assessing English proficiency for university study* (pp. 25–46). Palgrave Macmillan.

Freedle, R. O., & Kostin, I. (1993). The prediction of TOEFL reading item difficulty: Implications for construct validity. *Language Testing, 10*(2), 133–170. https://doi.org/10.1177/026553229301000203

Goldman, S. R., Braasch, J. L., Wiley, J., Graesser, A. C., & Brodowinska, K. (2012). Comprehending and learning from internet sources: Processing patterns of better and poorer learners. *Reading Research Quarterly, 47*(4), 356–381. https://doi.org/10.1002/RRQ.027

Grabe, W., & Stoller, F. L. (2013). *Teaching and researching reading*. Routledge.

Graesser, A. C., & Forsyth, C. M. (2013). Discourse comprehension. In D. Reisberg (Ed.), *The Oxford handbook of cognitive psychology* (pp. 475–491). Oxford University Press.

Harding, L., Alderson, J. C., & Brunfaut, T. (2015). Diagnostic assessment of reading and listening in a second or foreign language: Elaborating on diagnostic principles. *Language Testing, 32*(3), 317–336. https://doi.org/10.1177/0265532214564505

Hellekjær, G. O. (2009). Academic English reading proficiency at the university level: A Norwegian case study. *Reading in a Foreign Language, 21*(2), 198–222. http://www2.hawaii.edu/~readfl/rfl/October2009/articles/hellekjaer.pdf

Jang, E. E. (2009). Demystifying a Q-matrix for making diagnostic inferences about L2 reading skills. *Language Assessment Quarterly, 6*(3), 210–238. https://doi.org/10.1080/15434300903071817

Karakoç, A. I., Ruegg, R., & Gu, P. (2022). Beyond comprehension: Reading requirements in first-year undergraduate courses. *Journal of English for Academic Purposes, 55*, 1–12. https://doi.org/10.1016/j.jeap.2021.101071

Keenan, J. M. (2012). Measure for measure: Challenges in assessing reading comprehension. In J. P. Sabatini, E. R. Albro, & T. O'Reilly (Eds.), *Meauring up: Advances in how to assess Reading ability* (pp. 77–87) Rowman & Littlefield Education.

Khalifa, H., & Weir, C. J. (2009). *Examining reading: Research and practice in assessing second language reading*. Cambridge University Press.

Kintsch, W. (2013). Revisiting the construction-integration model of text comprehension and its implications for instruction. In D. E. Alvermann, N. J. Unrau, & R. B. Ruddell (Eds.), *Theoretical models and processes of reading* (6th ed., pp. 807–839). International Reading Association.

Lee, Y.-W. (2015). Diagnosing diagnostic language assessment. *Language Testing, 32*(3), 299–316. https://doi.org/10.1177/0265532214565387

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. https://doi.org/10.1080/15434300903079562

Leighton, J. P., & Gierl, M. (Eds.). (2007). *Cognitive diagnostic assessment for education: Theory and applications*. Cambridge University Press.

Leu, D. J., Gregory McVerry, J., Ian O'Byrne, W., Kiili, C., Zawilinski, L., Everett-Cacopardo, H., et al. (2011). The new literacies of online reading comprehension: Expanding the literacy and learning curriculum. *Journal of Adolescent & Adult Literacy, 55*(1), 5–14. https://doi.org/10.1598/JAAL.55.1.1

Li, H., & Suen, H. K. (2013). Constructing and validating a Q-matrix for cognitive diagnostic analyses of a reading test. *Educational Assessment, 18*(1), 1–25. https://doi.org/10.1080/10627197.2013.761522

Liu, X. (2018). *Establishing the foundation for a diagnostic assessment of reading in English for academic purposes*. PhD thesis, The University of Auckland, Auckland.

Liu, X., & Brown, G. T. L. (2019). Investigating students' perceived cognitive needs in university academic reading: A latent variable approach. *Journal of Research in Reading, 42*(2), 411–431. https://doi.org/10.1111/1467-9817.12275

Liu, X., & Read, J. (2020). General skill needs and challenges in university academic reading: Voices from undergraduates and language teachers. *Journal of College Reading and Learning, 50*(2), 70–93. https://doi.org/10.1080/10790195.2020.1734885

Liu, X., & Read, J. (2021). Investigating the skills involved in reading test tasks through expert judgement and verbal protocol analysis: Convergence and divergence between the two methods. *Language Assessment Quarterly., 18*, 357–381. https://doi.org/10.1080/15434303.2021.1881964

Lumley, T. (1993). The notion of subskills in reading comprehension tests: An EAP example. *Language Testing, 10*(3), 211–234. https://doi.org/10.1177/026553229301000302

Macedo-Rouet, M., Potocki, A., Scharrer, L., Ros, C., Stadtler, M., Salmerón, L., & Rouet, J. F. (2019). How good is this page? Benefits and limits of prompting on adolescents' evaluation of web information quality. *Reading Research Quarterly, 54*(3), 299–321. https://doi.org/10.1002/rrq.241

Marais, I., & Andrich, D. (2008). Formalizing dimension and response violations of local independence in the unidimensional Rasch model. *Journal of Applied Measurement, 9*, 200–215. http://www.curriculum.edu.au/verve/_resources/ARC-Report11DistinguishingMDandRD.pdf

McNamara, D. S., & Magliano, J. (2009). Toward a comprehensive model of comprehension. In B. Ross (Ed.), *The psychology of learning and motivation* (Vol. 51, pp. 297–384). Academic.

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). A brief introduction to evidence-centered design. *Educational Testing Service, 2003*, i–29.

Perfetti, C. A. (2007). Reading ability: Lexical quality to comprehension. *Scientific Studies of Reading, 11*(4), 357–383. https://doi.org/10.1080/10888430701530730

Read, J. (2015). *Assessing English proficiency for university study*. Palgrave Macmillan.

Richter, T., & Maier, J. (2017). Comprehension of multiple documents with conflicting information: A two-step model of validation. *Educational Psychologist, 52*(3), 148–166. https://doi.org/10.1080/00461520.2017.1322968

Riley, G. L., & Lee, J. F. (1996). A comparison of recall and summary protocols as measures of second language reading comprehension. *Language Testing, 13*(2), 173–189. https://doi.org/10.1177/026553229601300203

Rouet, J.-F., Britt, M. A., & Potocki, A. (2019). Multiple-text comprehension. In J. Dunlosky & K. A. Rawson (Eds.), *The Cambridge handbook of cognition and education* (pp. 356–380). Cambridge University Press.

Rupp, A. A., Ferne, T., & Choi, H. (2006). How assessing reading comprehension with multiple-choice questions shapes the construct: A cognitive processing perspective. *Language Testing, 23*(4), 441–474. https://doi.org/10.1191/0265532206lt337oa

Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications*. Guilford Press.

Urmston, A., Raquel, M., & Aryadoust, V. (2016). Can diagnosing university students' English proficiency facilitate language development? In J. Read (Ed.), *Post-admission language assessment of university students* (pp. 87–109). Springer.

Weir, C. J., Yang, H., & Jin, Y. (2000). *An empirical investigation of the componentiality of L2 reading in English for academic purposes*. Cambridge University Press.

Weir, C. J., Hawkey, R., Green, A., Unaldi, A., & Devi, S. (2012). The relationship between the academic reading construct as measured by IELTS and the reading experiences of students in their first year of study at a British university. In L. Taylor & C. J. Weir (Eds.), *IELTS collected papers 2: Research in reading and listening assessment* (pp. 37–119). Cambridge University Press.

Zheng, Y., & Cheng, L. (2008). Test review: College English test (CET) in China. *Language Testing, 25*(3), 408–417. https://doi.org/10.1177/0265532208092433

Pre-post Elicited Imitation: Documenting Student Gains in the Purdue Language and Cultural Exchange

Lixia Cheng and Xiaorui Li

Abstract This chapter provides an account of a locally-developed Elicited Imitation (EI) test used to gauge international undergraduate students' entry-level English proficiency and the use of the instrument to document end-of-year gains. Significant gain scores, with a strong associated effect, contributed to evidence in support of program effectiveness. A brief history of the Purdue Language and Cultural Exchange, an English for Academic Purposes program, is followed by a report on a three-stage quantitative analysis of the EI test, which has played an instrumental role in securing institutional support. Stage 1 examined the test's measurement properties through the application of Classical Test Theory. After the verification of technical qualities in Stage 1, Stage 2 focused on pre-post test score gains and the effect size. In Stage 3, pre-post score changes were evaluated at the item level through the computation of Hedge's *g*, an instructional sensitivity index for each item. Besides the pre-post score gains for program evaluation in Stage 2, this chapter also demonstrates the use of an instructional sensitivity index to complement the standard item-level analyses on difficulty and discrimination. Findings informed test revision and provided insights on program integration of assessment and instruction.

Keywords Local language testing · Elicited imitation · Program evaluation

1 Introduction: Context, Purpose, and Problem Encountered

Local tests are created, administered, and validated to fulfill specific functions for an associated language program in a specific local context. This chapter focuses on the elicited imitation (EI) section of the Assessment of College English–International

L. Cheng (✉)
Purdue Language and Cultural Exchange, Purdue University, West Lafayette, IN, USA
e-mail: clixia@purdue.edu

X. Li
Oral English Proficiency Program, Purdue University, West Lafayette, IN, USA
e-mail: li1828@purdue.edu

X. Yan et al. (eds.), *Local Language Testing*, Educational Linguistics 61,
https://doi.org/10.1007/978-3-031-33541-9_6

(ACE-In), a post-entry English proficiency test developed and used by the Purdue Language and Cultural Exchange (PLaCE), which is an English for Academic Purposes (EAP) program for international undergraduate students at Purdue University. Purdue is a large, public (state funded), R1[1] institution of higher education in the state of Indiana. Known for its strength in Science, Engineering, Technology, and Mathematics (STEM), the University has long ranked among the top five for international student enrollment at public universities in the United States. From 2004 to 2014, international undergraduate enrollment increased by an astonishing 10-year growth rate of 172%, peaking in 2014 with 5282 international undergraduates. This undergraduate total, combined with the international graduate class of 3798, meant that international students comprised 23.4% of all the 38,770 students enrolled at Purdue in 2014. Record increases in international undergraduate enrollment were common across U.S. universities at that time and were embraced by university administrators for at least three reasons: to increase revenue, replace in-state students lost to declining domestic birthrates, and increase and enhance campus internationalization (Fischer, 2014; Ginther & Yan, 2018). To compensate for decades-long reductions in state funding across all levels of public education, public universities have steadily increased enrollment, especially of international and out-of-state undergraduate students, as these students pay double the in-state, public tuition rate.

In addition to their potential contributions to revenue, international undergraduate students raised Purdue's academic profile (Ginther & Yan, 2018), another important consideration for market-driven university administrators who pay attention to college rankings published in magazines such as the *U.S. News* and *World Report's America's Best Colleges*. Rankings are greatly influenced by applicants' scores from standardized tests for college admissions, such as the SAT and ACT. At Purdue, "international undergraduates enter with higher SAT total scores (+100 points), graduate at a slightly higher four-year rate (51% vs. 49%), and obtain slightly higher GPAs than their domestic counterparts from the state of Indiana" (Ginther & Yan, 2018, p. 272). Indeed, international students' success as measured by the traditional standards such as GPA and the four-year graduation rate has undermined arguments for the need and development of academic support programs for international students.

2 Testing Problem Encountered

In 2014, the rapidly growing number of international undergraduate students attracted attention. Faculty and staff expressed concern about aspects of success that GPA and graduation rates cannot capture, particularly the development of language

[1] According to the Carnegie Classification of Institutions of Higher Education, Research I (R1) universities in the U.S. are the top-tier doctoral degree granting universities with very high research activities. On top of the common requirements for R1 and R2 (Research II) universities (i.e., 20+ research/scholarship doctoral degrees; at least $5 million in research expenditures), an aggregate research activity index and a per-capita research activity index are usually calculated to distinguish R1s from R2s. Various measures are included in the two research activity indices.

communication skills and intercultural competence (National Academy of Sciences, 2011) in an instructional environment increasingly dependent on students' collaboration and interaction with peers and faculty (e.g., IMPACT at Purdue, https://www.purdue.edu/impact/). Purdue University's Senate passed three resolutions calling for the development of an EAP program for incoming international undergraduates; consequently, in 2014, at the height of Purdue's international undergraduate enrollment boom, Purdue's administration funded PLaCE, an instructional and assessment program created to address these issues.

But there were conditions: PLaCE was initially established as a two-year pilot program, and the need for the program would be examined each year with respect to the number of incoming international undergraduate students, their first-language (L1) backgrounds, and their entry-level language proficiency scores. The selection criterion for PLaCE was set at 101 on the TOEFL iBT (or a comparable overall score on the IELTS); that is, international undergraduate students entering with a TOEFL iBT total score of less than 101 would be required to enroll in a two-semester language and intercultural support sequence: English 110 and 111, "American Language and Culture for International Students I & II" (PLaCE, n.d.). However, the administration suggested that the need for the program could be ameliorated, even eliminated, by admitting and matriculating students with higher TOEFL iBT scores. Enrollment management countered this suggestion with the fact that even with an annual applicant pool of more than 10,000, the annual yield (enrolled international undergraduate students at the beginning of the semester) remained normally distributed with a mean total score of 100. In fact, international undergraduate enrollment at Purdue over the years has stabilized to around 1000 each year, about 45% of whom remain PLaCE eligible.

The two-year pilot phase given to PLaCE, however, motivated program developers to look for ways in which the English 110 and 111 courses could be embedded in broader curricular requirements without adding to the credit hours required to complete the undergraduate degree. The University Undergraduate Curriculum Committee approved the program's application for English 110 to be accepted as meeting the foundational learning outcomes under the category of "Human Cultures: Humanities" (Purdue University, n.d.). Thus, PLaCE program administrators were able to lower anticipated resistance from students and their academic advisors by ensuring that English 110 and 111 were not only credit bearing, but English 110 would also contribute to the required credit hour total in students' degree plans.

Finally, an additional, critical requirement was imposed. In order for PLaCE to secure recurring funding after the pilot phase, the program must demonstrate proven added value, preferably, pre-post gains in language proficiency scores. The requirement that the program demonstrate effectiveness was embedded in the Daniels[2] administration's emphasis on assessment and accountability, as evidenced by a memo published on September 16, 2021:

[2] Mitchell E. Daniels Jr., the 49th governor of Indiana from 2005 to 2013, served as president of Purdue University from 2013 until the end of 2022.

> Over the next year we will work together to execute the Next Moves[3] plan to produce a positive return on investment, continue to ensure the Purdue experience is accessible, affordable and high quality for our students, and develop key performance indicators and metrics to enable a consistent and simple view of the efficiency and efficacy of University functions, operations and departments.

The Daniels administration's emphasis on assessment and accountability, as evidenced in the Next Moves as well as in the Purdue Moves launched in fall 2013, presented opportunities for an academic program such as PLaCE. At the practical and administrative levels, PLaCE was in an excellent position to negotiate for the inclusion of a full-time Assistant Director of Testing, two half-time graduate testing assistants, and a reduced teaching load for its 14–17 lecturers (from four courses per semester to three) to facilitate and support language assessment and program evaluation. As part of their required duties, lecturers would participate in rater training, exam ratings, test item writing, and scale development. Involvement in and responsibility for assessment was distributed throughout the program.

In December 2017, the Board of Trustees approved a slight increase in the international student fee, effective in August 2018, to provide a sustainable funding model for the continuation of PLaCE's English 110 and 111 course sequence as well as the development of a set of non-credit, advanced short courses for international undergraduate and graduate students. Also included in the approved proposal was an annual program evaluation comprising the demonstration of significant pre-post language proficiency gains.

As the most developed, examined, and revised section of the ACE-In, EI has been selected as a consistent, core method to demonstrate language proficiency gains of international undergraduates in the foundational English 110 and 111 course sequence. Evidence obtained from pre-post comparisons has been a central contributor to the PLaCE program's ability to secure continued financial support. Furthermore, these results have helped PLaCE administrators develop a better understanding of what gains are possible given our instructional approach and in which lexical, grammatical, and syntactic areas we can expect to see more noticeable gains. Our local test has successfully fulfilled these critical functions for PLaCE and will continue to do so.

3 Literature Review

Our literature review begins with a discussion of the construct.

[3] As continuation and enhancement of the Purdue Moves launched in 2013, Purdue's Next Moves announced in 2021 still represents four strategic imperatives: affordability and accessibility, STEM leadership, world-changing research, and transformative education (https://www.purdue.edu/purduemoves/).

3.1 Elicited Imitation: What It Measures

EI, a psycholinguistic task eliciting oral production of language chunks, originated in L1 developmental studies and was later adopted by the field of Second Language Acquisition (SLA). In an EI assessment task, the examinee is expected to repeat each prompt sentence as accurately as possible after hearing it once (Larsen-Freeman & Long, 1991). EI test performances are normally rated by human raters following a rating scale or by e-raters with certain parameters built into the e-rater program. In terms of designing an EI test, test developers have considered a wide range of factors such as whether to include formulaic sequences, conversational lexicon, or ungrammatical structures in the sentence stimuli; the syllable count, vocabulary profiles, and syntactic complexity of each prompt sentence; as well as the need to include a repetition delay (Erlam, 2006; Van Moere, 2012; Yan et al., 2016).

The resurgence of interest in EI since the early 2000s has contributed to guidelines for task design and validity evidence for using EI as a measure of second language (L2) proficiency (e.g., Baten & Cornillie, 2019; Davis & Norris, 2021; Erlam, 2006; Hsieh & Lee, 2014; Jessop et al., 2007; Suzuki & DeKeyser, 2015; Van Moere, 2012; Yan et al., 2016; Yan, 2020). EI is argued to measure L2 linguistic competence, by tapping into language learners' implicit knowledge and language processing automaticity when learners process the meaning and form of the prompt sentence during the aural comprehension and oral reproduction stages (Davis & Norris, 2021; Van Moere, 2012).

Despite ongoing discussions about the relationship between explicit knowledge and implicit knowledge, the argument that one's linguistic knowledge comprises both explicit and implicit knowledge has been widely accepted (Ellis, 2005, 2009a). Ellis (2009b) argued that various types of language tests can provide separate measures of explicit and implicit knowledge. For example, learners' explanation of specific linguistic features is often considered as the representation of their explicit knowledge, while the use of these features in oral and written language is believed to be associated with implicit knowledge (Ellis, 2009b). Measures of implicit knowledge, therefore, should elicit spontaneous performance that avoids learners' conscious use of linguistic knowledge. EI tasks constitute one type of instrument developed and used to measure implicit knowledge.

As to automaticity of language processing, a psycholinguistic trait involved in EI tests, Levelt's (1989) information processing model states that a speaker experiences three stages of information processing to produce a meaningful L2 utterance. These three stages are (1) conceptualizing input, (2) conducting grammatical and phonological encoding, and (3) articulating a meaningful utterance. The speaker's implicit knowledge plays a crucial role during the second stage, which directly affects the performance in the third stage. Based on Levelt's (1989) model, an instrument that enacts the information processing procedure and captures the L2 performance in the third stage is likely to tap into the learner's implicit knowledge. In an EI task, examinees start from listening to a sentence stimulus, which triggers

the information processing procedure. After conceptualizing the input, examinees need to decode, within a fleeting period of time, the grammatical and phonological features by using their implicit knowledge. The ultimate reconstruction and reproduction of the prompt sentence represents the outcome of this series of online processing. During an EI task, proficient examinees are more capable of using implicit knowledge to decode the input, reconstruct the sentence, and produce a satisfactory response, whereas low proficiency examinees often have difficulty completing the entire process and producing a grammatical response that retains the meaning of the prompt sentence.

To address concerns about some examinees "parroting" prompt sentences, several studies related to EI task design have been conducted to investigate the potential problem of examinees using rote repetition without processing the meaning of the sentence stimuli (e.g., Erlam, 2006; Kostromitina & Plonsky, 2021; Spada et al., 2015; Yan et al., 2016). The design factors most extensively studied include the length of the prompt sentence, grammatical structures in the sentence, delays in repetition, and task instructions about focusing on meaning. EI task design often incorporates features proposed by Ellis (2009b) for measures of implicit knowledge: for example, EI tasks should elicit performances under time pressure, and task instructions should lead examinees to focus on the meaning of sentence stimuli instead of the linguistic structure. An EI test including longer prompt sentences and sentences with a wide range of lengths seems more desirable (Kostromitina & Plonsky, 2021). With a careful test design, examinees would have a minimal possibility of leveraging explicit linguistic knowledge to decode and reconstruct the prompt sentences.

3.2 Pre-Post Score Gains: As Validity Evidence for Test and as Instructional Sensitivity Metric for Items

Test validation is an iterative process of accumulating evidence in support of a particular way of using and interpreting test scores. Part of this evidence can rightfully be related to test scores' sensitivity to changes in examinee abilities as a function of construct-associated teaching or learning (Cronbach, 1971; Messick, 1989). Chapelle et al. (2008) argued for the importance of using the evidence of test score gains after instruction to enhance the validity argument for the TOEFL iBT. Likewise, Ling et al. (2014) maintained that "one fundamental way to determine the validity of standardized English-language test scores is to investigate the extent to which they reflect anticipated learning effects in different English-language programs" (p. 1). They extended the validity argument for the TOEFL iBT by comparing the language learning effects—as reflected in TOEFL iBT score changes—of students enrolled in intensive English programs in the United States versus China. According to Ling et al. (2014), the TOEFL iBT score gains in both settings indicated a consistent improvement of students' language proficiency as a result of instruction, which

added a layer of validity evidence for the TOEFL iBT as a measure of English language proficiency. In relation to our local context in PLaCE, Cheng (2018) conducted a pilot study to explore using English 110 and 111 students' score gains in a pre-post design to obtain additional validity evidence in support of the locally-developed ACE-In.

Despite the challenges of student attrition and tracking students' pre- and post-test performances longitudinally, comparisons of trustworthy pre- and post-test performance data provide valuable information about the learner group's development and the instructional program's effectiveness, as well as help to extend the validity argument for the local test used. In a related vein, Polikoff (2010) argued from a psychometric point of view that an instructionally sensitive test should be able to capture the differences in examinees' performances as a result of the instruction they have received. While longitudinal tracking of pre-post score gains focuses on the entire sample of examinees, Polikoff's proposal of instructional sensitivity as a metric for test quality—especially pertaining to a local test embedded in an instructional program—takes an alternative approach to view score gains at the item level. The inclusion of instructional sensitivity in the routine suite of analyses on a local test's technical qualities can add a new layer of information, particularly useful for guiding test revision efforts. In other words, instructional sensitivity can be examined as an innovative, additional item property of a local test integrated into the curriculum of an instructional program. If instructional sensitivity was considered in test design to start with, highly sensitive test items based on examinees' performance data can suggest both good instruction and a high-quality test that matches the instruction, whereas insensitive items can result from poor integration of assessment and instruction (Polikoff, 2010).

4 Methods

In this study, we focused on the measurement qualities of the ACE-In EI items and the efficacy of using EI pre-post score changes to track PLaCE students' English as a Second Language (ESL) proficiency development. In addition, we examined the instructional sensitivity of all the ACE-In EI items. We adopted the Classical Test Theory (CTT) framework to examine the technical qualities of the EI items because CTT is a fairly straightforward and less technically demanding approach to analyzing item properties such as difficulty and discrimination to identify reliable as well as poorly functioning items (Bachman, 2004; Crocker & Algina, 1986; Moses, 2017). The following research questions guided our analysis:

1. Do the ACE-In EI items demonstrate acceptable measurement qualities? After the verification of interrater reliability in EI scores, the parameters to be examined would include test form comparability, item difficulty, item discrimination, section-internal consistency, and standard error of measurement.

2. Did PLaCE students achieve pre-post ACE-In score gains, which can provide evidence for improved L2 general proficiency resulting from ESL instruction and practice in the two-semester course sequence?
3. What are individual EI items' instructional sensitivity values, for the purposes of informing test revision and better integration of assessment and instruction?

4.1 Instrument: The ACE-In EI Module

The ACE-In EI section was designed to assess international ESL undergraduate students' ability to "understand and fluently use the spoken academic register" through two trained lecturer-raters' blind ratings of the grammatical accuracy and semantic retention in examinees' spoken reproduction of each prompt sentence (PLaCE, 2016). Appendix A presents the EI test instructions and a warm-up item; Appendix B shows the five-point 0–1–2–3–4 ACE-In EI rating scale adapted from published sources (e.g., Ortega et al., 1999). Our EI rating scale awards the highest score of "4" to exact repetition, the second highest score of "3" to a grammatical paraphrase, a "2" for minor grammar errors and/or minor meaning deviation in the reproduction, and a "1" for major grammar errors and/or major meaning deviation. A score of "0" is reserved for complete silence or just one or a few words repeated from the prompt sentence that do not make sense or even remotely approximate the original meaning of the prompt sentence.

Our EI rating scale is relatively easy to apply by raters. It has been associated with satisfactory interrater agreement and has not required extensive rater calibration. The efficiency and high-quality of EI rater training and ratings can be attributed to the narrower focus in these rating tasks where the rater would simply compare the examinee's reproduction with the prompt sentence to identify grammatical errors and meaning deviations. In contrast, rating free-response speaking tasks requires raters to attend to a list of distinct but interrelated concerns such as intelligibility, comprehensibility, coherence, content development, pronunciation, prosody, fluency, grammar, syntax, and lexical usage.

Whole-group and/or individual rater calibrations have been regularly conducted since August 2016, before every large round of ACE-In test administration, to improve rater alignment with the rating scale (see Appendix B) and consequently, enhance interrater reliability. These EI rater calibration sessions have focused on analyzing benchmarks per prompt sentence; discussing and trying to form a consensus about the severity of certain meaning deviations caused by lexical addition, omission, or substitution; and maintaining cumulative notes about rating decisions made with the entire rater-group. Besides, assumptions about what an EI test measures and the value of this assessment to our program are discussed and acknowledged by all lecturer-raters involved. Lecturer-raters are brought to the collective understanding that (1) EI measures learners' ability to reconstruct or reproduce an utterance, as facilitated by their implicit grammatical knowledge and language processing automaticity; (2) grammatical accuracy and meaning retention in the

reconstructed sentence are appropriate proxies for the state of an examinee's implicit knowledge store.

4.2 Participants

We analyzed the EI test performances of all the international ESL undergraduate students enrolled in the English 110 and 111 course sequence in the 2017–2018 academic year who had submitted their TOEFL iBT scores for university admissions and who completed the ACE-In pre- and post-test during their two semesters in the PLaCE program. These two participant-selection criteria—for the purposes of understanding and describing our study sample and comparing pre- and post-test scores—resulted in a sample including a total of 261 participants: 148 (56.7%) males and 113 (43.3%) females. The age range was 17–21 at the time of the pretest, with a median of 18. While our sample included students from 31 countries and regions, the four largest subgroups were 177 (67.8%) students from China, 18 (6.9%) from India, 12 (4.6%) from South Korea, and 11 (4.2%) students from Taiwan.

The TOEFL iBT subsection score means (see Table 1) indicate that on average, our participants had high-intermediate English proficiency in correspondence to CEFR Level B2, except that the TOEFL iBT listening mean score classified this group as one demonstrating advanced listening skills or CEFR Level C1 listening performances (Educational Testing Service, 2019). The slightly higher categorization of our participants' TOEFL iBT listening scores on the CEFR scale was most likely due to the fact that a combined total of 76.6% of our study sample composed of East Asian students who tend to perform well on language tests written in a multiple-choice question (MCQ) format and used to assess receptive language skills, such as the TOEFL iBT listening section.

4.3 Data Collection

PLaCE students complete the ACE-In pretest for 2% of their English 110 course grade, from late August through mid-September, due to limited availability of an on-campus lab. Their post-test is completed between mid-March and early April, in

Table 1 Descriptive statistics for participants' (N = 261) TOEFL iBT subsection and total scores

	Mean	SD	Min.	Max.	95% Confidence Interval for Mean
Reading	23.63	2.80	14	30	[23.29, 23.97]
Listening	23.69	2.59	18	30	[23.37, 24.00]
Speaking	22.16	1.74	18	28	[21.95, 22.37]
Writing	22.95	2.22	17	29	[22.68, 23.22]
Total	92.43	5.26	74	101	[91.78, 93.07]

the second half of the English 111 course, also for a 2% course grade. The gap between the pre- and post-test in our study was, on average, 198 days or approximately 10 days short of seven months.

The following data on the 261 participants were obtained through generating reports from either the ACE-In Admin application or Cognos, a university approved reporting tool: (1) Participants' demographic information, including sex, birth year, and native country; (2) EI pre- and post-test item ratings and section-total scores (i.e., sum of 12 item ratings) assigned by each PLaCE lecturer-rater; and (3) TOEFL iBT subsection and total scores. In this study, we used only rater-scores to represent our participants' EI test performances. At the time when scoring decisions were made in Academic Year 2017–18, if the section-total scores assigned by the first two raters differed by at least six points in the pretest or at least four points in the post-test, the particular EI test was assigned to a third trained rater. The finalized section-total score was, then, the average of the section-total scores given by the two raters with a smaller score difference. The arbitrary decision rule of when an EI test needed a third rater (i.e., whether the first two rater-scores differed by six or four points in the pre- or post-test) was made primarily in view of the PLaCE lecturer-raters' rating workloads.

4.4 Data Analysis

The EI test data was analyzed in three stages mirroring the three research questions. Statistical programs such as Excel, SAS, and SPSS were used for data mining, analysis, and visualization. In Stage 1 of the data analysis, EI pre- and post-test scores were used in the examination of test form comparability, item difficulty and discrimination, internal consistency, and the standard error of measurement. While the EI post-test data registered similar findings, this paper only reports the results from performing the above-mentioned analyses on the pretest scores. Both pre- and post-test scores were included in this paper for the report on the Stage 1 evaluation of interrater reliability in the EI ratings, on the Stage 2 examination of EI score changes, as well as on the Stage 3 computation of EI items' instructional sensitivity values.

CTT Analyses of Measurement Qualities

There were four steps in the analyses of ACE-In EI measurement qualities in Stage 1. First, to ensure the EI rater-scores were trustworthy indicators of participants' test performances, interrater reliability in the pre- and post-test ratings was assessed through a consensus estimate operationalized as the Pearson correlation coefficient between the two section-total rater-scores eventually used in deciding the finalized section-total score of a student (Stemler, 2004). In addition, a consistency estimate of interrater reliability was operationalized as the frequency count of agreement

levels (e.g., exact agreement, adjacent agreement, or discrepancy) between the two rater-scores. We modified the terms for the agreement levels and used 'excellent agreement' to denote circumstances where the two section-total scores assigned by the two raters differed by 0–3 points and 'good agreement' for cases with a difference of 4–6 points between the two section-total scores. A section-total score difference of three points means that the two raters likely assigned an identical score each to 9 of the total 12 items and an adjacent score to the remaining quarter of the section. A section-total score difference of six points means that the two raters possibly assigned an identical score each to half of the 12 items and an adjacent score to the other half.

Second, comparability of the four ACE-In operational forms was examined through a box plot depicting the mean and quartiles of the EI section-total score for each pretest form as well as a one-way ANOVA (Analysis of Variance) investigating the effect of the pretest form on the section-total score.

Third, the item difficulty of the polytomously scored EI items was operationalized as the mean item score averaged across the two raters' item ratings. Thus, theoretically, there were nine possible item difficulty values ranging from 0–4 with a half-point interval. The difficulty indices for EI items were then classified into three categories: difficult, average difficulty, and easy, with the mean item score falling into one of these ranges: Difficult [0, 1.50], Average Difficulty [1.51, 2.50], and Easy [2.51, 4.00] (Li, 2020). Item discrimination was examined through polyserial correlation coefficients computed between the item score and the section-total score when a latent correlation between a continuous variable (i.e., the section-total score) and an ordered categorical variable (i.e., the item score) could be reasonably assumed. The discrimination indices were classified into four categories: very poor (item to be removed), marginal (item to be revised), okay (little or no revision needed), and good discrimination (item functions satisfactorily) (Ebel, 1965). The four ranges for the discrimination index were: Very Poor [−1.00, 0.19], Marginal [0.20, 0.29], Okay [0.30, 0.39], and Good [0.40, 1.00] (Ebel, 1965).

Fourth, test reliability was examined through Cronbach's alpha as an internal consistency index for assessing how tightly the items in the ACE-In EI module hold together to measure the same construct. Because the variances of item scores vary somewhat widely, the overall standardized Cronbach's coefficient alpha—rather than the raw alpha—for each pretest form was used and compared against the suggested value of 0.70 for the threshold of satisfactory reliability (Nunnally & Bernstein, 1994). Test reliability was also studied via the standard error of measurement (SE), which estimates how repeatedly taking the ACE-In might end up distributing the EI section score around their 'true' score, which is a founding concept of CTT as CTT centers on using the observed scores obtained from a test to estimate examinees' true ability (Crocker & Algina, 1986; Fulcher, 2013). We used Eq. (1) to calculate the SE for the ACE-In EI module.

$$SE = Standard\ deviation\ of\ observed\ scores \times \sqrt{1 - Reliability} \tag{1}$$

Pre-post Score Comparisons

After examining the measurement qualities of the ACE-In EI module, we compared participants' pre- and post-test scores to determine if there were significant score gains. Because normality testing found that neither the pretest nor post-test scores were normally distributed, we conducted a non-parametric equivalent of the paired samples *t*-test, i.e., the Wilcoxon signed-rank test, to examine if the post-test EI score was a significant improvement over the pretest score. In the case of significant score gains, the effect size was then calculated using Eq. (2).

$$Effect\ size r = Wilcoxon\ signed\quad rank test\ statistic Z \div \sqrt{Number\ of\ pairs} \tag{2}$$

Item Analysis: Instructional Sensitivity

If significant improvement was found between the pre- and post-test scores of our participant group, we would then examine all the individual EI items more closely to determine which prompt sentences may have higher instructional sensitivity values than the others. For this purpose, we computed the Hedge's *g* statistic per EI item because Hedge's *g* is a measure of the effect size based on mean scores, while including adjustments for sample size (Hedges, 1981). The adjustments were necessary because the pre-post comparison per EI item across the four test forms usually did not land the same sample size. This was because ACE-In examinees are always assigned a brand-new test form when they take the test again.

5 Results and Discussion

Results are discussed in terms of reliability, test form comparability, CTT analyses of item difficulty and discrimination, pre-post gain scores, and instructional sensitivity.

5.1 Interrater Reliability in the EI Pre- and Post-test Section-Total Rater-Scores

The consensus estimates of interrater reliability were a significant, very high Pearson correlation coefficient ($r = 0.93$, $p < 0.0001$) for both the pre- and post-test. Owing to the mechanism of including third raters, there was no pre- or post-test with a section-total score difference of more than six points on a scale ranging from 0 to 48 points for the whole EI section. Of the 261 pretests, 227 of them (87.0%) registered excellent agreement between the section-total scores assigned by the two

raters, with a score difference of 0–3 points (e.g., for the 3-point section-total score difference, one possibility was that the two raters exactly agreed on 9 items and had adjacent agreements on the other 3 items); another 34 tests (13.0%) registered good agreement with a section-total score difference of 4–6 points (e.g., for the 6-point section-total score difference, one possibility was that the two raters exactly agreed on 6 items and had adjacent agreements on the other 6 items). Of the total 261 post-tests, 235 of them (90.0%) registered excellent interrater agreement and 26 other tests (10.0%), good agreement. These strong, positive evidence of consensus and consistency in the EI pre- and post-test ratings indicates that the rater-assigned section-total scores were reliable indicators of test performances.

5.2 Test Form Comparability

Table 2 presents the descriptive statistics of the pretest EI section-total scores. Across the four forms, the pretest section-total score had its mean concentrated around 24 points out of maximum 48 points and its standard deviation around 6 points. The boxplot in Fig. 1 shows the mean, quartiles, and outliers in the pretest section-total score distribution by test form. Besides the graphic indications of test form comparability in the boxplot, a one-way between-subjects ANOVA test shows that test form had no significant effect on the EI pretest section score ($F_{(3, 257)} = 1.3$, $p = 0.276 > 0.01$).

5.3 Item Analysis in a CTT Framework: Item Difficulty and Discrimination

Table 3 lists the difficulty and discrimination indices for each of the 48 EI items (four test forms with 12 prompt sentences in each form). The difficulty index for each EI item—operationalized as the mean item score averaged across the two raters' item ratings—ranged from 1.05 to 2.73. The lower bound of item difficulty at 1.05 reflects item performance with major grammar mistakes and/or major meaning deviation, according to the rating scale. The upper bound of item difficulty at 2.73 reflects item performance that almost qualified as a good paraphrase of the prompt

Table 2 Descriptive statistics for EI pretest section-total scores (max. 48 points)

Form	N	Mean	SD	Min	Max	95% Confidence Interval for Mean
EI form 1	64	23.78	6.08	12.5	37.5	[22.26, 25.30]
EI form 2	70	23.39	6.07	8.5	44.0	[21.94, 24.83]
EI form 3	68	24.35	5.92	14.0	39.5	[22.91, 25.78]
EI form 4	59	25.42	6.47	15.5	42.0	[23.73, 27.10]
EI overall	261	24.19	6.14	8.5	44.0	[23.44, 24.94]

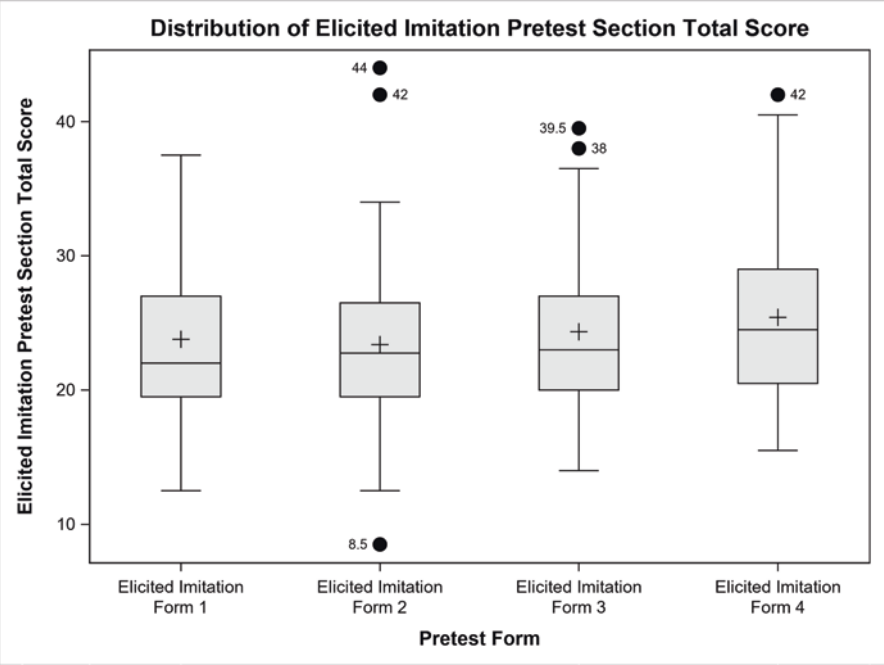


Fig. 1 Boxplot for EI pretest section-total score by test form

sentence. Eight to ten (67%–83%) of the total 12 EI items in each form were found to have average difficulty (i.e., mean item score between 1.51 and 2.50).

Regarding item discrimination, the highest polyserial correlation coefficient of any EI item was 0.78; the lowest discrimination index was 0.42, which was also the only discrimination index below 0.50. Thus, in addition to acceptable difficulty levels, all the 48 ACE-In EI items also had good discrimination, with their discrimination index above 0.40, including 47 items with an index well above the 0.40 threshold for good discrimination.

5.4 Item Analysis in a CTT Framework: Internal Consistency and Standard Error of Measurement

As shown in Table 4, the Cronbach's alpha for each EI pretest form was between 0.85 and 0.88, suggesting strong internal consistency. The standard error of measurement for the EI pretest section score was relatively small, approximately two points out of the maximum total of 48 points.

Table 3 EI pretest item difficulty and discrimination indices by form and item

	Form 1		Form 2		Form 3		Form 4	
	Difficulty	Discrimination	Difficulty	Discrimination	Difficulty	Discrimination	Difficulty	Discrimination
Item 1	1.83 (A)	0.68	1.14 (H)	0.67	1.40 (H)	0.65	1.89 (A)	0.67
Item 2	2.19 (A)	0.75	2.25 (A)	0.71	1.41 (H)	0.67	2.25 (A)	0.75
Item 3	2.61 (E)	0.57	2.17 (A)	0.55	2.18 (A)	0.55	1.68 (A)	0.69
Item 4	2.04 (A)	0.70	2.01 (A)	0.60	1.92 (A)	0.61	1.98 (A)	0.56
Item 5	1.73 (A)	0.56	1.81 (A)	0.63	2.06 (A)	0.60	2.33 (A)	0.77
Item 6	2.23 (A)	0.66	2.59 (E)	0.70	2.35 (A)	0.78	2.14 (A)	0.75
Item 7	2.56 (E)	0.65	2.45 (A)	0.70	2.54 (E)	0.61	2.06 (A)	0.72
Item 8	2.13 (A)	0.74	2.19 (A)	0.74	2.73 (E)	0.74	2.49 (A)	0.62
Item 9	1.80 (A)	0.72	1.56 (A)	0.72	2.11 (A)	0.61	1.49 (H)	0.57
Item 10	1.05 (H)	0.64	1.86 (A)	0.59	1.86 (A)	0.76	2.40 (A)	0.64
Item 11	1.55 (A)	0.42	1.49 (H)	0.65	2.15 (A)	0.78	2.15 (A)	0.71
Item 12	2.07 (A)	0.54	1.86 (A)	0.78	1.65 (A)	0.52	2.55 (E)	0.67

Note:
Hard items (H): mean item-score from two raters' item ratings = [0, 1.50]
Average difficulty (A): mean item-score from two raters' item ratings = [1.51, 2.50]
Easy items (E): mean item-score from two raters' item ratings = [2.51, 4.00]

Table 4 EI pretest reliability assessed with Cronbach's alpha and Standard Error of Measurement (SE)

	EI Form 1	EI Form 2	EI Form 3	EI Form 4
Cronbach's alpha	0.85	0.87	0.86	0.88
SE	2.35	2.19	2.22	2.24

Table 5 Descriptive statistics for EI pre- and post-test section scores (N = 261)

	Median	Mean	SD	Min	Max	95% Confidence Interval for Mean
EI pretest	23.0	24.19	6.14	8.5	44	[23.44, 24.94]
EI post-test	25.5	26.55	5.70	13.5	44.5	[25.85, 27.24]

5.5 *ACE-in EI Pre- and Post-test Score Comparisons*

The summary statistics about the EI pre- and post-test section scores in Table 5 suggest that the pretest had a mean section score of 24.19, which was lower than the post-test mean section score of 26.55. The pretest also displayed a wider spread in the score distribution than the post-test.

Results of Shapiro-Wilk tests indicate that neither of the pre- or post-test section-total scores was normally distributed: $W_{ei\_pre} = 0.96$, $p < 0.0001$; $W_{ei\_post} = 0.96$, $p < 0.0001$. Therefore, a Wilcoxon signed-rank test was conducted to examine if significant differences existed between the pre- and post-test section scores. The Wilcoxon test results indicate that the post-test EI section score was significantly higher than the pretest section score, with a large effect size ($Z = 8.324$, $p < 0.0001$; $r = 0.52$). To put this effect size into perspective, the mean effect size in education research is Cohen's $d = 0.4$ (Héroux, 2017) while Cohen's d is the effect size measure for paired samples t-tests. For our study, the score changes from the pretest to the post-test are also reflected in the overlaid histograms and fitted normal curves in Fig. 2.

5.6 *A Close-Up View of Individual EI Items' Instructional Sensitivity*

Given the large effect size of the improvement from the pretest EI section score to the post-test score for the whole group of participants when the entire EI section containing 12 items was considered, we then proceeded to investigate the instructional sensitivity of each EI item for the purposes of test review and possible revision. All 48 EI items were categorized into two groups based on the results: sensitive items (Hedge's $g \geq 0.30$) and insensitive items (Hedge's $g < 0.30$). The range of

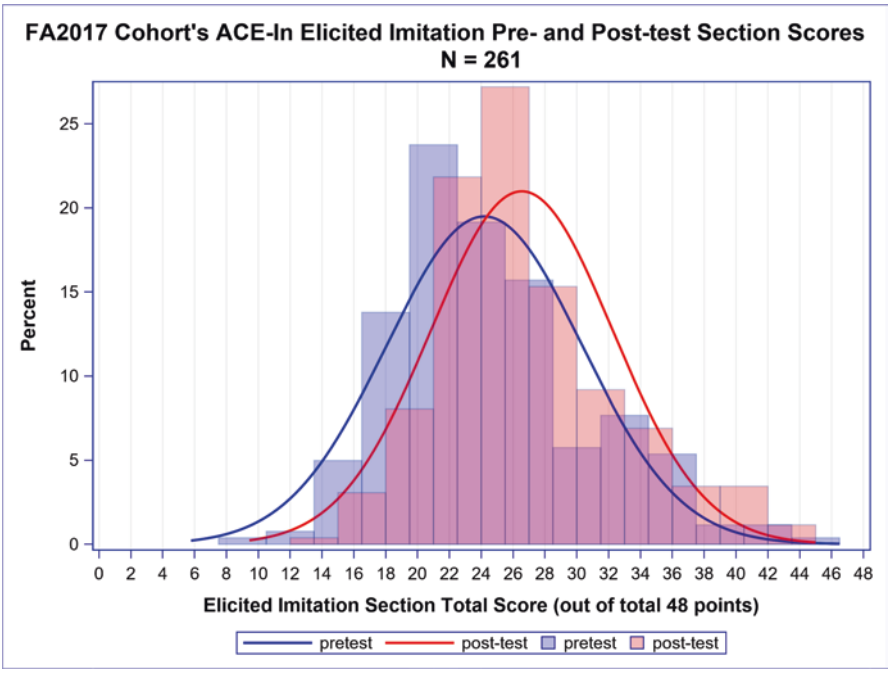


Fig. 2 EI pre- and post-test section score distributions

Hedge's *g* values was from 0.03 to 0.56 with only 17 (35.4%) items identified as instructionally sensitive. However, finding a relatively small proportion of sensitive items is not unusual; this pattern appears to be typical when tests are not completely aligned with any specific classroom curriculum (Polikoff, 2010). In relation to our local context, the 48 ACE-In EI items analyzed in this chapter were mostly developed before the PLaCE English 110 and 111 curricula were designed, implemented, revised, and stabilized (Table 6).

6 Implications for Test Developers and Users

Our study found that the ACE-In EI module has excellent measurement properties, including acceptable item difficulty, good discrimination, satisfactory test reliability, strong internal consistency, and a small error of measurement. In addition, the statistical testing and the graphical presentation in the second stage of our study suggest that our sample of 261 international first-year students made significant gains, with a large effect size, in their EI post-test performances as compared to

Table 6 EI instructional sensitivity (Hedge's *g*) by form and item

	EI Form 1	EI Form 2	EI Form 3	EI Form 4
Item 1	0.31 (S)	0.49 (S)	0.46 (S)	0.54 (S)
Item 2	0.11	0.17	0.34 (S)	0.17
Item 3	0.10	0.20	0.26	0.34 (S)
Item 4	0.15	0.24	0.36 (S)	0.27
Item 5	0.38 (S)	0.56 (S)	0.20	0.40 (S)
Item 6	0.16	0.30 (S)	0.16	0.23
Item 7	0.09	0.08	0.10	0.11
Item 8	0.19	0.17	0.24	0.31 (S)
Item 9	0.32 (S)	0.56 (S)	0.03	0.51 (S)
Item 10	0.15	0.42 (S)	0.23	0.49 (S)
Item 11	0.07	0.19	0.14	0.07
Item 12	0.08	0.18	0.11	0.08

Note:
Sensitive items (S): Hedge's $g \geq 0.30$
Insensitive items: Hedge's $g < 0.30$

their pretest performances nearly seven months earlier. Hence, the whole unit of 12 ACE-In EI items in each form was sensitive enough to capture the changes in student abilities as a function of the ESL support that students received in English 110 and 111.

However, when individual EI items were examined for pre-post score changes, only 17 (35.4%) out of the total 48 EI items were found to be *instructionally sensitive*, with their Hedge's *g* value equal to or above 0.30. A small Hedge's *g* value could have resulted from one or more of these reasons:(1) the particular EI item was insensitive to instruction-derived improvement of students' implicit knowledge about the specific lexical, grammatical, and syntactic points tested in the item; (2) the instruction in the English 110 or 111 course did not fully match this EI assessment item in terms of the lexical, grammatical, and syntactic aspects covered; and (3) the EI item was relatively easy for the high-intermediate proficiency (CEFR B2) learners in our study sample, allowing relatively little room for improvement in the post-test, and thus, there was a weak effect size on the score gains before and after instruction.

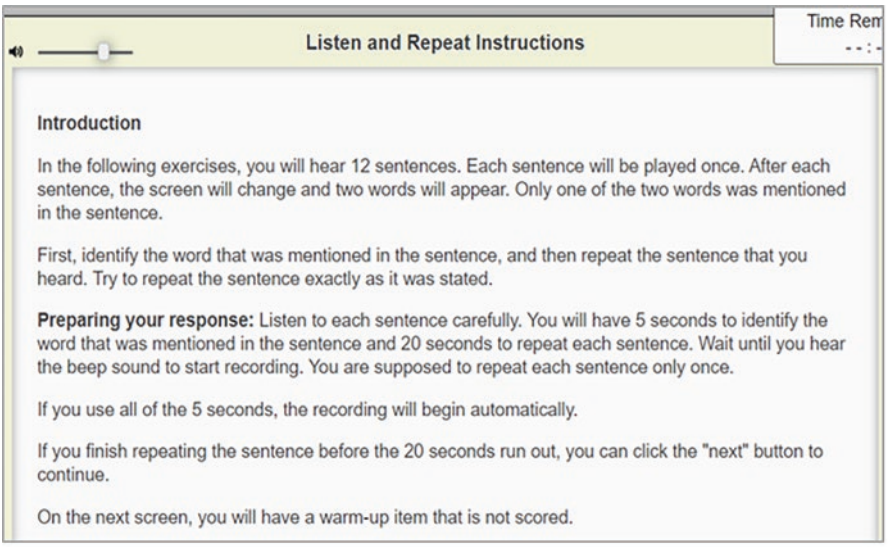
A closer examination of the individual EI prompt sentences found to have good instructional sensitivity led to the following speculations about which linguistic features may contribute to high instructional sensitivity in an EI item for examinees similar to our English 110 and 111 students. First, EI prompt sentences, with a complicated syntactic structure, tend to be associated with good instructional

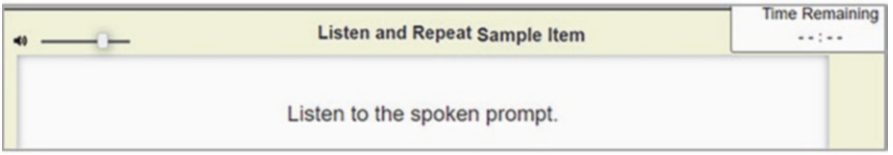
sensitivity when the effect size is calculated for the pre- and post-test item score changes. An example from our study is the prompt sentence, "The way that English classes are taught here might differ from the way in your country." A common type of grammatical error in our examinees' reproductions, especially during the pretest, is to compare "the way that English classes are taught here" with "your country." Second, certain formulaic sequences can cause difficulty, especially during the pretest when students may not have had enough exposure to or familiarity with the set phrase. Consequently, the inclusion of such formulaic sequences can contribute to good instructional sensitivity of an EI item. An example from our study is the prompt sentence, "Sometimes it's helpful to ask questions in class as opposed to keeping them to yourself." The formulaic sequence, "as opposed to" in this EI sentence, has clearly caused difficulty to many examinees, especially so during the pretest. Students' performance with this set phrase saw a significant improvement during the post-test. Third, parallel structures also can be associated with good instructional sensitivity, probably due to the cognitive demand required for processing and reconstructing such structures, in addition to the main syntactic components in the prompt sentence. A case in point is this prompt sentence: "Regular exercise is extremely important for long-term health and well-being." A deviant version for the parallel structure of "long-term health and well-being" is "long-term being and health" which has shown up in many English 110 or 111 students' reconstructed versions of the prompt sentence.

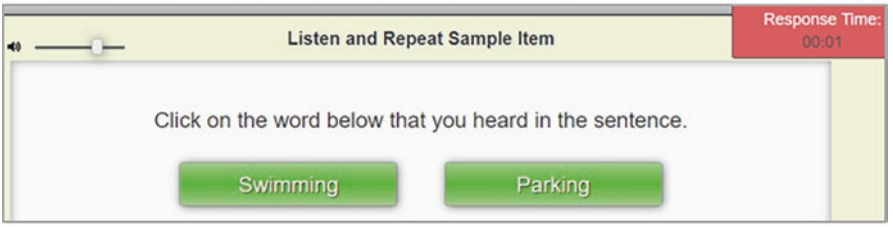
In this chapter, we presented a detailed account of using an in-house academic English proficiency test, i.e., the ACE-In EI module, to track international undergraduate students' ESL proficiency development over the course of nearly seven months when the students were engaged in a two-semester sequence of two language and cultural support courses. We believe that the CTT analyses of EI items' measurement qualities, the pre-post tracking of students' L2 general proficiency development, and the item-level analysis of instructional sensitivity demonstrated in this chapter will offer practical takeaways to language program administrators and testing specialists. These analyses, while not super complicated or technically demanding, have allowed us—program administrators, language educators and testers—to perform the following critical functions and responsibilities for the PLaCE program: (1) to review test quality and performance, (2) to gather evidence for program effectiveness to support and sustain our program, (3) to extend the validity argument for the ACE-In EI module, and (4) to collect detailed item-level information—such as difficulty, discrimination, and instructional sensitivity—to inform test revision, including on the areas that can be improved so that better integration of assessment and instruction will be demonstrated in the revised EI test.

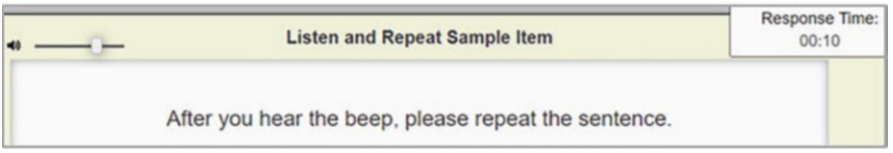
Appendices

Appendix A. Elicited Imitation ("Listen and Repeat" in ACE-In) Instructions and Warm-Up Item









Appendix B. Elicited Imitation (a.k.a. "Listen and Repeat" in ACE-In) Rating Scale

4 **Exact Repetition**	Repeating the prompt sentence word for word; however, contractions, expansions, substitutions for proper noun (e.g., names) are allowed.	**Prompt Example:** *Purdue students have free access to printing on campus.* **Performance Examples:** *a. Purdue students have free access to printing on campus.* *b. University students have free access to printing on campus.* *c. All students have free access to printing on campus.*
3 **Appropriate Paraphrasing**	Did not repeat word for word but paraphrase the prompt sentence in such a way that the response is grammatical and remains the same meaning as the prompt.	**Prompt Example:** *Purdue students have free access to printing on campus.* **Performance Examples:** *a. Purdue students can print their documents for free on campus.* *b. Purdue students enjoy free printing service on campus.* *c. You can have free access to printing on campus.* *d. Students print for free on campus.* *e. Students print free on campus.*
2 **Minor Deviation**	Repeating the prompt sentence with minor grammatical errors which do not distort meaning; or missing minor information which does not change the main idea of the prompt sentence; or a combination of both.	**Prompt Example:** *Purdue students have free access to printing on campus.* **Performance Examples:** *a. Purdue students have free access to printers on campus.* *b. Purdue students have free access to printing.* *c. Student has free access to printing on campus.*
1 **Major Deviation or Inadequate Response**	Repeating the prompt sentence with major grammatical errors which distort meaning; or missing substantial information which changes the main idea of the prompt sentence; or a combination of both.	**Prompt Example:** *Purdue students have free access to printing on campus.* **Performance Examples:** *a. Purdue students print on campus.* *b. Students have free printers on campus.* *c. Purdue students have free access on campus.*
0 **Omission**	Complete silence, or only one or few words repeated which do not make sense by itself.	**Prompt Example:** *Purdue students have free access to printing on campus.* **Performance Examples:** *a. Purdue* *b. Students* *c. Free printing* *d. Etc*
N/A	Technical Difficulty	

References

Bachman, L. (2004). *Statistical analyses for language assessment*. Cambridge University Press.

Baten, K., & Cornillie, F. (2019). Elicited imitation as a window into developmental stages. *Journal of the European Second Language Association, 3*(1), 23–34. https://doi.org/10.22599/jesla.56

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (2008). Test score interpretation and use. In C. A. Chapelle, M. K. Enright, & J. M. Jamieson (Eds.), *Building a validity argument for the test of English as a foreign language* (pp. 1–26). Routledge.

Cheng, L. (2018, September). *Extending the validity argument for an in-house ESL proficiency test through test score gains* [conference presentation abstract]. Midwest Association of Language Testers (MwALT), Madison, WI, United States. https://www.purdue.edu/place/resources/abstracts.html

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. CBS College Publishing.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443–507). American Council on Education.

Davis, L., & Norris, J. (2021). Developing an innovative elicited imitation task for efficient English proficiency assessment. *ETS Research Report, RR–21–24*. Educational Testing Service. https://onlinelibrary.wiley.com/doi/full/10.1002/ets2.12338

Ebel, R. L. (1965). *Measuring educational achievement*. Prentice-Hall.

Educational Testing Service. (2019). Performance descriptors for the TOEFL iBT® test. https://www.ets.org/s/toefl/pdf/pd-toefl-ibt.pdf

Ellis, R. (2005). Measuring implicit and explicit knowledge of a second language: A psychometric study. *Studies in Second Language Acquisition, 27*(2), 141–172. https://doi.org/10.1017/S0272263105050096

Ellis, R. (2009a). Implicit and explicit learning, knowledge and instruction. In R. Ellis, S. Loewen, C. Elder, R. Erlam, J. Philp, & H. Reinders (Eds.), *Implicit and explicit knowledge in second language learning, testing and teaching*. Multilingual Matters.

Ellis, R. (2009b). Measuring implicit knowledge and explicit knowledge of a second language. In R. Ellis, S. Loewen, C. Elder, R. Erlam, J. Philp, & H. Reinders (Eds.), *Implicit and explicit knowledge in second language learning, testing and teaching*. Multilingual Matters.

Erlam, R. (2006). Elicited imitation as a measure of L2 implicit knowledge: An empirical validation study. *Applied Linguistics, 27*(3), 464–491. https://doi.org/10.1093/applin/aml001

Fischer, K. (2014, November 17). *International-student numbers continue record-breaking growth*. Chronicle of Higher Education. https://www.chronicle.com/article/international-student-numbers-continue-record-breaking-growth/

Fulcher, G. (2013). *Practical language testing* (2nd ed.). Hodder Education.

Ginther, A., & Yan, X. (2018). Interpreting the relationships between TOEFL iBT scores and GPA: Language proficiency, policy, and profiles. *Language Testing, 35*(2), 271–295. https://doi.org/10.1177/0265532217704010

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics, 6*(2), 107–128. https://doi.org/10.3102/10769986006002107

Héroux, M. (2017, July 27). *Cohen's d: How to interpret it?* Scientifically sound: Reproducible research in the digital age. https://scientificallysound.org/2017/07/27/cohens-d-how-interpretation/

Hsieh, A. F. Y., & Lee, M. K. (2014). The evolution of elicited imitation: Syntactic priming comprehension and production task. *Applied Linguistics, 35*(5), 595–600. https://doi.org/10.1093/applin/amu036

Jessop, L., Suzuki, W., & Tomita, Y. (2007). Elicited imitation in second language acquisition research. *Canadian Modern Language Review, 64*(1), 215–238. https://doi.org/10.3138/cmlr.64.1.215

Kostromitina, M., & Plonsky, L. (2021). Elicited imitation tasks as a measure of L2 proficiency: A meta-analysis. *Studies in Second Language Acquisition*, 1–26. https://doi.org/10.1017/S0272263121000395

Larsen-Freeman, D., & Long, M. H. (1991). *An introduction to second language acquisition research*. Longman.

Levelt, W. J. M. (1989). *Speaking: From intention to articulation*. The MIT Press.

Li, X. (2020). *The technical qualities of the elicited imitation subsection of the assessment of college English, international (ACE-in)*. [Unpublished doctoral dissertation]. Purdue University.

Ling, G., Powers, D. E., & Adler, R. M. (2014). *Do TOEFL iBT® scores reflect improvement in English-language proficiency? Extending the TOEFL iBT validity argument. ETS Research Report, RR–14–09*. Educational Testing Service. https://doi.org/10.1002/ets2.12007

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher, 18*(2), 5–11. https://doi.org/10.3102/0013189X018002005

Moses, T. (2017). A review of developments and applications in item analysis. In R. Bennett & M. von Davier (Eds.), *Advancing human assessment* (pp. 19–46). Springer.

National Academy of Sciences. (2011). *Assessing 21st century skills: Summary of a workshop*.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). McGraw-Hill.

Ortega, L., Iwashita, N., Rabie, S., & Norris, J. (1999). *A multilanguage comparison of measures of syntactic complexity*. University of Hawai'i, National Foreign Language Resource Center.

PLaCE. (n.d.) *Courses*. Purdue language and cultural exchange. Retrieved March 17, 2022, from https://www.purdue.edu/place/courses/index.html

PLaCE. (2016). *ACE-In Elicited Imitation Test Specs*.

Polikoff, M. S. (2010). Instructional sensitivity as a psychometric property of assessments. *Educational Measurement: Issues and Practice, 29*(4), 3–14. https://doi.org/10.1111/j.1745-3992.2010.00189.x

Purdue University. (n.d.) *Course listing*. Purdue University Office of the Provost. Retrieved March 17, 2022, from https://www.purdue.edu/provost/students/s-initiatives/curriculum/courses.html

Spada, N., Shiu, J. L. J., & Tomita, Y. (2015). Validating an elicited imitation task as a measure of implicit knowledge: Comparisons with other validation studies. *Language Learning, 65*, 723–751. https://doi.org/10.1111/lang.12129

Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research & Evaluation, 9*, Article 4. https://doi.org/10.7275/96jp-xz07

Suzuki, Y., & DeKeyser, R. (2015). Comparing elicited imitation and word monitoring as measures of implicit knowledge. *Language Learning, 65*(4), 860–895. https://doi.org/10.1111/lang.12138

Van Moere, A. (2012). A psycholinguistic approach to oral language assessment. *Language Testing, 29*(3), 325–344. https://doi.org/10.1177/0265532211424478

Yan, X. (2020). Unpacking the relationship between formulaic sequences and speech fluency on elicited imitation tasks: Proficiency level, sentence length, and fluency dimensions. *TESOL Quarterly, 54*(2), 460–487. https://doi.org/10.1002/tesq.556

Yan, X., Maeda, Y., Lv, J., & Ginther, A. (2016). Elicited imitation as a measure of second language proficiency: A narrative review and meta-analysis. *Language Testing, 33*(4), 497–528. https://doi.org/10.1177/0265532215594643

Exploring the Alignment of Test Tasks with Curriculum Goals among Finnish 6th Graders

Marita Härmälä and Raili Hilden

Abstract This study investigates the alignment of the written, official curricula and a national test of learning outcomes at the end of 6th grade of basic education in Finland. The research questions address (1) the alignment of content standards regarding categorical thematic concurrence, (2) the complexity of knowledge required for solving the item, and (3) the proficiency level needed for correct answers and attainment of the prescribed target levels. The test items were analysed against the thematic categories in the curricula, against the knowledge dimensions of the revised Bloom's taxonomy, and by determining their CEFR levels through a standard-setting procedure. The results suggest that the curricular objectives of text interpretation skills were richly represented in the test items in reading and listening, while all the five objective domains were evenly covered by items measuring text production skills. Contrary to expectations about digital delivery, even interaction could be measured. The most frequent content areas were the student him/herself, family, friends, hobbies, and free time. For the alignment in terms of proficiency level, the task difficulty across the four skills was appropriate. Future prospects for designing digital test tasks in accordance with the communicative objectives found in the official curricula are discussed.

Keywords Local language testing · Common European framework of reference (CEFR) alignment · Standard setting · Difficulty · Bloom's Taxonomy · Finland

M. Härmälä (✉)
Finnish Education Evaluation Centre, Helsinki, Finland
e-mail: marita.harmala@karvi.fi

R. Hilden
Department of Education, University of Helsinki, Helsinki, Finland
e-mail: raili.hilden@helsinki.fi

X. Yan et al. (eds.), *Local Language Testing*, Educational Linguistics 61,
https://doi.org/10.1007/978-3-031-33541-9_7

1 Introduction

In Finland, comprehensive school education starts when the child is seven and continues for 9 years. Basic education is free for everyone, regardless of a family's socio-economic background and place of residence. Every child has equal opportunities to learn and achieve at his/her full potential. Gratis education and equal opportunities for everyone are examples of values, which manifest in learning objectives and teaching activities described in the national core curriculum. The core curriculum also defines the key content for all school subjects. The national core curriculum is a framework around which local curricula are designed (https://minedu.fi/en/comprehensive-school). The latest core curriculum was revised in 2014, but it has been updated recently by including descriptive criteria for all different subjects and grades.

English is the most frequently studied foreign language, starting at the age of 9 (for this cohort). By the end of grade six, the pupils have received 16–18 yearly hours of instruction in English language meaning 608–684 class lessons taught by a class teacher or a specialised subject teacher (https://www.finlex.fi/fi/laki/alkup/2018/20180793).

In contrast to many other countries, there has been no school inspection system in Finland since the 1990s. Consequently, there are no final exams at the end of compulsory education. The attainment of curricular goals is evaluated in predetermined intervals for the various school subjects. For foreign languages, the cycle is somewhat longer than for L1 and mathematics, but recently a more regular longitudinal administration has been established. The evaluations are state-mandated, and their purpose is to gather information for educational policymaking, legislation, and for revising the national core curriculum. The education providers and schools also get their own results compared with the national averages to be used for self-evaluation purposes. In Finland, all the municipalities are education providers as it is their responsibility to organize compulsory education for their inhabitants. Consequently, one education provider may have several schools on its responsibility.

The curricular goals of language education comprise general and subject-specific aims (NCC, 2014). The general aims draw on global, European and national frameworks for desired human and social development, alongside the transversal competences required for coping with the challenges of a sustainable future (OECD, 2021; Council of Europe, 2018). The subject-specific aims are derived from the ideas of communicative language teaching, most tangibly embodied in the CEFR (2001) and its companion volume (2018).

In this study, we investigate to what extent the content aims and outcome statements for good skills at the end of 6th grade, align with the tasks used in a national evaluation of learning outcomes. Contrary to many other countries, Finland has not had school supervisors since the 1980s. In addition, textbooks and study materials are no longer controlled. Finnish teachers are virtually autonomous and deeply trusted (Salminen, 2021) partly because they have an academic degree in the subject

they teach. Consequently, the national evaluations of learning outcomes are the only way to gather data on how well the goals of the core curriculum are achieved in different parts of the country and if there are any differences in attainment of outcomes across genders, language groups, and family backgrounds.

The analysis sheds light on the alignment of the test tasks and the written core curricula with respect to the categorical concurrence and the complexity of knowledge, but it also gives certain insight in the balance of representation (Webb, 2007, p. 14). Furthermore, we explore the difficulty levels of the test items as compared to the target levels defined in the curricula.

Despite the pivotal role of the curricular alignment of a nationwide assessment, research initiatives in this line have been scarce to date, especially those addressing young learners. Moreover, to our knowledge there are no studies applying the Bloomian taxonomy in language education in Finland despite its popularity in other school subjects. In the alignment process, we explore how usable the Bloom's taxonomy (1956) revised by Anderson and Krathwohl (2001) would be for analysing the cognitive content and knowledge dimensions of foreign language test tasks. This endeavour is motivated by a current Finnish trend of non-language school subjects to describe the multiple levels of knowledge, skills and abilities required for different school grades to design, for example, test tasks.

1.1 Research Questions

This chapter uses both qualitative and quantitative data to address three research questions:

RQ1 What are the links between the test tasks/items and the learning goals and content areas of the national core curriculum? The first research question investigates to what extent the test tasks and items correspond to the learning goals and performance criteria (can do statements) set for good skills in the NCC. There are 10 teaching goals in the NCC (Appendix A): four clearly linguistic (goals 7–10) and other six more general, such as growing into cultural diversity and language awareness (goals 1–4) and language learning skills (goals 5–6). The non-linguistic goals were not explicitly included in the test and were therefore excluded from the present analysis.

RQ2 What are the links between the cognitive processes and knowledge dimensions required by the test tasks? The second research question investigates whether Bloom's revised taxonomy is applicable to defining the cognitive processes and knowledge dimensions included in a test of language proficiency. To do this, the cognitive processes and knowledge dimensions required by the test tasks were first analysed and then related to the NCC learning goals.

RQ3 To what extent does the difficulty of the items correspond to the CEFR level for good skills in the NCC? To answer the third research question, a standard setting procedure for defining the cut-scores in reading and listening was organised.

It was possible to set two cut-scores for both skills. The method used was the 3 DC (Keuning et al., 2017) and IRT. Additional support for judging the appropriateness of the difficulty level of the tasks was gathered from the seventh grade teachers (Härmälä et al., 2019).

In sum, our study examines the alignment of the English test tasks for Finnish 6th graders with the learning goals and content areas of the national core curriculum. In doing this, we gather evidence for the content validity of the test to be used in the future development of the national evaluations of learning outcomes.

2 Alignment

Alignment in general terms refers to a match between two categories. In a school context, it means the links between the course objectives and the teachers' instructional activities on the one hand, and between the course objectives and assessment on the other. One way to explore the content validity of a test is to investigate its compliance with teaching of a specific curriculum. (Bachman & Palmer 2010; Kane 2001, 2013; Weir, 2005). A reverse strand of inquiry is to investigate the washback effect of testing on teaching (Beikmahdavi, 2016; Chaofeng, 2019). In the current study, we primarily address the content validity of the national evaluation by exploring its relationship with the core curriculum.

In recent decades, alignment of test items with the curricular content has gained increased attention (Papageorgiou et al., 2020). This tendency has been intensified by the accountability movement, but also in pursuit of identifying areas that need improvement in schools or school districts (Papageorgiou, 2016). The importance of alignment is widely acknowledged today as a prerequisite for fair assessment (Weideman, 2019), as a way to promote positive washback (Sultana, 2018) and to narrow the gap between the official, written document, its classroom implementation (the taught curriculum), and the assessment of the particular content (the tested curriculum) (Squires, 2012).

The correspondence between written curriculum and assessment can be judged in four major aspects (Webb, 2007). Categorical concurrence determines whether the curricula and the assessment address the same content categories, such as themes, topics or functions. The second category describes the complexity of knowledge required by the curricula on one hand, and by the assessment, on the other. The third angle is the breadth of knowledge, and ultimately, the balance of representation "indicate the degree to which one objective is given more emphasis in the assessment than another" (Webb, 2007, p. 14). The current study primarily focuses on categorical concurrence and the complexity of knowledge (Webb, 2007), but also gives certain insight in the balance of representation.

Webb's complexity of knowledge is closely comparable with the concept of cognitive level. The most famous classification of cognitive objectives of school learning is the Bloom's taxonomy, revised by Anderson and Krathwohl

(Krathwohl, 2002; Seaman, 2011). The core of the taxonomy is to rate cognitive processes along four knowledge dimensions: from the lower process of remembering, understanding, applying, and analysing to the higher processes of evaluating and creating new knowledge. The knowledge dimension covers four types of content: factual, conceptual, procedural and metacognitive knowledge. In this study we use the two-dimensional grid as an analysis tool to classify the objectives stated in the core curricula.

The revised Bloom's taxonomy has seldom been applied to language subjects, but its orientation in general knowledge domains, deserves increased attention in language education. In a recent study on 8th grade English curriculum goals, Kozikoğlu (2018) observed the priority on low order thinking skills and the application of declarative knowledge compared with higher order thinking skills.

In the language education field, a number of studies have been carried out with students at the basic education age. Papageorgiou et al., (2020) explored the alignment of the TOEFL Primary language tests and an online course in English as foreign language. Their findings suggest a high correspondence of the learning activities offered by the course and the content of test-tasks. Pourdana and Rajeski (2013) explored the use of the revised Bloom's taxonomy in determining the difficulty level of reading comprehension items for EFL university students and found that the taxonomy was "compatible with the currently used models of material development" (p. 206). Samira et al. (2021) scrutinized IELTS listening and reading tests through the lens of the revised Bloom's taxonomy. Their results suggested a considerable misalignment between the test items and the learning objectives with regard to the higher levels.

3 Assessing Young Learners' Language Skills

The focus of this study is on 12–13 years old students who are still in the midst of development in many ways. The young age of the students set certain limits on the tasks used i.e., in relation to their cognitive demands and level of abstraction. There are four stages in the development of memory, understanding and verbal reasoning and the kind of task that are "doable" for adolescents aged 12/13 (Hasselgreen & Caudwell, 2016). Teenagers (13–17 years) are able to sort essential information from non-essential information, tackle abstract problems whereas problem-solving is still reliant on real-life experiences and world-knowledge often remains limited (ibid, 5). Fifteen-year-old students are able to critically evaluate texts, recognize conflicting viewpoints and facts from fiction (ibid, p. 10). Their literacy development is, however, influenced by environmental conditions, such as family, peer group, and classroom, causing great individual variation in literacy.

Age can affect both the abilities and the domains of language use of young learners (Hasselgreen & Caudwell., 2016, p. 22). Abilities depend on how concrete or abstract, immediate or distant the concepts are, and also on the ability to link ideas to coherent texts, to see the world from the perspective of others. The domain of use

expands and is impacted by the growing understanding of the world and the context of learning. In the context of our study, the learning target was expressed as the CEFR level for good skills in the NCC. The good level corresponds to the grade 8 on a scale from 4 to 10 and it is defined as CEFR level A2.1 for all the major skills.

4 The National Core Curriculum as a General Guideline

The NCC is a general guideline for all teaching and learning in Finland. It is a steering document for the schools and their activities as well as a framework where the goals and contents of different school subjects are described to be used by the teachers of different subjects. The Finnish national core curricula for basic and general upper secondary education cover, in addition to all the other school subjects, both foreign languages as well as the first languages for bilingual children (NCC, 2014). The CEFR levels are split in three bands at A1 level and in two at all the other levels thus allowing to distinguish even smaller developmental steps. The current NCC from 2014 defines the level of good skills for the transition phases of basic education, that is 6th and 9th grade. The corresponding CEFR levels are summarized in Table 1. Table 1 also gives as an example the content areas for English from grades 1 to 9 in relation to the knowledge and skills required. A general guideline for all age groups is to use everyday texts that are age-relevant and of interest for the students. The texts may be chosen together with the students.

For grades 3–6 and 7–9, interaction skills, text interpretation skills, and text production skills are combined. The more precise definitions of good language skills can be found in Appendix B.

According to Hasselgreen & Caudwell (2016), teenagers between 8/9 years and 12/13 years are able to attain CEFR level B1 while for the older age group (13–17) even B2 would be attainable. However, there are large individual and cultural differences both downwards and upwards. The Finnish National Agency for Education has defined the level for good skills and knowledge as CEFR A2.1 as this level has been considered appropriate in the Finnish context for the students who have studied English on average 3 years and have received approximately 304 h of instruction (NCC, 2014).

5 Methods

The evaluations of learning outcomes are organised in different school subjects. Mathematics and L1 are evaluated every 4th year, the other subjects every 10th or 15th year. The latest evaluation in English was organised in spring 2021 and administered to 9th graders. It was the second phase of a longitudinal study; the first measurement having taken place in 2018 at the beginning of 7th grade. Both evaluations included tasks in all the major skills and the results were reported in terms of the Finnish application of the CEFR scales.

Table 1 Content areas and CEFR levels for good skills in the NCC for long syllabus English

Content areas related to teaching goals	Grades 1–3	Grades 4–6	Grades 7–9
1. Interaction skills	greeting, saying goodbye, thanking, asking for help, telling about oneself practicing polite language through songs, play, drama and gaming practicing to interfere the meaning of words on basis of the context, world knowledge and skills in other languages practicing to cope with limited range of language reacting in a natural way in interactive situations	learning to listen, speak, read and write about many kinds of topics, such as me and myself, family, friends, school, hobbies and freetime, living in an English speaking environment actual, current topics greeting, asking for help, expressing opinions practicing vocabulary by means of different text types: stories, plays, interviews, song lyrics giving opportunities to practice more demanding language use situations learning to find materials in English, i.e., from the surroundings, internet, library observing and practicing pronunciation, word/sentence stress, intonation practicing to recognize phonetic symbols	functioning in English in different communities actual topics preparing to continue to upper secondary education getting familiar with the language skills young people need in working life and studies participating as an active agent both locally and globally learning vocabulary and structures in many kinds of texts such as narrative, descriptive and instructive texts observing and practicing different kinds of interaction situations by using various communication channels
2. Text interpretation skills	getting familiar with spoken texts first, then written texts practicing understanding through songs, plays, stories and pictures listening and observing pronunciation, word/sentence stress, intonation		
3. Text production skills	abundant practicing of pronunciation, word/sentence stress, intonation practicing vocabulary and structures as part of interaction with the help of pictures, songs, plays, stories, gaming providing gradually opportunities to get familiar with writing		

(continued)

Table 1 (continued)

Content areas related to teaching goals	Grades 1–3	Grades 4–6	Grades 7–9
4.Assessment	to guide learning with the help of supportive feedback practice peer and self-assessment various ways of showing own learning	the pupil becomes aware of his/her own skills, develops them by emphasizing the most natural modes of expression supportive, versatile methods (i.e., ELP) grade given in relation to the goals set in the local curriculum and the national assessment criteria peer and self-assessment all major skills taken into account	versatile methods (i.e. ELP), peer and self-assessment all the 4 major skills supportive and formative feedback all the national assessment criteria taken into account possibility to compensate for not achieving the goal in some of the content areas
5.CEFR level	No CEFR level given	A2.1	B1.1

The data were gathered by using stratified random sampling to ensure regional representativeness. The sample consists of 4633 pupils from 132 schools, of which 116 were Finnish-speaking and 16 Swedish-speaking. The pupils were 12/13 years old, and they had just started the 7th grade which is a transition point between the lower and higher phase of the comprehensive school. Due to this, some of the students had recently changed schools and teachers.

The material consists of tasks measuring the students' language proficiency (28 items in listening and 27 in reading, 3 in speaking and 2 in writing). The tasks were designed by five English teachers working in comprehensive schools or general upper education. Additionally, two university professors specialising in English language teaching and learning participated in the evaluation group. The first author of this article was responsible for the item writing process as well as the entire evaluation project.

The tasks were designed on the basis of the NCC objectives and assessment criteria for good skills. They were piloted among 404 6th graders from both Finnish and Swedish speaking schools (n = 8) and the items with a reasonably high item-rest correlation and an adequate fit with 3-parameter IRT-model were chosen for the final test. The test was administered digitally via a net browser. The national evaluations are not high stakes for the students as the major aim of the evaluations is to produce data for national and local decision making. However, the teachers are encouraged to use the results to support their student assessment.

The quantitative data were analysed on school and student level by making use of multilevel modelling (SPSS). For the alignment purposes, the texts of the national core curriculum were analysed qualitatively by means of content analysis (RQ1) and by applying a grid based on the cognitive processes and knowledge dimensions of Bloom's taxonomy (RQ2). The CEFR levels were set by using the 3 DC method (Keuning et al., 2017) and IRT analysis (RQ3).

The 3 DC procedure divides the complete test into a number of clusters and uses empirical data and an item response model to relate the scores of the clusters to the scores of the complete test. The relationships between the clusters and the complete test are presented to the subject-area experts on a specially designed assessment form. The experts were asked to use the assessment form to indicate the score that students would be expected to achieve in each cluster if they were exactly on the borderline of proficiency. Because of the design of the assessment form, the assessment is associated with both content information and empirical data. In the present study, the panel comprised 10 teachers of English and other experts in English language teaching, learning and assessment.

The links between the 60 test tasks/items and NCC's learning goals and content areas were investigated item by item. The knowledge dimension covers factual, conceptual, procedural and metacognitive types of content. We used a two-dimensional grid as an analysis tool to classify the NCC objectives. We included the following contents in the knowledge dimension:

- factual knowledge: alphabet, vocabulary, grammatical terms
- conceptual knowledge: parts of phrases, principles of grammar
- procedural knowledge: subject-specific skills, building up texts and phrases, functioning in a cultural context, registers, reading strategies
- metacognitive knowledge: strategic knowledge, appropriate contextual knowledge, self-knowledge, setting learning goals, self-assessment

The cognitive process dimension, we defined as follows:

- to remember: to retrieve from long term-memory
- to understand: to construct meaning from texts
- to apply: to carry out a procedure in a situation
- to analyse: break into parts and determine how the parts relate to one another
- to evaluate: to make judgements
- to create: to put elements together to form a coherent whole

6 Results and Discussion

The findings are reported by answering each of the research questions in turn.

RQ1. What are the links between the test tasks and the learning goals and contents of NCC?

First, the descriptors were cut in shorter sentences. Then, each item's correspondence with the descriptions of the Evolving language proficiency scale (CEFR level A2.1) and the content goals of the national core curriculum was defined individually by the authors. The classifications were 99% unanimous, the only differences being in one of the speaking tasks. When a consensus was found, the final occurrences were counted item wise. Almost 90% of the items matched the descriptors in reading and listening (Table 2, Chap. 8). The least frequent text interpretation skill was

"the student is capable of very simple reasoning supported by the context", which is appropriate given the young age of the students and the relatively low CEFR level targeted.

For text production skills, all the five dimensions of the proficiency scale were fairly evenly represented in the items (from 5.0% to 8.3%). For the interaction skills, the results were interesting as it was thought at the test design phase that interaction would be only with difficulty suited for the digitised test format. However, it was possible to measure at least to some extent how the students exchange thoughts and information in familiar, everyday situations and use most polite common greetings and forms of address. The most challenging skills were "asking for clarification" and "applying the expressions used by the communication partner", which is understandable again considering the test format and channel. Summary of the analysis is in Table 2.

Table 2 Test items in relation to the Evolving language proficiency scale descriptions for good skills

Evolving language proficiency scale (level A2.1) P1–P19		N	%
Interpretation skills			
P1	Understands written texts	55	92
P2	Understands clear speech that include simple, familiar vocabulary and expressions	56	93
P3	Understands the core contents of short and simple messages that are of interest to him/her	56	93
P4	Understands the main points of a predictable text containing familiar vocabulary	55	92
P5	Is capable of very simple reasoning supported by the context	16	27
Production skills			
P6	Is able to describe every day and concrete topics using simple sentences and concrete vocabulary	4	77
P7	Is able to describe topics important to him/her using simple sentences and concrete vocabulary	4	77
P8	Masters an easily predictable basic vocabulary	5	8
P9	Masters many key structures	4	77
P10	Knows how to apply some basic rules of pronunciation, also in expressions that have not been practised	3	5
Interaction skills			
P11	Exchange thoughts or information in familiar, everyday situations	4	77
P12	Can occasionally maintain a communication situation	2	3
P13	Participates increasingly in communication, resorting to non-verbal expressions less often	2	3
P14	Needs to ask for clarification or repetition quite frequently	0	0
P15	Is somewhat able to apply the expressions used by the communication partner in his/her communication	1	22
P16	Can manage short social situations	2	3
P17	Is able to use the most polite common greetings and terms of address	3	5
P18	Is able to politely address requests, invitations, proposals, apologies	2	3
P19	Is able to respond to requests, invitations, proposals, apologies	2	3

The content areas of the NCC were equally parsed comparably into shorter phrases. Then, the occurrence of each was explored in the 60 items. Here, the classifications were approximately equivalent. The most frequent content areas were the student him/herself, family, friends, hobbies and free time. These contents were extremely well presented in the tasks. Additionally, there were numerous texts on actual, current events. In sum, the students were able to practice the basic vocabulary by means of different text types as stated in the NCC content goals. What was totally missing in the tasks was "finding materials in English from the internet etc." and this was again because of the limits of the platform. Another totally missing skill was "practising to recognize phonetic symbols" as there was no item measuring this. Summary of the NCC content areas and their coverage by the items is in Table 3.

In summary, it can be concluded that both the categorical concurrence and skills coverage was as intended in the test tasks and items.

RQ2. What are the relations among the cognitive processes and knowledge dimensions required by the test tasks?

The complexity of knowledge was investigated by applying the cognitive process and knowledge dimensions of the Bloom and Krathwohl extended taxonomy. The processes range from remembering through understanding, applying and analysing to the higher processes of evaluating and creating new knowledge. The cognitive dimension is illustrated by a verb and the knowledge dimension by a noun. In our study, the learning aims of the Evolving language proficiency scale at level A2.1 (Appendix B) are summarized in Table 4. Numbers in the parenthesis indicate their occurrence.

We then placed the verbs and nouns in a two-dimensional grid by choosing a box that presents the highest order skills for the goal in question. For instance, "understand texts" would be placed in the cognitive dimension "understand" and knowledge dimension "procedural".

Table 3 The content areas of the NCC for grades 3–6 in long syllabus English

Content areas of the NCC		N	%
CA	Learning to listen, speak, read and write about many kinds of topics, such as me and myself, family, friends, school, hobbies and free time, living in an English-speaking environment	60	100
CB	Actual, current topics	60	100
CC	Greeting, asking for help, expressing opinions	4	77
CD	Practicing vocabulary by means of different text types: Stories, plays, interviews, song lyrics	56	93
CE	Giving opportunities to practice more demanding language use situations	21	35
CF	Learning to find materials in English, i.e., from the surroundings, internet, library	0	0
CG	Observing and practicing pronunciation, word/sentence stress, intonation	3	5
CH	Practicing to recognize phonetic symbols	0	0

Table 4 The verbs and nouns in the proficiency scale

Evolving language proficiency scale						
	VERB	NOUN				
Interaction skills	1. Exchange	Thoughts	Information			
	2. Maintain	Communication				
	3. Participate in	Communication				
	4. Ask for	Repetition	Clarification			
	5. Apply	Expressions				
	6. Manage	Situation				
	7. Use (2)	Greetings	Terms of address			
	8. Address	Requests	Invitations	Proposals	Apologies	
	9. Respond to	Requests	Invitations	Proposals	Apologies	
Interpretation skills	10. Understand (4)	Texts (2)	Speech	Messages	Expressions	Vocab
	11. Reason					
Production skills	12. Describe (2)	Topics (2)				
	13. Master (2)	Vocabulary	Structures			
	5. Apply	Rules				

Table 5 Cognitive processes and knowledge dimensions of the NCC proficiency scale

	Cognitive Process Dimension (verb)					
Knowledge dimension (noun)	1. Remember	2. Understand	3. Apply	4. Analyze	5. Evaluate	6. Create
A. Factual knowledge			13, 5[a]			
B. Conceptual knowledge						
C. Procedural knowledge		10, 11	12			1, 2, 3, 4
D. Metacognitive knowledge						5, 6, 7, 8, 9

[a]The numbers here refer to the verbs listed in Table 4

As Table 5 shows, the NCC proficiency scale weighs heavily on creating. This finding is not in line with Samira et al. (2021) but clearly understandable considering the relative emphasis of the NCC scale on text production skills where the students need to produce written or spoken text themselves instead of only copying or remembering things they have studied.

Table 5 is in line with the relative importance accorded to production and interaction skills in the proficiency scale (14 statements out of 19, see Table 2). However,

in the test tasks, interpretation skills clearly outnumbered tasks requiring interaction and production. This is not very surprising considering the long tradition in Finland of testing interpretation skills instead i.e., of speaking. The lack of balance is largely because of the testing mode and large number of participants. A digitally administered test has certain limits for testing i.e. speaking and interaction. Another explanation is that in a national low-stakes test like ours, the schools and individual teachers cannot be expected to use lots of time for assessing the performances if not compensated for the extra work. However, we believe that emphasising the communicative purposes and real-world values of the test could contribute to valuing speaking and writing in local instructional programs, too (see discussion in Dimova et al., 2020). Involving the stakeholders (here: teachers, principals) during the various phases of the evaluation process could also allow for the teachers to develop their skills in criterion-based assessment and thereby increasing the transparency of pupil assessment both locally and nationally.

RQ3. To what extent does the difficulty of the items correspond to the CEFR level for good skills and knowledge in the NCC?

In the standard setting, two cut scores were set (A1.3/A2.1 and B1.1/B1.1 and above.) For the purposes of the present article, the category A2.1/A2.2 is critical as it includes the level A2.1 which stands for good skills. In setting the cut-scores we used the 3 DC method (Keuning et al., 2017).

The cut-scores for CEFR levels A2.1–A2.2 ranged in listening from 12 to 22 and in reading from 13 to 22 points. After the standard setting, the items were assigned to CEFR levels through IRT-analysis and counting probabilities. We decided to use 50% probability, which means that the student has 50% probability of answering the item correctly and consequently, the item is at the same level as the student's skills. For example, in the task "Emma and Tom" (Fig. 1) a student attaining in total 15 points in listening, has answered this task correctly with 50% probability. In 2-point tasks, the 50% probability stands for 1 point.

In Fig. 1, the items are ordered in relation to their difficulty. The vertical lines stand for the items falling between levels A2.1/A2.2.

Fourteen of the items fall within the levels A2.1/A2.2 and 11 under A2.1.[1] Two items were at level B1.1 or above. These items were open-ended questions with 2 points where the student needed to be capable of very simple reasoning. It can be concluded that the listening comprehension test was at an appropriate difficulty level and also included easy items. 52% of the items were at levels A2.1/A2.2, which was the same percentage that had been targeted during the item writing phase. The listening tasks succeeded thus measuring the targeted CEFR level well. This is confirmed by the teachers (n = 238), 66% of whom considered the difficulty of the tasks to be well or extremely well suited for the age-group. The corresponding percentage for reading was a bit lower, that is 59% (Härmälä et al., 2019).

[1] Camden market: 1 item deleted; Emma and Tom included 4 T/F items that were put together.

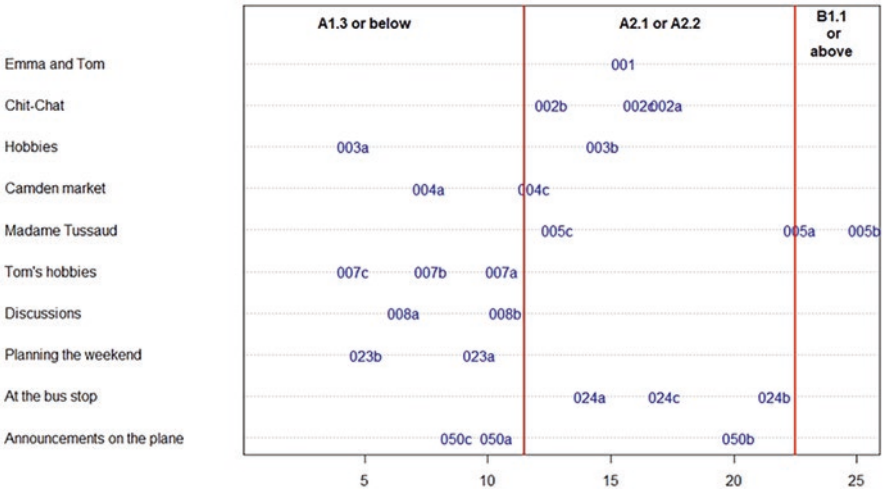


Fig. 1 Listening items ordered by difficulty

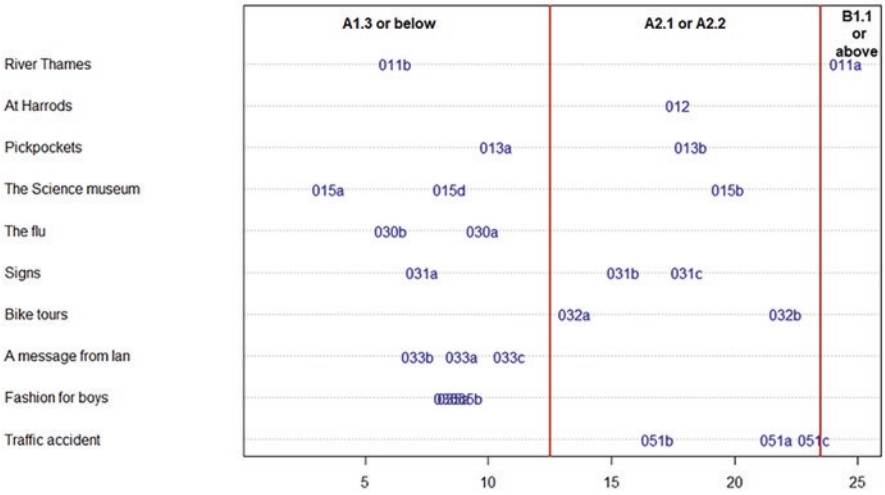


Fig. 2 Reading items ordered by difficulty

In Fig. 2, the tasks and items in reading are presented. The vertical lines stand for the items falling between CEFR levels A2.1/A2.2.

Thirteen items were below A2.1 and one at B1.1 or above. Consequently, 12 items (46%) fell within the targeted proficiency levels.[2] The difficulty level of the reading test could therefore be considered appropriate. However, 34% of the teachers thought that the difficulty of the tasks suited the age-group only satisfactory (Härmälä et al., 2019).

[2] Science Museum: 1 item deleted.

The most difficult item (River Thames) was again an open-ended question requiring basic reasoning supported by the context. The pupils were invited to deduce the answer from the implicit information given in the text. Why this item was so difficult for the pupils could be because of their young age and also their relatively low proficiency level. At the same time, short answer questions tend to be the best items in terms of discrimination and allow for the best pupils to show how much they actually understand in a text.

In writing and in speaking, the students' proficiency level was an average of the tasks for those students who completed both the writing tasks and at minimum two of the three speaking tasks. In writing, the most frequent proficiency level attained was A2.1 20%, the next frequent A1.3 (18%). In total 52% of the students achieved A2.1 or higher. In speaking, the results were quite the same: the most frequent CEFR level was A2.1 (24%), and then A1.3. In total 53% of the students attained A2.1 or higher. Seventy-five percent of the teachers (n = 238) considered the writing tasks well or extremely well suited for the age-group, the corresponding percentage for speaking being 62 (Härmälä et al., 2019).

Task difficulty in all of the four skills corresponded well to the targeted CEFR level. However, the number of items at B1.1 or higher was modest. This was understandable as the aim was to investigate to what extent the students attain the level for good skills and not the levels above it. In the future, however, including more difficult items could be considered. At the same time, care should be taken to keep a certain balance between the very easy and difficult items to make the test motivating for different kinds of pupils.

7 Insights Gained

The study explored the alignment of a national evaluation test with a core curriculum. The primary focus was on the link between the official curriculum document and the assessment tasks designed to measure the level of attainment of the objectives. The links were established first, by examining the tasks and items in relation to the learning goals described in the scale of Evolving language proficiency. Second, categorical concurrence was established by analysing the themes and topics of the tasks. Additionally, item difficulty was investigated based on the cut-scores set in a standard setting panel. A tentative aim was to apply the Bloom & Krathwohl taxonomy to analyse the complexity of knowledge required by the tasks. Based on the results, the overall appropriateness of the taxonomy appeared to be satisfactory, both as a tool for categorisation of content substance, and for checking for the alignment between curriculum goals and test items. The taxonomy fits for item writing purposes, but similar analyses need to be replicated across multiple levels of proficiency and age cohorts.

The content of the tasks and items corresponded extremely well to the learning goals and content areas described in the national core curriculum. The difficulty level of the tasks was aimed at CEFR level A2.1 and as a whole, the items were of

appropriate difficulty for the students. However, there is an imbalance between the relative weight that the core curriculum puts on productive and interactive skills versus the test content that currently emphasises interpretative skills. However, putting more weight on receptive skills is a matter of practicality. Multiple-choice items can be scored automatically by machine whereas speaking and writing require - at least for the time being - human raters, that is, teachers. In the future, human and machine ratings could be combined to lessen the workload for the schools. There could also be a more flexibly accessible test bank for the productive skills and the number of items in productive skills could then alter between evaluations and regions.

The digital test format enables multiple new ways of addressing interaction and integrative skills. There could be, for instance, more tasks in which interpretation, production and interaction were combined, for instance having tasks where the student first listens to a text and then summarizes its main points in speaking or writing. Additionally, the evaluation could focus more on one or two sets of skills at a time. And we could also reiterate using Bloom's taxonomy in a more systematic way already during the item writing phase because our analysis showed that it worked also for a foreign language. Re-considering the structure of the test would allow for the test developers to put more weight on real-life language skills, such as speaking, in line with the communicative ethos of the core curriculum.

Appendices

Appendix A: Objectives of Instruction in the A-Syllabus in English in Grades 3–6 (NAE, p. 241)

Teaching goal		Learning goals and contents derived from the teaching goals[a]	Knowledge and skills for grade 8:
Growing into cultural diversity and language awareness			
O1	To guide the pupil to notice the linguistic and cultural richness of their surroundings (11) To guide the pupil to notice the linguistic and cultural richness of the world (12) To guide the pupil to notice the status of the language s/he studies (13)	The pupil learns to notice the linguistic and cultural richness of his/her surroundings (11) The pupil learns to notice the linguistic and cultural richness of the world (12) The pupil learns to notice the status of the language s/he studies (13)	**The pupil is able...** To describe in general terms the language spoken in his/her surroundings To list the most widely spoken languages in the world To quantify the distribution of English

(continued)

Teaching goal		Learning goals and contents derived from the teaching goals[a]	Knowledge and skills for grade 8:
O2	To motivate the pupil to value his/her own linguistic and cultural background (21) To motivate the pupil to value the linguistic and cultural diversity of the world (22) To motivate the pupil to encounter people without prejudices (23)		Not used
O3	To guide the pupil to observe phenomena that are common to and different in languages (31) To guide the pupil to support the development of his/her linguistic curiosity and reasoning (32)	The pupil learns to observe phenomena that are common to and different in languages (31) The pupil learns to support the development of his/her linguistic curiosity and reasoning (32)	To make observations on the differences and similarities related to structures To make observations he the differences linguistic similarities related to vocabulary To make observations he the differences linguistic similarities related to other features of English and his/her own mother tongue or other language s/he knows
O4	To guide the pupil in finding material in the target language (41)	The pupil finds material in the target language (41)	To describe the kind of available English material that promotes his/her learning
Language learning skills			
O5	To explore the objectives of the instruction jointly (51) To create a permissive classroom atmosphere in which getting the message across has the most important role (52) To create a permissive classroom atmosphere in which encouraging learning together has the most important role (53)	The pupil learns to explore the objectives of the instruction jointly (51) The pupil learns to create a permissive classroom atmosphere in which getting the message across has the most important role (52) The pupil learns to create a permissive classroom atmosphere in which encouraging learning together has the most important role (53)	**The pupil is able** To describe the study goals The pupil participates in completing group assignments

(continued)

Teaching goal		Learning goals and contents derived from the teaching goals[a]	Knowledge and skills for grade 8:
O6	To guide the pupil to take responsibility for his/her language learning (61) To encourage the pupil to practice his/her language proficiency confidently, also using ICT (62) To encourage the pupil to experiment to find the ways of learning that are the best suited for him/her (63)	The pupil learns to take responsibility for his/her language learning (61) The pupil learns to practice his/her language proficiency confidently, also using ICT (62) The pupil learns to find the ways of learning that are the best suited for him/her (63)	The pupil sets goals for his/her language learning The pupil practices different ways of learning languages, also using ICT The pupil also improves and assesses his/her skills
Evolving language proficiency, interaction skills			
O7	To arrange opportunities for the pupil to practice spoken communication using different communication channels (71) To arrange opportunities for the pupil to practice written communication using different communication channels (72) To arrange opportunities for the pupil to practice interaction using different communication channels (73)	The pupil learns to practice spoken communication using different communication channels (71) The pupil learns to practice written communication using different communication channels (72) The pupil learns to practice interaction using different communication channels (73)	**The pupil...** Is able to exchange thoughts/ information in familiar situations Is able to exchange thoughts/ information in everyday situations Is able to occasionally maintain a communicative situation
O8	To support the pupil in using linguistic communication strategies (81)	The pupil learns to use linguistic communication strategies (81)	Increasingly participates in communication Resorts less often to non-verbal expressions Needs to ask for clarification or repetition quite frequently Is somewhat able to apply the expressions used by the communication partner to his/her own communication
O9	To help the pupil expand his/her knowledge of phrases (91) To help the pupil expand his/her knowledge of phrases that are part of respectful language use (92)	The pupil learns to expand his/her knowledge of phrases (91) The pupil learns to expand his/her knowledge of phrases that are part of respectful language use (92)	Can manage short social situations Is able to use the most polite greetings Is able to use terms of address Is able to politely express requests, invitations, proposals, apologies etc. Is able to respond to requests, invitations, proposals, apologies etc.

(continued)

	Teaching goal	Learning goals and contents derived from the teaching goals[a]	Knowledge and skills for grade 8:
Evolving language proficiency, text interpretation skills			
O10	To encourage the pupil to interpret spoken texts that are age-appropriate (101) To encourage the pupil to interpret written texts that are age-appropriate (102) To encourage the pupil to interpret spoken texts that are interesting to him/her (103) To encourage the pupil to interpret written texts that are interesting to him/her (104)	The pupil learns to interpret spoken texts that are age-appropriate (101) The pupil learns to interpret written texts that are age-appropriate (102) The pupil learns to interpret spoken texts that are interesting to him/her (103) The pupil learns to interpret written texts that are interesting to him/her (104)	**The pupil...** Understands texts that contain easy and familiar vocabulary Understands clear speech Understands the core contents of short and simple messages that are of interest to him/her Understands the gist of a predictable text containing familiar vocabulary Capable of very simple reasoning supported by the context
Evolving language proficiency, text production skills			
O11	To offer the pupil abundant opportunities for practising age-appropriate small-scale **speaking**, also paying attention to **pronunciation** that is essential in terms of the content of the text in question (111) To offer the pupil abundant opportunities for practising age-appropriate small-scale **speaking**, also paying attention to **structures** that are essential in terms of the content of the text in question (112) To offer the pupil abundant opportunities for practising age-appropriate small-scale **writing**, also paying attention to **structures** that are essential in terms of the content of the text in question (113)	The pupil learns to practice age-appropriate small-scale **speaking**, also paying attention to **pronunciation** that is essential in terms of the content of the text in question (111) The pupil learns to practice age-appropriate small-scale **speaking**, also paying attention to **structures** that are essential in terms of the content of the text in question (112) The pupil learns to practice age-appropriate small-scale **writing**, also paying attention to **structures** that are essential in terms of the content of the text in question (113)	**The pupil...** Is able to describe everyday topics Is able to describe concrete topics Is able to describe topics important to him/her using simple sentences Is able to describe topics important to him/her using concrete vocabulary Masters an easily predictable basic vocabulary Masters many key structures Knows how to apply some basic rules of pronunciation also in expressions that have not been practiced

[a]These are not yet in the NCC but written here by the authors

Appendix B: The Evolving Language Proficiency Scale for Level A2.1

THE EVOLVING LANGUAGE PROFICIENCY SCALE, second national language and foreign languages, National core curriculum for basic education 2014					
Interaction skills				Text interpretation skills	Text production skills
Proficiency level	*interacting in different situations*	*using communication strategies*	*cultural appropriateness of communication*	*text interpretation skills*	*text production skills*
A2.1 **First stage of basic proficiency**	The student is able to exchange thoughts or information in familiar, everyday situations and can occasionally maintain a communication situation.	The student increasingly participates Increasingly in communication, resorting to non-verbal expressions less often. The student needs to ask for clarification or repetition quite frequently and is somewhat able to apply the expressions used by the communication partner in his or her own communication.	The student can manage short social situations. The student is able to use the most common polite greetings and terms of address as well as to politely express requests, invitations, proposals, apologies etc. and to respond to these.	The student understands written texts and clear speech that include simple, familiar vocabulary and expressions. The student understands the core contents of short and simple messages that are of interest to him or her and the main points of a predictable text containing familiar vocabulary. The student is capable of very simple reasoning supported by the context.	The student is able to describe everyday and concrete topics and those important to him or her using simple sentences and concrete vocabulary. The student masters an easily predictable basic vocabulary and many key structures. The student knows how to apply some basic rules of pronunciation, also in expressions that have not been practised.

References

Anderson, L. W., & Krathwohl, D. R. (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. Longman.

Bachman, L., & Palmer, A. (2010). *Language testing in practice*. Oxford University Press.

Beikmahdavi, N. (2016). Washback in language testing: Review of related literature first. *International Journal of Modern Language Teaching and Learning, 1*(4), 130–136.

Bloom, B. S. (1956). *Taxonomy of educational objectives*. Longman.

CEFR. (2001). *Common European Framework of Reference for languages: Learning, teaching, assessment*. Council of Europe. Council for Cultural Co-operation. Education Committee. Modern Languages Division. Cambridge University Press.

Chaofeng, Z. (2019). *Literature review on washback effect of the language testing*.

Council of Europe. (2018). *Common European Framework of reference for languages: learning, teaching, assessment. Companion volume with new descriptors*. Council of Europe. Available: https://rm.coe.int/cefr-companion-volume-with-new-descriptors-2018/1680787989

Dimova, S., Yan, X., & Ginter, A. (2020). *Local language testing* (*Design, implementation, and development*). London and New York.

Finnish National Agency of Education. (2014). National Core Curriculum for Basic Education. Publications 2016:5.

Härmälä, M., Huhtanen, M., Puukko, M. & Marjanen, J. (2019). *A-englannin oppimistulokset 7. vuosiluokan alussa 2018*. Kansallinen koulutuksen arviointikeskus. Julkaisut 13:2019.

Hasselgreen, A., & Caudwell, G. (2016). *Assessing the language of young learners*. British Council Monographs.

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement, 38*, 319–342.

Kane, M. T. (2013). Validation as a pragmatic, scientific activity. *Journal of Educational Measurement, 50*(1), 115–122.

Keuning, J., Straat, J. H., & Feskens, R. C. W. (2017). The Data-Driven Consensus (3DC) procedure: A new approach to standard setting. In S. Teoksessa Blömeke & J.-E. Gustafsson (Eds.), *Standard setting in education. The Nordic countries in an international perspective* (pp. 263–278). Springer.

Kozikoğlu, İ. (2018). The examination of alignment between national assessment and English curriculum objectives using revised Bloom's Taxonomy. *Educational Research Quarterly, 41*(4), 50–77.

Krathwohl, D. R. (2002). A revision of Bloom's Taxonomy: An overview. *Theory Into Practice, 41*(4), 212–218. https://doi.org/10.1207/s15430421tip4104_2

OECD. (2021). *Education at a Glance 2021: OECD indicators*. OECD Publishing. https://doi.org/10.1787/b35a14e5-en

Papageorgiou, S. (2016). 20. Aligning language assessments to standards and frameworks. In *Handbook of second language assessment* (pp. 327–340). De Gruyter Mouton.

Papageorgiou, S., Xu, X., Timpe-Laughlin, V., & Dugdale, D. M. (2020). *Exploring the alignment between a curriculum and a test for young learners of English as a foreign language*.

Pourdana, N., & Rajeski, J. S. (2013). Estimating the difficulty level of EFL texts: Applying Bloom's taxonomy of educational objectives. *International Journal of Applied Linguistics & English Literature, 2*(6), 202–211. https://doi.org/10.7575/aiac.ijalel.v.2n.6p.202

Salminen, J. (2021). *Trust in the Finnish Education System - historical explanation*. Lecture given at the SICI workshop in Finland 14–15 October 2021.

Samira, B., Mohammad Sadegh, B., & Mortaza, Y. (2021). Learning objectives of IELTS listening and reading tests: Focusing on revised Bloom's taxonomy. *Research in English Language Pedagogy, 9*(1), 182–199. https://doi.org/10.30486/relp.2021.1916940.1244

Seaman, M. (2011). Bloom's taxonomy: Its evolution, revision, and use in the field of education. *Curriculum and Teaching Dialogue, 13*(1–2), 29.

Squires, D. (2012). Curriculum alignment research suggests that alignment can improve student achievement. *The Clearing House, 85*(4), 129–135. https://doi.org/10.1080/00098655.2012.657723

Sultana, N. (2018). Investigating the relationship between washback and curriculum alignment: A literature review. *Canadian Journal for New Scholars in Education/Revue canadienne des jeunes chercheures et chercheurs en éducation, 9*(2), 151-158.

Webb, N. L. (2007). Issues related to judging the alignment of curriculum standards and assessments. *Applied Measurement in Education, 20*(1), 7–25. https://doi.org/10.1207/s15324818ame2001_2

Weideman, A. (2019). Definition and design: Aligning language interventions in education. *Stellenbosch Papers in Linguistics Plus, 56*(1), 31–46.

Weir, C. (2005). *Language testing and validation. An evidence-based approach* (Research and practice in applied linguistics). Palgrave Macmillan.

When Student Background Overrides Score-Based Placement: Tension Between Stakeholders' Conceptualizations of Fairness

Ivy Chen, Ute Knoch, and Annemiek Huisman

Abstract Language placement tests should be connected to the language curricula into which they place learners, reflective of local needs and student populations, and fair to all stakeholders. What constitutes fairness within a given context may differ – language testers and teaching staff may have different ideas on how to achieve fairness. This chapter draws on data from a review of placement testing procedures for two languages (French and Chinese) at a large Australian university. We examine the use of overriding rules that consider students' prior experience with the target language, in addition to their test scores when making placement decisions. We show the impact of these rules through the analysis of enrolment numbers and final subject grades. We also report students' and academic staff's perceptions of the accuracy of placement decisions. The use of overriding rules led to lower enrolment numbers amongst weaker students affected by the rules, more (successful) requests from affected students to enrol at a lower and more appropriate level, and often negatively impacted student achievement. The chapter has implications for the development of local tests, as it serves as an example of how placement rules can be adjusted to accommodate test-taker populations, with fairness being an important consideration.

Keywords Local language testing · Foreign language placement tests · Fairness

1 Test Purpose and Testing Context

Students entering university language programs need to be placed in classes appropriate for their language ability, taking into account the myriad of previous language learning experiences and family backgrounds that they may enter with.

I. Chen (✉) · U. Knoch · A. Huisman
Language Testing Research Centre, University of Melbourne, Melbourne, Australia
e-mail: ivy.chen@unimelb.edu.au; uknoch@unimelb.edu.au; annemiek.huisman@unimelb.edu.au

X. Yan et al. (eds.), *Local Language Testing*, Educational Linguistics 61,
https://doi.org/10.1007/978-3-031-33541-9_8

To be at their most effective, language placement tests should be deeply connected to the language curricula into which they place learners, be reflective of local needs and student populations, and result in efficient and accurate decision-making about student levels (Bernhardt et al., 2004). By doing so, placement tests can potentially also have a positive effect on professionalizing the language departments by building understanding of test constructs and of language assessment more broadly and, correspondingly, instructors can raise test developers' awareness of curricular and instructional issues and values about specific features of their student populations. The placement test policy that supports the implementation and use of such tests should therefore ideally be co-created by language departments and language testers. Local tests embedded within these specific contexts are likely to be more effective than existing off-the-shelf products (Dimova et al., 2020; O'Sullivan, 2019).

Such tests, however, often do not feature in the literature, as they are frequently developed locally on restricted budgets, with limited language testing expertise, and, once in operation, are not necessarily sufficiently evaluated to ensure test use and subsequent decision-making result in desired consequences (Plakans & Burke, 2013). We evaluated a suite of foreign language placement tests (and their related procedures) which were developed locally at a large Australian university. In this chapter, we aim to explore one aspect of test fairness, namely the impact of overriding rules put in place by the language teaching staff of some languages to place students based on their previous experience with the target language. We report findings from the French and the Chinese placement tests, as their overriding rules were put in place mainly to deal with two very different issues perceived by the staff: keeping a sense of cohort within students who studied French in high school and discouraging dishonesty in the Chinese heritage-learner students.

To begin, we describe the situation before our suite of placement tests were developed, briefly cover the placement testing procedure, and explain how the language subjects (that students are placed into) work. Prior to the development of these tests, for some languages, students could choose their own subject level by reading descriptions in the course handbook or teaching staff would place them into a level based on certain background characteristics (e.g., taking the target language in high school). For other languages, the teaching staff conducted placement days in which they would spend full days conducting interviews with students and collecting writing samples.

The university leadership team decided that placement procedures should be relatively uniform across languages (but at the same time allowing for placement policy differences across languages). They also stipulated that the placement procedure needed to be administered online, automatically scored and integrated with the local institutional systems for subject enrolment (i.e., automatic placement of students into a subject level). A team of language test developers within the university (of which the authors are a part) was tasked with developing the placement tests for the nine languages offered and worked closely in collaboration with the teaching staff of each language department, who differed in terms of their support for a

placement test and/or their ideas on what this should look like. The language testing team was careful never to *push* its views onto the languages staff but to work with the staff members through the development, trialling and standard setting for the tests. It was hoped that the joint development process would create *buy-in* from the various groups and some sense of ownership of the procedure.

1.1 Placement Test Procedures

The placement test procedure was designed to place students into general language subjects, with the number of placement levels corresponding to the number of levels established in the curriculum, which were language specific. While language subjects are classified as undergraduate level, everyone wanting to take one must go through the placement test procedure, including graduate students, whether they are foreign language majors or those taking the language as an elective. Generally, placement outcomes are followed, and requests from students to move levels up or down are considered only when students have attended some classes at the beginning of the semester.

The tests were designed to be computer-administered and scored, with tasks suitable for a range of proficiency levels. While tasks were kept as similar as possible across languages, language-specific changes were made. Each test consists of three sections: (1) a listening section (partial dictation for most languages; click on incorrect characters for character-based languages), (2) a cloze elide section (i.e., click on superfluous words/characters randomly inserted in texts), and (3) a text completion section (C-test for most languages; multiple-choice C-test for character-based languages).

New students first complete a placement questionnaire. Based on the responses to this questionnaire, students are placed either directly into a level, or they are directed to the placement test and then placed based on their test score, or a combination of test score and background questionnaire responses (i.e., previous learning experience). Cut-scores were set using the analytic judgement method (e.g., Plake & Hambleton, 2001) in standard-setting workshops with the language teaching staff (French $N = 5$, Chinese $N = 4$) using writing samples from test takers from a range of proficiency levels (French $N = 52$, Chinese $N = 55$).

Overriding rules determine when test takers are directly placed into a level based on answers in the questionnaire or are excluded from certain lower levels (i.e., placing some students higher than their test score would indicate). These rules are absolute, hence the term *overriding*, as there was no opportunity for manual placement in the placement test process, which the leadership team required to be automatic (i.e., from the background questionnaire and placement test to linking placement outcomes with the local institutional systems for subject enrolment). The university supports students wishing to extend themselves, so test takers can enrol at a level higher than the one they were placed into, but not into a lower level, unless they gain permission from the teaching staff.

1.2 How the Content Subjects Work

A subject level reflects the *proficiency level* of the subject being taught rather than when the student takes the subject (e.g., French 1 is for absolute beginners and not necessarily for first-year students, who may start at a higher level), and this is shown on the academic transcript. French has seven levels (French 1 to French 7) while Chinese has 10 (Chinese 1 to Chinese 10). The odd-numbered levels for each language (e.g., French 1, 3, 5) are available in semester 1 and the even-numbered levels (e.g., French 2, 4, 6) are available in semester 2 of each academic year. Placement into these levels allows test takers to begin in either semester (e.g., a certain test score allows a student to take either French 3 in semester 1 or take French 2 in semester 2; students can then decide which semester they would like to start), though for most languages, including French and Chinese, absolute beginners must begin at the lowest level.

When taken as an elective unrelated to their major, students can select the language, the semester they start, and how many semesters they wish to study (provided they have enough electives). When taken as part of a degree program (major, minor, diploma), students must begin in semester 1 but can start at any subject level depending on their proficiency at entry (e.g., start at French 1, 3, or 5). These students are required to complete at least the sixth level (e.g., French 6); students starting at a higher level (e.g., French 3 or 5) may exit at a higher level.

2 Testing Problem Encountered

Placement tests affect several key factors for a foreign language program: sufficient enrolment numbers are essential for maintaining the program, student success should not be negatively affected by placement, and class homogeneity in terms of learning background may be important for a better learning experience. In other words, it would be ideal if students perceive their placement to be appropriate and feel confident enough to enrol, placements are accurate enough in terms of proficiency level so that the subject is not too difficult (or too easy) for students, and classes are homogenous enough in learning background to build a sense of cohort. Examples of homogeneity in this sense include absolute beginners in a class with other beginners (i.e., no unconfident or weak false beginners), students who studied the language in high school study with their peers at university no matter how weak they are, and students mostly have similar learning profiles (e.g., heritage learners with jagged profiles speaking well in class may demotivate other learners).

As the placement tests and associated procedures were developed together with the teaching staff, there was inevitably some tension due to differences in

perspectives. While we (the language testing team) were concerned about fairness in terms of accurately placing individual test takers (and consequently maximizing enrolment numbers), the teaching staff focused on fairness for the students in general and thus placed great emphasis on discouraging dishonesty in students (i.e., lying about their background in the language and/or deliberately performing badly on the placement test) and prioritized more homogenous classes to improve the classroom experience. We supported the use of student background questionnaire answers as extra information to aid the placement process when necessary, but felt that blanket rules may be off-putting for students (and may result in more work for staff), despite the possibility for students to move levels if their tutor later determines that they have been placed incorrectly. Since the placement tests served the language programs, we deferred to the teaching staff and put overriding rules in place for their languages if the staff felt strongly about them.

French and Chinese were selected as case studies for this chapter, as they represent very different language groups, student populations and program-specific needs; hence, they provide insights also found for our other language tests. Overriding rules were put in place by the two language programs for very different reasons, trying to address different perceived *problems* with their respective student populations. During the development and trialling of the testing procedure, language departments were the participants in standard-setting workshops and were able to formulate placement policies for specific student groups within their cohorts, which provided programs with the opportunity to take ownership of perceived or real language-specific challenges. French staff perceived problems placing Australian school-leavers who had taken French in their final year of school (e.g., perceived unfairness if a test taker did poorly in the placement test and began in French 1 with absolute beginners), and Chinese staff identified that most placement issues were related to placing heritage learners (a hugely diverse group) and L1 speakers. The specific rules put in place by each language are discussed in more detail later in this chapter.

To explore the effect of language staff's idea of achieving fairness through implementing these overriding rules, we asked three research questions:

RQ1. Are enrolment numbers different for students affected by overriding rules? This covers the proportion of enrolling students and the proportion of students enrolling at the level they were placed into (i.e., not moved to a higher or lower level).

RQ2. For enrolling students, are final subject marks different for students affected by overriding rules?

RQ3. What are staff perceptions of the effect of overriding rules, and do students feel that the overall placement procedures (including overriding rules) result in homogenous classes as intended?

3 Literature Review

In the past two decades, fairness has been increasingly recognized as being a crucial consideration in the field of language testing. Test fairness refers to the internal, or technical, quality of a test and is related to all phases of assessment, from design and development to administration and to use (e.g., McNamara et al., 2019; McNamara & Ryan, 2011; Xi, 2010), and in our case, encompasses not just the placement test itself but also the rules governing its implementation. In other words, comparability for relevant test-taker groups is key: there should be no systematic bias against any group in test items and coverage, test access and administration, and test interpretation and use. For placement testing, this entails equitable treatment of all test takers throughout the placement procedure, resulting in comparable placement outcomes for students at similar levels of proficiency.

In the education literature, the more common but increasingly less supported view is that of equality of treatment, where students receive the same treatment and resources no matter their membership to different linguistic, ethnic, or socio-economic groups (Center for Public Education, 2016; OECD, 2006). In contrast, equitable treatment of students is defined as giving students individualized treatment and sufficient resources for success based on individual needs; this involves providing extra resources to support those from disadvantaged groups, for instance, which may run counter to equality (Center for Public Education, 2016; OECD, 2006).

To take the notions of equality and equity from the school context where learning takes place over time and apply them to the language placement context in a tertiary setting where placement is a one-off occurrence, some analogies need to be drawn. Treatment at school is analogous to placement at a language level. When striving for equity, disadvantaged students are treated differently and are allocated more resources at school to support and prepare them for success; this is akin to using a placement test to accurately gauge the level of proficiency of all students, to recommend a suitable language level that is not too difficult and not too easy for each, so they will be sufficiently but not overly challenged in class. Then, and this is particular to the placement testing context, by also giving students the choice to enrol at a higher level that they feel is more appropriate for their previous background in the language, highly motivated students are given the option to extend themselves, especially if this means a weaker student can study alongside their high school peers, for example.

Students being grouped by year level (i.e., based on age and number of years in school) is analogous to potential language students being grouped by their language learning background. In other words, under an equal system, Grade 8 students would be a population to be given equal treatment in class, which is comparable to treating students who took Year 12 French in high school as one incoming student population to be equally placed at the same level coming in to tertiary education, even though some would have performed poorly in their high school French exams and others would have completed high school eight years ago rather than be an assumed recent graduate. Similarly, this entails treating students who completed

Year 5 in a Chinese-speaking country or region as another homogeneous incoming student population, even though some would have stopped using Chinese (or at least writing and perhaps reading in Chinese) altogether since then, while others would have had continuous private tutoring. This would thus forcibly keep these student populations within the same cohorts, even if this means that weaker students may struggle at university or be put off by the difficulty of the subject and not enrol.

Placement into language programs may be difficult in multicultural contexts because different student subgroups present different language learning profiles. In the literature, two groups of students are often differentiated: heritage language learners (HLLs) and foreign language learners (L2 learners). In our case, many of those who go through the Chinese placement procedure are HLLs (or even international students from Chinese-speaking countries) while almost all of those who go through the French placement procedure have a more traditional L2 learner background.

HLLs have been defined as speakers who come from homes where the target language is spoken (Lowe, 1998) or more broadly from a background (e.g., ancestral) where there is a connection to a language, even if it is not spoken at home (Fishman, 2001). HLLs present as a very diverse group, with proficiencies ranging from barely receptive to completely productive (Fairclough, 2005). The sources of language acquisition of HLL have been presented as a triad, comprising of (1) family, (2) community/community schools, and (3) formal education (Kagan, 2005), where each of these elements can have greater or lesser relevance depending on a number of factors (including the language, the history of migration and attitudes to language preservation). The major difference between HLLs and L2 learners is that for the former, exposure and acquisition begins in the home, while for the latter it usually starts in classroom settings.

A number of authors have described typical differences in the language proficiency profiles of heritage language learners and foreign language learners (Ilieva & Clark-Gareca, 2016; Malone et al., 2014; Kondo-Brown, 2003). HLLs are generally described as having better oral skills, including advanced phonology, extensive everyday vocabulary, some knowledge of grammar and considerable fluency (in particular, speech rate – Kagan & Friedman, 2003; Polinsky, 2008), while literacy skills may lag behind (e.g., Ilieva & Clark-Gareca, 2016; Malone et al., 2014). L2 learners usually have more even profiles and are often weaker in their productive skills when compared to their receptive skills, which means that simply taking a test designed for L2 learners and using them to test HLLs may be problematic, especially if those tests do not cover a wide enough range of proficiencies and are thus unable to discriminate between the high-scoring HLLs (e.g., Elder, 1996).

Due to the differing nature of these two groups, it has been suggested that it is best practice in placement testing for university courses to draw on both background and performance data, and that tests developed for foreign language learners may not be the best at measuring proficiency of HLLs (Kagan, 2005; Ilieva & Clarke-Gareca, 2016); similar practices of including learner characteristics in placement decisions exist and are common in contexts without HLLs (e.g., Dimova et al., 2020; Plakans & Burke, 2013). Fairclough (2012) notes that programmatic

considerations need to also be taken into account before designing a test, including (a) the mission of the program, (b) the program and learner characteristics, and (c) the course content.

In the literature on HLL language education and assessment, a number of studies have been published detailing placement testing procedures (e.g., Sekerina, 2013; Beaudrie & Ducar, 2012; Burgo, 2013; Fairclough, 2006, 2012). This literature is limited in scope, in that it mostly focussed on school contexts in the United States and is concentrated on the placement of Spanish heritage learners. There is much less literature available focussing on other contexts, especially on languages taught outside the United States; also, local tests in general are often not evaluated or that evaluation is not reported. As a result, our understanding of how best to assess HLLs within higher education settings remains low.

4 Methods

The data, both quantitative and qualitative, are described by data source and their respective participants, and how they relate to the evaluation of the impact of overriding rules.

4.1 Placement and Subject Marks Data

Placement test data from tests accessed between April 5, 2018 and Mar 17, 2019 were collected. Final subject marks were supplied by the language programmes and matched to test takers' test data and background data collected in the questionnaire. These grades were from test takers who had enrolled in a French or Chinese language subject for Semester 2 in 2018 or Semester 1 in 2019. 1001 students accessed the French placement procedure during that time, and all took the background questionnaire. Out of these, 512 were then required to move on to the placement test, while the rest (N = 489) were placed directly into a French level based on their questionnaire answers (either because they were complete beginners or were subject to one of the overriding rules described in more detail below). 1319 students took part in the Chinese placement procedure. Following the questionnaire, 684 were directly placed into a Chinese level, while the other half (N = 635) went on to take the test.

To evaluate the impact of overriding rules, the data were analysed in terms of their effect on enrolment numbers and final subject marks through comparison between those affected by an overriding rule (i.e., subject to an overriding rule and either did not take the placement test or were *pushed up* into a higher level than expected from their test scores) and those who were not (i.e., placement using test scores or subject to an overriding rule but with placement congruent with test scores).

4.2 Student Views

Student views were gathered through two instruments: surveys and interviews. Since students were not usually told if they had been affected by an overriding rule, in this chapter, we could not ask directly about the impact of these rules and can only report on more general perceptions.

Survey of Placement Experience

An online survey gathering students' general feedback on the placement test was administered using SurveyMonkey. The survey questions were broad ranging and covered five themes: (a) test taking experience, (b) placement test result, (c) enrolment experience, (d) class experience, and (e) overall experience. Invitations were sent to all students who had taken the placement test for one (or more) of the languages included in the review. In this chapter, we report the results of three yes/no questions: (a) if most students in their class were of roughly the same level of language proficiency, (b) if there were any students far stronger than the average level, and (c) if there were any students far weaker than average. 102 French students and 78 Chinese students responded, with 58 French students and 33 Chinese students providing answers to these particular questions.

Focus Groups

A sub-group of students who indicated in the survey their willingness to take part in an interview, were invited to take part in focus groups. The focus groups were semi-structured around the same five themes as the survey and were organised by language. Eight students who had taken French and 17 who had taken Chinese took part. We report their perceptions of the homogeneity of their classes and their own experiences or what they had heard from their fellow students.

4.3 Academic Staff Focus Groups

Focus groups were held with academic staff, both lecturers and tutors, to gather information on their experience with the placement test. The focus groups were semi-structured, and questions centred around the volume and nature of student enquiries, perceived accuracy of placement test results, and overall satisfaction with the placement test procedures. The focus groups were organised by language, with a total of 29 participants; this included two focus groups with eight academic staff members for French and one focus group with four academic staff members for Chinese. Findings are reported for perceived accuracy of placement, with a focus on overriding rules.

5 Results and Discussion

The evaluation of the impact of overriding rules covered three areas: their effect on (a) enrolment numbers and on (b) final subject marks for those who did enrol, and (c) perceived accuracy of placement using overriding rules. Both French and Chinese had rules limiting test takers with previous L2 study experience and Chinese had additional rules for HLLs and L1 speakers. These rules for placing specific test-taker populations were, in some cases, put in place despite advice otherwise from the language testing team. Before presenting the findings of the review of the overriding rules, we briefly describe the rules for both languages.

Although the French background questionnaire had questions capturing information on test takers who had studied French in French-speaking countries and those who had studied through French as medium of instruction, the overriding rules only cover a subset of the L2 learners who had taken French as a subject in Australian high schools. As shown in Table 1, there was a general rule about completion of any Year 12 (final year of high school) French in Australia, and there were more specific overriding rules for the two school-leaving certificates common to students applying to the university, which overrode the general rule. It was noted that while most rules were consistent, certain IB students were directly placed into a French level without taking the placement test while other IB students and all other test takers were given the chance to take the placement test. We recommended that IB students take the placement test and be given the opportunity to do better than expected on the test and be placed into a higher (and more appropriate) level; the teaching staff concurred.

The Chinese background questionnaire covered a similar range of student backgrounds as the French one, but since a much larger proportion of test takers were HLLs, there was a larger number and range of associated overriding rules, as shown in Table 2. With the HLL rules, if someone self-identified as an L1 speaker, they were automatically placed and did not see the question on schooling in Chinese. With the prior L2 learning rules, the Year 12 rules overrode the general rules for study in a non-Chinese-speaking country.

Table 1 French overriding rules: high school French

Test-Taker Background			Test?	Possible subject placement levels					
				1	2	3	4	5	6
Completion of any year 12 French in Australia			Y	–		✓			
Victorian certificate of education (VCE) above a certain score			Y	–				✓	
International baccalaureate (IB)	Some prior knowledge of French: Higher level (HL)/ standard level (SL)	Most	N	–		✓		–	
		Above a certain score	N	–				✓	
	No prior knowledge of French (ab initio)	Most	N	–		✓		–	
		Above a certain score	Y			✓			

Table 2 Chinese overriding rules

Test-Taker Background			Test?	Possible subject placement levels									
				1	2	3	4	5	6	7	8	9	10
L1 speaker of Mandarin/Cantonese			N	–									
Some schooling in a Chinese-speaking country/region: highest year level completed	1/2		Y	–				✓					
	3/4		Y	–						✓			
	5		N	–								✓	
	6–12		N	–									
Studied Chinese in a non-Chinese-speaking country: length of study (years)	<2		Y					✓					
	2–4		Y	–				✓					
	4–6		Y	–				✓					
	>6		Y	–						✓			
Completion of year 12 Chinese in an Australian school	VCE	First language	N	–									
		Second language	Y	–				✓					
	IB	A (L1 speakers)	N	–									
		B HL	Y	–						✓			
		B SL	Y	–				✓					
		Ab initio	Y	–				✓					

Note: Exclusion from Chinese 1–10 = test takers directed to Chinese specialist subjects

5.1 RQ1 Enrolment Numbers

Evaluating the impact of overriding rules on enrolment numbers involved comparing the enrolment numbers of students who had been placed into a level congruent with their placement test score and those of students who had been pushed by an overriding rule into a higher level than that suggested by their test score. Tables 3 and 4 show enrolment numbers by placement type, for French and Chinese, respectively. For both languages, test takers affected by an overriding rule were significantly less likely to enrol in a language subject than normally-placed students (i.e., those placed by test score only or whose placement was not affected by overriding rules): for French, this was 38.0% ($N = 41$) vs. 50.0% ($N = 447$), $\chi^2(11001) = 5.63$, $p = .02$ for all students and 40.3% ($N = 25$) vs. 57.3% ($N = 258$), $\chi^2(1512) = 6.37$, $p = .01$ when only looking at those who took the placement test; for Chinese, this was 6.5% ($N = 51$) vs. 54.1% ($N = 288$), $\chi^2(11319) = 377.22$, $p < .001$ for all students and 31.7% ($N = 33$) vs. 54.1% ($N = 288$), $\chi^2(1636) = 17.44$, $p < .001$ for those who took the test. This suggests that students who were pushed into a higher level may have been concerned about their ability to cope in the higher level (Tables 3 and 4).

Rough estimates of how many more students would have enrolled if they had not been affected by an overriding rule were calculated. For French, assuming that the

Table 3 Enrolment numbers for French (not enrolled vs. enrolled, by enrolment level vs. placement level), by placement type (placement using test scores only vs. an overriding rule, separated by congruency with test score where applicable)

				Enrolment (vs. placement)				
	Placement	Affected by overriding rule?	Not enrolled	Higher level	Same level	Lower level	Sum	Total
No test	Beginner	N/A	254	0	189	n/a	189	443
	Overriding rule	Y	30	0	16	0	16	46
	Sum		284	0	205	0	205	489
Test	Test score	N	90	1	86	0	87	177
	Overriding rule	N	102	7	160	4	171	273
		Y (pushed up)	37	0	21	4	25	62
	Sum		229	8	267	8	283	512
Total		N or N/A	446	8	435	4	447	893
		Y	67	0	37	4	41	108
	Sum		513	8	472	8	488	1001

Table 4 Enrolment numbers for Chinese (not enrolled vs. enrolled, by enrolment level vs. placement level), by placement type (placement using test scores only vs. an overriding rule, separated by effect of overriding rule where applicable)

				Enrolment (vs. placement)				
	Placement	Affected by overriding rule?	Not enrolled	Higher level	Same level	Lower level	Sum	Total
No test[a]	Overriding rule	Y	665	0	0	18	18	683
Test	Test score	N	154	11	167	9	187	341
	Overriding rule	N	90	20	81	0	101	191
		Y (pushed up)	71	2	24	7	33	104
	Sum		315	33	272	16	320	635
Total		N	244	31	248	9	288	532
		Y	736	2	24	25	51	787
	Sum		980	33	272	34	338	1319

Note: [a] = Unlike for French (and the other languages), absolute beginners were directed to take the test. As specialist subjects ($N = 677$) are not language subjects, enrolment into those is not counted (only 16 were given permission to enrol in a language subject)

percentage of enrolling students for those not affected by the overriding rules (50.0%) would hold for all students, there would be an expected increase of 13 enrolling students out of 1001 students who go through the placement procedure, which translates to a 2.7% increase in enrolment numbers (from 488 to 501) and an overall increase in percentage of enrolments from 48.8% to 50.0%. Similarly for Chinese (but excluding those who were directly placed outside the Chinese language subjects because almost all of them were L1 speakers) and assuming that the

percentage of enrolling students would remain at 54.1%, there would be an expected increase of 23 enrolling students out of 1319 students who go through the placement procedure, which translates to a 6.9% increase in enrolment numbers (from 338 to 361) and an overall increase in percentage of enrolments from 25.6% to 27.4%.

Turning to the students who enrolled in a language subject, the number of test takers who had been affected by an overriding rule was disproportionately large in the group that gained permission to enrol at a level lower than the one they had been placed into, as compared with the group that enrolled at the level they had been placed into. Table 5 shows how these differences are significant for French and Chinese, for all students and when only looking at those who took the placement test. While these numbers are small, they suggest that students who did poorly in the placement test were aware of their true lower level and more actively sought to enrol at a lower level and/or were able to make a more convincing argument to the language teaching staff for them to be placed at a lower level.

5.2 RQ2 Final Subject Marks

Final subject marks were used as an indication of how well students coped with the subject. For French, we examined each overriding rule's effect on the students' final subject marks separately, by comparing them with the final subject marks of students not affected by the rules. Due to space constraints, we only provide figures for the overriding VCE score rule where test takers reported a scaled VCE score above 35 ($N = 43$; others reported raw scores, no scores, or impossible scores). As shown in Table 6, 22 test takers (51.1%) were affected by the overriding rule and only 14 enrolled, with three enrolling at a lower level congruent with their test-score placement (French 3) and 11 enrolling into French 5 as suggested. These 11 students had lower final subject marks (score-based placement French 2/3: 62.00

Table 5 Proportion of students affected by overriding rules for those who enrolled at a lower level than they were placed into compared to those who enrolled at the same level

		Enrolment (vs. placement)				Chi-square test (Difference in proportion of students affected by rules)		
		Lower level		Same level				
		%	*N*	%	*N*	df	χ^2	*p*
French	All students	50.0%	4	7.8%	37	1480	17.86	< .001
	Subset who took test	50.0%	4	7.9%	21	1275	16.63	< .001
Chinese	All students	73.5%	25	8.8%	24	1306	93.78	< .001
	Subset who took test	43.8%	7	8.8%	24	1287	19.12	< .001

Table 6 French final subject marks for test takers potentially affected by the VCE Score Rule (>35), with comparison marks for normally placed test takers

Student group	Placement level based solely on test score	*N*			Descriptive statistics (enrolled in French 5)					
		Did not enrol	Enrolled, but not French 5	Enrolled in French 5	*M*	Median	Min.	Max.	Range	*SD*
VCE score rule: Affected (pushed up to French 5)	French 1/2	1	0	0	–	–	–	–	–	–
	French 2/3	1	1	1	62.00	62.00	62	62	0	–
	French 3/4	6	2	10	74.10	73.50	68	79	11	3.57
VCE score rule: Not affected (score high enough for French 5)	French 4/5 or 5/6	7	0	14	74.14	75.50	50	84	34	8.13
Comparison: Not affected (and scored high enough for French 5)	French 4/5 or 5/6			96	77.00	78.00	50	89	39	5.84

[N = 1], French 3/4: 74.10 [SD = 3.57, N = 10]) when compared with the other students with VCE scores above 35 who had congruent test scores (M = 77.00, SD = 5.84), though the mean final subject mark of the score-based placement French 3/4 group was similar to that of the French 4/5 or 5/6 group subject to but not affected by the VCE score rule (M = 74.14, SD = 8.13, N = 14). Overall, in the case of the two overriding rules relating to Year 12 French (other than IB, where we were unable to determine impact as many overriding rules did not give test takers a chance to take the placement test), final subject marks showed that students may struggle more than would be expected if placement were based solely on the overriding rules, especially if they score much lower than would be expected of someone with their learning background, which led to us recommending these rules be abolished.

For Chinese, as very few students affected by overriding rules enrolled at the suggested level, there were not enough data to analyse each overriding rule individually. Instead, the overall effects of the rules were explored for combined rules (e.g., exclusion from Chinese 1–2, exclusion from Chinese 1–4), as shown in Table 7, though even then, comparisons for exclusion from Chinese 1–2 (N = 0) and from Chinese 1–6 (N = 1) were impossible because affected students either did not enrol or enrolled at a different (primarily lower) level. As for exclusion from Chinese 1–4, out of the 65 affected test takers, only 21 students (32.3%) enrolled in Chinese 5 as suggested, and of them, one withdrew. The average final subject marks were lower than that of others who enrolled in Chinese 5 but scored as expected on the placement test (i.e., those not subject to overriding rules and those not affected by overriding rules): score-based placement of Chinese 1/2: 74.00 (N = 1), Chinese 2/3: 59.57 (SD = 9.16, N = 7), Chinese 3/4 M = 69.42 (SD = 7.22, N = 12), compared with normally-placed students M = 77.38 (SD = 6.95). Overall, in addition to lower enrolment levels and a higher likelihood to enrol at a lower level already explored above, the test takers who were pushed into a higher Chinese level due to overriding rules were more likely to receive lower final subject marks compared to other students, though limited data meant that we recommended the overriding rules be further investigated or abolished.

5.3 RQ3. Perceived Accuracy of Placement

Perceived accuracy of placement is reported separately for each language, beginning with student perceptions of the overall placement procedures in terms of the homogeneity of their classes, followed by staff perceptions of the effect of overriding rules. Overall, for both languages, the teaching staff decided to retain the overriding rules (with the amendment to the French IB rules so that students could take the test and have the opportunity to do better than expected or at least provide additional information on their proficiency level). They believe the overriding rules to positively contribute to accurate placement and lead to more homogenous classes.

Table 7 Chinese final subject marks for test takers affected by overriding rules and enrolled at the suggested level, with comparison marks of normally placed test takers

Student group	Placement level based solely on test score	*N*			Descriptive statistics (enrolled at suggested level)					
		Did not enrol	Enrolled, but not at the suggested level	Enrolled at the suggested level	*M*	Median	Min.	Max.	Range	*SD*
Excluded from Chinese 1–2	Chinese 1/2	2	1	0	–	–	–	–	–	–
	Chinese 2/3	0	1	0	–	–	–	–	–	–
Excluded from Chinese 1–4	Chinese 1/2	8	1	1	74.00	74.00	74	74	0	–
	Chinese 2/3	14	2	7	59.57	60.00	43	71	28	9.16
	Chinese 3/4	17	2	13 (1 withdrew)	69.42	68.50	60	82	22	7.22
(Comparison group: Chinese 5)	Chinese 4/5 or 5/6			50	77.38	77.00	56	91	35	6.95
Excluded from Chinese 1–6	Chinese 1/2	5	0	1	80.00	80.00	80	80	0	–
	Chinese 2/3	5	1	0	–	–	–	–	–	–
	Chinese 3/4	5	0	0	–	–	–	–	–	–
	Chinese 4/5	4	1	0	–	–	–	–	–	–
	Chinese 5/6	9	1	0	–	–	–	–	–	–
(Comparison group: Chinese 7)	Chinese 6/7 or 7/8			14	73.29	75.5	60	80	20	6.742

Note: Exclusion from Chinese 1–8 only applies to those who do not take the test ($N = 4$), and of those, two did not enroll and two enrolled at a lower level

In the survey, most French students reported feeling that most students in the class were of roughly the same level of language proficiency (N = 41, 70.7%), though more than half also noticed at least one student in their class obviously stronger than the average level (N = 37, 63.8%) and at least one who was obviously weaker (N = 30, 51.7%). Focus group students agreed that proficiency levels were generally similar in their classes, apart from one specific comment about a student from Mauritius who seemed to be more advanced than the others.

French staff generally found the placement test to be a useful tool that works for most students and determined that the disadvantage of overriding rules for enrolling students was slight enough to ignore, given that in the focus groups, staff reported having more homogenous classes (especially French 1 and 2) after the implementation of these rules. They acknowledged that they sometimes still needed to move a few students between French 3 and 5, in some cases undoing the effect of an overriding rule.

The large proportion of HLLs made appropriate placement more difficult for Chinese than for French. In the survey, a slightly (but not significantly lower) proportion of students reported feeling that most students in the class were of roughly the same level of language proficiency (N = 21, 63.6%) and many noticed at least one student in their class obviously stronger than the average level (N = 24, 72.7%) and fewer than half noticed obviously weaker students (N = 14, 42.4%). Focus group students perceived high levels of *sandbagging* the test or lying in the background questionnaire. For instance, one student, who started out in Chinese 1, explained:

> It's like a lot of the people in the class, they already knew Chinese. And I felt like a massive disadvantage, 'cause they're like "Oh, I already know Chinese, but I managed to like flunk the test or something".

This made them feel unmotivated and gave the tutors the false sense of belief that the students had all successfully learned the material, so tutors went through the material too quickly.

On the other hand, as heritage learners explained, they struggled in class as their writing skills were generally much weaker than their speaking skills; this was made more difficult since the target language is character-based (i.e., with less of a relationship between phonology and the written form). Typical comments from focus groups include: "I can recognize words, but like I can't write it"; "I'm really good at listening and reading in Chinese, but I'm not quite good at writing"; and "A lot of us will come from backgrounds where we speak it at home, but there's no writing involved, and I feel like that's a really big problem". This is arguably less of an issue with the placement test and more of a curricular issue, as curricula were primarily designed with L2 learners in mind: many HLLs placed at the right level are still likely to be seen by other students as being dishonest because of their much stronger speaking and listening skills.

Chinese staff taking part in the focus groups noted that the placement test is a useful tool that works and places honest students at the right level and that overriding rules are useful, especially for heritage learners, to discourage dishonest students. They explained, based on their previous years of teaching, that these dishonest students (many South-East Asians, often of Chinese heritage) start at low levels on

purpose to get high marks. They pointed to a possible issue with how degrees are structured at the university, where a large proportion of subjects taken can be elective (25%): many use Chinese as an easy way to increase their overall GPA. This was why the Chinese staff did not implement our recommendation to open another stream of subjects for HLLs; they believed that there would be low interest in a heritage learner stream, as HLLs would continue to try to enrol at lower Chinese subject levels instead. A related issue is that the staff are unable to force students out of the subjects even if it is clear they are at the wrong level; they can only recommend that the students move up. Even though the issue of cheating cannot be easily solved, staff were positive about the introduction of the additional permission form students need to complete and sign when requesting a level change that had been implemented sometime after the placement procedure was first put in place. The form points out very clearly that dishonesty is academic misconduct, and staff say that students are being more honest than before.

6 Insights Gained

Although most of the test development process went relatively smoothly for all languages despite some initial resistance from some staff from some language departments, one point of difference remained in our perspectives of what was fair for students: the use of overriding rules in some languages, including French and Chinese, which may result in systematic bias against the students affected by these rules (McNamara et al., 2019; McNamara & Ryan, 2011; Xi, 2010). Although we, the language testing team, supported the use of student background questionnaire information as part of the placement process (e.g., used in fine-tuning placement levels when students or tutors were concerned or where students were placed very close to cut-scores), we were worried that blanket rules may have a negative impact on students, in the form of lower enrolment numbers and poorer performance in class for those who enrolled. In other words, the resulting placement testing procedures for the languages with overriding rules followed an equal rather than equitable approach (Center for Public Education, 2016; OECD, 2006). The teaching staff's perception of fairness meant maximising student morale through an avoidance of unfair advantage of false beginners (including HLLs and those who took the language in high school, no matter how long ago) over true beginners and maintaining a sense of cohort among students with similar learning backgrounds (which only affected weaker students, as stronger students were placed into a higher level).

As expected, the evaluation did not support all the overriding rules, as they lead to different outcomes for the students affected. For both languages, test takers who were pushed up by overriding rules into a higher-level subject than their test score would suggest were more likely to either not enrol (RQ1), successfully dispute their placement with teaching staff (RQ1), or score lower in the subject (RQ2), which suggests that an equitable approach to placement would be better for affected students. While students reported feeling that classes were mostly homogenous, some

noticed classmates who were clearly stronger and fewer perceived weaker classmates (RQ3). Notably, for Chinese, L1 speakers and HLLs (mostly South-East Asians, often of Chinese heritage) were not always caught when deliberately *sandbagging* the test or lying in the background questionnaire. This has been more recently addressed through *scaring* students by making them sign an official document with an academic integrity clause kept in their student file; Chinese staff noted a corresponding decrease in such cases.

Nonetheless, as the placement tests were ultimately for the language programs and it was important for the language teaching team to take ownership of perceived or real language-specific challenges, these rules were put in place. Overriding rules arose from the teaching staff's concern about fairness for students in terms of their learning experience, which meant prioritizing more homogenous classes and, in the case of Chinese, also discouraging dishonesty. Teaching staff were generally positive about the effectiveness of overriding rules in homogenizing their classrooms (RQ3).They noted that students could always be moved to a different level once classes started, if it was obvious that they had been placed incorrectly, though this could not solve the issue of lower enrolment numbers.

The staff were happy for high-scoring students to enter higher subject levels, so the French staff took up our suggestion to let all IB students take the placement test while retaining their respective limitations on the lowest possible subject level, so high-scoring students could potentially start at a higher subject level than expected by their IB scores. Having stronger students placed at the appropriate subject level would lead to more homogenous classes. Conversely, teaching staff were cautious of letting low-scoring students, both honest and dishonest, enter lower subject levels if students had a certain background in learning the target language regardless of how long they had last learned the language. They explained that the other students in the class would perceive unfairness in the placement process if students from certain learning backgrounds were placed at an unexpectedly low level, with French staff being most concerned about Australian school leavers who had taken French in their final of school being placed into French 1 with absolute beginners if they did poorly in the placement test.

For Chinese, most issues were related to placing the hugely diverse group of heritage learners and L1 speakers, with much of it due not so much to placement issues but to curricular issues. Therefore, we recommended opening a heritage learner stream to improve the classroom experience for both HLLs and L2 learners. HLLs noted that the subjects they had been placed in were not quite appropriate for their stronger oral skills but weaker literacy skills, which has been found to be typical of HLLs (e.g., Ilieva & Clark-Gareca, 2016; Kagan & Friedman, 2003; Malone et al., 2014; Polinsky, 2008). While our placement procedure draws on both background and performance data for placement, as advised (Kagan, 2005; Ilieva & Clarke-Gareca, 2016), as tests developed for L2 learners may not accurately measure HLL proficiency, the curricular issue the HLLs raised could not be avoided. For the L2 learners, students reported feeling demotivated because the HLLs (both honest and dishonest) seemed too advanced when speaking up in class, which led to tutors getting the false sense that the students had successfully learned the material and moved

through the class materials at a fast pace. Since the Chinese teaching staff deemed that interest in a heritage learner stream would be low, this suggestion was not taken up.

7 Implications for Test Developers and Score Users

Although local tests co-created by language departments and language testers are likely to be more effective than existing off-the-shelf products (Dimova et al., 2020; O'Sullivan, 2019), these tests are underrepresented in the literature. This chapter reported on part of an evaluation of a suite of locally developed placement tests to place students into university language programs. In particular, we evaluated the impact of overriding rules put in place by the language teaching staff to aid the placing of specific student populations, meaning that certain background variables determined or limited their placement, even if this differed from the test score. These overriding rules were not necessarily supported by the test development team at the development stage due to concerns with unequal treatment of particular test-taker groups leading to issues with test fairness, but they were nevertheless implemented as part of the collaborative approach to test development and local policy formation.

When all the findings of the placement test evaluation had been analysed, a report was created for each language detailing the findings and recommendations. Similarly, a more general report was written, which focused on findings and recommendations applicable across languages. These reports were sent to the language departments for their consideration along with an oral presentation to each department and an overview of the findings and recommendations were also presented to a faculty-level board representing all languages. After that, the language testing team met with each language department separately for their feedback and for discussing how they wanted to proceed with the recommendations; changes to the placement test procedure were made after these discussions. This close coordination and collaboration between stakeholders within the local testing context is a major advantage of locally developed tests, allowing the possibility to review local cut-scores and adjust placement policies to account for (changes in) local test populations.

This chapter has two major implications for developers of local tests. First, it is important to keep in mind potential stakeholder differences in perspectives/considerations. In our context, we (the language testing team) were concerned about fairness in terms of accurately placing individual test takers (and consequently maximizing enrolment numbers), while the teaching staff focused on fairness for the students in general and thus discouraged dishonesty in students and prioritized more homogenous classes to improve the classroom experience. With overriding rules, while they may serve to discourage weaker students from continuing with their studies or have a negative impact on student achievement (i.e., supporting an equitable rather than equal approach to placement), and therefore problematic rules were recommended to be abolished as part of the program evaluation, they were

ultimately retained. Language staff argued that they were necessary for fairness, to keep false beginners from beginner-level subjects or L1-speakers from taking a low-level subject.

Second, we have described how the collaborative test development and evaluation process has created buy-in from the language departments, who had initially been highly sceptical of the need of a test (see also Beaudrie & Ducar, 2012) and were generally quite reluctant to make changes to their existing processes at the outset of the curriculum reform. All teaching staff commented on the value of the test, and how much they had needed this procedure and appreciated the collaboration. The process has increased the language assessment literacy (Berry et al., 2019; Malone, 2013; Taylor, 2009) of teaching staff, who now regularly call on the language testing team with their own suggestions on how to use and improve the placement tests. Without the availability of this language testing expertise, the language programs would likely still draw on either time-consuming or rather limited placement practices (e.g., by using background questionnaires only). The language testing team has also greatly benefited from the collaboration by learning about the specific challenges to placement encountered by the different language programs. Such knowledge expands the current literature on language assessment, which is mostly focussed on the assessment of English in different contexts, to provide more detailed information about specific challenges in assessing certain languages, but also specific local challenges encountered in the context in question.

References

Beaudrie, S. M., & Ducar, C. (2012). Language placement and beyond: Guidelines for the design and implementation of a computerized Spanish heritage language exam. *Heritage Language Journal, 9*(1), 77–99. https://doi.org/10.46538/hlj.9.1.5

Bernhardt, E. B., Rivera, R. J., & Kamil, M. L. (2004). The practicality and efficiency of web-based placement testing for college-level language programs. *Foreign Language Annals, 37*(3), 356–365. https://doi.org/10.1111/j.1944-9720.2004.tb02694.x

Berry, V., Sheehan, S., & Munro, S. (2019). What does language assessment literacy mean to teachers? *ELT Journal, 73*(2), 113–123. https://doi.org/10.1093/elt/ccy055

Burgo, C. (2013). Spanish in Chicago: Writing an online placement exam for Spanish heritage speakers. *Borealis–An International Journal of Hispanic Linguistics, 2*(1), 199–207. https://doi.org/10.7557/1.2.1.2496

Center for Public Education. (2016). *Educational equity. What does it mean? How do we know when we reach it?* Research brief. https://files.eric.ed.gov/fulltext/ED608822.pdf

Dimova, S., Yan, X., & Ginther, A. (2020). *Local language testing: Design, implementation, and development*. Routledge. https://doi.org/10.4324/9780429492242

Elder, C. (1996). The effect of language background on "foreign" language test performance: The case of Chinese, Italian, and modern Greek. *Language Learning, 46*(2), 233–282. https://doi.org/10.1111/j.1467-1770.1996.tb01236.x

Fairclough, M. (2005). Spanish and heritage language education in the United States: *Struggling with hypotheticals*. Iberoamericana Libros–Vervuert. https://doi.org/10.31819/9783865278944.

Fairclough, M. (2006). Language placement exams for heritage speakers of Spanish: Learning from students' mistakes. *Foreign Language Annals, 39*(4), 595–604. https://doi.org/10.1111/j.1944-9720.2006.tb02278.x

Fairclough, M. (2012). A working model for assessing Spanish heritage language learners' language proficiency through a placement exam. *Heritage Language Journal, 9*(1), 121–138. https://doi.org/10.46538/hlj.9.1.7

Fishman, J. (2001). 300-plus years of heritage language education in the United States. In J. K. Peyton, D. A. Ranard, & S. McGinnis (Eds.), *Heritage languages in America: Preserving a national resource* (pp. 81–89). Delta Systems Company.

Ilieva, G. N., & Clark-Gareca, B. (2016). Heritage language learner assessment: Toward proficiency standards. In M. Fairclough & S. M. Beaudrie (Eds.), *Innovative strategies for heritage language teaching: A practical guide for the classroom* (pp. 214–236). Georgetown University Press.

Kagan, O. (2005). In support of a proficiency-based definition of heritage language learners: The case of Russian. *International Journal of Bilingual Education and Bilingualism, 8*(2–3), 213–221. https://doi.org/10.1080/13670050508668608

Kagan, O., & Friedman, D. (2003). Using the OPI to place heritage speakers of Russian. *Foreign Language Annals, 36*(4), 536–545. https://doi.org/10.1111/j.1944-9720.2003.tb02143.x

Kondo-Brown, K. (2003). Heritage language instruction for post-secondary students from immigrant backgrounds. *Heritage Language Journal, 1*(1), 1–25. https://doi.org/10.46538/hlj.1.1.1

Lowe, P. (1998). Keeping the optic constant: Framework of principles for writing and specifying the AEI definitions of language abilities. *Foreign Language Annals, 31*(3), 358–380. https://doi.org/10.1111/j.1944-9720.1998.tb00582.x

Malone, M. E. (2013). The essentials of assessment literacy: Contrasts between testers and users. *Language Testing, 30*(3), 329–344. https://doi.org/10.1177/0265532213480129

Malone, M., Peyton, J. K., & Kim, K. (2014). Assessment of heritage language learners: Issues and directions. In T. G. Wiley, J. K. Peyton, D. Christian, S. C. Moore, & N. Liu (Eds.), *Handbook of heritage, community, and native American languages in the United States* (pp. 349–358). Routledge. https://doi.org/10.4324/9780203122419.ch33

McNamara, T., & Ryan, K. (2011). Fairness versus justice in language testing: The place of English literacy in the Australian citizenship test. *Language Assessment Quarterly, 8*(2), 161–178. https://doi.org/10.1080/15434303.2011.565438

McNamara, T., Knoch, U., & Fan, J. (2019). *Fairness, justice & language assessment: The role of measurement*. Oxford University Press. https://doi.org/10.1080/15366367.2020.1739796

O'Sullivan, B. (2019). Localisation. In L. I. Su, C. J. Weir, & J. R. Wu (Eds.), *English language proficiency testing in Asia: A new paradigm bridging global and local contexts* (pp. xiii–xxviii). Routledge. https://doi.org/10.4324/9781351254021

OECD. (2006). *Ten steps to equity in education.* Policy brief. http://www.oecd.org/education/school/39989494.pdf

Plakans, L., & Burke, M. (2013). The decision-making process in language program placement: Test and non-test factors interacting in context. *Language Assessment Quarterly, 10*(2), 115–134. https://doi.org/10.1080/15434303.2011.627598

Plake, B. S. & Hambleton, R., K (2001). The analytic judgment method for setting standards on complex performance assessments. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 283–312). Routledge. https://doi.org/10.4324/9781410600264.

Polinsky, M. (2008). Heritage language narratives. In D. M. Brinton, O. Kagan, & S. Bauckus (Eds.), *Heritage language education: A new field emerging* (pp. 149–164). Routledge. https://doi.org/10.4324/9781315092997-11

Sekerina, I. (2013). A psychometric approach to heritage language studies. *Heritage Language Journal, 10*(2), 203–210. https://doi.org/10.46538/hlj.10.2.4

Taylor, L. (2009). Developing assessment literacy. *Annual Review of Applied Linguistics, 29*, 21–36. https://doi.org/10.1017/S0267190509090035

Xi, X. (2010). How do we go about investigating test fairness? *Language Testing, 27*(2), 147–170. https://doi.org/10.1177/0265532209349465

Local Tests, Local Contexts: The Italian Language Testing Scenario

Sabrina Machetti and Paola Masillo

Abstract This chapter aims at reporting the development of policies and practices designed to support the linguistic integration of adult migrants in Italy (Machetti S et al., Language policy and linguistic justice. Springer, 2018). The chapter builds on a report on language assessment practices at the local level carried out within the framework of current legislation with the purpose of focusing on the use of language requirements for the issuance of a long-stay permits for non-EU citizens (Masillo P, La valutazione linguistica in contesto migratorio: il test A2. Pacini Editore, 2019). The chapter presents the results of an impact study conducted to monitor local procedures for the teaching and assessment of L2 Italian. A repertoire of local good practices was collected at national level to stimulate and support the constant commitment of teachers indirectly involved in language policies. Through both qualitative and quantitative research methodology, the chapter addresses how the promotion of a research project involving the systematic and structured teacher training in the LAL field led to an improvement in terms of test development and to a more general (positive) impact on L2 Italian learning and teaching processes.

Keywords Local language testing · Migration · Residence requirements · Italy · Common European Framework of Reference

This chapter is a work of shared reflection of both authors who together developed paragraph 6 and References. However, author1 developed the paragraphs 3, 4, 5; author2 developed the paragraphs 1, 2.

S. Machetti (✉) · P. Masillo
Certification of Italian as Second/Foreign Language Centre,
University for Foreigners of Siena, Siena, Italy
e-mail: machetti@unistrasi.it; masillo@unistrasi.it

X. Yan et al. (eds.), *Local Language Testing*, Educational Linguistics 61,
https://doi.org/10.1007/978-3-031-33541-9_9

1 Introduction: Test Purpose and Testing Context

Over the past two decades, a significant number of European countries have introduced linguistic requirements for the purposes of migration, such as first entry, permanent residency, and/or citizenship (ALTE, 2016; Machetti et al., 2018; Rocca et al., 2020; Masillo, 2021; Carlsen & Rocca, 2021). In 2009, Italy aligned its language policies with those of other European countries. Until that moment, migration policies had considered the Italian language as an essential tool for integration but had never used it as an explicit indicator of migrants' possibilities of integration in the country. Such a choice in terms of language policy fostered the creation of an ideological association between migrants' proficiency in the Italian language and their right (and willingness) to remain in the country. Considering the provisions of the current legislation, according to Law no. 94/2009 (*Provisions on public security*; Ministerial Decree of June fourth, 2010 – partially updated by the Ministerial Decree of December seventh, 2021), the issuance of a long-term residency permit is subject to the achievement of a certain level of language proficiency in L2 Italian, which must be certified by passing a language test. Therefore language – or rather the possession of a certain level of language proficiency – may represent an obstacle to a migrant's integration into the labour market. The Italian language thus loses its nature as an inalienable right that should be defended and promoted by the Italian Government. This clearly conflicts with what the Italian Constitution itself states, promoting a view of language as a barrier.

In other words, Italy entered the panorama of the so-called *Testing Regimes*, in which the assessment of proficiency in the host country's official language is a requirement for obtaining a long-term residence permit, as well as a requisite for the mobility of citizens within the labour and training systems (Extra et al., 2009; Van Avermaet, 2010; Extramiana et al., 2014).

The content of Law no. 94/2009 – enacted at a time of shift in Italy's consensus towards the ideas of a traditionalist and reactionary Right party – is not so surprising when analysed in the light of what was already happening or beginning to happen in most European countries.

Our chapter places the Italian scenario within the broader European panorama of integration and language policies implemented for migrant citizens. The main objective of the chapter is to discuss the implementation of language tests as Italian policy tools in the educational practices carried out by the public adult educational centres throughout the country (namely CPIA – Provincial Adult Education Centres). CPIAs, as we will see in greater detail in the next paragraphs, are educational institutions that provide first and second cycle instruction for adults (aged 16 and over). In addition, CPIAs provide literacy and Italian language courses for foreigners, especially to adult migrants coming from extra-European countries.[1]

[1] https://www.miur.gov.it/i-centri-provinciali-per-l-istruzione-degliadulti#:~:text=L'istruzione%20degli%20adulti%20%C3%A8,263%20del%2029%20ottobre%202,012

According to the Ministerial Decree of June 4, 2010, it was decided not to use a centralised test at national level, but to assign the development and administration of new tests to the above-mentioned local public adult educational centres. The Ministerial Decree defines the test procedure and states that the test must be at the A2 level of the *Common European Framework of Reference for Languages. Learning, Teaching, Assessment* (Council of Europe, 2001). However, the Ministry does not justify the selection of a required proficiency level (Masillo, 2019), in fact slavishly following the determinations already made at European level. Thus, it is established that the long-stay permit is linked to a proficiency in the language of the host country that is not only 'unprofiled', but also not homogeneous among the countries concerned (Barni, 2010, 2012, 2013; Machetti, 2016; Rocca et al., 2020).

The various reports promoted over time by the Council of Europe (Extramiana & Van Avermaet, 2011; Extramiana et al., 2014; Rocca et al., 2020) in the different member States have highlighted how the imposition of containment and control policies – e.g., by measuring the level of proficiency in the host country's official language – has resulted not only from the increase in the number of migrants, but also from the fact that the policy-makers themselves have considered it as one of the indispensable tools to guarantee the security of the various member States and indigenous citizens (Extramiana et al., 2014; Barni, 2012). Even the most recent Council of Europe's report (Rocca et al., 2020) shows that only 7 countries out of the 40 analysed have no language requirements: a significantly higher trend than in previous reports. Although the use of a test as a tool to measure the linguistic competence of migrants at different stages of their migration process is a widespread and consolidated practice in Europe, it is a procedure that is still questionable (McNamara, 1998, 2005, 2011; Shohamy, 1997, 2001). Admittedly, the situation in Italy is aggravated by the fact that Italian politicians, the civil society and academics are unable to critically oppose the enactment of such practices (Barni & Machetti, 2005; Machetti, 2012).

This chapter examines the language testing practices carried out at local level within the framework of the current legislation, focusing on (1) the role played by CPIAs teachers in these practices and (2) the teacher training programmes offered to support more valid and reliable assessment processes and tools. The local tests we will discuss belong to the language testing practices implemented for migrants applying for the long-term residence permit. These local tests are assumed to represent the values and priorities arising from the Italian Law no. 94/2009 and are supposed to address the problems that derive from the needs within the local context in which the assessment procedure is used (Dimova et al., 2020).

The chapter discusses the results emerging from an impact study (CLIQ 2017–2021 – http://www.associazionecliq.it/2019/01/16/studio-e-igliamento-dellimpatto-dei-percorsi-formativi-e-valutativi-presso-i-cpia-2017-2020/) conducted in Italy to monitor procedures for L2 Italian teaching and assessment within the context of CPIAs. The aim is to identify and isolate a set of good practices, replicable also outside the Italian context. Through both qualitative and quantitative data, we discuss how the promotion of a research project including LAL training for teachers has led to a more general (positive) impact on language learning and teaching processes. A repertoire of local good practices was also collected at national

level to stimulate and support the constant commitment of teachers indirectly involved in language policies. Since the introduction of the legislation in force, teachers' experience with the target audience, i.e., adult migrants, has allowed them to reach a higher level of familiarity with the procedures of measuring and assessing language competence.

The Ministerial Decree of June 4, 2010 has stipulated that CPIAs must manage the new tests, specifically CPIAs that have participated in national and international research projects in the field of learning and teaching Italian as a second language and whose teacher have attended refresher and training courses in Italian as a second language. Our findings stem from the first difficulties encountered by teachers of CPIAs because of their involvement in the current legislation. These issues have made teachers aware of the implications of designing a language test that has not only a diagnostic value but also leads to the issuance of an official document, whose socially recognised value is embodied in a long-term residence permit (Masillo, 2019).

1.1 Integration: Basic Principles and Action Strategies

The characteristics of the socio-cultural reality introduced by migration phenomena have developed and diversified over time. They have been raising needs in the management of cultural diversity by the host country, especially in terms of reception and integration (Council of Europe, 2008). Below, we analyse three documents that provide us with a better understanding of the issue.

As stated in the *Resolution 1437 (2005), Migration and integration: a challenge and an opportunity for Europe*:

> The concept of integration aims at ensuring social cohesion through accommodation of diversity understood as a two-way process. Immigrants have to accept the laws and basic values of European societies and, on the other hand, host societies have to respect immigrants' dignity and distinct identity and to take them into account when elaborating domestic policies.

The Resolution clearly considers integration as a "bilateral process" (Niessen & Huddleston, 2010, p. 78) of mutual accommodation, which should be included in the goals and topics of interest of a national policy. The same proposal is made by the *White Paper on Intercultural Dialogue: Living together as equals in dignity*, issued by the Council of Europe in 2008: integration should be understood as a two-way process, as the ability of individuals to live together in full respect of individual dignity, thus involving common welfare, pluralism and diversity, non-violence, solidarity and participating in social, cultural, economic and political life (Beacco et al., 2014, p. 9).

The concept of a 'culture of integration' is introduced in a more recent *European Parliament Resolution 2013* (2012/2131(INI)) addressing the integration of migrants and its effects on the labour market and social security, thus reiterating the two-way nature of the integration effort on the part of migrants and the host country. In point 1 of the *Resolution*, the European Parliament argues that:

> integration into the labour market and into society requires commitment on both sides, on the one hand especially in relation to language learning, familiarity with and respect for the legal, political and social systems, customs and usages, and patterns of social interaction in the host country, and on the other hand by building an inclusive society, granting access to the labour market, institutions, education, social security, healthcare, access to goods and services and to housing, and the right to participate in the democratic process […].

Finally, as stated in the Council of Europe's LIAM (*Linguistic Integration of Adult Migrants*) project website, the definition of the integration of foreign citizens – not only newcomers but also naturalized third-country nationals – is described as a theoretically multifaceted process, to the extent that various indicators contribute to defining the process of inclusion in the host society.

1.2 The Role of Language in Integration Procedures

According to a commonly shared perception, integration mainly depends on culture, education, and in this case, language. Learning the language of the host country has long been a key step in the integration process, presumably grounded on the belief that the knowledge of the official language and cultural values of the host country can automatically ensure that they are shared by the "newcomers" (Van Avermaet, 2010). The presence of foreign citizens' competence in the language of the host country is now consolidated in the political discourse on integration, because it is seen as a clear sign of a perceived and desired integration (Niessen & Huddleston, 2010).

According to European directions (cf., e.g. *Resolution 68 (18), On the teaching of languages to migrant workers*), knowledge of the language of the host country is recognized as one of the indispensable requirements for foreign citizens to be able to express themselves and carry out their rights and duties, as well as for participating in the integration process (including linguistic integration) in which they are involved. Learning the language of the host country, therefore, is a fundamental step in the integration process and this consideration now seems to be widely shared even among those who are primarily involved in reception and integration projects.

Regarding the Italian context, the most recent *Rapporto SIPROIMI*[2] (Giovannetti & Somai, 2020, p. 74) states that Italian language proficiency has become one of the requirements on which the process of migrants' inclusion in the local community is based. It is the essential basis for the construction of social relations, for training

[2] Law No. 173 of 18 December 2020 renames the Protection System for Persons with International Protection and Unaccompanied Foreign Minors (SIPROIMI) to *SAI - Sistema di accoglienza e integrazione* (Reception and Integration System). The new law provides for the accommodation of applicants for international protection as well as holders of protection, unaccompanied foreign minors, and foreigners in administrative proceedings entrusted to social services, upon reaching the age of majority. (https://www.retesai.it/la-storia/)

and work paths, for the use of local services and, more generally, for the acquisition of a sense of community belonging and for the exercise of active citizenship. In the *Rapporto SIPROIMI* (*ivi*, p. 104), it is reported that 92.6% of the local educational projects provide Italian language courses for more than 10 hours per week. Within this process and within the reception system set up by the host country, the role of language teachers becomes central in supporting and guiding the learner towards the achievement of his/her goals. A great deal of flexibility has been required in order to provide a widespread supply of facilitating tools for the access of migrants to the language courses.

Language learning, as stated in the *Piano d'azione per l'integrazione e l'inclusione 2021–2027* (*Action Plan for Integration and Inclusion 2021–2027*, Commissione Europea, 2020), should not end within the few months of arrival in the host country and should not be limited to basic levels. It is necessary to extend the course offer to intermediate and advanced levels as needed (*ivi*, p. 10). This could be combined with civic education courses, which allow the understanding of a legal text, for example, so that foreign citizens can become fully involved in their host society.

However, as underlined by Carlsen and Rocca (2021), the multiple roles of language – on one hand, the ability to communicate efficiently and adequately in a context, and on the other, as marker of individual and social identity – promoted the passage from language as a right to language as a barrier, then to test misuse. Through this transition, the multiple roles of language have paved the way towards defining the linguistic requirement as the sole means of access to social and civil rights. The issue is very delicate as language, from being an added value for social inclusion and integration in the host community, is sometimes reduced to a barrier for access to civil rights, and becomes a political instrument as well as a barrier for the granting of residence or citizenship rights in the national territory (Shohamy, 2001, 2009).

It is worth mentioning that in the late 1990s the Association of Language Testers in Europe (ALTE) was asked by a significant number of European member States to develop language tests for migration, residency, and citizenship purposes (ALTE, 2016). This request led to the establishment of the Language Assessment for Migration and Integration (LAMI) Group. The LAMI Group has been working very hard to support decision makers of countries where mandatory tests of this kind are in place or under consideration. The activity of the LAMI Group includes providing reliable and relevant information, recommendations and good testing practices aimed to support data-driven decision-making. Its work is completely devoted to the promotion of test fairness, as unfair tests "may result in migrants being denied civil or human rights, as underlined by the Parliamentary Assembly of the Council of Europe in Recommendation 2034" (p. 7).

1.3 Ministerial Guidelines for Test Development

In Italy, the Ministerial Decree of June 2010 was followed by the signing, in November of the same year, of an agreement between the Ministry of the Interior and the Ministry of Education, University and Research. The latter also combined the agreement with the drafting of official *Guidelines*[3] (MIUR, 2010) for the design of the A2 test required to obtain a long-term residence permit in Italy and the related assessment procedures. As explained in the official Guidelines, which contain test specifications and assessment criteria, the test consists of three sections: Listening Comprehension, Reading Comprehension, and Written Interaction.

The test is designed on the comprehension of short texts and the candidate's ability to interact, in accordance with level A2 of the CEFR (oral comprehension, written comprehension, and written interaction). The Listening Comprehension section is divided into two parts, consisting of two short texts that are representative of the listening comprehension sub-skills of a conversation between L1 speakers, announcements and instructions, radio and audio recordings, and TV programmes. The Reading Comprehension section is also structured in two parts, consisting of two short texts that are representative of the following sub-skills: reading correspondence, reading instructions and understanding how to orient oneself, inform oneself and argue. Finally, the Written Interaction section refers to the sub-skills of interaction through correspondence, replying to messages, filling in forms.

By means of special tables modelled on those proposed for level A2 of the CEFR (Council of Europe, 2001), then adapted to the Italian language and to the linguistic-communicative needs of adult migrants in Italy, the Guidelines specify – the test content, the reference vocabulary, the socio-communicative functions. The Guidelines also indicate the test format, the duration, and the assessment criteria. The entire test lasts 60 min. The Listening Comprehension and Reading Comprehension sections last 25 min respectively and consist of 10 items each (multiple-choice, true/false, matching, etc.). The items must be based on the requirements defined by the CEFR for the A2 profile regarding the characteristics and the genres of the written and audio input. The Written Interaction section lasts 10 min and can be either composed by short responses or extended writing.

As regards the assessment criteria, the candidate must achieve at least an overall score of 80% to pass the test. The weight of each test section is distributed as follows (Table 1):

Table 1 Percentage score distribution

Test section	%
Listening comprehension section	30
Reading comprehension section	35
Written interaction section	35

[3] From now on: Ministerial Guidelines.

Based on the weight attributed to the individual test sections, scores are given as follows:

1. Listening Comprehension section:
   - For each correct answer to an item, 3 points are given.
   - No points are given for missing or incorrect answers.
   - Total maximum score for the test: 30 points.
2. Reading Comprehension section:
   - For each correct answer to an item, 3.5 points are given.
   - No points are given for missing or incorrect answers.
   - Total maximum score for the test: 35 points.
3. Written Interaction section:
   - Test fully and correctly completed (consistent and appropriate responses are given to the information requested or the form is completed in full): up to 35 points.
   - Test partially completed (answers are not always consistent and appropriate to the information requested or the form is partially filled in): up to 28 points.
   - No evidence (no answers are given to the information requested or the form is not completed): zero points.

2 Problem Encountered

The possibility for CPIAs to work in adult education was established by Presidential Decree No. 263 of 2012, but the tradition of adult education in Italy dates back to the late 1990s with the establishment of the Permanent Territorial Centres (CTP). Since 2010, literacy and Italian language learning courses for adult migrants have aimed at attesting the students' proficiency in Italian at CEFR level A2.

The core staff of each Centre consist of secondary and primary school teachers of different subjects: Italian, history and civic education, geography, math, natural sciences, foreign languages, and technical education. Only from 2016 is the presence of at least two L2 Italian teachers in each educational centre envisaged. The number of these teachers could be increased on the basis of projects run by the Centres, taking into account the specific needs of the context, i.e., target groups, migratory flows, labour market flows, etc. Therefore, in most cases teachers should have received specific training in teaching Italian to adult migrants, including on language assessment matters. Unfortunately, the reality is very different. After the 2010 legislation – which entrusts CPIA teachers with the design of the A2 tests for the obtainment of the long-stay permit – a problem emerged: most of these teachers are not sufficiently trained, either pre-service or in-service, in the management of students who have different levels of literacy, and different migration histories. Lack of language testing and assessment literacy also emerged among teachers.

In this scenario, according to the Ministerial Guidelines, each teacher develops his/her own test according to the characteristics of the local context. Consequently, each local test takes into account the specific characteristics of the local context: the students' profiles, the variety of Italian proposed in second language courses, etc.

In the years immediately following the 2010 legislation, this trait of specificity – commonly considered a strength of local tests as they could meet the test takers' specific needs (Dimova et al., 2020) – turned out to be a potentially limiting factor. The reason must be found in the lack of teacher training in this field. Initially, the result of this local language testing practice was a lack of fairness and comparability among the tests designed. Each teacher, in his or her role as an inadequately trained test developer, worked on tests in a very impressionistic or traditional manner, e.g. "recycling" assessment tools used for other purposes. As a matter of fact, the results obtained by the administration of A2 tests in Italy were very different from region to region (Masillo, 2019).

When considering the fact that local rating committees could adapt the rating scale and criteria to each relevant reality, the problems described above affected the whole assessment process. This was clear from the major critical issues of the employed assessment procedures: problems in the formulation of the band descriptors in the rating scale, limits in their implementation, difficulties in their interpretation as well as in the interpretation of scores (Masillo, 2019).

To summarise, the use of tests no longer intended solely to measure language proficiency for educational purposes, but as a political instrument and a gatekeeper to a social system (McNamara, 1998, 2005, 2011; Shohamy, 1997, 2001) had a strong impact on the work of the teachers involved. Theoretically, this could and should have been positive; pushing teachers to broaden their horizons by rethinking their profession. In practice, as trainers supporting language development, found themselves in the role of language testers, without adequate prior training.

3 Methods

This section is dedicated to the illustration of the research methods used in the impact study (*Studio e analisi dei percorsi formativi e valutativi,* FAMI 1603, 2017–2020). The study has been conducted in Italy between 2017 and 2021 (June 2021), in CPIAs that had developed and administered the A2 test necessary for the issuance of a long-stay permit.

The study is part of a larger study supported by European funds and managed by the Italian Government conducted by L2 Italian Certification Centres in collaboration with experts and trainers in language testing and assessment working along with the CLIQ Association (Barni & Machetti, 2019). The purpose of this study was monitoring

1. the quality of the training offered to adult migrants in Italy;
2. the assessment procedures; and
3. tools that are part of the training courses, including A2 tests.

Therefore, this study relates to the national actions of language policy with the purpose of creating a bottom-up process supporting the linguistic integration of adult migrants in Italy. The study involved 27 CPIAs, 25 school principals, teachers, and migrants through interviews and questionnaires.

The interviews, which lasted about 40 min each, were conducted with the school principals and were aimed at surveying the characteristics of the Italian language courses organised in their centres and the figures involved (teachers, intercultural mediators, etc.). In addition, school principals were interviewed on the possible complementary services to the courses, such as baby-sitting or free transport, often realised in collaboration with local institutions.

The questionnaires were administered online to teachers, with the goal of detecting the characteristics of the following activities organized by the CPIAs:

– Italian language courses (67 respondents)
– Knowledge of Society courses (43 respondents)
– Assessment practices and tools adopted in the CPIAs, also with reference to the A2 tests (83 respondents).

The study was conducted 6 years after the entry into force of the A2 test regulations. Therefore, it was aimed at understanding whether the actions implemented over these years to improve the situation created by the regulations had had a positive impact, primarily among teachers. Among the main actions was the provision to CPIA teachers of new guidelines to develop the A2 tests, elaborated by the L2 Italian Certification Centres. These guidelines present a central innovation: the recommendation to introduce a Speaking Task in the A2 test. This choice improves the face and content validity of the test but may pose additional administrative and developmental challenges for teachers. For this reason, another important innovation concerned the launch of training courses aimed at teachers and designed to increase their competence in the area of speaking assessment. We will discuss these courses and, more generally, the contents and structure of the teacher training programmes in § 4. Here it is sufficient to underline how these programmes have been decisive for the professionalisation of the teachers themselves in the field of language testing and assessment. At the same time, the training has proven to be crucial for the teachers' initiation of more valid, reliable, and fair local language testing practices and tools.

4 Results and Discussion

In this section we discuss the data obtained from the identification of the most useful actions for increasing the quality of A2 tests for a long stay in Italy.

The data coming from the administration of teacher questionnaire enabled us to profile the teachers involved in language teaching and assessment practices at the CPIAs. The majority of teachers (81%) were women. Half of them (51%) were between 41–55 years old; 27% of teachers were between 30 and 40; 21% were older

than 55; and only the 2% were younger than 30. In addition to a university degree, 27% of teachers held a specific certification for teaching L2 Italian, and 17% had similar postgraduate qualifications. Among them, 34.3% had been working in a CPIA for more than 6 years, 22% for less than a year, 19% for 3–4 years, 14% for 5–6 years, and 10% for 1–2 years. A total of 33% of the teachers had a 2–5 years' experience as teachers of Italian to foreigners, and only 6% of the respondents had less than 2 years' experience.

The aim of the study was also to 1) identify a set of indicators aimed primarily at detecting good practice regarding A2 test design and assessment criteria in public education centres, and 2) subsequently provide guidelines regarding the standardisation of various assessment procedures and instruments. The aim was thus to increase the validity and reliability of the tests. Good practices were selected based on their compliance with 44 indicators, relating to the type and quality of teaching and assessment procedures and tools. The most relevant indicators for the purposes of this study concern the following elements:

– the use of both Ministerial and CLIQ Guidelines;
– the use of test design and test administration methods;
– assessment procedures and criteria;
– information and training activities carried out prior to test administration, aimed at test-takers;
– the provision of teacher training activities on language learning, teaching and assessment;
– methods for sharing results.

Based on the data collected through interviews and questionnaires between 2017 and 2021, we found that among the teachers (employed in the participating public educational centres), only a fairly small number did not develop tests on an impromptu basis and without reference documents. First, 93% of the teachers involved in the project claimed to have at least 1 year of experience in developing, administering and assessing A2 tests (29% had 1–3 years of experience; 11% 4–5 years; 53% more than 5 years). Their assessment experience was mainly related to test administration. With regard to test development, 84% of the teachers stated that they had considered both Ministerial and CLIQ Guidelines, and various materials developed by the L2 Italian Certification Centres/ as reference (more specifically, the reference syllabi for L2 Italian in migration context developed for literacy levels, levels A1, A2, B1, and B2 since 2010 (http://www.associazionecliq.it/sillabi/). As for the Ministerial Guidelines, although 88% of teachers completely supported them as they are, 45% of them would be in favour of revising them, particularly with regard to the inclusion of the Speaking task.

The teachers in favour of a revision stated that the Ministerial Guidelines should be clearer regarding the following:

– procedures for candidates with little or no previous education;
– cut-off points, which are essential to establish the threshold for passing the test;
– the assessment criteria for the writing test.

A small percentage of teachers (less than 15%) proposed to replace the test instituted from 2010 with a speaking-only test; to add an oral interaction test to the planned tests; to replace the entire test with a test on reading texts aloud.

Teachers involved in the testing process had drawn up test development documents such as technical manuals for test specifications designed precisely for their local context. According to them, in these documents, teachers had defined the test format based on both the Guidelines and the syllabi produced and made available by the Certification Centres. Since 25% of teachers believed that the test level needed constant adaptation to be better related to the A2 level, not surprisingly, they also stated that they had been working on the redefinition of the test level. According to teachers, this process was supported by the development of transparent explanatory sub-descriptors for: the A2 level, the criteria used in the selection of input texts, and the type of tasks and socio-pragmatic activities proposed in the test. As for the last two elements, 20% of the teachers highlighted several critical issues. In more detail, 84% of the teachers who stated that they were working on the development of the tests by using the Ministerial Guidelines, declared that they had elaborated concise guidelines for the development of items and sample tasks, which were then made available to all the teachers of the CPIAs involved in the project.

With regard to the test administration procedures, also considering test accommodations for candidates with reading and/or writing difficulties or other needs, teachers stated that they followed the macro-indications of the Ministerial Guidelines, but only 1 out of 27 CPIAs (CPIA#04) states that they had drawn up a document defining the possibilities of exemption from the test for candidates with disabilities (e.g., severe language learning limitations). This document follows the guidelines drawn up by the health authority of the region in which the CPIA is located, and has also been acknowledged by the authorities responsible for issuing long-stay permits (police headquarters and prefectures).

Finally, teachers developed criteria for the written and spoken interaction sections that focused on grammatical accuracy, while only one CPIA developed rating criteria that focused on written interaction. Again, one CPIA out of 27 (#CPIA019) stated that they have worked with a specific level of detail on the rating scales and criteria, with particular reference to the rating scale for the written interaction test.

With regard to information and training activities for test-takers carried out prior to the administration of the test, 36% of the teachers stated that they systematically provided test-takers with information on the test concerning its structure, the assessment criteria adopted, how the test itself is administered and the length of each test section. This was further supported by the production of a video, edited by the teachers of CPIA#010, describing the procedures related to test administration (http://www.associazionecliq.it/wp-content/uploads/2021/05/Repertorio-BUONE-PRASSI-TEST-DM-Procedure-e-svolgimento.mp4). In addition to the importance of the above-mentioned findings, the study highlights that 77% of the teachers had collected samples of written and spoken performances, calibrated on level A2. They also made them available - via paper or digital archives - to colleagues and students (in only 16% of the participating CPIAs) for the identification of critical issues

concerning the assessment of the Written and Oral Interaction test performances, both in terms of test format and rating scales (http://www.associazionecliq.it/wp-content/uploads/2021/04/5-Test-DM-Esempi-di-prove.pdf). For instance, the rating scale already proposed in the Ministerial Guidelines for Written Interaction is holistic (Weigle, 2002), i.e., requires assignment of a global score to the performance. According to the teachers, a holistic scale can represent a potential obstacle because the outcome of the assessment (test fully and correctly completed; test partially completed; test with no evidence) may lead to imprecise and vague score interpretations. More precisely, 9% of the teachers pointed out that the rating scales proposed in the various ministerial documents present excessively vague descriptors; 7% of the teachers stated that the descriptors are subject to a subjective interpretation; 6% of the teachers claimed that it is extremely difficult to grasp the difference between the various score bands proposed. Finally, 3% of the teachers complained that the writing section has much more weight in the final score than the other sections of the test.

Given this result, the research project was combined with the design and delivery of teacher training course. All the teachers involved in the project and also other teachers who had participated in projects similar to this one in 3 different regions of Italy participated in the course. The teacher training activities primarily addressed the basic concepts and techniques of language testing and assessment: the ABCs of Assessment (purposes of assessment; types of assessment; recording and reporting outcomes; quality of assessment); Assessing reading skills; Assessing writing skills; Assessing listening skills; Assessing speaking skills; Providing feedback; Test impact. These activities were intended for and were carried out by teachers with limited knowledge and experience in the field of language testing and assessment. The activities consisted of face-to-face lectures, conducted in the classroom, and seminar activities, also conducted in the classroom. The in-presence seminar activities conclude with the assignment of tasks to the teachers, which are carried out remotely by the teachers either in pairs or in small groups. Afterwards, the completed activities are returned to the trainers using a platform, are evaluated by the trainers themselves who then return the outcomes of the activities to the participants in the next in-presence training session. Finally, the activity is further discussed and its outcomes are shared by the training participants.

The remaining teacher training activities were designed and intended for teachers who already possessed knowledge and skills in language testing and assessment and had already undergone training on these topics. These activities have as their main objective the standardisation of assessment and thus provide teachers with the knowledge and tools to design valid and reliable items and to validly and reliably assess their students' written and oral performances. The training was designed and carried out taking into account what is recommended for this purpose by the *Manual* for Relating Language Examination to the CEFR (Council of Europe, 2009), a Council of Europe document whose primary purpose is to help different users - and thus also teachers at local level – to develop, apply and report practical and transparent procedures in a cumulative process of continuous improvement in order to relate

their language tests to the CEFR. The approach developed in the *Manual* offers guidance to users in describing examination coverage, administration and analysis procedures; in correlating the results reported by the examination with the Common Reference Levels of the CEFR; and in providing supporting evidence reporting the procedures followed for this purpose. Following the Manual, the teachers training activities included 3 distinct phases. The first phase was the familiarisation phase: this phase is carried out by presenting teachers with activities designed to ensure that they have a detailed knowledge of the CEFR. The second phase was aimed at getting teachers to work on the description of the tests they had created themselves, whose content and task types they were required to specify. This phase ensures that the definition and production of the test have been undertaken carefully, following good practices. The third phase focuses on the standardisation of judgement: this is done by offering teachers standardised samples of oral and written performance. In this stage teachers rate the samples, discuss why certain samples are one level rather than another and acquire experience in relating descriptors to particular levels of performance. In this phase, teachers also work on the standardisation of local performance samples and then get to the moment of the actual standard-setting, which consists of the process of defining 'minimum scores' for the different test scores.

The study highlighted the positive impact of teacher training in that a large percentage of teachers (82%) identified the need for continuous and dedicated training in the future on more specific topics, for example on the pretesting stage, which according to the data collected is currently carried out by only 18% of teachers in A2 test development and administration.

5 Insights Gained

The study shows that from 2017 to the present day in CPIAs (i.e. at local level) there have been several changes that have directly affected assessment procedures and tools. It also highlights how these changes are mostly related to the positive impact of continuous teacher training on these processes and issues conducted by language assessment experts. One of the most significant changes concerns gaining awareness of the importance of the local test being representative of local contexts and needs, particularly those of adult migrants. The local tests are embedded in the local adult instructional system and are administered on national territory within the framework of a selection of schools. Despite the testing issues that emerged right after 2010, eventually teachers have shown a greater awareness and an emerging assessment literacy. The teachers' involvement in the above-mentioned procedures has also led to a wide-ranging debate about the importance of regular training and constant updating on language testing and assessment.

Unfortunately, the data collected in Italy is aligned to most European countries, where assessment literacy remains a very weak point of the entire assessment process. The models on which teacher training programs could be inspired (a good

example is what is proposed within the European project TALE, https://taleproject.eu)[4] certainly need to be expanded, particularly in the direction of the needs of low-literate migrants and test misuse, which we have seen to be central in the tests for long-term residence permit in a country.

6 Implications for Test Developers and Users

The inclusion of knowledge of the language of the host country among the requirements for issuing a long-term residence permit reveals an approach that is the result of a conception of the integration process that places the foreign national in the main position of being responsible for the process itself (Masillo, 2019). Following the reflection of Beacco, Little and Hedges in their *Guide to policy development and implementation* (2014, p. 10),

> if the priority of member states really is to achieve the effective linguistic integration of new arrivals and not to control migration flows by monitoring their language skills, then the language programs offered must be of high quality because only that will truly help adult migrants to adapt to a new linguistic and cultural situation.

ALTE's 2016 document can be considered as one of the starting points for our case study, as there it is stated that the role and value of an adequate language training and monitoring system is fundamental in order to guarantee the conditions for achieving the necessary linguistic requirement and to be consistent with the above-mentioned integration objectives, understood as mutual commitment of the parties involved (ivi, 2014).

In the bilateral spirit of the integrative process defined at the beginning, it is desirable that foreign immigration continues to be perceived both as an (educational) resource and as an opportunity, which can encourage a project of linguistic planning and education aimed at a broad plurilingual perspective. This perspective should not only encourage and support language training courses in the language of the host country, but should also give due recognition, at institutional but also at social level, to the identities and cultures of which citizens of foreign origin are bearers (Vedovelli, 2017). In conclusion, we can argue that the involvement of teachers by using the local test under consideration could increase their literacy in language assessment through participation. A higher level of teacher involvement would also allow for the integration of teaching and assessment procedures, which could be beneficial for migrants.

[4] Tsagari et al. (2018).

References

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

Barni, M. (2010). La valutazione della competenza linguistico-comunicativa in italiano L2 e le politiche europee: considerazioni e prospettive. In L. Marigo, C. Capuzzo, & E. Duso (Eds.), *L'insegnamento dell'Italiano L2/LS all'Università. Nuove sfide e opportunità* (pp. 397–417). Il Poligrafo.

Barni, M. (2012). Diritti linguistici, diritti di cittadinanza: l'educazione linguistica come strumento contro le barriere linguistiche. In S. Ferreri (Ed.), *Linguistica educativa: atti del XLIV Congresso Internazionale di Studi della Società di Linguistica Italiana (SLI). Viterbo, 27–29 settembre 2010*. (pp. 213–223). Bulzoni.

Barni, M. (2013). Competenza linguistica, test e integrazione degli immigrati: il caso italiano. *Paradigmi, 1*, 139–149.

Barni, M., & Machetti S. (2005, June 2–5). The (lack of) professionalism in language assessment in Italy. In *Poster abstract presented at the 2nd EALTA conference*. Voss.

Barni, M., & Machetti, S. (2019). Le certificazioni di italiano L2. In P. Diadori (Ed.), *Insegnare italiano L2* (pp. 122–133). Le Monnier.

Beacco, J. C., Hedges, C., & Little D. (2014). *Linguistic integration of adult migrants: Guide to policy development and implementation*. Retrieved from: http://www.coe.int/t/dg4/linguistic/liam/search_category/category_en.asp?

Carlsen, C., & Rocca, L. (2021). Language test misuse. *Language Assessment Quarterly., 18*(5), 477–491. https://doi.org/10.1080/15434303.2021.1947288

Commissione Europea. (2020). *Piano d'azione per l'integrazione e l'inclusione 2021–2027*.

Council of Europe. (2001). *The common European framework of reference for languages: Learning*. Cambridge University Press.

Council of Europe. (2008). *White paper on intercultural dialogue – Living together as equals in dignity*. Council of Europe Publishing.

Council of Europe. (2009). *Relating language examinations to the common European framework of reference for languages: Learning, teaching, assessment (CEFR). A Manual*. Language Policy Division. Retrieved from: www.coe.int/t/dg4/linguistic/manuel1_en.asp.

Dimova, S., Xun, Y., & Ginther, A. (2020). *Local language testing. Design, implementation, and development*. Routledge.

Extra, G., Spotti, M., & Van Avermaet, P. (2009). Testing regimes for newcomers. In G. Extra, M. Spotti, & P. Van Avermaet (Eds.), *Language testing, migration and citizenship. Cross-national perspectives on integration regimes* (pp. 3–33). Continuum.

Extramiana C., & Van Avermaet P. (2011), *Language requirements for adult migrants in Council of Europe member states: Report on a survey*. Council of Europe.

Extramiana, C., Pulinx, R., & Van Avermaet, P. (2014). *Linguistic Integration of Adult Migrants: Final Report on the 3rd Councilof Europe Survey*. Council of Europe.

Giovannetti, M., & Somai, A. (2020). *Rapporto Annuale SIPROIMI/SAI*. Tipografia Grasso Antonino s.a.s.

Machetti, S. (2012). Tratti della competenza linguistico-comunicativa e pratiche di valutazione scolastica. Un confronto tra i diversi ordini di scuola. In R. Grassi (Ed.), *Nuovi contesti d'acquisizione e insegnamento: l'italiano nelle realtà plurilingui* (pp. 175–188). Guerra Edizioni.

Machetti, S. (2016). Test e certificazioni linguistiche: tra eticità, equità e responsabilità. In A. De Meo (Ed.), *Professione Italiano. L'italiano per i nuovi italiani: una lingua per la cittadinanza* (pp. 139–148). Il Torcoliere.

Machetti, S., Barni, M., & Bagna, C. (2018). Language policies for migrants in Italy: The tension between democracy, decision-making, and linguistic diversity. In M. Gazzola, T. Templin, & B. A. Wickström (Eds.), *Language policy and linguistic justice*. Springer.

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

Masillo, P. (2021). Lingua e cittadinanza italiana: uno studio sulla validità della valutazione linguistica per la cittadinanza. In *Studi Italiani di Linguistica Teorica e Applicata, anno L, 2021, numero 1*. Pacini Editore.

McNamara, T. (1998). Policy and social considerations in language assessment. *Annual Review of Applied Linguistics, 18*, 304–319. https://doi.org/10.1017/S0267190500003603

McNamara, T. (2005). 21st century shibboleth: Language tests, identity and intergroup conflict. *Language Policy, 4*(4), 351–370. https://doi.org/10.1007/s10993-005-2886-0

McNamara, T. (2011). Managing learning: Authority and language assessment. *Language Teaching, 44*(4), 500–515. https://doi.org/10.1017/S0261444811000073

Ministero dell'Istruzione, dell'Università e della Ricerca - Direzione generale dell'istruzione e formazione tecnica superiore e per i rapporti con i sistemi formativi delle regioni – Ufficio IV, 2010, Vademecum (ai sensi della nota n 8571 del 16 dicembre 2010 del Ministero dell'Interno). (2010). *Indicazioni tecnico-operative per la definizione dei contenuti delle prove che compongono il test, criteri di assegnazione del punteggio e durata del test.* Retrieved from: http://hubmiur.pubblica.istruzione.it/alfresco/d/d/workspace/SpacesStore/d6686cab-4f36-4c32-acb3-97f2090ead92/vademecum.pdf

Niessen, J., & Huddleston, T. (2010). *Manuale sull'integrazione per i responsabili delle politiche di integrazione e gli operatori del settore*. Unione Europea.

Rocca, L., Carlsen, C., & Deygers, B. (2020). *Linguistic integration of adult migrants: Requirements and learning opportunities*. Council of Europe.

Shohamy, E. (1997). Testing methods, testing consequences: Are they ethical? Are they fair? *Language Testing, 14*(3), 340–349. https://doi.org/10.1177/026553229701400310

Shohamy, E. (2001). *The power of tests: A critical perspective on the use of language tests*. Pearson.

Shohamy, E. (2009). Language tests for immigrants. Why language? Why tests? Why citizenship? In G. Hogan-Brun, C. Mar-Molinero, & P. Stevenson (Eds.), *Discourse on language and integration* (pp. 45–59). John Benjamins Publishing Company.

Tsagari, D., Vogt, K., Froelich, V., Csépes, I., Fekete, A., Green A., Hamp-Lyons, L., Sifakis, N., & Kordia, S. (2018). *Handbook of assessment for language teachers*. Retrieved from: http://taleproject.eu/.

Van Avermaet P. (2010). Language assessment and access. A climate change. In ALTE and Centro Valutazione Certificazioni Linguistiche – CVCL – Università per Stranieri di Perugia (Eds.), *Valutazione Linguistica ed Integrazione nel contesto italiano: Atti del Convegno LAMI. Language assessment for integration in the Italian contexts: Proceedings* (pp. 16–24).

Vedovelli, M. (2017). Le lingue immigrate nello spazio linguistico italiano globale. In M. Vedovelli (Ed.), *L'italiano dei nuovi italiani* (pp. 27–48). Aracne.

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

Identifying the Phonological Errors of Second-Modality, Second-Language (M2-L2) Novice Signers Through Video-Based Mock Tests

Luigi Lerose

Abstract This study investigates how video-based mock tests can reveal the types of phonological errors that novice British Sign Language (BSL) learners make most often, and how teachers might help learners increase their phonological accuracy. The testing need that this local language test was designed to address was a predominance of phonological errors among 10 level-1 BSL learners at the author's university, with relatively little progress being made in enabling them to target their phonological skills effectively. The mock test devised to meet this need, required that students send a video clip of themselves signing about an assigned topic. To categorise the errors, the five phonological parameters of BSL, handshape, location, orientation, movement, and non-manual features, were applied. The results of the mock test showed that most common errors were in the handshape and movement parameters, but that movement errors were repeated more often between signers. Some movement errors seemed to result from confusion between signs in the same semantic domain. Deploying and analysing mock tests in this way can meet the local needs of learners while giving teachers insights that they can use to inform their lessons, and opportunities to increase their own meta-linguistic awareness and analytical skills.

Keywords British sign language · Local language testing · Sign language testing · Sign language teaching · Sign language learning · Sign language phonology

L. Lerose (✉)
British Sign Language and Deaf Studies, University of Central Lancashire (UCLan), Preston, UK
e-mail: llerose@uclan.ac.uk

X. Yan et al. (eds.), *Local Language Testing*, Educational Linguistics 61,
https://doi.org/10.1007/978-3-031-33541-9_10

1 Introduction: Test Purpose and Testing Context

British Sign Language (BSL) is a natural language, used by many deaf people in the UK as their "preferred language" (Ladd, 2003). It is also a second language (L2) for both hearing and deaf people and is taught in several universities as well as other organisations and deaf centres. 'L2 learning' and 'second language acquisition' (SLA) refer to the study of the processes through which individuals and groups learn a language subsequent to their first one, even though this might actually be their third language, or fourth, or fifth, etc. (Saville-Troike & Barto, 2016). Most L2 learners of BSL are hearing people whose first language is a spoken language and first modality is the oral-aural modality. So, the majority of L2 BSL learning involves not only the acquisition of a new language but also that of a new modality, the visual-gestural modality (Meier, 2012). This is sometimes called 'M2 learning' or 'second modality acquisition'. Students for whom BSL is both a L2 and M2, need to learn to use their hands, arms, bodies, head, and face to express the language, and become familiar with concepts related exclusively to the signed modality such as visual iconicity, simultaneity (articulating two signs at the same time), and harnessing the signing space in front of the body (Chen Pichler & Koulidobrova, 2015; Holmström, 2019).

Signature, the main awarding body for BSL qualifications,[1] estimate that since 1982 they have supported more than 400,000 learners (Signature, 2019). Phonology is central to the development of BSL fluency, but is often a difficult area to target in the classroom (Chen Pichler & Koulidobrova, 2015). It was observed that when a particular group of students were learning new signs, they watched each other and produced the signs correctly under the teacher's supervision. But after a certain time, their production seemed to be "transformed" into an incorrect version of the form, as depicted in Fig. 1 (cf. Humphreys et al., 2010).

Learners often identified their production as correct, even when it was not, after viewing their video clips. Signing for them is a new language mode, because their language experience is based chiefly on the vocal-auditory modality. Observation suggests that many of their production errors are phonological. Previous research has found that beginner learners have trouble articulating signs with multiple phonological parameters, and find the parameters of handshape and movement especially difficult (cf. Ortega & Morgan, 2015 on BSL, and Williams & Newman, 2016, on ASL). Existing sign language tests were not designed to target these skills specifically and were not being used to gather data that would support learners and their teachers beyond the testing situation. Therefore, the purpose of this research

[1] The content of sign language courses in the UK is chiefly determined by Signature, the national awarding body for BSL qualifications. The beginner level under Signature includes their modules BSL101, BSL102, BSL103, BSL201, BSL202, and BSL203. At the University of Central Lancashire, where the present study took place, these six modules are amalgamated into two large modules, namely BSL100 and BSL150, which are each taught for one semester (i.e. a full academic year is required to complete both). These modules correspond to level A1-A2 on the Common European Framework of Reference for Languages (CEFR).

Fig. 1 The standard form of the BSL sign STAR (top) and an articulation wherein the handshape is incorrect (bottom)

has been to apply video-based mock tests to meet this need and look into the kinds of phonological errors that are the most common for first-time learners of BSL.

The paper is organised as follows. Section 2 sets out the testing problem and rationale for the study and discusses some studies on sign language assessment. Section 3 comprises a literature review looking into the BSL teaching context and some research on phonological parameters and errors, as well as the concept of mock tests. Section 4 describes the study's participants, context, and method that was used to tabulate and analyse the phonological errors. Section 5 gives an overview and discussion of the findings, followed by Sect. 6 which explores some potential insights and recommendations arising from the findings. A conclusion is provided in Sect. 7.

2 Testing Problem Encountered

Testing is perceived as a key way of measuring a person's capacity to enter into and function within a particular environment, and language testing is an especially crucial factor in people's lives in the domains of work, study, health, and migration (McNamara, 2000). Because of this, language testing can be a form of 'institutional control' over individuals (McNamara, 2000, p. 4). It is therefore essential for the methods of language testing to be scrutinised (ibid.).

However, there is little research on sign language assessment and its connection to sign language teaching and research. Notably, several studies on assessing sign language vocabulary have been led by Tobias Haug (e.g. Haug, 2005; Haug et al.,

2019, 2022). The study explored in Haug et al. (2019) focusses on the development and evaluation of two Swiss German Sign Language (*Deutschschweizer Gebärdensprache,* DSGS) vocabulary-size tests: one relies on self-reporting by learners, where learners click 'yes' or 'no' to indicate whether or not they know each sign; and one uses a verifiable format where learners translate items from written German into DSGS. The authors find that both tests could be employed in different DSGS learning contexts as placement and/or diagnostic instruments for beginners. Haug et al. (2022) discusses the process of selecting vocabulary items for a DSGS assessment and feedback system that uses automatic sign language recognition. The aim of this system is to provide adult L2 learners with guidance on the manual parameters of the signs they articulate, including handshape, hand position, location, and movement. It can be used for summative testing where learners undertake a sentence-level exam as part of their module, and also receive automated feedback on their performance.

BSL tests are often not as effective as they could be in terms of informing teaching, because most of the time teachers do not carry out empirical analyses of learners' production. Consequently, there is a need to examine systematic ways in which sign language instructors can inform their teaching practice, increase the efficacy of their assessments, and improve their learners' outcomes by using video-based mock tests to analyse specific aspects of their BSL production.

The testing need that this local language test was designed to address was a predominance of phonological errors among level 1 BSL learners at my university, with relatively little progress being made in enabling them to target their phonological skills effectively. Many BSL learners have never had to distinguish between various possible finger, wrist and hand movements, for example keeping the hand and wrist still whilst fluttering the fingers and moving the arm forward (as in the sign for 'spider', see Fig. 2). To develop a system flexible enough to tune into both languages and both modalities, the learner needs explicit and direct instruction to

Fig. 2 'Spider' in BSL

learn L2 perception and production, under optimal conditions for general and functional phonological learning (cf. Bradlow, 2008). However, providing this kind of instruction is more difficult if the teacher has not empirically evaluated learners' target language output.

So far, there are few studies that have used video-based mock tests to categorise and analyse errors in BSL learners' production and address the local testing need of improving phonological awareness. Addressing this gap in the literature could benefit sign language teachers and learners by giving them evidence-based approaches to try in the classroom.

3 Literature Review

This literature review provides a background to the study by highlighting some extant research on phonological parameters in sign language linguistics (Sect. 3.1); exploring what is known about the causes of learners' phonological errors (Sect. 3.2); and looking at the concept of mock tests and how they are used in language learning (Sect. 3.3).

3.1 Phonological Parameters of BSL

Most current literature (i.e. Brien, 1992; Fenlon et al., 2015, 2018; Sutton-Spence & Woll, 1999) on sign language phonology assumes that there are five phonological parameters: handshape, location, orientation, movement, and non-manual features (see Fig. 3).

Fig. 3 The five phonological parameters of the BSL sign HAPPY

The five parameters can be studied through looking at 'minimal pairs' wherein only one parameter differs between signs (Bochner et al., 2011; Chen Pichler, 2011; Williams & Newman, 2016). For example, the BSL signs KNOW and DENTIST differ mainly in their location, and the signs PASTA and DNA differ mainly in their orientation. Figure 4 shows a selection of other minimal pairs found in BSL.

When students make a phonological error, it is possible that they are confusing two signs in a minimal pair. Williams and Newman (2016) looked at the phonological substitution errors of hearing M2-L2 learners of ASL. They found that most (64%) were movement errors, followed by errors related to handshape (31%) and location (5%). They suggest that movement is acquired later in the learning process than other parameters. Research on beginner BSL learners' articulation of phonological parameters is still limited. Ortega and Morgan (2015) found that handshape and movement were the most and second-most difficult parameters to articulate respectively. They also found that learners were less accurate in their articulations of signs with high levels of iconic motivation, probably because of these signs' similarities to common gestures. This gave the learners a sense of familiarity which caused them to pay less attention to these signs' precise phonology compared to more arbitrary signs (Ortega & Morgan, 2015).

3.2 Causes of Phonological Errors

This section briefly looks at some phenomena that have been identified in the literature as potential causes of phonological errors. The first is language transfer, and the second and third are the language universals known as motor skills and perceptual bias.

Spoken language users learning a sign language for the first time (who may be referred to as second-modality, second-language, or M2-L2, learners) may draw on their previous experience with common hearing gestures, such as co-speech gestures and facial gestures (Chen Pichler & Koulidobrova, 2015). It may be difficult for them to notice the phonological features that differentiate these gestures from the signs they are learning (ibid.). This might lead them to incorporate or assimilate the sign into their existing mental categories of gestures, so that they substitute gestural handshapes that are similar to sign language handshapes, resulting in inaccurate articulation (ibid.). This 'assimilation' phenomenon is known generally as 'language transfer' (Benson, 2002; Lado, 1957; Whitney, 1881).

Chen Pichler (2011) studied errors in the handshape parameter in four hearing non-signers' articulation of 10 gestures and 38 signs from ASL. She found that their handshape errors came from markedness and transfer. Evidence of language transfer affecting beginner learners of BSL was found by Ortega-Delgado (2013). The learners in his study did a sign repetition task prior to undertaking their first 11-week BSL course. The task required them to produce a range of individual BSL signs: iconic signs, less iconic signs, and arbitrary (non-iconic) signs. The researcher

Handshape - WORK / TALK (B handshape / G handshape)

Movement: LIFE / FEEL (up and down repeatedly / up once only)

Location: KNOW / DENTIST (forehead / mouth)

Orientation: PASTA / GENETICS (right hand palm orientated down and left hand up / right hand orientated right and left hand left)

NMF: APPOINTMENT / AGREEMENT (mouthing / mouthing and head nod)

Fig. 4 Examples of minimal pairs in BSL for each of the five parameters

examined their accuracy across four parameters: handshape, movement, location, and orientation. He found that their articulation of the parameters was less accurate for the more iconic signs. Ortega-Delgado surmised that this could be due to the participants recognising these signs as similar to gestures that they already knew, leading to assimilation of the target sign into their existing mental categories of iconic gestures. Because this assimilation may be less feasible for the more arbitrary signs, the participants were more likely to produce arbitrary signs correctly (ibid.).

In addition to transfer, the concept of 'language universals' can shed light on why signers produce phonological errors (Chen Pichler & Koulidobrova, 2015). Universals are factors that affect language acquisition cross-linguistically and cross-modally, and these include two that might be particularly likely to contribute to phonological errors: motor skills and perceptual bias (ibid.). An example of the former according to Chen Pichler and Koulidobrova (2015) is that beginner signers might use locations and movements that are too close to the body (see Fig. 5), because they have not developed the motor skills required to articulate signs at the correct distance. As an example of the latter, M2-L2 signers commonly produce 'mirror errors' involving use of the wrong hand or direction (ibid.). This is because they are articulating a 'mirror image' instead of the correct sign, due to their perception of the sign (ibid.).

An L1 learner 'tunes' their phonetic system to the patterns of sounds in their language, but for L2 language acquisition, this is different (Bradlow, 2008). L2 learners must progress from monolingual to bilingual. This means an adult L2 language learner must shift from tuning into the L1 sound structure to using a flexible system that can be tuned into the sound structure of both L1 and L2 (Iverson et al., 2003). It can be challenging for teachers to support the development of this flexible system in the classroom.

Fig. 5 The standard form of the sign EVENING (left) and an articulation wherein the hands are too close to the body (right)

Another challenge is targeting both comprehension and intelligibility. Derwing (2008) points out that there are important differences between these, and finding increases in comprehension does not necessarily reveal meaningful information about intelligibility. She mentions a study by Couper (2003) who analysed the pre-tests and post-tests of students learning L2 English pronunciation, and found fewer errors in both reading and production in the post-tests. Derwing (2008, p. 351) emphasises that while this supports the value of teaching pronunciation, it does not shed light on which errors, and which changes in production, have affected intelligibility. Similarly, fluency and accuracy are sometimes conflated by teachers even though they are two different concepts. A fluent signer may be able to produce language and convey meaning in real time but still have inaccurate articulation (Pallotti, 2009). For the purposes of this research, 'accuracy' can be generally defined as the correct pronunciation of every sign produced at the phonological level, without the focus on the semantic level as is often found in translation and interpreting studies.

Inaccurate articulation may involve one or more phonological parameters, such as movement or handshape. Several studies show that one of M2-L2 signers' most accurate parameters is location (Bochner et al., 2011; Corina, 2000; Marentette & Mayberry, 2000; Ortega-Delgado, 2013; Williams & Newman, 2016). Emmorey and Corina (1990) report that signers identify the location and orientation of a new sign first, and movement last.

Errors can be related to the use of inappropriate classroom resources such as still drawings that inaccurately reflect the sign and sometimes to learners confusing a sign with a similar gesture that they already know. For example, Fig. 6 shows the sign for 'thank you' with the correct location at the chin (top row) and incorrect location at the lips (bottom row). This particular error might be due to

Fig. 6 The standard form of the sign THANK-YOU (top) and an articulation wherein the location is incorrect (bottom)

classroom handouts that do not clearly show the correct location, and it may also stem from the fact that the sign looks similar to the common gesture for 'kiss' or 'blow a kiss'.

Sometimes resources are created without the expertise of L1 and fluent deaf signers, and unfortunately, they can remain in use for many years, hampering efforts to teach and learn correct phonology. Because of the brain's 'systemic plasticity', when a learner makes an error during language production, they might become predisposed to make the same mistake again, i.e. 'the error itself may be learned' (Humphreys et al., 2010, p. 151). So, the articulation of errors can have long-lasting effects on learners.

Deaf teachers can often discern a "hearing accent" (Chen Pichler, 2011) in non-deaf signers, even those who have been signing for many years (Ferrara & Lerose, 2017). This is often because of non-standard phonological features, similar to foreign accents among speakers. This can be a source of frustration, as teachers may work intensively on phonology and help a learner achieve accurate articulation through consistent and repeated language modelling, only to see their production regress, with previous mistakes reappearing. This is perhaps due to a substantial and additional difficulty that sign language learners must cope with, beyond what their spoken-language counterparts encounter; they must acquire a new modality at the same time as a new language. This can be termed as a new "channel" in Jakobson's (1966) theory of communication functions, which also include the levels of "code," "message" and "context," as well as the interactants, "sender" and "receiver."

Hearing speakers learning a spoken L2 can rely on their knowledge of sounds, words, inflections, intonations, etc. in their L1. They already have experience of acquiring a spoken language, and can sound out words and compare them to others. They just need to develop a more flexible system enabling them to tune into both their L1 and L2 sound structures, as they progress from monolingual to bilingual (Bradlow, 2008). However, a hearing speaker learning a signed L2 must not only learn a new language but also acquire the ability to communicate in the visual-gestural modality for the first time.

Absorbing utterances through the eyes only, and producing utterances with the hands and face, is often not intuitive for this learner group and requires them to build a complex set of modality skills almost from scratch, whilst also developing L2 knowledge, i.e. lexicon and grammar. Most of the language theories or models that they rely on cognitively, even unconsciously, are based on sound and speaking. Whilst they may have had their spoken pronunciation corrected by parents, teachers or friends on countless occasions since they first learned to talk, most will have very little if any experience of their handshape being corrected. They also lack experience with simultaneity (articulating two signs at the same time), exploiting the signing space, sustaining eye gaze, using one versus two handed signs, and attending to differences in active and passive hands and phonological parameters such as orientation, direction, location, and non-manual features (such as eyebrow movements).

3.3 Concept of Mock Tests

Mock tests are used for a variety of reasons and aim to reap many benefits such as to support learning by verifying what students have learned; give students the chance to practise and validate what they think they know; provide immediate knowledge of results and feedback; give students a chance to monitor their progress; and also allow tutors to create the materials needed to track students' improvement (González Romero, 2016).

Research shows that students who take mock or practice tests on the same material as the real test often score better than students who do other things to prepare such as re-studying or filler activities (Adesope et al., 2017). This concept of 'testing effect', a term from cognitive psychology (Roediger & Karpicke, 2006), means increases or gains in learning and retention that can happen when students take mock or practice tests. The term 'testing effect' is sometimes used as a reference to the assumption that mock or practice tests have greater learning benefits than other common study strategies (Adesope et al., 2017). Naujoks et al. (2022) found that when students took part in mock or practice tests, their exam performance was better and they made more accurate metacognitive judgements. The authors conclude that participation in practice tests is a valuable learning strategy that stimulates students' ability to expand on the learning content and contributes to their ability to estimate their own personal performance (see also Barenberg & Dutke, 2018; Schwieren et al., 2017). The literature on the use of mock tests to meet local language-testing needs is rather scarce, but practically speaking, the flexibility and immediacy of mock tests makes them seem an ideal tool for addressing specific local needs on the fly.

In addition, practising the physical aspects of the assessment provides an opportunity for students to test out the assessment environment and this can reduce the anxiety that comes with the pending marked and graded assessment (Pérez Castillejo, 2019; Turner, 2009). Research by Jenkins (2008) also shows that students find it useful to have a mock exam in the curriculum to better prepare them for the final exam, and that mock exams held a few weeks before the end of the semester reduce their anxiety.

4 Methods

The data for the research is the BSL production of 10 participants, all higher education students aged 18 and over who had no prior experience either with a sign language or with the Deaf community. This is the group of learners whose local testing need was identified. During the first didactic session (week 1), the students were informed that their video-based mock tests would be used by the teacher not only to give them feedback on an individual basis immediately afterward, but also for analytical purposes in order to improve his teaching strategies. They were told that only the teacher would have access to their tests, and that none of their names or any

other identifying information would be published. The students were advised that to safeguard everyone's privacy, they would not be able to watch each other's video clips or each other's feedback sessions with the teacher.

Individualised feedback from the teacher immediately after the mock test is always provided to give students an idea of how well the quality and quantity of their signed output will meet the criteria for the real test.

By the time the video-based mock tests were carried out, the students had been attending their module for 11 weeks, which amounted to about 44 h of face-to-face instruction in class. (The full semester is 12 weeks, involving 48 h of instruction.) During these 11 weeks, linguistic theory and meta-linguistic concepts were not always covered in class (though some teachers may discuss them to a limited extent), as these components are only included in the curriculum in years 2 and 3, not in year 1.[2]

Each student had been sending a video clip every week through the FastFolder, which is software created by the University of Central Lancashire (UCLan) that allows teachers to create files for collecting the clips that students submit after completing their various assignment tasks. Every week there was a different assigned topic that corresponded to what was taught in the lecture, for example family, food and drink, transport, clothes, shopping, and descriptions of people. In weeks eight and nine respectively, they had their first mock test and real test. This first assessment was different to the mock and final test in weeks 11 and 12, as it covered both comprehension and production instead of just production, and the production component was more constricted and did not ask students to sign about a topic of their choice. However, this meant that by week 11, the students had already had the experience of filming themselves and participating in the sequence of a BSL mock and real assessment.

For the mock test in week 11, they were invited to produce a 3-min video clip of themselves signing about any or all of the topics of the prior weeks. A typical clip included students introducing themselves, listing their family members, explaining what transport they used to get to university each day, etc. The teacher-researcher was in his office while the students were recording themselves, to avoid causing them undue anxiety. Sometimes students came to the office to ask a question about the test, in which case the teacher had to go back into the room to answer the question, for example clarifying some steps of the assessment or confirming if their linguistic production is clear. This may seem unusual to teachers in contexts where mock tests do not allow conversation with the teacher, but in sign language mock tests, it is common to permit this, because one of the aims is to give learners the maximum amount of opportunity to improve their linguistic production before the assessment, regardless of the number of requests. (In the real test, 1 week after the mock, the students were only allowed to ask the teacher about technological issues with the computers or cameras). The students had 2 h to record and re-record as

[2] This is because to a large extent, university and other courses in the UK that teach sign language follow the requirements and curriculum of Signature, which oversees the certification and regulation of BSL courses. Because most of the linguistics content within the Signature curriculum is at level 3 and above, level 1 and 2 students tend to have minimal exposure to linguistics concepts.

much as they needed to. They were asked to save the final clip in FastFolder at the end of the 2 h. They were not using any new signs that they had just acquired that week, because they had learnt them all in prior weeks. But they were still free to use their own preferred style and format, as they had in the previous weeks, like a sequential narrative or a summary.

A few days after the mock test, the teacher spent 15–30 min watching each student's clip and discussing it with them in a one-to-one tutorial focussing on the local testing need of increasing phonological awareness. At a later date, after the cohort had finished the course and received their marks, the teacher went through the clips from the mock exam to identify, categorise and count the phonological errors that appeared. Because the time required for this exceeded what the teacher was able to accomplish in 1 week, it was unfortunately not possible to show the aggregated results to the learners before their mock test. However, about 1 year later, the teacher presented the aggregated results at a seminar to which all BSL students were invited, and as this cohort were in their first year at the time of the mock test, they were all still at the university at that time of the seminar, and several of them attended it.

To categorise the errors, the researcher used the five phonological parameters that are discussed in foundational works on BSL (i.e. Brien, 1992; Fenlon et al., 2015, 2018; Sutton-Spence & Woll, 1999): handshape, location, orientation, movement and non-manual features. The researcher viewed each clip and identified all of the occurrences of errors including the sign involved and the type of error, such as FAMOUS (see Fig. 7) located at the mouth instead of the chin (location error).

Each error identified by the researcher was also checked with a sign language teacher colleague, who was asked to either verify or reject it; all of the errors were verified. A spreadsheet was created showing the signs for which participants produced an error (see Table 1). Only those errors that were produced by at least three participants were included in the analysis, to avoid focussing on errors that may be the result of individual idiosyncrasies (Yu & Zellou, 2019), and highlight errors that are more likely to be the result of some inherent difficulty with an aspect of a sign's phonology.

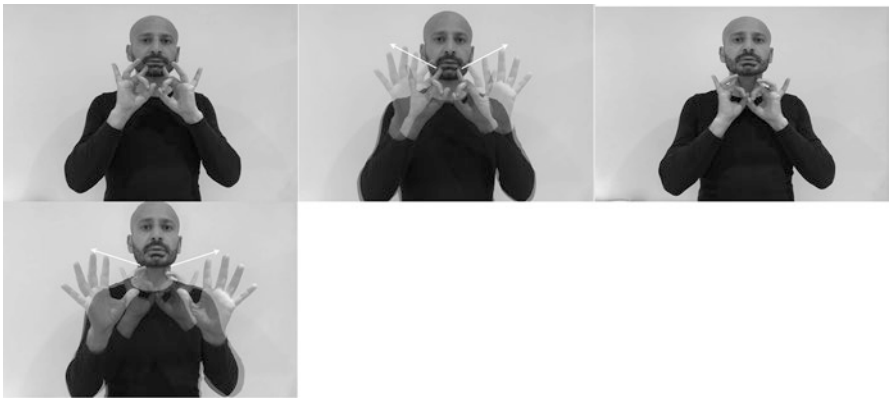


Fig. 7 Correct articulation of the BSL sign FAMOUS

Table 1 Numbers of errors in each phonological parameter produced by each of the 10 participants

	Total	H	M	L	O	NMF
Participant-1	7	2	3	1	1	
Participant-2	8	3	2	1	2	
Participant-3	3	1	1		1	
Participant-4	2	2				
Participant-5	3	1	2			
Participant-6	5	2	1	1	1	
Participant-7	13	5	5	1	1	1
Participant-8	5	2	3			
Participant-9	8	5	2	1		
Participant-10	4	2	2			
	58	25	21	5	6	1

5 Results and Discussion

The pie chart (see Fig. 8) is a summary of the results that are listed in more detail in Table 1. These results support those of Ortega and Morgan (2015), who found that handshape and movement errors were substantially more numerous than errors in the other parameters. Interestingly, movement errors were repeated across signers much more than handshape errors. Further research is needed to determine whether this finding can be replicated in a larger-scale study and propose more precise reasons for it. Occurrences of handshape errors were numerous but idiosyncratic, i.e. produced only by one signer and not repeated across multiple signers. This may be related to the fact that distinct handshapes have been investigated in the literature to a far greater extent than distinct movements. This had an effect on the analysis, because the researcher was able to identify differences between handshapes more easily than differences between movements. This may have contributed to the perception that movement errors were repeated more often.

There were also mirror errors in the production of fingerspelling and numbers, but these two categories of forms were omitted from the data. This is because these are probably much more likely to involve handshape and location errors and thus skew the results toward these types of errors. In addition, it was important to group like with like, i.e. lexical signs with other lexical signs. Fingerspelled words are not normally considered to be signs in the same way as lexical forms are. In fact, the inclusion of fingerspelling as an integral part of BSL is sometimes controversial in the literature (Brown & Cormier, 2017; Sutton-Spence, 1994).

The same errors were found in the same group of signs repeatedly. Some of the signs for which movement errors were recorded belonged to the same semantic domain, such as money. Many of the participants produced RENT, BUY, SELL, PAY, and MONEY (see Fig. 9) with movement errors. Sometimes, it appeared that the signers were confusing the movement of one sign with that of another, which perhaps suggests that they were cognitively grouping these signs together into one

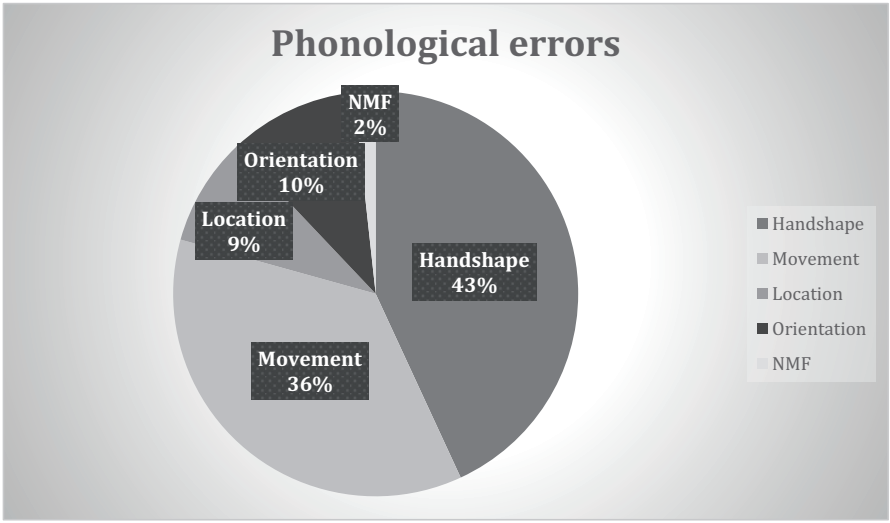


Fig. 8 Division of the 58 total errors from the study's 10 participants into the five phonological parameters

Fig. 9 The BSL signs RENT, BUY, SELL, PAY, and MONEY

semantic domain but confusing their articulations. This is similar to what Williams and Newman (2016) found for hearing learners of ASL, who produced more errors in the movement parameter than other parameters, possibly due to organising their mental lexicon into minimal pairs and semantically related signs.

For example, the sign RENT was produced with one forward movement instead of the correct two, which may be due to confusing the form of RENT with that of PAY (which does have one forward movement). Indeed, another error was that participants signed PAY with two forward movements.

Fig. 10 The BSL signs NAME and COUNCIL

Future research could perhaps examine signs of similar frequency and semantic complexity, so that infrequent and/or complex or abstract signs are not being compared with common and simple ones. For example, in these data, the sign NAME has few errors and COUNCIL has many. This could be due to something inherent in the phonology of these signs (see Fig. 10), for example the wrist twist in COUNCIL being difficult to articulate, but could also be due to 'council' being a less frequent concept than 'name', especially in the context of the sign language classroom.

6 Insights Gained

The findings suggest that teachers can use video-based mock tests to meet local testing needs, and identify signs that may cause particular problems for students, and concentrate more explicitly on the phonology of such signs to make sure that each student is able to produce them. In doing so, teachers may be able to "determine the relative gravity of various pronunciation errors and set up a system of teaching priorities" (Chan & Li, 2000, p. 83).

The method is relatively accessible and straightforward to implement in a teacher's own action-based research (Rosen et al., 2015) to inform their reflection or even as a classroom exercise for students. For example, a teacher might ask their students to watch clips of themselves from the mock test and identify different types of errors, and tabulate the results in the manner described above. Students might benefit from identifying the most common phonological errors and developing skills in self-analysis using their knowledge of phonological parameters.

Feedback is a crucial part of the language development process and can aid learners to develop elements of BSL grammar that are difficult to master, such as the use of aspect. This feedback is better when delivered formatively, throughout the course, rather than after a mock assessment that is so close to the actual end assessment in terms of time. The local testing needs can be assessed again during the mock, and further feedback can be given, but there are certainly additional advantages in providing guidance much earlier in the course.

Due to the impossibility of analysing, cross-checking and aggregating the data from the mock test in the space of 1 week, it is difficult to envisage how the results could be fully deployed to inform the test preparation of the same cohort from whom they were gathered. In fact, the cohort whose local testing need was identified may be in a different year and have a different teacher by the time the analysis is complete. Nonetheless, presenting the results to students and teaching colleagues at a seminar was still a fruitful way to disseminate the findings to the students. Other ways might include showing the results to the next cohort of first year students when they are preparing for their mock test even if their mock test is targeting a different local testing need. This could be used as a motivational tactic, i.e. inspiring the students to aim for fewer errors as a class in a particular category than the previous year's cohort. Individual tutors or groups of teaching colleagues could perform these analyses on a longitudinal basis to see if it is possible to generate comparisons and identify trends. At the same time, the results suggest that it is possible for teachers to use video-based mock tests to target local testing needs for specific groups of learners.

7 Conclusions: Implications for Test Developers and Users

This paper has described a small-scale study on addressing the local testing need of a group of beginner BSL learners who needed to increase their phonological awareness, and discussed how video-based mock tests might be used in the sign language classroom. Existing sign language tests were not designed to target phonology specifically and were not being used to gather data that would support learners and their teachers beyond the testing situation. Therefore, this research sought to apply video-based mock tests to meet this need and look into the kinds of phonological errors that are the most common for first-time learners of BSL. The vast majority of the common errors produced by these level 1 students were in the handshape and movement parameters, in accordance with Ortega and Morgan (2015), but movement errors were repeated across signers more than handshape errors, which were more idiosyncratic. Several of the movement errors appeared to be caused by confusion between signs in the same semantic domain, such as RENT and PAY. Other potential causes of these errors were considered, including the challenge of acquiring a new modality (Chen Pichler & Koulidobrova, 2015).

The findings suggest that using video-based mock tests may provide teachers with an avenue for meeting the local needs of specific cohorts of learners, such as raising awareness of phonological strengths and weaknesses. Devising, administrating and analysing these tests could help teachers better support their learners as they progress towards the next level of BSL study.

Acknowledgements This study was developed based on discussions at the third LESICO conference (the International Conference of Sign Language Teachers) in 2017; the first Sign Cafè event (international workshop on cognitive and functional approaches to sign language linguistics) in July 2018; and the third PRO-Sign Conference (a series focussing on sign language teaching in Europe, funded by the European Center for Modern Languages at the Council of Europe) in October 2018. The author is grateful to the colleagues that have so generously offered their insights and expertise at these events, which have inspired this strand of work. The author was provided with a grant from the University of Central Lancashire's Arts and Humanities Research Academy (AHRA) to support work on the images for this paper. Finally, the author would like to thank Jenny Webster for providing English language support.

References

Adesope, O. O., Trevisan, D. A., & Sundararajan, N. (2017). Rethinking the use of tests: A meta-analysis of practice testing. *Review of Educational Research, 87*(3), 659–701. https://doi.org/10.3102/0034654316689306

Barenberg, J., & Dutke, S. (2018). Testing and metacognition: Retrieval practise effects on metacognitive monitoring in learning from text. *Memory, 27*(3), 269–279. https://doi.org/10.1080/09658211.2018.1506481

Benson, C. (2002). Transfer/cross-linguistic influence. *ELT Teaching, 56*(1), 68–70. https://doi.org/10.1093/elt/56.1.68

Bochner, J. H., Christie, K., Hauser, P. C., & Searls, J. M. (2011). When is a difference really different? Learners' discrimination of linguistic contrasts in American sign language. *Language Learning, 61*, 1302–1327. https://doi.org/10.1111/j.1467-9922.2011.00671.x

Bradlow, A. R. (2008). Training non-native language sound patterns: Lessons from training Japanese adults on the English /p/–/l/ contrast. In J. G. Hansen Edwards & M. L. Zampini (Eds.), *Phonology and second language acquisition* (pp. 287–308). John Benjamins.

Brien, D. (Ed.). (1992). *Dictionary of British sign language/English*. Faber and Faber and the British Deaf Association.

Brown, M., & Cormier, K. (2017). Sociolinguistic variation in the nativisation of BSL fingerspelling. *Open Linguistics, 3*, 115–144. https://doi.org/10.1515/opli-2017-0007

Chan, A., & Li, D. (2000). English and Cantonese phonology in contrast: Explaining Cantonese ESL learner's English pronunciation. *Language, Culture and Curriculum, 13*(1), 67–85. https://doi.org/10.1080/07908310008666590

Chen Pichler, D. (2011). Sources of handshape error in first-time signers of ASL. In G. Mathur & D. J. Napoli (Eds.), *Deaf around the world: The impact of language* (pp. 96–121). Oxford University Press.

Chen Pichler, D., & Koulidobrova, H. (2015). Acquisition of sign language as a second language. In M. Marschark & P. E. Spencer (Eds.), *The Oxford handbook of deaf studies in language* (pp. 218–230). Oxford University Press.

Corina, D. P. (2000). Some observations regarding Paraphasia in American sign language. In K. Emmorey & H. Lane (Eds.), *The signs of language revisited: An anthology to honor Ursula Bellugi and Edward Klima* (pp. 493–507). Lawrence Erlbaum.

Couper, G. (2003). The value of an explicit pronunciation syllabus in ESOL teaching. *Prospect, 18*(3), 53–70. http://hdl.handle.net/10292/1524

Derwing, T. M. (2008). Curriculum issues in teaching pronunciation to second language learners. In J. G. Hansen Edwards & M. L. Zampini (Eds.), *Phonology and second language acquisition* (pp. 347–369). John Benjamins.

Emmorey, K., & Corina, D. (1990). Lexical recognition of signed languages: Effects of phonetic structure and morphology. *Perceptual and Motor Skills, 71*, 1227–1252. https://doi.org/10.2466/PMS.71.8.1227-1252

Fenlon, J., Cormier, K., & Schembri, A. (2015). Building BSL SignBank: The lemma dilemma revisited. *International Journal of Lexicography, 28*(2), 169–206. https://doi.org/10.1093/ijl/ecv008

Fenlon, J., Cormier, K., & Brentari, D. (2018). The phonology of sign languages. In S. J. Hannahs & A. R. K. Bosch (Eds.), *The Routledge handbook of phonological theory* (pp. 453–475). Routledge. https://doi.org/10.4324/9781315675428

Ferrara, C., & Lerose, L. (2017). *Teaching L2 and developing M2*. Presentation at LESICO, the 3rd international congress of sign language teachers, Basel, Switzerland, 13–15 October 2017. Available at: https://www.youtube.com/watch?v=46B3oXVXQuA. Accessed 25 Jan 2018.

González Romero, R. (2016). The implications of business English mock exams on language progress at higher education. In A. Pareja-Lora, C. Calle-Martínez, & P. Rodríguez-Arancón (Eds.), *New perspectives on Teaching and working with languages in the digital era* (pp. 293–302). Research-publishing.net. https://doi.org/10.14705/rpnet.2016.tislid2014.442

Haug, T. (2005). Review of sign language assessment instruments. In A. Baker & B. Woll (Eds.), *Sign language acquisition* (pp. 51–86). John Benjamins.

Haug, T., Ebling, S., Boyes Braem, P., Tissi, K., & Sidler-Miserez, S. (2019). Sign language learning and assessment in German Switzerland: Exploring the potential of vocabulary size tests for Swiss German sign language. *Language Education & Assessment, 2*(1), 20–40. https://doi.org/10.29140/lea.v2n1.85

Haug, T., Ebling, S., Tissi, K., Sidler-Miserez, S., & Boyes Braem, P. (2022). Development of a technology-assisted assessment for sign language learning. *International Journal of Emerging Technologies in Learning (iJET), 17*(06), 39–56. https://doi.org/10.3991/ijet.v17i06.26959

Holmström, I. (2019). Teaching a language in another modality: A case study from Swedish sign language L2 instruction. *Journal of Language Teaching and Research, 10*(4), 659–672. https://doi.org/10.17507/jltr.1004.01

Humphreys, K. R., Menzies, H., & Lake, J. K. (2010). Repeated speech errors: Evidence for learning. *Cognition, 117*, 151–165. https://doi.org/10.1016/j.cognition.2010.08.006

Iverson, P., Kuhl, P. K., Akahane-Yamada, R., Diesch, E., Tohkura, Y., Kettermann, A., & Siebert, C. (2003). A perceptual interference account of acquisition difficulties for non-native phonemes. *Cognition, 87*, B47–B57. https://doi.org/10.1016/s0010-0277(02)00198-1

Jakobson, R. (1971[1966]). *Selected writings II: Word and language*. Mouton.

Jenkins, S. J. (2008). Is there value for administering a mock corroborative clinical practicum prior to the final exam? A two-year perspective for 1st year clinical dental hygiene students. *Journal of Dental Hygiene, 82*(5). Poster presented at the American Dental Hygienists' Association October 2008.

Ladd, P. (2003). *Understanding deaf culture*. Multilingual Matters.

Lado, R. (1957). *Linguistics across cultures*. University of Michigan Press.

Marentette, P. F., & Mayberry, R. I. (2000). Principles for an emerging phonological system: A case study of early ASL acquisition. In C. Chamberlain, J. P. Morford, & R. I. Mayberry (Eds.), *Language acquisition by eye* (pp. 71–90). Lawrence Erlbaum.

McNamara, T. (2000). *Language testing*. Oxford University Press.

Meier, R. P. (2012). Language and modality. In R. Pfau, M. Steinbach, & B. Woll (Eds.), *Sign language: An international handbook* (pp. 574–601). Mouton de Gruyter.

Naujoks, N., Harder, B., & Händel, M. (2022). Testing pays off twice: Potentials of practice tests and feedback regarding exam performance and judgment accuracy. *Metacognition and Learning, 17*, 479–498. https://doi.org/10.1007/s11409-022-09295-x

Ortega, G., & Morgan, G. (2015). Phonological development in hearing learners of a sign language: The influence of phonological parameters, sign complexity, and iconicity. *Language Learning, 65*(3), 660–688. https://doi.org/10.1111/lang.12123

Ortega-Delgado, G. (2013). *Acquisition of a signed phonological system by hearing adults: The role of sign structure and iconicity*. PhD dissertation, University College London.

Pallotti, G. (2009). CAF: Defining, refining and differentiating constructs. *Applied Linguistics, 30*(4), 590–601. https://doi.org/10.1093/applin/amp045

Pérez Castillejo, S. (2019). The role of foreign language anxiety on L2 utterance fluency during a final exam. *Language Testing, 36*(3), 327–345.

Roediger, H. L., & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science, 17*, 249–255. https://doi.org/10.1111/j.1467-9280.2006.01693.x

Rosen, R. S., Turtletaub, M., DeLouise, M., & Drake, S. (2015). Teacher-as- researcher paradigm for sign language teachers: Toward evidence-based pedagogies for improved learner outcomes. *Sign Language Studies, 16*(1), 86–116. https://doi.org/10.1353/sls.2015.0026

Saville-Troike, M., & Barto, K. (2016). *Introducing second language acquisition*. Cambridge University Press.

Schwieren, J., Barenberg, J., & Dutke, S. (2017). The testing effect in the psychology classroom: A metaanalytic perspective. *Psychology Learning & Teaching, 16*(2), 179–196. https://doi.org/10.1177/1475725717695149

Signature. (2019). Who we are. *Signature: Excellence in communication with deaf people* [online]. Available at: https://www.signature.org.uk/who-we-are. Accessed 17 Jan 2019.

Sutton-Spence, R. (1994). *The role of the manual alphabet and fingerspelling in British sign language*. University of Bristol dissertation.

Sutton-Spence, R., & Woll, B. (1999). *The linguistics of British sign language: An introduction*. Cambridge University Press.

Turner, C. E. (2009). Examining washback in second language education contexts: A high stakes provincial exam and the teacher factor in classroom practice in Quebec secondary schools. *International Journal of Pedagogies and Learning, 5*(1), 103–123. https://doi.org/10.5172/ijpl.5.1.103

Whitney, W. (1881). On mixture in language. *Transactions of the American Philological Association, 12*, 5–26. https://www.jstor.org/stable/i348238

Williams, J., & Newman, S. (2016). Phonological substitution errors in L2 ASL sentence processing by hearing M2-L2 learners. *Second Language Research, 32*(3), 347–366. https://doi.org/10.1177/0267658315626211

Yu, A. C. L., & Zellou, G. (2019). Individual differences in language processing: Phonology. *Annual Review of Linguistics, 5*, 131–150. https://doi.org/10.1146/annurev-linguistics011516-033815

An Indigenous Rubric for Assessing Dispensing Drugs in English of Thai Pharmacy Students: Reliability and Perception Issues

Sasithorn Limgomolvilas and Jirada Wudthayagorn

Abstract Creating a rubric to assess English for language specific purposes, where specific and indigenous contexts need to be well-balanced, is challenging. In a pharmacy course, dispensing drug is one of the most important skills required of pharmacy students. Drawing on the indigenous criteria of domain experts, criteria for assessing Thai pharmacy students' dispensing drug skills in English were developed. The rubric consists of three areas: pharmaceutical science skills, language use, and strategic competence. The study set out to provide validity evidence for the indigenous rubric through Many-Facet Rasch Measurement (MFRM) and interviews with the raters. Results showed that while the raters could reliably assess skills related to pharmaceutical knowledge, students were rated more harshly on language use and strategic competence. Pharmaceutical knowledge was rated more leniently. Follow-up interviews with raters supported the usefulness of the rubric in assessing pharmacy students' performances. The rubric proved useful for language instructors as it is easy to use and reliable. The students also have been able to better understand what language instructors and content experts expect with regard to content knowledge and their language performances.

Keywords Indigenous rubrics · Dispensing drugs · Role-play · Reliability · Perception

1 Introduction

English has long been an important core subject for Thai students at all levels: from primary to higher education. When Thai students study at universities, they are required to complete at least 12 English credits (3 credits for each course). The first 6 credits are devoted to general English skills, and the remaining six are for English

S. Limgomolvilas · J. Wudthayagorn (✉)
Chulalongkorn University Language Institute, Chulalongkorn University, Bangkok, Thailand
e-mail: sasithorn.li@chula.ac.th; jirada.w@chula.ac.th

X. Yan et al. (eds.), *Local Language Testing*, Educational Linguistics 61,
https://doi.org/10.1007/978-3-031-33541-9_11

for Academic Purposes (EAP) or English for Specific Purposes (ESP). English for Specific Purposes centers on many fields including that of medicine, aviation, and tourism, to name a few. In Thailand, however, few studies have looked at the integration of disciplinary knowledge and local language assessment of EFL pharmacy students.

As requested by the Pharmacy Faculty at a university, an English course that can equip Thai students with skills used at the workplace was developed. To this end, this study investigates an English for Pharmaceutical Science Professional course at one prestigious Thai university. The selected course involves fifth-year pharmaceutical students, who are taking their final English class immediately before being placed in internships at hospitals, affiliated pharmaceutical companies or organizations. The aim of this course is to provide English skills to students to be prepared for the work situation. It can be said that this course is delineative of English to Pharmacy courses in Thailand.

To derive areas of teaching and assessing students, we have considered Nawanopparatsakul et al. (2009–2010) who have established in their research that by the time of graduation, pharmacy students must be able to "provide patient care in cooperation with patients, manage and use resources of the health care system, and promote health improvement, wellness, and disease prevention in cooperation with patients." (p. 28). Drawing upon the work of Nawanopparatsakul et al. (2009–2010) in conjunction with a needs analysis conducted by Limgomolvilas and Wudthayagorn (2019), which aggregated viewpoints shared by the teaching staff from the Faculty of Pharmaceutical Science (who designed the curriculum for the Pharmacy program), the instructors from the English Language Institute (who were in charge of the English courses in the curriculum), and additional pharmacists (who earned the Bachelor of Pharmacy) (Limgomolvilas & Wudthayagorn, 2019), the authors conclude that language knowledge, communication skills, and content knowledge are three pivotal areas for teaching and assessing the students in this course.

In terms of the content for the English for Pharmaceutical Science Professional course, the curriculum was built upon a needs analysis study (Limgomolvilas & Wudthayagorn, 2019). The teaching staff at the Faculty of Pharmaceutical Science where the study took place and a group of licensed pharmacists have found that dispensing drugs must be included in this course (Limgomolvilas & Wudthayagorn, 2019). This is because dispensing drugs is a specialized skill, which cannot be assessed solely on language knowledge and communication skills. It is also important to note that, in addition to the English for Pharmaceutical Science Professional course in this study, the Pharmacy program has embedded two communication skills courses in Thai to prepare students for their internships, including preparation classes for taking the license examination and for working in a drugstore. It can be summarized that language knowledge, communication skills, and content knowledge are crucial for the students to be able to dispense drugs effectively.

In regards to assessment, there is an opportunity to develop a rubric that combines both language skills and content knowledge. However, for language

instructors who are not experts in the pharmaceutical content, the extent to what level and how much of the content to be applied is unclear. To fix this issue, the authors believe that developing a rubric that combines key features (i.e., language knowledge, communication skills, and content) is warranted. Wudthayagorn (2015) suggested that in order to help pharmacy students master dispensing skills, performance-based assessment, such as role-play, should be encouraged. The practice requires students to integrate their content knowledge, language knowledge, and communication skills to perform a role-play test task effectively. To this end, an indigenous criteria of Thai pharmacy dispensing skills were developed accordingly so that English instructors could use them to assess student performance and offer essential and relevant feedback.

2 Test Purpose and Test Context

To fulfill one of the learning objectives, stated in the English for Pharmaceutical Science Professional course, which emphasizes the students' ability to dispense drugs effectively, a role-play test task was developed to determine whether students had achieved the learning objective. Two important elements for the test task were given to the students so that they could prepare and practice. The first aspect was a pharmaceutical case covering simple details of a patient's symptoms and information which students could elicit before dispensing drugs. The second element was an indigenous rubric with detailed criteria. After the teaching unit on pharmacist-patient communication had been covered during a period of two weeks, the students had one week to practice the dispensing skills learned from the class.

Because the test could not be administered to every student simultaneously, time allotment and administration needed to be well organized. A total of five raters (that is, all class instructors) were employed to assess the cohort group of 145 students. The role-play test was administered in two time slots, 9:00–12:00 and 13:00–16:00. Eight pharmaceutical cases were divided according to the time slots as four cases in the morning paralleled the four cases in the afternoon.

A role-play test manual was distributed to all parties, composing of three sections, which are guidelines for administrators, raters, and students. The detailed information in the manual dictates the assessment process. This guides all parties to conduct the role-play similarly and accommodates those who were new to the process.

When doing role-plays, students acted as pharmacists while raters, who were not their class instructor, acted as patients and rated the students simultaneously. To ensure reliability, all role plays were videotaped so that another instructor could double rate the performances. After the rating process, students received their test results as well as feedback. This would be beneficial for students to further develop their dispensing skills when they have another internship at the drug stores after finishing this course.

3 Testing Problem Encountered

In 2013, when the authors first taught the students, the following rubric was used to assess the students' ability in dispensing drugs (Table 1).

Four criteria, covering content, language, voice and non-verbal communication, were assessed. The scores of each criterion ranged from one (or poor) to five (or excellent). For example, if the students provided relevant content to the case, they would get scores of four or five, based on the raters' impression. Alternatively, if their language was very clear, fluent and appropriate, they would get scores of four or five. Instructors could add a comment if necessary. After employing this rubric over several semesters, the raters concluded that based on their observation from the scores and raters' reports, the adjectives in the parenthesis were somewhat vague and could lead to unreliable judgments from one rater to another. The raters reported that they were unsatisfied with how students' dispensing performance could not be accurately reflected with this rubric, especially for the content criterion, an important aspect in judging the aptitude for drug dispensation. In addition, language instructors may have found it difficult to provide meaningful formative feedback to the students due to the fact that the rubric appeared too simple and general. Since the raters were not specialists in the pharmaceutical science field, specific details to advise students to perform better at drug dispensation could not be provided effectively unless a Thai standard similar to that found on the national license examination were deployed. It was thought that such modification would benefit the students greatly in terms of validity and positive washback as it was the same standard that had been conventionally found in internships and on pharmaceutical license examinations in Thai. In brief, the rubric previously utilized (as shown in Table 1) could neither benefit language instructors when giving feedback to the students, nor students when using the rubric as a guideline to practice dispensing drugs. Thus, this particular local problem motivated the authors to design an indigenous rubric for specific learning outcomes.

4 Literature Review

Before we present the methods used to answer the research questions, we provide an overview of literature related to indigenous criteria in ESP and research related to other ESP tests in medical English.

Table 1 Scoring rubric for dispensing drugs

Criteria	Comment
Content (relevant)	1 2 3 4 5
Language (clear, fluent, appropriate)	1 2 3 4 5
Voice (correctly pronounced, appropriate for the situation)	1 2 3 4 5
Non-verbal communication (appropriate, professional)	1 2 3 4 5

Wudthayagorn (2015, p. 132)

4.1 Indigenous Criteria in English for Specific Purposes

The term "indigenous criteria" was first coined by Jacoby (1998), who first described indigenous criteria as the values expressed by a group of physicists when critiquing each other's conference presentations. Pill (2013) defines indigenous criteria as "criteria applied to performance in a particular context by insiders who share a common perspective," and was based on Jacoby and McNamara's work (1999), whose indigenous criteria were established collaboratively by consulting with informants, pharmacists working in the target field, and pharmacy students. The necessity for collaborative work to include specific background knowledge has been recognized (Elder & McNamara, 2016); likewise, professional knowledge should also be adequately represented in collaborative work (Elder et al., 2012). Their concepts correspond to the three qualities that all ESP assessments consider crucial when creating an indigenous rubric; that is, that they cover language use in a specific context, that they use precise specific purpose content, and that there is an interaction between specific purpose language and specific purpose background knowledge (Douglas, 2001).

4.2 Tests for Language for Specific Purposes

The Occupational English Test (OET*)* is used to assess and certify health professionals who wish to work in Australia, New Zealand, and Singapore. It is worthwhile to review the OET because the OET is performance-based and has been established as authentic by extensive literature. The criteria used in the OET follows the laws of its respective country and only measures test takers' linguistic ability without assessing their professional competence. In the literature, numerous studies have looked at various ways to improve the quality of the OET. For example, in the area of authenticity, recent qualitative studies (O'Hagan et al., 2016; Pill, 2016; Woodward-Kron & Elder, 2016) have adjusted the OET, including the approved and attested indigenous criteria on communication, which is derived from expert judgment with the use of partial credit scale: *clinician engagement* and *management of interaction*. In the OET, two new criteria were added in order to increase the authenticity of the test, including criteria relevant to professionals while allowing for judgment by language experts. Other sets of studies (Elder & McNamara, 2016; Macqueen et al., 2016; Pill & McNamara, 2016; Woodward-Kron & Elder, 2016) were conducted in order to improve the quality of the OET and relied on professional perspectives on the scale criteria.

Apart from the OET, two other performance-based assessments aimed at pharmacist-patient communication are the Objective Structured Clinical Examination (OSCE) and the Objective Structured Pharmacy Examination (OSPE). Given that these tests have been extensively reviewed for validity and reliability, classroom assessment research might be able to adopt these well-established international and national tests into the curriculum readily; however, it is likely that

none of them are suitable options when the attributes of individual language usage and the characteristics of target language usage situations in Thailand are considered. Although several international English tests have been developed, they do not respond directly to indigenous assessment/criteria (Jacoby & McNamara, 1999). Therefore, an indigenous assessment using local contexts was developed as this study focuses on Thai pharmaceutical students, an aspect of assessment that matters most in this local context. The focus of this paper is on the reported observation that Thai pharmacists often resorted to quick consultation and dispensation of drugs in English. This is based on feedback that the Faculty of Pharmaceutical Science at the university of the present study had previously received from some supervisors overseeing students at their internships, criticizing the students' inability to communicate effectively in English when dispensing drugs to foreigners. Such observations are in line with what has been reported in the literature—in one study, Byrnes (2008) notes that an impetus for developing locally pertinent content arose from demands by professionals and educational programs calling for further skill development.

When examining the methods of validating a rubric for ESP assessment, most of the research was found to have applied both qualitative and quantitative methods, while relying heavily on the latter. An example of such application is the study by Hyvärinen et al. (2012) whose team utilized an analytic rubric in order to collect a large amount of data on pharmacy students' practical training during their peer and self-evaluation. To validate the analytic indigenous rubric, the study employed role-play scores obtained from raters for MFRM, comments from pharmacy teaching staff and from students' self-evaluation for content analysis. As a matter of fact, such conduct can attest for the relevance of the test content (Dimova et al., 2020).

Pill (2016) investigated indigenous criteria through qualitative research. His participants' comments – based on general scopes of stronger and weaker performance – were placed into categories. This provided great insight into the health professionals' judgment by avoiding guidance to language or communication skills. Other similar health-professional-related studies were conducted with qualitative methods. To validate the authenticity of the OET test, Woodward-Kron and Elder (2016) compared the similarities between the OET and OSCE by exploring the discourse structure and the management of communication tasks based on an analytical approach. Another study by Macqueen et al. (2016) employing thematic analysis, explored the OET's effect in terms of stakeholders' perception and language representation. After interviewing three stakeholder groups and dividing the results into categories representing "boundary objects," the researchers attested the positive washback of the test on both test takers and their society.

As shown in previous studies (Macqueen et al., 2016; Pill, 2016; Woodward-Kron & Elder, 2016), the interviews were mostly organized with categorical analysis. The perceptions of the raters and the informants were deemed necessary in validating the rubric since they were engaged in the practice and recognized the aspects that mattered to their real-world judgments; this could also offer some positive washback. Unfortunately, studies of local tests designed for local contexts,

especially those found in a Thai classroom, are evidently scarce. A few studies by Thai scholars (Sinwongsuwat, 2012; Wudthayagorn 2015) have mentioned the importance of performance-based assessment with particular focus on rubric development. To bridge the gap, one consideration is to develop a local test that can facilitate the students from their entry point to the point of achievement as a course learning objective (Dimova et al., 2020). The same can be said for this indigenous rubric – with regards to the dispensing drugs task – and usage of said rubric among language instructors to guide students to the learning outcome of effectively and professionally dispensing drugs to patients.

5 Methods

This study addresses two research questions as follows.

5.1 Research Questions

1. Can language instructors reliably use the revised rubric, developed using indigenous criteria, to rate student's dispensing performance?
2. What are the language instructors' perceptions of the rubric?

5.2 Participants

There were two groups of participants who took part in the study. The first group comprised 145 fifth-year pharmacy undergraduate students, who took English for Pharmaceutical Science Profession in the 2017 academic year. Prior to these courses, they completed two fundamental English courses and one ESP course. They signed consent forms to participate in this research. The second group of participants were six raters, five of which were classroom instructors of the students, while one had previously taught the course. Two raters had an international background (i.e., Great Britain and the Philippines), and the rest were Thai instructors of English. All raters had more than two years of experience in teaching and rating the students. Among them, four of the raters used and reported the problem on the old rubric (as shown in Table 1). In a one-hour training session, all raters were given details of the role-play task so that they could practice rating the samples by using the given rubric. The rubric was also presented and explained to students in class along with the drug labels that they needed to write on. The pharmaceutical task stems from real-life cases created in consultation with pharmacists and teaching staff of the Faculty of Pharmaceutical Science at the university of the present study. Each student was allowed roughly five minutes to role-play dispensing drugs based

on written information of patient profiles that varied in age, gender, and initial symptom, for example:

- 23-year-old female with headache
- 70-year-old male with heartburn
- 10-year-old male with diarrhea

5.3 Indigenous Rubric Development

Bridging the learning outcomes and instructional practice in the English for Pharmaceutical Profession course cannot be successful without the assistance of the content experts like pharmaceutical teaching staff and licensed pharmacists. Likewise, Dimova et al. (2020) recommend involvement of content experts when integrating content and language. In this study, three groups of experts worked collaboratively to create the indigenous rubric: the pharmaceutical teaching staff, licensed pharmacists who have experience working in drug stores, and Thai instructors of English for the English for Pharmaceutical Science Professional course. This collaboration was essential to reflect on the indigenous criteria because their expertise is local and relevant to the students. However, the issue does not involve only which content to be used, but also how integral each part can be and to what extent. As Byrnes (2008) suggested, collaborative work between the language test developers and the experts is the key to balance the score weight between language and professional knowledge. In fact, O'Hagan et al. (2016) agreed that some involvement from the experts in the field can add a greater range of authenticity to the rubric.

The indigenous rubric was developed with reference to a Thai dispensing rubric (See Appendix) used locally at the university of the present study, along with consultation with content experts. The drug dispensing assessment in Thai previously developed and used by the Pharmaceutical teaching staff was given to all students at the beginning of their first internship course. As every pharmacy student needs to pass the national examination to obtain their license, the teaching staff offer courses (in Thai) that help students pass the requirements. Although the Pharmacy Council of Thailand does not supply a standard rubric nor force pharmacy schools and students to use the same rubric, the Pharmacy Council provides core guidance and notice of the standard requirements. In this way, the Faculty of Pharmaceutical Science developed the drug dispensing rubric in Thai appropriate for their own teaching and assessment. It is, thus, essential for the authors to develop an indigenous rubric to be used in this course by referring to the Thai rubric already available for the students. As a result, the role-play test task and the indigenous rubric value the local criteria as required by the Faculty of Pharmaceutical Science and the Pharmacy Council of Thailand.

5.4 Criteria Selection

Criteria selection without the eyes of experts and people in the community can be perplexing. To look at how the criteria was viewed by different stakeholders, Limgomolvilas and Wudthayagorn (2019) investigated the agreement on criteria among students and pharmacists. The questionnaire asked them to respond to two sections, namely, 'Overall Communication Skills' and 'Pharmaceutical Knowledge' on a four-point Likert scale. The results of the questionnaire suggested that four items should be excluded from the rubric: *introducing your name to patients, asking for the patient's name, calling patients by name*, and *working with patients to schedule doses*. These items were not criteria listed in the Thai dispensing assessment but were listed in assessments conducted in other countries. In terms of linguistics, pharmacy experts did not view using correct English language and pronunciation as important as did the students. However, it was decided that language skills should be included, as this was the central aim of the course for the students and therefore a crucial aspect of the underlying construct of the rating scale.

After consulting with experts and language instructors regarding the results from the questionnaire, three main sections were found: pharmaceutical science skills, language use, and strategic competence. A total of 17 criteria were included, consisting of the three foregoing sections: nine criteria for pharmaceutical science skills, four criteria for language use, and four criteria for strategic competence. The number of assessment criteria represented the diagnostic focus of the assessment, which was used in a formative manner as part of the English course the students attended. As shown in Table 2, the criteria were grouped into three sections and the scores differed based on various criteria, depending on their focus and difficulties.

5.5 Weighted Scores

The score weight was divided among the three sections, with the most weight assigned to language use since it is meant to assess the student's ability to applying their language skills. The total possible score a student could achieve was 50 points, divided across three subsections: pharmaceutical science skills (15 points), language knowledge (20 points), and strategic competence (15 points). As stated, pharmaceutical science skills comprised nine criteria, utilizing a dichotomous scoring system: zero or one, and zero or two. The suggested nine pharmaceutical skills criteria – endorsed by informants – were listed according to the dispensing process: *patient awareness, allergy, underlying disease, impression, dispensing, reason for dispensing, instruction, caution*, and *verifying understanding*. Likewise, the score weight of one and two for each criterion was determined by the experts when considering the effort put in completing the role-play performance.

Table 2 Criteria and score range

Criteria	Score range
I: Pharmaceutical science skills (15 points)	
Patient awareness	0–2
Allergy	0–1
Underlying disease	0–1
Impression	0–1
Dispensing	0–2
Reason for dispensing	0–2
Instruction	0–2
Caution	0–2
Verifying understanding	0–2
II: Language use (20 points)	
Grammar	2–3–4–5
Pronunciation	2–3–4–5
Question type	2–3–4–5
Word choice	2–3–4–5
III: Strategic competence (15 points)	
Voice	1–2
Initiate communication	1–2–3–4
Conclude the encounter	2–3–4–5
Non-verbal communication	1–2–3–4

For language use and strategic competence, the score range was discussed and assigned by language instructors who appeared to have the fine-tuned ability to rate language use and strategic competence. At first, the criteria in both sections were assigned a 4-point scale with different score weights for each criterion. However, the 4-point scale of the voice criterion using decimal places could not be analyzed with MFRM as such analysis accepts only whole numbers. Thus, the researchers decided to change the 4-point scale to a dichotomous system by rounding the numbers from 0.5 to 1 and from 1.5 to 2.

5.6 Data Collection and Analysis

As part of a classroom assessment, 145 students took the test with a rater who was not their classroom instructor. The students gathered in a preparation room where all of their communication tools were retained until they had finished the exam. The students were given patient cases consisting of the three pieces of information immediately before they began the assessment (i.e., age, gender, and initial symptoms). Examples of this were “23-year-old female with headache” and “70-year-old male with heartburn.” The students had five minutes to finish the role-play test task. Once finished, they turned in their consent forms, which were signed before they started the role-plays. They also evaluated their own performance before collecting

their belongings and leaving, without discussing details with other students who had not yet taken the exam. The role-play was recorded so that a second instructor could rate the performance. The interrater reliability between the first and the second rater was also calculated. In general, the interrater reliability was higher than 0.80 which was above satisfactory.

In this study, four instructors (out of six) who took the role of raters were available for phone interviews (by the first author) for approximately 15 min each. The set of questions were sent to them one week before the appointment. The interviews were recorded, transcribed, and analyzed, with consideration taken towards certain areas, such as confidence and comfort with the discussion, sections that were easier or harder to rate, and effectiveness of the problems and recommendations. This part of the data was matched to the quantitative results of rating to explain the findings in more detail.

Although the criteria can be grouped into three sections in conformity with experts' judgments, literature review, and previous research done by the authors (Limgomolvilas & Wudthayagorn, 2019, 2022), it was necessary to examine whether performing to receive score from such criterion is relevant in such context. Nonetheless, the study should demonstrate that this rubric was deemed suitable for the judging dispensing performance. As a result, the scores of each criterion were recorded separately and gathered in an excel spreadsheet for MFRM analysis through the FACETS software (Linacre, 2018). Based on this set of students' scores with information of raters and criteria, the specification file was analyzed to investigate three facets. The first facet considered the scores that reflected each candidate's dispensing performance compared to the group. The second facet examined the raters' performance, which was used to calculate rater's severity in scoring the test takers. The third investigated the rubric criteria, specifically determining the difficulty level of each criterion. The scores were analyzed to reflect the three facets in a Wright map and a rater measurement report.

6 Results and Discussion

6.1 RQ1 Can Language Instructors Reliably Use the Indigenous Rubric to Rate Student's Dispensing Performance?

This part is divided into two parts, which includes the Wright map and the rater measurement report. The Wright map in Fig. 1 presents a general picture of the finding while the rater measurement examines each criterion in detail. The content analysis depicts the raters as a rubric user.

In the Wright map of Fig. 1, the first column, Measr, shows the logit scale of all candidates, raters, and criteria. The second column, labeled Candidate, shows the number of the students on the scale using stars and dots. A star represents two

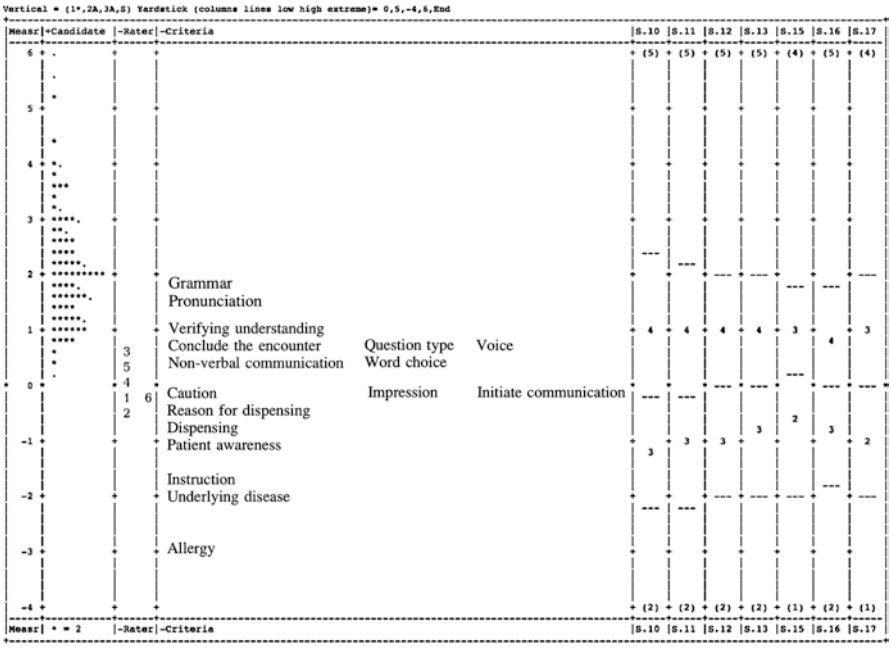


Fig. 1 Wright map – Speaking scale

candidates while a dot represents one candidate. The candidates are placed along the scale according to their score; the higher they are, the higher the score they received. None of the candidates are below zero logits, which is the risk of receiving less than 50% score from the average severe raters on the criterion of average difficulty (McNamara et al., 2019). The third column, Rater, indicates the position of their judgment; the higher the scale, the harsher the rater is. The fourth column, labeled Criteria, shows the position of how difficult each criterion was rated by the raters, with the highest as the most difficult and the lowest as the least difficult. The rest of the columns show the scale category of each criterion from one level to another. In addition, content analysis was employed to analyze qualitative data resulting from the interview. The data were transcribed and analyzed according to categories and sub-categories.

When taking a closer look at the raters, it can be seen in Fig. 1 that rater number 3 is the harshest rater, while rater number 2 is the most lenient. However, all six raters are homogenous considering that the raters are within ±1 logits, which is a fair level for performance assessment according to Fairbairn and Dunlea (2017). Among the criteria in the fourth column, *Grammar* is placed highest followed by *Pronunciation*, which indicates that these two are the most difficult criterion for the candidate to achieve. At the same time, *Allergy* is the lowest, meaning that it is the easiest criterion for students. For *Initiate communication*, the Wright map shows

that its digits of level 3 are the smallest scale in comparison to the other criteria. This means that for this criterion, it is not difficult for the candidates to achieve level 4 when the scale of level 3 is quite narrow. In other words, not much difference was identified between students in level 3 and 4.

The location of most criteria in the Wright map was shown to lie around the middle, with some floating at the bottom. The fact that the criteria positions were not too far apart indicates that students' performance on most of the criteria are comparable. Among them, asking about a patient's allergy was the easiest criterion.

Table 3 describes the judgment of the raters, presenting each individual and the related positions of their judgment to other raters. The raters were arranged according to their severity logits (see also Fig. 1). The severity logits, assigning the position of the rater on the logit scale, showed that the logit scale difference between the most severe rater, Rater 3, at .56 and the most lenient rater, Rater 4, at −.31 is .87, which is considered low. In addition, the standard error is rather low, with the highest on Rater 6 at .11, meaning that the raters' severity is in the same acceptable range. Another statistic to consider is the infit range. As recommended by Bond and Fox (2007), for rater consistency, a range of infit should be between 0.7 and 1.3. The result indicates that the infit score of Rater 4 is higher than the whole group at 1.38. The normalized distribution, which is shown as Zstd, is less compatible to the expected model response at over 2.

Based on the Wright map shown in Fig. 1 and the rater measurement report (extract) displayed in Table 3, the raters were reliable when using the rubric to assess the students' performance, except for one misfit rater, Rater 4. However, they were more likely to assign higher scores (i.e., 4 and 5) to examinees. This behavior can be seen from Fig. 1, as the highest score bands (i.e., 4 and 5) which align with the range of abilities for most examinees, are wider than the other score bands (i.e., 3 and 2). As shown in the Wright map, examinees rarely received a score of 1 on all criteria. Despite the slight misalignment between examinee ability and criterion difficulty, raters seemed capable of applying the rating scale, and the psychometric quality of the scale is overall satisfactory.

Table 3 Rater measurement report (extract)

Raters	Measure	S.E.	Infit MnSq	ZStd	Estim. Discrm	Correlation		Upper CI	Lower CI
						PtMea	PtExp		
3	.56	.06	.76	−4.8	1.26	.58	.57	0.68	0.44
5	.27	.06	.86	−2.3	1.05	.52	.52	0.39	0.15
1	−.06	.07	1.01	.1	.96	.45	.48	0.08	−0.45
2	−.16	.07	1.09	1.3	1.02	.47	.47	0.06	−0.44
6	−.30	.11	1.19	1.6	.86	.50	.52	−0.16	−0.2
4	−.31	.07	1.38	5.1	.77	.54	.52	−0.17	−0.38

6.2 RQ2 What Are the Language Teachers' Perceptions of the Rubric?

Language teachers' perceptions towards the rubric can be analyzed into the following themes.

Comfort and Confidence in Using the Rubric Based on interviews, most language teachers who rated the students felt comfortable using the rubric, although one rater did not. This stems from the fact that it was the rater's first time teaching the course, performing the task, and using the rubric.

> Yes, you'll feel pretty comfortable after you've done a couple as well. It's quite straightforward making sure that you cover everything on the rubric. I didn't think there were any major problems. (Rater 1)

In terms of confidence, all raters agreed that they were confident rating all sections, especially for the dichotomous system of the pharmaceutical science skills.

> Yes, for the most part, because the criteria were quite clearly labeled. So even if I didn't have the subject knowledge, I was still able to assess them based on the criteria. I would say 8 out of 10. (Rater 2)

Ease of Use For the level of ease and difficulty, they believed that pharmaceutical science knowledge was the easiest part because they only needed to check yes or no, as shown in the statement below.

> Pharmaceutical science skills were the easiest to rate as it required me to choose only yes or no. Also, the students have been through similar tests in Thai so it's not hard for them to follow the criterion. (Rater 3)

However, while one rater believed that the rubric can be easy to use, some of the criteria seemed subjective to reflect pharmaceutical knowledge accurately.

> Word choice can appear easy, but I sometimes wondered whether some words were included in layman terms. Students might assume that the instructor understands everything and not switch medical terms to layman terms. In addition, when I talked to students after the test, most of them tended to think that their word choice was easy enough to comprehend. One student asked whether the word "hypertension" is easy to understand. Although I used this word regularly, the word "high blood pressure" is easier to understand. I'm used to rating the students in this section because I'm a language instructor. (Rater 4)

6.3 Problems and Suggestions

One question from the language instructors was this, *"if the students were not capable of talking about their subject matter, should they be scored, since they have performed the criterion?"* Generally speaking, the raters were hesitant to give full scores to students because some students could not perform as well as the suggested answers Some criteria were viewed as redundant or needed more description to

differentiate the levels, such as that of *silence* and *posture*. Lastly, raters complained about the time constraints placed on the test administration, since some raters experienced stress and weariness from role-playing with a class of about 30 students in a three-hour class time.

> The time constraint is an issue for me. I needed some time to rate students, but the next student was pushed in before I could finish. I feel like I needed more time in between the students. This issue didn't happen when I was a second rater. It was easier for me since I didn't need to worry about necessary details I needed to cover while acting as a patient. (Rater 3)

The MFRM results were collinear to the positive impressions reported in rater interviews. Despite the difficulties in some criteria, the raters believed that this rubric was easy to use especially when assessing pharmaceutical science skills as binary decisions. Similar to the binary decisions found on a performance tree decision (Fulcher et al., 2011), this study has shown that the binary system is beneficial for language instructors in terms of making better judgments on students' applied professional knowledge. In contrast, this method is used by Thai pharmacy teaching staff when assessing Thai pharmacy student communication skills, because such skills are not their specialty.

Unlike in previous studies (Elder et al., 2012; O'Hagan et al., 2016; Pill & McNamara, 2016) which focused on establishing criteria that could be a threshold of language ability and content knowledge, the raters in this study viewed this rubric as efficient in assessing a test taker's performance. They felt confident that the rubric could guide them through justifying the student's pharmaceutical science skills even if these skills were not in their area of expertise. For ease of use, these skills were categorized separately as 'language use' and 'strategic competence' and were provided to raters on a binary scale, as either yes or no options. In fact, the raters commented on how quickly their decision could be made on scoring pharmaceutical science skills criteria when the content matched the guided answers. Sadler and Donnelly (2006) stated that a fundamental threshold of content knowledge is necessary for the test taker to perform the task and the quality of a performance can be observed when the knowledge threshold is acquired. Thus, an indigenous rubric that clearly identifies pharmaceutical science skills, such as the one developed, can facilitate the rating process, while inherently having score justifications for the test taker's performance.

7 Insights

Based on this study, five major insights are discussed below.

First, the authors agree with Dimova et al. (2020) that local language tests can clearly benefit a classroom assessment system even though their natures are not typical to a classroom assessment method. The role-play test task and the indigenous rubric proved helpful for all parties, especially for language instructors who are not experts at the content, but are knowledgeable of language use and strategic

competence. The students can also utilize the indigenous rubric as a guideline to improve their dispensing performance while conforming to the Thai pharmacy professional standard. In addition, the learning outcomes meet the requirements of the study program. Thus, developing an ESP classroom assessment as well as creating an indigenous rubric for a local context is possible, when language instructors work collaboratively with content teachers, who contribute their expert knowledge on what is essential to be included.

Second, language instructors are able to reliably rate student performance. This observation holds true when they rate student performance against pharmaceutical science skills. However, the observation of a slight misalignment between the examinee's abilities and the criterion suggests a need to continue refining the rating scale. To improve the rubric, explanations for some criteria should be added, especially towards criteria that is difficult to understand at a higher level (Bijani & Mona, 2017). Following this recommendation, our indigenous rubric can be further revised under the Grammar and Pronunciation criteria. As French (2003) suggested, choosing suitable descriptive words for each criterion and example performances can resolve obscure interpretations of the rubric. French (2003), however, noted that descriptions should not be too lengthy to allow more effective rating when assessing students in real time. This parallels Purpura et al. (2015)'s notion that relying on scores only cannot lead to rating consistency. Some explanations to further clarify understanding for raters and test takers should be supplied.

Third, training raters is a necessity, and the use of certain tools, such as interrater agreement (Hauenstein & McCusker, 2017), is recommended. This is to ensure that the scores are reliable across all raters. The results of this study showed that one of the raters – who had been teaching for more than thirty years – was the harshest rater. Such rater matched what Eckes (2009) found in their study of senior raters – that raters who have decades of in-service teaching may rate harsher than their younger peers, as they may want to "set the standard" for others. Similarly, the more experience the raters had, the higher their standards were (Bonk & Ockey, 2003; Hauenstein & McCusker, 2017). In addition, experienced raters tended to have more bias than inexperienced raters, even after rater training (Bijani & Mona, 2017). However, the harsh rater in this study conformed to the group agreement and is not considered a threat towards the assessment's results overall. Following such results, training is needed to ensure that all raters conform, especially for those who are senior or experienced raters. Training and norming sessions can help a rater obtain self-consistency (Avineri et al., 2010; Elder et al., 2005). Likewise, Johnson and Riazi (2017) recommended that rater training and norming sessions should be held at a different time in order to aid rater consistency. To sum up, the rater in this study, who was misfit, could receive more training sessions to better align their judgment within the group, while the harsh rater could be matched with a more lenient rater in order to balance their scores.

Fourth, sample performances of all score levels should be supplied to raters during training sessions. Faez et al. (2011) insisted that sample performance for each level is necessary for facilitating the rater's understanding of the scales and levels. In addition, Stahl and Lunz (1992) and Lumley and McNamara (1993) endorsed

providing feedback to raters regarding how they each assessed in order to decrease the rating inconsistency. Similarly, utilizing results from the MFRM to improve each rater's performance could lessen the raters' judgment effect (Bonk & Ockey, 2003). Based on previous studies (Bijani & Mona, 2017; Bonk & Ockey, 2003; Eckes, 2009), good training sessions should offer raters details that describe their rating performance. This can be obtained from MFRM, as Schaefer (2008) stated that MFRM could effectively help enhance a rater's fairness and accuracy in performance-based assessment. Discussion sessions about the results between raters may also facilitate understanding.

Finally, the practicality of assessing 30 student role-plays, and recording them for a second rater within three hours should be addressed. More staff might be needed, both for assessing the students and for administering the test task. More raters can shorten the time frame for the whole test administration, and also decrease the weariness of the raters.

So far, the indigenous rubric used in the role-play test task created in this course has proved to be useful and effective for both language instructors and students. The development "is designed to represent the value and the priorities" (Dimova et al., 2020, p. 1) in a Thai pharmaceutical context. Based on the rubric, three components – pharmaceutical science skills, language use, strategic competence – are emphasized. With a specific and clear list of criteria, the raters feel at ease when rating students while being able to produce reliable scores. The students know exactly what the language instructors and the content experts are expecting. The students can also use this rubric as a guideline when practicing which, in turn, helps them gain greater confidence when being assessed. Their improvement may not only be based on teachers' feedback, but also on their own self-reflection, guided by an indigenous rubric such as the one developed.

Acknowledgements This study was supported by the Thailand Research Fund through the Royal Golden Jubilee (RGJ) Ph.D. program (Grant No. PHD/0208/2558) under the Royal Thai Government.

Appendix

Questions for Raters

Overall Rubric Usage

- Did you feel comfortable using the rubric?
  - Which section was the easiest to rate?
  - Which section was the hardest to rate?
  - Does the rubric specify adequate information for the rating? If not, what should be added in?

– Can this rubric be used to assess and differentiate the student's performance in dispensing drugs in English?
– Are there any criteria you think should be added in the rubric?
– Are there any criteria you think it's unnecessary?

Topic		Yes	No
Patient awareness	*Investigate and respond to patient's concerns and needs* (*Elicit patient questions, concerns, reasons for visit, current health condition and drugs currently being taken*)	2	0
Allergy	*Ask patient about their allergies*	1	0
Underlying disease	*Ask patient about their underlying diseases*	1	0
Impression	*State the possible disease*	1	0
Dispensing	*Provide the name of the drugs(s)*	2	0
Reason(s) for dispensing	*Provide reason(s) for dispensing the drugs*	2	0
Instruction	*Provide instruction of the drugs(s) dispensed*	2	0
Caution	*Provide detail on caution*	2	0
Verifying understanding	*Verify patient understanding of the drugs usage*	2	0

Pharmaceutical Science Knowledge

• Were you confident rating the students in this section?
– Do you think the scale for each criterion is appropriate?
– Were there any criteria you found hard to rate for students?
– Do you think the score level is appropriate? If not, how would you like it to be?
– For each criterion, please specify your confidence rating it as a non-pharmacist.
– Any recommendations?

Language Use

• Were you confident rating the students in this section?
– Do you think the scale for each criterion is appropriate?
– Were there any criteria you found hard to rate for students?
– Any recommendations?

Strategic Competence

• Were you confident rating the students in this section?
– Do you think the scale for each criterion is appropriate?
– Were there any criteria you found hard to rate for students?
– Any recommendations?

Rubric

I: Pharmaceutical science skills (30% – 15 points) Part I score: _____

Topic		Yes	No
Patient awareness	*Investigate and respond to patient's concerns and needs* (*Elicit patient questions, concerns, reasons for visit, current health condition and drugs currently being taken*)	2	0
Allergy	*Ask patient about their allergies*	1	0
Underlying disease	*Ask patient about their underlying diseases*	1	0
Impression	*State the possible disease*	1	0
Dispensing	*Provide the name of the drugs(s)*	2	0
Reason(s) for dispensing	*Provide reason(s) for dispensing the drugs*	2	0
Instruction	*Provide instruction of the drugs(s) dispensed*	2	0
Caution	*Provide detail on caution*	2	0
Verifying understanding	*Verify patient understanding of the drugs usage*	2	0

II: Language Use (40% – 20 points) Part II score: _____

Topic	Need improvement	Fair	Good	Excellent
Grammar (*intelligible*)	2	3	4	5
Pronunciation (*intelligible*)	2	3	4	5
Question type (*use appropriate open-ended and close-ended questions and not leading the patients*)	2	3	4	5
Word choice (*layman terms: easy-to-understand word choice for patient and clarify medical words if needed*)	2	3	4	5

III: Strategic Competence (30% – 15 points) Part III score: _____

Topic	Need improvement	Fair	Good	Excellent
Voice: *tone, volume, pace, silence*	1		2	
Initiate communication: *greet patient warmly, identify patient's identity*	1	2	3	4
Conclude the encounter: *summarize information, ask if any questions arise, thank patient*	2	3	4	5
Non-verbal communication: *eye contact, gesture, posture, professional manner*	1	2	3	4

Total score: _____

Comments: _____
_____
_____
_____
_____

References

Avineri, N., Londe, Z., Hardacre, B., Carris, L., So, Y., & Majidpour, M. (2010). Language assessment as a system: Best practices, stakeholders, models, and testimonials. *Issues in Applied Linguistics, 18*(2), 251–265. https://doi.org/10.5070/L4182005334

Bijani, H., & Mona, K. (2017). Investigating the effect of training on raters' bias toward test takers in oral proficiency assessment: A FACETS analysis. *Journal of Asia TEFL, 14*(4), 587–836. https://doi.org/10.18823/asiatefl.2017.14.4.7.687

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences*. Lawrence Erlbaum Associates Publishers.

Bonk, W. J., & Ockey, G. J. (2003). A many-facet Rasch analysis of the second language group oral discussion task. *Language Testing, 20*(1), 89–110. https://doi.org/10.1191/0265532203lt245oa

Byrnes, H. (2008). Assessing content and language. In *Encyclopedia of language and education* (pp. 2182–2197). Springer. https://doi.org/10.1007/978-0-387-30424-3

Dimova, S., Yan, X., & Ginther, A. (2020). *Local language testing*. Routledge.

Douglas, D. (2001). Language for Specific Purposes assessment criteria: where do they come from? *Language Testing, 18*(2), 171–185. https://doi.org/10.1177/026553220101800204

Eckes, T. (2009). Many-facet Rasch measurement. In S. Takala (Ed.), *Reference supplement to the manual for relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment (Section H)*. Council of Europe/Language Policy Division. https://rm.coe.int/CoERMPublicCommonSearchServices/DisplayDCTMContent?documentId=0900001680667a23

Elder, C., & McNamara, T. (2016). The hunt for "indigenous criteria" in assessing communication in the physiotherapy workplace. *Language Testing, 33*(2), 153–174. https://doi.org/10.1177/0265532215607398

Elder, C., Knoch, U., Barkhuizen, G., & Randow, J. (2005). Individual feedback to enhance rater training: Does it work? *Language Assessment Quarterly, 2*(3), 175–196. https://doi.org/10.1207/s15434311laq0203_1

Elder, C., Pill, J., Woodward-Kron, R., McNamara, T., Manias, E., Webb, G., & Mccoll, G. (2012). Health professionals' views of communication: Implications for assessing performance on a health-specific English language test. *TESOL Quarterly, 46*(2), 409–419.

Faez, F., Majhanovich, S., Taylor, S. K., Smith, M., & Crowley, K. (2011). The power of "can do" statements: Teachers' perceptions of CEFR-informed instruction in French as a second language classrooms in Ontario. *Canadian Journal of Applied Linguistics, 14*(2), 1–19. https://journals.lib.unb.ca/index.php/CJAL/article/view/19855/21653

Fairbairn, J., & Dunlea, J. (2017). *Aptis speaking and writing rating scales revision*. https://www.britishcouncil.org/sites/default/files/aptis_scale_revision_layout.pdf

French, A. (2003). The development of a set of assessment criteria for speaking tests. *Research Notes, 13*, 8–16. https://www.cambridgeenglish.org/Images/23128-research-notes-13.pdf

Fulcher, G., Davidson, F., & Kemp, J. (2011). Effective rating scale development for speaking tests: Performance decision trees. *Language Testing, 28*(1), 5–29. https://doi.org/10.1177/0265532209359514

Hauenstein, N. M., & McCusker, M. E. (2017). Rater training: Understanding effects of training content, practice ratings, and feedback. *International Journal of Selection and Assessment, 25*(3), 253–266.

Hyvärinen, M.-L., Tanskanen, P., Katajavuori, N., & Isotalus, P. (2012). Evaluating the use of criteria for assessing profession-specific communication skills in pharmacy. *Studies in Higher Education, 37*(3), 291–308. https://doi.org/10.1080/03075079.2010.510183

Jacoby, S. (1998). *Science as performance: Socializing scientific discourse through the conference talk rehearsal. (Doctor of Philosophy)*. University of California.

Jacoby, S., & McNamara, T. (1999). Locating competence. *English for Specific Purposes, 18*(3), 213–241. https://doi.org/10.1016/S0889-4906(97)00053-7

Johnson, R. C., & Riazi, A. M. (2017). Validation of a locally created and rated writing test used for placement in a higher education EFL program. *Assessing Writing, 32*, 85–104. https://doi.org/10.1016/j.asw.2016.09.002

Limgomolvilas, S., & Wudthayagorn, J. (2019). Needs analysis of dispensing assessment for Thai pharmacy students. *NIDA Journal of Language and Communication, 24*(36), 46–63. https://lcjournal.nida.ac.th/main/public/jn_pdf/journal_v24_i36.pdf

Limgomolvilas, S., & Wudthayagorn, J. (2022). Designing a drug-dispensing test task using the SPEAKING grid. *REFLection, 29*(1), 232–246. https://so05.tci-thaijo.org/index.php/reflections/article/view/258954/174239

Linacre, J. M. (2018). *Facets computer program for many-facet Rasch measurement (Version 3.81.0)*. Beaverton, Oregon.

Lumley, T., & McNamara, T. F. (1993). *Rater characteristics and rater bias: Implications for training*. https://files.eric.ed.gov/fulltext/ED365091.pdf

Macqueen, S., Pill, J., & Knoch, U. (2016). Language test as boundary object: Perspectives from test users in the healthcare domain. *Language Testing, 33*(2), 271–288. https://doi.org/10.1177/0265532215607401

McNamara, T., Knoch, U., & Fan, J. (2019). *Fairness, justice and language assessment*. Oxford University Press.

Nawanopparatsakul, S., Keokitichai S., Wiyakarn S., & Chantaraskul C. (2009–2010). Challenges of pharmacy education in Thailand. *Silpakorn University International Journal*, *9–10*, 19–39. https://www.thaiscience.info/journals/Article/SUIJ/10813225.pdf

O'Hagan, S., Pill, J., & Zhang, Y. (2016). Extending the scope of speaking assessment criteria in a specific-purpose language test: Operationalizing a health professional perspective. *Language Testing, 33*(2), 195–216. https://doi.org/10.1177/0265532215607920

Pill, J. (2013). *What doctors value in consultations and the implications for specific-purpose language testing*. Doctor dissertation. The University of Melbourne.

Pill, J. (2016). Drawing on indigenous criteria for more authentic assessment in a specific-purpose language test: Health professionals interacting with patients. *Language Testing, 33*(2), 175–193. https://doi.org/10.1177/0265532215607400

Pill, J., & McNamara, T. (2016). How much is enough? Involving occupational experts in setting standards on a specific-purpose language test for health professionals. *Language Testing, 33*(2), 217–234.

Purpura, J. E., Brown, J. D., & Schoonen, R. (2015). Improving the validity of quantitative measures in applied linguistics research. *Language Learning, 65*(S1), 37–75. https://doi.org/10.1111/lang.12112

Sadler, T. D., & Donnelly, L. A. (2006). Socioscientific argumentation: The effects of content knowledge and morality. *International Journal of Science Education, 28*(12), 1463–1488. https://doi.org/10.1080/09500690600708717

Schaefer, E. (2008). Rater bias patterns in an EFL writing assessment. *Language Testing, 25*(4), 465–493. https://doi.org/10.1177/0265532208094273

Sinwongsuwat, K. (2012). Rethinking assessment of Thai EFL learners' speaking skills. *Language Testing in Asia, 2*(4), 1–11. https://doi.org/10.1186/2229-0443-2-4-75

Stahl, J., & Lunz, M. (1992). *A comparison of generalizability theory and multi-faceted Rasch measurement.* Paper presented at the Midwest Objective Measurement Seminar, University of Chicago.

Woodward-Kron, R., & Elder, C. (2016). A comparative discourse study of simulated clinical roleplays in two assessment contexts: Validating a specific-purpose language test. *Language Testing, 33*(2), 251–270. https://doi.org/10.1177/0265532215607399

Wudthayagorn, J. (2015). Implementing outcome-based assessment: Lessons learned from an English for Pharmacy course. In P. Darasawang & H. Reinders (Eds.), *Innovation in language learning and teaching: The case of Thailand* (pp. 126–140). Palgrave Macmillan.

Using Spoken Dialog Systems to Assess L2 Learners' Oral Skills in a Local Language Testing Context

Yasin Karatay

Abstract The assessment of oral proficiency in instructional contexts presents interesting opportunities when designed to foster communicative performance in a specific target language use domain. A task-based Tourism English oral performance assessment was designed using a specialized spoken dialogue system (SDS) in which the computer is programmed to act as a hotel guest and examinees respond as a hotel employee. Based on a mixed-method research design, this study examined whether task administration conditions and the rubric for scoring performance assessment are appropriate for providing evidence of Tourism English ability. The analysis of 30 L2 students' oral performances, their post-test questionnaire responses, and semi-structured individual interviews indicates that they considered the tasks engaging and relevant to their future profession. Similarly, four raters' verbal reports from semi-structured individual interviews suggest that the test tasks effectively elicited ratable speech samples that can be argued to represent students' oral communication skills in a hospitality setting.

Keywords Local language testing · Spoken dialog systems · Tourism English · English for specific purposes

1 Introduction: Test Purpose and Testing Context

According to the Council of Higher Education in Turkey, there are currently approximately 100,000 hospitality students in over 250 Tourism and Hotel Management (THM) programs in Turkey (https://istatistik.yok.gov.tr/). As more than 40 million tourists visit Turkey every year (https://www.ktb.gov.tr/EN-249283/tourism-statistics.html), there is an increasing need for employees who can communicate with the growing tourist population. However, there is a discrepancy

Y. Karatay (✉)
Department of English, Iowa State University, Ames, IA, USA
e-mail: ykaratay@iastate.edu

X. Yan et al. (eds.), *Local Language Testing*, Educational Linguistics 61,
https://doi.org/10.1007/978-3-031-33541-9_12

between what is expected in the workplace and classroom practices in THM programs (Abdel-Ghany & Abdellatif, 2012; Zahedpisheh et al., 2017).

In this regard, a computer-based tourism English oral communication (CTEOC) task was designed to assess tourism English students' oral communication skills in a THM program at a Turkish state university. Currently, all THM students in the local university are required to take one tourism English course each semester, a total of eight courses before graduation. The primary purpose of these courses is to help students gain the knowledge and skills necessary to start a career in tourism. Students' performances are evaluated based on summative assessments (i.e., two midterms and a final exam) designed and created by the testing unit in the School of Foreign Languages at the local university. However, traditionally a discrete-point assessment approach has been taken to assess THM students' tourism English performances. In other words, similar to other English for Specific Purposes (ESP) assessment contexts (Aysu & Ozcan, 2021; Zahedpisheh et al., 2017), the tests used in these tourism English courses are multiple-choice items or target only receptive skills, grammar, or vocabulary, which imposes a crucial limitation in the local language testing context: underrepresentation of tourism English oral communication. Therefore, CTEOC was designed to assess students' oral communication skills through dialogic conversations in a virtual environment that simulates the target language use (TLU) domain through a set of scenarios described in the methods section.

2 Testing Problem Encountered

In order to reveal the language skill needs of students studying in the local THM program, a rigorous domain analysis based on Evidence-centered Design (ECD) model (Mislevy et al., 2003) was carried out in the local institution. A task-based needs analysis survey for nine tourism English course instructors, semi-structured interviews with four course instructors and three THM faculty members, and document analysis on Tourism English III and IV syllabi and course books that were used in the THM program were used in the domain analysis. A detailed description of the data collection, data analysis, and findings for the investigation of domain experts' judgments and materials relevant to the target tourism English oral communication skills and potential test tasks can be found in Karatay (2022). To summarize the findings, discrepancies were found between the actual language teaching practices in the tourism English classes and (1) THM faculty members' and course instructors' expectations and (2) the course objectives and textbook activities. For example, despite the variety of oral communication skills in coursebooks and the focus on these skills in the course objectives, the instructors spent most of their time on vocabulary and grammar skills, sparing little or no time for speaking activities, primarily because of crowded classrooms. The findings also suggested that the lack of a summative assessment to assess students' oral communication skills pushed the course instructors to sacrifice classroom practices that target these skills. However,

they also reported that, in an ideal world, oral communication skills such as dealing with guest problems, asking and answering questions in a pragmatically appropriate manner, and using a comprehensible pronunciation while doing so were among the skills that graduates with a tourism major would need to acquire. Similar expectations regarding the skills, knowledge, abilities, and processes required for a successful tourism English oral communication were also apparent in the course objectives and THM faculty members' interview data.

Overall, the findings suggested that THM students are expected to perform oral communication tasks competently, giving almost always appropriate responses at an appropriate speech rate (without unnatural language-related pauses) and effective pronunciation with only minimal errors in production, as well as a high degree of grammatical accuracy in both simple and complex structures and domain-specific vocabulary use. However, these sub-skills are not targeted in the current assessment practices because of several logistical issues such as crowded classrooms and the resource-intensive nature of administering an oral communication assessment in the local context.

3 Literature Review

The literature review addresses language characteristics of the hospitality domain, Role-play as a Test Task in ESP Assessment, and SDS Tasks for Oral Proficiency Assessment.

3.1 Oral Communication of ESP Learners in the Hospitality Context

Hospitality in this study refers to the array of activities to satisfy guests, which depends substantially on staff's ability to accommodate ever-evolving customer needs. Students majoring in hospitality programs are expected to be employed in different jobs in hospitality settings. They can work in services such as reception counter staffing, hotel management, and hotel restaurant management. Depending on the job setting, they are expected to use English in situations such as checking guests into a hotel, giving information about hotel facilities, giving directions, and requesting tourist information (Zahedpisheh et al., 2017). In several needs analysis studies regarding the language needs of tourism students, researchers have reported that employees are expected to communicate fluently with linguistically diverse tourists while performing any of these tasks (Abdel-Ghany & Abdellatif, 2012; Leslie & Russell, 2006; Zahedpisheh et al., 2017). Thus, of critical importance to staff is their ability to manage effective communication with customers, which often requires competency in a second language. Undoubtedly, as the *lingua franca* of

tourists worldwide, English is the most commonly used language of hospitality (Blue & Harun, 2003). As Leslie et al. (2004) underline, effective communication involves "to be able to hold a conversation and listen and comprehend customer needs in English." Therefore, students taking any tourism English classes need to be highly motivated to be accurately fluent in a high level of professional service language (Zahedpisheh et al., 2017).

One seminal work in defining the construct of tourism English is that of Blue and Harun (2003). Exploring the sociolinguistic aspect of hospitality in receptionist-guest exchanges at a hotel front desk, Blue and Harun (2003) identified "hospitality language" as "loosely constituted but identifiable linguistic specialism," emerging from "a combination of procedural, behavioral and linguistic acts, verbal and non-verbal, direct and indirect" (p.89). They suggest that with underlying distinctive linguistic patterns that characterize it, hospitality language has its own register. Front desk interactions usually generate predictable and dyadic utterances with mostly adjacency pairs such as greeting– greeting, question–answer, apology–acceptance, and request–assent.

In this study, the focus of tourism English oral communication is restricted to the interaction between front desk hotel employees and customers. Based on the definition by Blue and Harun (2003), tourism English oral communication in a hotel setting, specifically the front desk talk, is information and service-driven, dealing predominantly with various services. It is often formal, sequential, and balanced, with few interruptions, mainly in the form of either one-word utterances or complete sentences. Except for rare cases, fillers are unusual, which demonstrates the promptness and efficiency of front desk staff. Overall, the structure of hospitality language is relatively straightforward, with a type of brief, balanced communication in terms of turn-taking and similar keywords repeated to carry out different actions with different people.

3.2 Role-Play as a Test Task in ESP Assessment

Role-play has been reported as a test task that assesses a test taker's ability to converse in naturally-occurring communication more effectively than oral interviews in which the examiner and test taker maintain their normal roles (Van Batenburg et al., 2019). Thanks to its affordances such as generating interactional behavior (i.e., turn-taking, requesting, complaining, suggesting) and affording some measure of control over that interaction, role-play has been predominantly considered as an appropriate test format in operationalizing the construct of interactional competence (Kasper & Youn, 2018; Youn, 2015). While role-play in language assessment is widely understood, it should not be unquestionably considered an authentic task if it is not well designed without a clear communicative goal (i.e., small talk or scripted role-play tasks). One source of inauthenticity in role-play assessment tasks is the behavior of the interlocutor. For example, interlocutors in a role-play task can use unnatural silences and defer suggestions to elicit particular responses from test

takers (Okada & Greer, 2013), help test takers by solving complications and facilitating certain interactional moves, and dominate the interaction (Manias & McNamara, 2016). To prevent this type of variability and ensure standardization in role-play assessments, the interlocutor's part can be scripted and controlled (Van Batenburg et al., 2019). However, the nature of its resource-intensive process requiring one-on-one interaction with an interlocutor or a human rater available in the room makes role-play tasks challenging (Roever, 2011). One way to solve the issues with standardization caused by undesirable variability of interlocutors and practicality is using a spoken dialogue system (SDS) in role-play assessments. The following section will explore these systems in more detail.

3.3 SDS Tasks for Oral Proficiency Assessment

The pervasiveness of speech technologies in every part of our daily lives will inevitably empower the field of the second language (L2) oral proficiency assessment (Litman et al., 2018). One such technology is an SDS, which can potentially enable test developers to assess some aspects of a test taker's interactional competence (Ockey & Chukharev-Hudilainen, 2021). Language learners have found the tasks developed in an SDS engaging and motivating (Evanini et al., 2018; Timpe-Laughlin et al., 2020). Also, as opposed to a human interlocutor, interacting with an SDS in interactive speaking tasks may reduce users' anxiety and increase their willingness to engage in oral interaction (Timpe-Laughlin et al., 2020). While the research on SDS-based tasks yielded promising results with regards to improvement in L2 learners' speaking proficiency (Timpe-Laughlin et al., 2020), they have also been used to assess the interactive speech of L2 learners (Chukharev-Hudilainen & Ockey, 2021; Gokturk, 2020; Timpe-Laughlin et al., 2017). Based on the design of the SDS, intended use, and target construct, these studies yielded both promising results and suggestions to improve the systems. One recent example is Chukharev-Hudilainen and Ockey' study (2021), which investigated a specialized SDS, Interaction Competence Elicitor (ICE), to deliver a paired oral discussion task and elicit discourse to assess the Test taker's oral proficiency, including interactional competence. They found that ICE could generate task-appropriate, ratable speech samples to assess interactional competence in 90% of cases. Using the same SDS (i.e., ICE), Ockey and Chukharev-Hudilainen (2021) compared the effectiveness of the SDS partner and human partner in eliciting discourse in a paired oral task. They found that the raters assigned similar scores for fluency, pronunciation, and grammar and vocabulary in human-human and computer-human interactions, whereas score assignment for interactional competence was favored in the second condition. Even though the raters in their study thought human-human interaction to be more authentic, the authors suggest that using an off-the-shelf, cloud-based ASR is cost-effective and robust in eliciting speech samples regarding L2 speakers' oral proficiency skills including interactional competence.

Although the problem with authenticity and naturalness of the interaction between an SDS and human was also reported in other studies (i.e., Litman et al., 2018; Timpe-Laughlin et al., 2017), test developers can integrate this technology into their assessments, especially designed for eliciting speech samples from constrained interactions such as reading aloud and elicited imitation (Litman et al., 2018). While the status of SDS technology today may not be useful for fully unconstrained tasks yet because of the unlimited number of paths that test takers can take to respond to such tasks, they can be useful in scenario-guided, constrained conversations in specific domains (Litman et al., 2018).

Tourism English, as described above, is one such specific domain where scenario-guided constrained conversations are exactly what students need to be able to participate in. However, as found in the domain analysis conducted in the local institution, there is a gap between classroom practices in ESP courses in these programs and employers' expectations. Despite the attractiveness of such technologies, there have been limited applications of SDSs for ESP assessments so far, and traditional delivery methods are still very popular in local language assessment practices because of limited local resources (Dimova et al., 2020), especially in the language tests in the target institution in this study. Therefore, a practical, efficient oral proficiency assessment is needed relevant to tourism students' English learning goals. To meet this need, a task-based tourism English oral performance assessment was designed based on the ECD model (Mislevy et al., 2003) using a specialized SDS in which the computer acts as a customer and examinees act as a hotel employee. As suggested by Dimova et al. (2020), "if you want to collect evidence about the relevance of the test content, you may take the opportunity to obtain subject specialists', instructors', and students' evaluation of the test content" (p. 28). Within this context, the present study set out to explore whether task administration conditions of the prototype SDS for tourism English are appropriate for providing evidence of targeted tourism English ability and whether the rubric for scoring test takers' responses is appropriate for providing evidence of targeted language abilities. To achieve these purposes, the following research questions were addressed:

1. What are test takers' (i.e., students) perceptions of the task administration conditions?
2. What are raters' (i.e., course instructors') perceptions of the rating scale for scoring test takers' responses in an SDS-based role-play assessment?

4 Methods

A mixed-method research design was adopted in this study (Creswell & Plano Clark, 2012). Using quantitative and qualitative data analysis methods, the researcher analyzed the two data sets separately and independently. The data were analyzed qualitatively and quantitatively to make an interpretation by identifying how the two sets of results were related to, converged, or diverged from each other under the interpretive argument framework.

4.1 Participants

A total of 30 English as a Foreign Language (EFL) students majoring in the THM program at a large Turkish university participated in the study. All the participants were junior students in the program taking Tourism English III with a mixed level of English language proficiency. The majority of the participants considered their English proficiency level as pre-intermediate (33%), intermediate (40%), or upper-intermediate (17%), while only 7% of them perceived themselves as elementary and 3% as advanced in English language proficiency. The participants in this study were selected on a voluntary basis and shared the characteristics of the intended test takers for the test.

Four raters with extensive ESP teaching backgrounds in the EFL setting were selected for the rating process on a voluntary basis in the School of Foreign Languages in the local university. All raters shared the same second language background (i.e., Turkish) with a highly advanced level of English proficiency and had EFL teaching experience of 10–14 years. They had extensive experience in test development processes and assessing general oral communication of EFL learners. However, since there had never been a tourism English oral proficiency assessment before, none of them had any experience assessing the construct of interest in this dissertation study. Therefore, they went through 2.5 h of rigorous rater training and calibration session before the data collection process.

4.2 Materials and Instruments

During the domain analysis stage in the ECD framework, suggestions for potential test tasks were elicited through interviews with three THM experts, nine ESP course instructors, ESP course syllabi and textbooks in the local institution, and a literature review based on the ECD framework (Mislevy et al., 2003). The domain analysis yielded hotel guest complaints as the most common task that a hotel employee is likely to encounter. Building on this finding, one interactive SDS-based task (i.e., dealing with complaints) with three different questions was designed for intermediate-level ESL/EFL learners.

The CTEOC test task, *Dealing with Guest Complaints,* simulates three situations in a hotel setting: a smell problem in a non-smoking room (the Smell task), noise coming from the street (the Noise task), and a broken air-conditioner (the Broken AC task). A prototype SDS was designed based on these test tasks in which test takers are expected to act as a receptionist at a hotel interacting with the computer that acts as a customer (see Appendix A for a sample test task). In the CTEOC test, test takers are provided with a URL that directs them to the test webpage. On the first screen, students enter their username and password that they obtain from the test administrator. Then, when test takers access the webpage in Fig. 1, they first read the context and communicative objective of the task. A visual representing a receptionist talking on the phone at the front desk is presented to help them visualize the

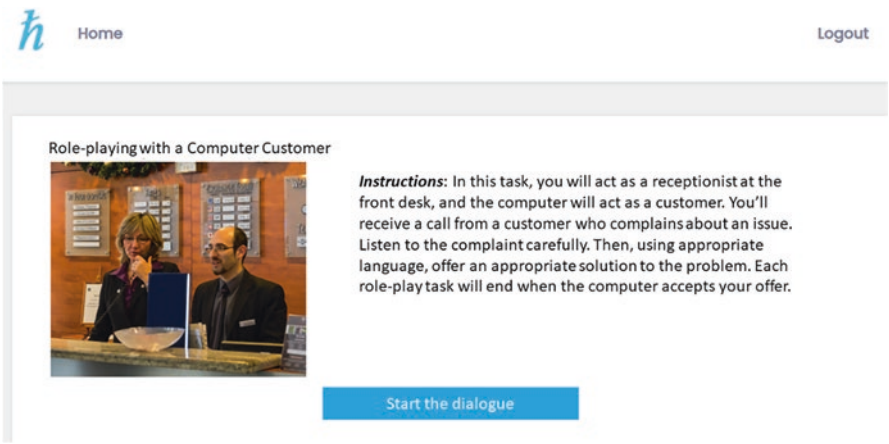


Fig. 1 A screenshot of the user interface in the CTEOC (Dealing with complaints)

role they are expected to play. Then, they click the "Start the dialogue" button, which initiates the SDS. At this stage, a virtual speaker starts telling the room number and then the specific problem with the room. The test taker is expected to apologize and find a solution to the associated problem, such as offering a room change, complimentary breakfast, a technician to fix the problem, etc. The SDS responds to what the test taker says by detecting keywords and phrases in the user's utterance (see Karatay, 2022 for a detailed description). To keep the conversation short as in real hospitality setting but long enough to elicit a ratable discourse from the test taker, the system is set to run for ten turns in total by rejecting each offer from the test taker and eventually accepting the offer. A sample interaction between a student and the SDS on a CTEOC test task can be found in this link.

4.3 Rating Scale and Scoring

Under the ECD framework, a rating scale was developed, piloted, and revised based on experts' consensus and feedback to assess test takers' oral performances on the CTEOC (Appendix B). The four-point analytic rating scale consists of four dimensions.

Following guidelines for rater training by Knoch et al. (2021), four raters completed a 2.5-h online rater training. Then, the researcher uploaded the audio files of each completed performance (i.e., three for each student) and sent them to the raters via the university's secure box system with a scoring sheet created for each rater. Each folder that contained the audio files and the scoring sheet was shared by each rater so that they could not access each other's ratings. Also, the order of the students and prompts were counterbalanced by re-numbering the audio

filenames and encouraging the raters to follow the order in the scoring sheet while rating.

Scores were calculated based on ratings from the three ratable speech samples produced by each student. Speech samples were rated using the 4-point rating scale by four experienced raters. Each rater assigned a score out of four for each category in the rating scale and took the average of their scores. Then the researcher took the average of the scores received from four different raters, which constituted the overall score of a student. All the ratings were carried out remotely and independently, and each rater was asked to enter the ratings into a scoring sheet specifically created for them and discouraged from discussing their ratings with each other. Therefore, the ratings provided by the raters were considered independent. Overall, the average length of speech samples was 1 min and 52 s. There were 30 audio files for each prompt (3 × 30), making 90 speech samples to be rated. It took approximately 6–10 h for each rater to complete the ratings.

4.4 Test Administration

Each participant took the test individually at a time when both themselves and the researcher were available. On the scheduled testing day, the researcher sent a detailed email to each participant to explain what they were expected to do and shared the link to the test webpage and their username that was assigned by the researcher. In the instruction email, the test takers were told to ensure that all the microphones and headphones were working and that there was a good quality of internet in the location. They were allowed to take the test via their cell phones, laptops, or in a computer lab. The test takers first had a chance to experiment with the SDS on their first try, in which they heard the conversational agent and attempted to talk with it for the first time. This practice attempt was excluded from the data analysis. Then, they took the following three questions (i.e., Broken AC, Noise, and Smell tasks). It took about 15 min on average for participants to complete the test.

Surveys and Semi-structured Interviews

This study employed an online questionnaire for test takers and individual interviews for both test takers and raters to collect data. A 5-point Likert scale questionnaire with six questions for background information and 15 questions for the test experience was used to elicit the test takers' perceptions of their interaction with the SDS. Then, a set of semi-structured interview protocols were also prepared for test takers to elicit their perceptions of the task administration conditions. As for the raters, a semi-structured individual interview protocol was developed to investigate the extent to which they perceive the rating as reflecting test takers' actual performance.

Data Analysis

The quantitative analysis of test takers' perceptions toward the SDS and raters' perception of the rating process were analyzed by examining the descriptive statistics for the responses to the 5-point Likert scales. Frequency distributions for each survey item in the post-task and post rating questionnaires were tallied to determine any common pattern or discrepancy in opinion among participants. The audio files of each interview were transcribed verbatim. Following Grounded Theory, which allows themes to emerge from the data (Glaser & Strauss, 1967), a Ph.D. student in an Applied Linguistics program and the researcher developed an initial coding scheme by going through the transcripts and coding line-by-line to identify the common themes within each of the areas of interest in this study: (a) for test takers, overall opinions about the user experience, task-specific perceptions towards SDS-based speaking activities, and possible improvements (b) for raters, general perceptions towards using the rating scale in an SDS-based speaking activity and overall opinions about the effectiveness of the rating scale.

After the coding scheme was confirmed by the two coders, one-third of the transcripts were coded to check the reliability of both the test taker data (Krippendorff's alpha = .81) and the rater data (Krippendorff's alpha = .89), and the rest was split by the two coders. Then, the findings were interpreted with the results of the quantitative data to shed light on test takers' and raters' perceptions of the innovative test task condition.

5 Results and Discussion

Table 1 shows a substantially high rater consistency ($\alpha = .94$) for the scores assigned by the four raters to 90 speech samples produced by 30 students. Also, Table 2 indicates that the mean score was 2.57, the minimum score was 1.2, and the maximum score was 3.6 on a four-point rating scale for 30 students. Overall, there were five Level 4, 11 Level 3, eight Level 2, and four Level 1 students based on the four-point rating scale.

The results indicated that both test takers and raters had positive perceptions toward the SDS-based tourism English assessment tasks. The following sections report findings based on the categories and themes from the survey and individual interview response data of test takers and raters.

Table 1 Interrater reliability of the ratings

Intraclass Correlation Coefficient						
	ICC	95% CI [LB, UB]	F	df1	df2	Sig
Average measures	.935	[.908,.955]	15.302	83	249	.000

Note. Two-way mixed effects model where people effects are random and measures effects are fixed

ICC Intraclass correlation coefficient, *CI* confidence interval, *LB* lower bound, *UB* upper bound

Table 2 Descriptive statistics for the CTEOC scores

	N	Mean	SD	Min	Max	Range
CTEOC	30	2.57	0.69	1.2	3.6	1–4

Table 3 Summary of 30 test-takers' comments on the CTEOC

	Positive	Mixed	Negative	Total comments
Overall perceptions of the CTEOC	17 (17)*	0 (0)	0 (0)	17
Potential test impact	26 (26)	0 (0)	0 (0)	26
Task authenticity	21 (21)	4 (4)	1 (1)	26
Preference (computer interlocutor)	9 (8)	4 (4)	10 (10)	23
Self-evaluation of performance	4 (4)	12 (12)	11 (10)	27

* The number in parenthesis indicated the number of the participants who gave the comments

5.1 Test Takers' Perceptions of the CTEOC Tasks

Table 3 below shows the categories of 119 unique comments provided by 30 test takers during the individual interviews, and Appendix C shows the percentages of the survey responses. In order to make more meaningful interpretations, while discussing each category, representative excerpts from students' individual interviews were presented along with the level of the students who provided each comment. These levels range from 1 (i.e., the lowest) to 4 (i.e., the highest).

Overall Opinions About the Tasks

As can be seen in the response to item 1 shown in Appendix C, overall, the majority of the test takers (91%) indicated that they enjoy the SDS-based speaking test tasks. Also, the coders identified 17 positive comments related to test takers' overall perceptions of the CTEOC tasks (Table 3). Twelve out of 17 students who positively commented on the CTEOC seemed to have contributed to their positive perception of it resembling a real-life interaction. A representative comment is that of Student 11, a Level 1 student, who stated:

> This was my first time experiencing such a thing. I took the test in my dormitory room with my roommates in the same place. They thought I was talking to a real person from a different country. (Student 11, Level 1, Post-test interview)

The realistic nature of the conversational agent was also described by 14 other students who used a variety of adjectives or expressions that are attributed to people, such as "annoyed," "stubborn," "difficult to convince," and "a good actor." An interesting metaphor was made by Student 17, who concluded:

> The computer was a little bit stubborn, but I am guessing that's how the system was designed. So, yes, she played her role very professionally, pushing the limits of my problem-solving skills. (Student 17, Level 4, Post-test interview)

Potential Test Impact

One of the desirable outcomes of the CTEOC is its potential impact on language learning and teaching practices in tourism English classes in the local institution. This issue was identified through the third question in the survey and an interview question asking explicitly about the potential test impact. According to the survey, among the 30 test takers, clear majorities thought that they would try to improve their speaking skill if they had a similar dialogue-based speaking test in the future (81%), which refers to a potential positive washback (Bachman & Palmer, 1996). During the interviews, all the students who commented on the possible impact of the CTEOC on their attitude towards tourism English oral communication predicted a potential positive impact (n = 26). To exemplify, Student 22 suggested:

> With only grammar, you can do little. If I knew that I would have a test like this, I would find all possible ways to practice my oral skills. (Student 22, Level 3, Post-test interview)

Task Authenticity

As the second and fourth items in the questionnaire indicate (Appendix C), the test takers found the tasks relevant to their future profession (81%) and thought that the interaction they had with the SDS was natural overall (81%). The interview data revealed that meaningful responses that the SDS generated during the interactions were the primary reason for 21 students who positively commented on task authenticity. For example,

> At one point, when I was stuck with offering a solution, she said, 'You've ruined my vacation.' This was pretty real and impressive. I felt bad. (Student 1, Level 3, Post-test interview)

Along similar lines, another utterance by the SDS made Student 10 feel that the interaction between her and the SDS was a real one rather than an artificial one.

> I was fascinated when it said, 'I can't move to another room because my kids are sleeping.' This is so relevant, logical, and realistic. I felt like I was talking to a friend in a classroom activity. (Student 10, Level 2, Post-test interview)

In addition to these positive opinions about the authenticity and the naturalness of the CTEOC tasks, four students had mixed opinions, and one student provided a negative opinion about these concepts regarding the test (Table 3). The four uncertain evaluations all seemed to be related to the limited corpus size of the SDS, as indicated in the following remark:

> After taking it several times, I realized that the number of sentences was not infinite as in other AI systems, and it repeated the same sentences on a couple of occasions in each trial. (Student 29, Level 3, Post-test interview)

Another confusion seems to be about the lag time of the system. While the participants thought that the test required them to respond promptly as in a real-life telephone conversation with a customer (86%), three were unsure whether the computer

needed to allow them to think about what to say next. It should be noted that all the three students who complained about the lag time between turns or frequent interruptions received an overall score of Level 1 and Level 2 on the CTEOC.

Overall Preference of a Partner

The participants do not appear to have a clear preference for a human partner (14%), whereas more than half of them (62%) preferred a computer partner in a dialogue-based assessment. Also, they expect to see more dialogue-based, computer-delivered speaking tasks in the future (62%). Test takers' responses to the interviews further explained their preference for a partner in role-play tasks as in the CTEOC. A total of 23 comments identified in students' interviews suggest that the students do not have a clear preference for a computer (n = 9) or a human partner (n = 10). The students who favor a human partner to interact with on a hospitality-related role-play task are mostly Level 1 and 2 students (n = 7) who expected to be accommodated because of the task demand or language proficiency.

The students who had mixed opinions about their preference for a partner pointed to the task relevance issues. For example:

> Normally I'd like to talk to a human rather than a computer because I can use my gestures and mimes to convince the other side. But in these situations, the customer will call you, not come to the front desk. So, you can't practice this with a human in front of you. (Student 10, Level 2, Post-test interview)

Self-Evaluation of Performance

In terms of the participants' self-evaluations of the effectiveness of test tasks in general and their performances, the students thought the SDS-based speaking test provided them with an opportunity to demonstrate their oral communication skills in a hotel situation (71%), the interaction between them was meaningful (57%), and the task difficulty was appropriate (62%). Also, most participants did not think that the computer's speech was fast (86%). However, when asked about their performances on the CTEOC during the interviews, the students had mostly mixed opinions (n = 12) or negative opinions (n = 11) out of 27 comments in this category. While the perception of the tasks was positive in general, some key aspects were elicited from the interview data, which provided more distinct insights into negative responses. One of these issues is the communication breakdowns caused by the lack of content knowledge. Out of 11 students who considered their performance unsuccessful, eight mentioned that since they did not have the internship or work experience yet, they were not prompt when thinking about how to solve the customer complaint. For example, Student 8 said,

> I don't have a problem with understanding the other side, but the problem is I don't' have any job or internship experience, and I don't know how to solve a problem like this. To be

> honest, I even googled how to solve an AC problem in a hotel room. (Student 8, Level 1, Post-test interview)

In particular, seven students complained about the lack of speaking practices in their tourism English classes.

Suggestions for Potential Task Improvements

In the survey and interviews, the students were asked whether they had any suggestions for improving the CTEOC. Findings point to one primary category of recommendations to advance the dialogue system. Eight out of 30 students advocated increasing the authenticity and complexity by improving the corpus size of the SDS. Even though most of the participants felt that the tasks were quite authentic insofar as they had to engage in a similar task in real hospitality settings, some (n = 5) suggested that new tasks could be integrated other than customer complaints at the front desk. In addition to the task variety, five students, who got at least a score of Level 3, commented on increasing the level of complexity of the current tasks by improving the ASR to have fewer generic statements and more meaningful turns. Even though the SDS in the CTEOC can maintain a conversation for ten turns, four turns higher than *Lisa Green* in Timpe-Laughlin et al. (2022) and one turn higher than *Ellie* in Kim et al. (2022), two students asked for longer dialogues.

Raters' Perceptions of the Rating Scale

The second research question investigated the raters' opinions about the use of the rating scale for scoring students' performances on SDS-based test tasks. To address this research question, the transcriptions of the semi-structured interviews were coded and analyzed, which yielded four themes identified in a total of 28 unique comments provided by four raters: overall usefulness of the rating scale, overall rating process (e.g., computer-based rating, rater training), rating human-computer interaction, and self-judgment of ratings. Table 4 below summarizes the number of comments provided by the raters under four categories.

Overall, the results indicated that all four raters had positive opinions about evaluating students' oral communication skills in a dialogue-based oral communication test. When asked whether they think the test and rating scale were effective in eliciting the examiners' oral communication skills in a hospitality setting, all four raters underlined the overall usefulness of both the test and the rating scale in assessing students' oral communication skills in an ESP context.

Since the raters were not involved as an interlocutor or test administrators during the process and were allowed to rate students' performances at their convenience, their opinions about the overall rating process were solicited during the interview.

Table 4 Summary of four raters' comments on the rating scale developed for the CTEOC

	Positive	Mixed	Negative	Total comments
Overall usefulness of rating scale	4 (4)*	0 (0)	0 (0)	4
Overall rating process	7 (4)	0 (0)	0 (0)	7
Rating human-computer interaction	6 (4)	3 (2)	0 (0)	9
Self-judgment of ratings	5(4)	3 (3)	0 (0)	8

*The number in parenthesis indicated the number of the participants who gave the comments

All four rates underlined the impact of the rater training on the process and the affordances of rating SDS-based interactions. Two raters specifically underlined the usefulness of the sample audio tracks that they listened to and the mock evaluating process they went through during the training. It appears that the quality of the rater training significantly impacted the ratings of the interactional competence category since none of the raters had an experience with the rating that specific subconstruct before. For example, Rater 1 said:

> This was my first time evaluating interactional competence since we don't have this criterion in our speaking exams here. So, I tried to remember our rating training sessions and referred to the descriptors in the rating scale while assessing that part. (Rater 1, Post-rating interview)

As for the convenience of the rating process in general, Rater 4 said:

> I did my ratings while having my coffee in my living room. And I stopped whenever I wanted. This comfort is not comparable to our face-to-face exams. (Rater 1, Post-rating interview)

Regarding the preference for computer-human interaction over human-human interaction, all four raters highlighted such affordances as the main reason for their evaluation by referencing the workload of a rater in the human-human oral exams. For example,

> In our exams, you are not just a rater. You are the one who administers the test in your own test room, asks the questions, be the interlocutor, rates the student, and calls the next one in. There are several redundant tasks in this process that the SDS handles in a very standardized way so that I can only focus on rating. (Rater 3, Post-rating interview)

The raters were also asked if the examinees would receive a higher/lower grade if they interacted with a human rather than a computer. Three out of four raters sympathized especially with low-level students and reported that they would assign slightly higher scores for those examinees, a finding that is in line with students who also thought that they would receive a higher score in a human-to-human oral exam. This finding was also found by Ockey and Chukharev-Hudilainen's (2021) raters, who thought that the computer partner was too harsh on low-level students. They thought they would be compensated and assigned a higher score with a human partner but also questioned the appropriacy of this kind of rater behavior.

In terms of the raters' judgments of their own ratings, while the interview data pointed to the five positive comments from four raters (Table 4) in general, three of the raters were unsure about the pronunciation and fluency ratings because of their familiarity with the examinees L1 background. For example, Rater 1 expressed:

> The easiest part to rate was interactional competence and grammar/vocabulary because the descriptors in those sections are pretty clear. It's easy to be objective. But for example, pronunciation and fluency are difficult for me because it is up to my understanding. (Rater 3, Post-rating interview)

Likewise, questioning the clarity of the test takers' pronunciation, Rater 4 was unsure whether a student's pronunciation was clear to her because she was sharing the same L1 background with them or because their pronunciation was actually clear.

6 Insights Gained

This study investigated the potential of using SDS-based role-play tasks to assess tourism English students' oral communication skills in an attempt to offer a solution to an educational problem: the lack of focus on oral communication skills in the local tourism English classrooms and assessment practices resulting from logistical challenges. The primary motivation behind this study was to create avenues for positive washback in language instruction in tourism English classrooms in the target context by offering a practical, reliable, and authentic oral communication test. As such, the findings obtained from the students' interview and survey data suggest that the students had an overall positive opinion about the authenticity and effectiveness of the SDS-based test tasks in eliciting their oral communication skills in simulated hospitality-related situations. They also believe that integrating the CTEOC into tourism English assessment practices would positively impact their learning processes and approach to oral communication in their discipline. Even though they did not show a clear preference for having a human or a computer partner for similar tasks as the CTEOC, they expect more SDS-based role-play tasks to prepare them for their future professions.

Similarly, data from the interviews revealed that all four raters were confident that the SDS was consistent across all test takers, eliminating the effect of partner variability on their performance. Based on their self-evaluations, this helped them consistently rate all of the categories on the rating scale. However, they commented hesitantly, especially on their ratings for low-level students, arguing that they could have been accommodated otherwise if the interlocutor were themselves, which was also found in Ockey and Chukharev-Hudilainen (2021). Also, some of the raters were suspicious about the effect of familiarity with the participants' L1 on their pronunciation ratings. Other than these issues, overall, raters seemed to think that

the rating scale was effective in eliciting the examiners' oral communication skills in a hospitality setting, and the convenience of rating SDS-based performances and the quality of the rating training increased their confidence in their ratings.

As revealed in the domain analysis, the course instructors in this study occasionally find themselves in a dilemma about the need to spend class time on oral communication practices but the impracticality of doing so. They end up sacrificing their teaching philosophies because of insufficient class hours, lack of an oral communication assessment, or resistance from students, all of which push teachers to a more written grammar-based instruction. This research demonstrated the promise for an assessment like CTEOC to offer solutions to the logistic problems to assess students' oral proficiency and motivate students to improve their oral communication skills to prepare for the CTEOC in their final exam, thus positively impacting language instruction in these classes leading to more communicative language instruction. Such a potential use of the CTEOC could even facilitate more tourism major graduates who have better English language oral communication skills and increase their chances of finding well-paid jobs and getting a promotion.

7 Implications for Test Developers and Users

This study elicited several implications for system designers and test developers who plan to develop SDS-based oral communication test tasks and test users. The outstanding suggestions elicited from the students were including fewer generic utterances and more meaningful turns, increasing the lag time between turns for low-level students, and expanding the corpus size of the SDS to have more task variety. The study utilized the keyword-spotting approach to increase the flexibility and robustness of the system for constrained dialogue situations as in hotel settings. Depending on the scope of the construct that is being assessed, this approach may serve its purpose well.

Since an SDS was used as an interlocutor in this local language test, this innovative approach to ESP assessment might inspire researchers, even local language testing teams, in the field. As Dimova et al. (2020) suggest, hiring software developers to outsource the design of digitally delivered tests is an alternative when the test development unit of a local institution does not include a software developer but has enough funding opportunities (p.93). The SDS used in this test based on an off-the-shelf, cloud-based ASR and a dialog manager script developed with a rule-based approach demonstrated how the cost of deployment and maintenance of such a system could be within reach of local institutions.

As data from students indicate, students are willing to improve their speaking skills if they have an interactive oral proficiency test similar to CTEOC. Therefore, a dialogue-based oral assessment in ESP programs such as the one in this study might have a positive washback on language instruction leading to more communicative language instruction. Such a test might facilitate more graduates with better

English language oral communication skills for their future profession, providing an advantage for finding a job or getting a promotion (Chapelle, 2021). Also, such tests as CTEOC can potentially be used for diagnostic purposes to identify weaknesses in L2 speakers' interactional competence.

These implications point to reasons that this study should inspire language testers, researchers interested in computer-assisted language learning, and classroom teachers to push apparent limits in finding solutions to language learning and teaching problems by utilizing cutting-edge technology in their own local contexts.

Appendices

Appendix A

Sample Transcript of a Test Task

1 c Hi, I am calling from room 263. It is very hot in my room, the A/C is not working.
2 u Oh I'm sorry to hear that how can I help you?
2 c You are not helping me.
3 u Okay I can change your room if you'd like.
3 c That won't work for me. Just do something about the A/C.
4 u Okay I'm.. I am sending a technician to fix the problem.
4 c You should have considered getting this fixed before you assigned me to this room.
5 u Okay. Sorry. I can give you a refund if you like.
5 c Giving my money back won't fix the problem.
6 u I know. But I am trying to find a solution.
6 c Look, I really appreciate you trying to help me. Is there a chance you can suggest a reasonable solution?
7 u Okay. Then I can move you to a new room.
7 c That's too much work for me. My luggage is very heavy.
8 u All right. Let me see.
8 c Are we still talking about the A/C?
9 u Yes.
9 c I don't think you are even trying to help me.
10 u if you don't want to move, then I can.. I can send a technician to fix the problem but you don't want it either.
10 c Alright, if you say so.

Note: c = Computer, U = user.

Appendix B

Rating Scale

Skills	Rating			
	4	3	2	1
Interactional competence	Task performed competently with almost always appropriate responses given in each task	Task performed generally competently with usually appropriate responses given in each task	Task performed somewhat competently and with somewhat appropriate responses given in each task	Task not completed due to limitations in responses given in each task
Fluency	Appropriate speech rate (i.e., no unnatural language -related pauses)	Mostly appropriate speech rate (i.e., few unnatural language -related pauses)	Somehow appropriate speech rate (i.e., some unnatural language -related pauses)	Inappropriate speech rate (i.e., too many unnatural language -related pauses)
Pronunciation	Very effective pronunciation and prosodic patterns with only minimal errors in production	Above average pronunciation and prosodic patterns but with occasional errors in production interpretation	Somehow average range of pronunciation and prosodic patterns but with errors in production affecting the delivery	Limited range of pronunciation and prosodic patterns with errors in production highly affecting the delivery
Grammar/ vocabulary	A high degree of grammatical accuracy in both simple and complex structures and domain-specific vocabulary use	Sufficient grammatical accuracy in both simple and complex structures and domain-specific vocabulary use.	Somehow sufficient grammatical accuracy in both simple and complex structures and domain-specific vocabulary use	Insufficient grammatical accuracy in both simple and complex structures and limited domain-specific vocabulary use.

Appendix C

Test Takers' Questionnaire Responses in 'Overall' and 'Preference' Categories

Category	Questionnaire item	1*	2	3	4	5
Overall	1. I really enjoyed the tasks.	0.0%	0.0%	9.5%	61.9%	28.6%
	2. The topic was relevant to real-life hotel situations.	4.8%	4.8%	9.5%	14.3%	66.7%
	3. If I know I'll have a speaking test like this, I might study to improve my speaking.	0.0%	0.0%	19.0%	33.3%	47.6%
Authenticity	4. I think the interaction I had with the computer was natural.	0.0%	4.8%	14.3%	28.6%	52.4%
	5. The computer did not allow me to think about what to say.	0.0%	9.5%	42.9%	33.3%	14.3%
	6. I needed to respond quickly as in real life telephone conversation with a customer.	0.0%	4.8%	9.5%	33.3%	52.4%
	7. There were instances when the computer did not respond to what I said.	0.0%	0.0%	19.0%	47.6%	33.3%
Preference	8. I would prefer to discuss with a human partner next time.	23.8%	57.1%	4.8%	9.5%	4.8%
	9. I would prefer to discuss with a computer partner next time.	4.8%	14.3%	19.0%	28.6%	33.3%
	10. I'd like to see more tasks like this in the future.	4.8%	4.8%	28.6%	33.3%	28.6%
Self-evaluation	11. The test allowed me to demonstrate my oral communication skills at a hotel situation.	0.0%	0.0%	28.6%	38.1%	33.3%
	12. The computer understood me well.	0.0%	4.8%	38.1%	47.6%	9.5%
	13. I was able to understand the computer well.	4.8%	9.5%	28.6%	28.6%	28.6%
	14. The difficulty of the task was appropriate for me.	0.0%	19.0%	19.0%	42.9%	19.0%
	15. The computer's speech was fast for me.	0.0%	4.8%	9.5%	19.0%	66.7%

*1=Strongly Disagree, 2=Disagree, 3=Neutral, 4= Agree, 5= Strongly Agree

References

Abdel Ghany, S. Y., & Abdel Latif, M. M. (2012). English language preparation of tourism and hospitality undergraduates in Egypt: Does it meet their future work place requirements? *Journal of Hospitality, Leisure, Sport & Tourism Education, 11*, 93–100. https://doi.org/10.1016/j.jhlste.2012.05.001

Aysu, S., & Ozcan, F. H. (2021). Needs analysis in curriculum design: Language needs of tourism students. *Sakarya University Journal of Education, 11*(2), 305–326. https://doi.org/10.19126/suje.854993

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford University Press.

Blue, G. M., & Harun, M. (2003). Hospitality language as a professional skill. *English for Specific Purposes, 22*(1), 73–91. https://doi.org/10.1016/S0889-4906(01)00031-X

Chapelle, C. A. (2021). *Argument-based validation in testing and assessment*. Sage Publishing.

Chukharev-Hudilainen, E., & Ockey, G. J. (2021). *The development and evaluation of interactional competence elicitor (ICE) for oral language assessments*. ETS Research Report. Educational Testing Service. https://doi.org/10.1002/ets2.12319

Creswell, J. W., & Plano Clark, V. L. (2012). *Designing and conducting mixed methods research*. Sage Publications, Inc.

Dimova, S., Yan, X., & Ginther, A. (2020). *Local language testing: Design, implementation, and development* (1st ed.). https://doi.org/10.4324/9780429492242

Evanini, K., Timpe-Laughlin, V., Tsuprun, E., Blood, I., Lee, J., Bruno, J., & Suendermann-Oeft, D. (2018). Game-based spoken dialog language learning applications for young students. In *Proceedings from the 2018 Interspeech conference* (pp. 548–549).

Glaser, B. G., & Strauss, A. L. (1967). *The discovery of grounded theory*. Aldine.

Gokturk N. (2020). *Development and evaluation of a spoken dialog system-mediated paired oral task for measuring second language oral communication ability in English* (Publication No. 28153696) [Doctoral dissertation, Iowa State University]. ProQuest Dissertations Publishing. https://doi.org/10.31274/etd-20210114-53

Karatay, Y. (2022). D*evelopment and Validation of Spoken Dialog System-Based Oral Communication Tasks in an Esp Context* (Order No. 29165842). Available from Dissertations & Theses @ Iowa State University; ProQuest Dissertations & Theses Global. (2725255274). https://www.proquest.com/dissertations-theses/development-validation-spoken-dialog-system-based/docview/2725255274/se-2

Kasper, G., & Youn, S. (2018). Transforming instruction to activity: Role-play in language assessment. *Applied Linguistics Review, 9*(4), 589–616. https://doi.org/10.1515/applirev-2017-0020

Kim, H., Yang, H., Shin, D., & Lee, J. H. (2022). Design principles and architecture of a second language learning chatbot. *Language Learning & Technology, 26*(1), 1–18. http://hdl.handle.net/10125/73463

Knoch, U., Fairbairn, J., & Jin, Y. (2021). *Scoring second language spoken and written performance: Issues, options and directions*. Equinox Publishing.

Leslie, D., Russell, H., & Govan, P. (2004). Foreign language skills and the needs of the UK tourism sector. *Industry and Higher Education, 18*(4), 255–266.

Leslie, D., & Russell, H. (2006). The importance of foreign language skills in the tourism sector: A comparative study of student perceptions in the UK and continental Europe. *Tourism Management, 27*(6), 1397–1407. https://doi.org/10.1016/j.tourman.2005.12.016

Litman, D., Strik, H., & Lim, G. S. (2018). Speech technologies and the assessment of second language speaking: Approaches, challenges, and opportunities. *Language Assessment Quarterly, 15*(3), 294–309. https://doi.org/10.1080/15434303.2018.1472265

Manias, E., & Mcnamara, T. (2016). Standard setting in specific-purpose language testing: What can a qualitative study add? *Language Testing, 33*(2), 235–249. https://doi.org/10.1177/0265532215608411

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). A brief introduction to evidence-centered design. *ETS Research Report, 03-16*. https://doi.org/10.1002/j.2333-8504.2003.tb01908.x

Ockey, G. J., & Chukharev-Hudilainen, E. (2021). Human versus computer partner in the paired oral discussion test. *Applied Linguistics, 1-21*. https://doi.org/10.1093/applin/amaa067

Okada, Y., & Greer, T. (2013). Pursuing a relevant response in oral proficiency interview role plays. In S. Ross & G. Kasper (Eds.), *Assessing second language pragmatics* (pp. 288–310). Palgrave Macmillan. https://doi.org/10.1057/9781137003522_11

Roever, C. (2011). Testing of second language pragmatics: Past and future. *Language Testing, 28*, 463–481. https://doi.org/10.1177/0265532210394633

Timpe-Laughlin, V., Evanini, K., Green, A., Blood, I., Dombi, J., & Ramanarayanan, V. (2017). Designing interactive, automated dialogues for L2 pragmatics learning. In V. Petukhova & Y. Tian (Eds.), *Proceedings of the 21*st *workshop on the semantics and pragmatics of dialogue* (pp. 143–152).

Timpe-Laughlin, V., Sydorenko, T., & Daurio, P. (2020). Using spoken dialogue technology for L2 speaking practice: What do teachers think? *Computer Assisted Language Learning, 1-24*. https://doi.org/10.1080/09588221.2020.1774904

Timpe-Laughlin, V., Sydorenko, T., & Dombi, J. (2022). Human versus machine: Investigating L2 learner output in face-to-face versus fully automated role-plays. *Computer Assisted Language Learning*. https://doi.org/10.1080/09588221.2022.2032184

Van Batenburg, E. S. L., Oostdam, R. J., Van Gelderen, A. J. S., Fukkink, R. G., & De Jong, N. H. (2019). Oral interaction in the EFL classroom: The effects of instructional focus and task type on learner affect. *The Modern Language Journal, 103*, 308–326. https://doi.org/10.1111/modl.12545

Youn, S. J. (2015). Validity argument for assessing L2 pragmatics in interaction using mixed methods. *Language Testing, 32*, 199–225. https://doi.org/10.1177/0265532214557113

Zahedpisheh, N., Bakar, A., Zulqarnain, B., & Saffari, N. (2017). English for tourism and hospitality purposes (ETP). *English Language Teaching, 10*(9), 86–94. https://doi.org/10.5539/elt.v10n9p86