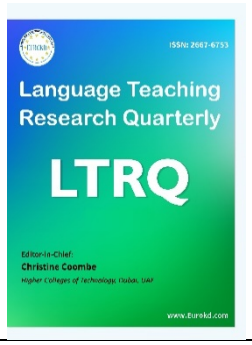

Language Teaching Research Quarterly

2022, Vol. 29, 120–133

Performance-based Speaking Tests: Possibilities in Local Language Testing

Slobodanka Dimova

University of Copenhagen, Denmark

Received *03 January 2022* ***Accepted*** *29 April 2022*

Abstract

Drawing on Glenn Fulcher's extensive work in performance-based language assessment of speaking, this paper explores the assessment of L2 speaking ability in local language testing contexts. For that purpose, I review Fulcher's influential work that highlights the relationship between the speaking construct, the task, the performance, and the scale, and I provide an overview of various task types, approaches to scale development, and rater-training programs. Then, I argue that local language testing provides a wide range of possibilities for task development, scale design, and rater training due to the opportunity for collaboration with local expertise (students, instructors, policy-makers) and the ability to keep up with the evolving speaking construct in the local context.

Keywords: *Performance-based Speaking Test, Local Language Testing*

Introduction

Evidence about performance-based speaking assessment can be found since the turn of the 20th century (Fulcher, 2012), but this type of assessment gained increasing attention with the introduction of communicative language teaching in the 1970s. At that time, the development of fluent speech production and engagement in communicative activities became essential instructional and learning goals, so inclusion of tasks that authentically represent real-world communicative situations became a staple in language teaching and assessment. Given that these tasks tend to elicit open-ended speech performance with various structural, functional, discursive, and content characteristics, there has been a growing interest in describing these characteristics in order to understand the nature of speaking ability and develop and use task-based instruction and performance-based speaking assessment (Fulcher, 2015; Nakatsuhara et al., 2020).

slobodanka.dimova@hum.ku.dk

doi:10.32038/ltrq.2022.29.08

Glenn Fulcher's contributions in the area of performance-based assessment of speaking have had a major impact in the field and inspired numerous research studies that have focused on identifying tasks that elicit a range of language skills related to various communicative purposes, as well as theoretical discussions regarding the different components of speaking ability and the various models of performance-based assessment of L2 speaking. Research in language assessment, in particular, has centered on task features and design, performance characteristics, scales and scoring, and rater training and behavior, mostly in relation to large-scale, international language tests or classroom assessment.

This paper draws on Fulcher's work to spotlight research and practice in local language testing of L2 speaking. Dimova et al. (2020) define local test "as one whose development is designed to represent the values and priorities within a local instructional program and designed to address problems that emerge out of a need within the context in which the test will be used" (p.1). Local language tests can be placed between large-scale standardized tests and classroom assessment and can serve various purposes (e.g., screening, placement, diagnostic, certification). For instance, many higher education institutions have developed local language tests for diagnostic or placement purposes, especially in the new non-English dominant contexts in which English has been introduced as a medium of instruction (EMI) (Dimova, 2020c).

I argue that local language testing allows for a wide range of possibilities when it comes to development of speaking tasks, rating scales, and rating procedures, because of the possibility to involve local expertise (students, instructors, policy-makers) and to engage in continuous test validation and development in order to keep up with the changing context. In the review, special attention is placed on tasks, rating scales, and raters, as Upshur and Turner (1999) suggested that the scoring process is not only influenced by test-takers' performance, but also by the test-takers' engagement with the task, as well as the raters' interaction with the scale, the task, and the test-takers' response.

Speaking Tasks

Twenty years ago, Fulcher (2003) pointed out that despite the growing research that improved our understanding of language variation across contexts, the different speaking genres, and the role and characteristics of interaction, the field lacked methodological research that identified the most effective approaches and techniques to teaching and assessing speaking. Despite the continuing research that focuses on describing new domains of language use (e.g., English uses in EMI programs at non-English dominant universities (Dimova & Kling, 2020), the emerging norms in English as a lingua franca (ELF) contexts, and the rising interest in translanguaging and plurilingual communication (Dimova, 2020a), today we still lack a common understanding (or a unified construct) of speaking (Xi et al., 2021). Although a unified construct might allow us to compare speaking performances across contexts, target language use domains, and assessment purposes, defining the construct in relation to the test purpose and the needs of the stakeholders remains a guiding principle in assessment of speaking. Therefore, Fulcher's argument that it is important to select the best speaking tasks in order to elicit relevant speaking performance that can be scored by a rating instrument that operationalizes the construct for the testing purpose maintains

its relevance. Given the possibility that the speaking performance may vary based on task features, difficulty, or conditions, and, therefore, affect our inferences about the test-taker's speaking ability, the task can play a key role in the construct definition (Fulcher & Reiter, 2003).

Performance-based speaking tasks follow the communicative legacy by eliciting open-ended responses through which test-takers demonstrate their ability to achieve a particular communicative purpose (Fulcher, 2000). Based on the assessment purpose and the inferences to be drawn based on assessment results, test developers can select from tasks that are characterized as monologic, interactional, or integrated (Dimova et al., 2020). Monologic tasks elicit individual test-taker production of longer responses based on narration or argumentation. For instance, narrative tasks require test-takers to describe a visual (e.g., picture, figure, cartoon, photo), construct a story (or a chronology) based on a series of pictures, or narrate personal experiences (Fulcher, 2003). Argumentative tasks, on the other hand, require that test-takers present their opinions on a general topic or a topic related to a particular disciplinary domain. Monologic speech production is usually elicited through a short prompt that introduces one or more questions or a topic, and, in some cases, information about how to structure the response. However, speech production can also be based on additional input (written or audio-visual), i.e. test-takers are asked to construct an oral response that draws on textual or auditory information. Due to their complexity and requirement for activation of not only test-takers' speaking skills but also their reading and/or listening comprehension skills, such tasks are referred to as "integrated tasks".

In the recent years, interactional and pragmatic competences have gained attention as essential components of the L2 speaking ability construct, and discussions regarding the validity of interactional tasks have become increasingly present in the language testing literature (Roever, 2011; Taguchi, 2018; Taguchi & Roever, 2017). Interactional tasks could involve a conversation with a trained interlocutor, who could also be an examiner, or a conversation between two (paired) or more (grouped) test-takers. Oral interviews are a typical type of an interactional task, in which a trained interlocutor follows a scripted set of questions and a standard protocol in order to elicit the test-taker's speaking performance. Paired (and group) tasks have traditionally been viewed as challenging due to the potential variability of interlocutors (age, gender, proficiency level) and the possibility for one interlocutor to dominate the conversation (O'Sullivan, 2000). However, recent research suggests that such variability in interactional tasks provides opportunities for co-construction of the interaction and elicitation of various competencies, such as conversational management, turn taking, topic initiation and ending, sequencing, and clarification request (Ducasse & Brown, 2009; May, 2009, 2011; Taylor & Wigglesworth, 2009; Youn, 2020).

Simulations and role-plays have the capacity to extend the elicitation of interactional competence by introducing a number of contextual variables, such as, but not limited to, the relationship between interlocutors, pragmatic actions, and communicative settings (Kasper & Rose, 2002). Although these task-types are not fully authentic, they allow test-takers to demonstrate whether they can fulfill social actions through meaningful interaction (Seedhouse & Nakatsuhara, 2018). Conversation analysis has been commonly applied as a research method to analyze the interactional features of tasks, be they interviews, paired tasks, or role-plays. Studies have repeatedly confirmed that unlike interviews, paired speaking tasks and role-plays exhibit the

potential to elicit authentic interactional features found in ordinary conversation (Al-Gahtani & Roever, 2012, 2018; Okada, 2010; Okada & Greer, 2013; Seedhouse & Nakatsuhara, 2018; Stokoe, 2013).

Based on Clark's (1979) classification of testing methods as indirect, semi-direct, and direct, both monologic and interactional tasks that are performed live are considered a *direct* testing method, in that they elicit a performance that requires use of the skills that we intend to assess (Hughes, 2003). These tasks are associated with higher level of authenticity and contextual representation, and, therefore, believed to be a more valid measure of speaking ability (Qian, 2009; Luoma, 2004). The drawback of performance-based speaking tasks is that, when administered live, they tend to be extremely time consuming. In order to overcome these practicality issues, digital delivery of L2 speaking tests first became an attractive option, but now it has become a necessity due to the COVID-19 restrictions and the inability to administer L2 speaking tests live. Digital delivery means that speaking tests are administered on a computer, or another digital platform, without examiners on site. This allows for simultaneous test administration to a large number of test-takers (e.g., in a large computer lab) and/or distant test administration, i.e. test-takers can complete the test at home or another certified location. Given that digitally-administered tests still elicit open-ended performances (usually recorded on the local machine, a server, or cloud storage) despite the lack of physical interaction with an examiner, these tests are referred to as *semi-direct* tests.

Semi-direct speaking tests have been criticized for their lack of authenticity and the interference of the construct of digital literacy with the construct of speaking ability. However, research suggests that in addition to being efficient and cost-effective, semi-direct tests tend to be more reliable than direct tests because they eliminate interlocutor variability (Qian, 2009). Although online oral communication has existed for a number of years, the increased online interaction due to COVID-19 (e.g., meetings, teaching, learning, socialization) has highlighted the extended domains of speaking ability that include various online contexts and genres. These developments may lead to an enhanced justification of digital test delivery in performance-based testing.

Rating Instruments

Development

Tasks elicit performance-based speaking responses that are multifaceted and require a holistic or a detailed perspective on the different levels of speaking ability. For that reason, development of a rating scale that guides the principled judgement of proficiency levels is central. Rating scales can be considered instruments that operationalize the measured construct (Davies et al., 1999), so they often tend to draw on existing theoretical frameworks that outline the various components of speaking ability. While theoretical and experiential approaches have led to intuitive and measurement driven scale development, Fulcher (1997, 2003) highlighted the importance of developing scales based on performance data analysis.

The intuitive approach to scale development requires that experts from the area (e.g., experienced language teachers, language testers, applied linguistics researchers) decide on the scale levels and verbalize descriptors based on their familiarity with language instructional

curricula, understanding of the characteristics of various language learners, and previous experiences with speaking tests (Fulcher, 2018).Through active engagement with the scale, those who use it have the opportunity to improve the scale descriptors over time (Kuiken & Vedder, 2020). These scales have been criticized for lack of specificity and theoretical underpinnings, which could lead to diverging scale interpretations and inconsistency among raters (Knoch, 2009).

Sometimes, the expert group in charge of scale development consults existing speaking scales and uses them as a starting point for development of e new one by drawing on scale descriptors relevant for their own testing purposes. For instance, the Common European Framework of Reference (CEFR) can offer numerous descriptors of different aspects of L2 speaking ability from which scale developers can select (Council of Europe, 2001). The process of selection, adaptation, and operationalization of scale descriptors to the local testing context has been termed as localization of CEFR (Council of Europe, 2001; 2020).

In order to validate the scales developed through the intuitive approach and reject the accusations of being abstract, imprecise, and ineffective, scale developers usually perform supplementary measurement analyses (e.g., multi-facet Rasch measurement). The purpose of such analyses is to establish the psychometric qualities of the scale and confirm that the scale can reliably differentiate actual test performances across the established scalar levels. Intuitive scales that have been finalized based on a measurement model, rather than performance data, are referred to as measurement-driven scales (Fulcher et al., 2011).

Fulcher (1987, 2003) argued that scale design that is driven by theoretical descriptions of the construct may lack precision and that using performance data to identify the construct allows for an improved understanding of human communication in a particular context. He applied conversation and discourse analysis of a corpus of transcribed speaking performances in order to detect the main performance features across the different proficiency levels and develop the scale descriptors. With this approach, the number of levels can be established based on a discriminant function analysis, which shows the ability of the scores to distinguish the proficiency levels based on the scale (Fulcher, 1993; 1996; 2003).

Types of Rating Scales

The different approaches to scale development, be they experiential or data-driven, can be applied to the development of different types of rating scales for measuring speaking performance, the most prevalent of which are holistic and analytic scales. The choice of rating scale type depends on the definition of the underlying construct and the assessment purpose (Fulcher, 2003). The holistic scales are based on one underlining construct, so one score is assigned based on the overall L2 speaking performance. On the other hand, analytic scales, or multiple-trait scoring rubrics, are defined sub-constructs of L2 speaking ability, so multiple scores are assigned in relation to different aspects (sub-constructs) of the L2 speaking performance (e.g., fluency, coherence, grammar).

Despite the simplicity of application associated with generic holistic scales, they have been criticized for being unreliable, broad, and inadequate for diagnostic feedback (Bachman & Palmer, 1996; Fulcher, 2003; Knoch, 2009; Ohta et al., 2018). Therefore, analytic scales are commonly recommended in local language testing situations, especially when the speaking test performances

are used for diagnostic and feedback purposes. The drawback of analytic scales is, however, their tendency to compartmentalize the overall speaking ability construct leading raters to focus on specific aspects of the performance and ignore the effects of the interaction between different speaking characteristics on the overall comprehensibility or communicative effectiveness of the L2 speaking performance.

Rating scales could be designed for uses across different tasks, or they could be task-specific. The purpose of a task-specific scale is to capture the salient trait of the speaking performance that is particular for the communicative context of the task, and therefore such scales are referred to as primary-trait scales. These scales also lend themselves more to the local language testing context because their development and use is time-consuming due to the expected provision of thick, prompt-specific descriptions of expected performances at each level.

One type of analytic scale that is based on a data-driven development approach is the Empirically Based Boundary scales (EBBs) (Turner, 2000; Turner & Upshur, 2002; Upshur & Turner, 1995, 1999). The development of EBB scales requires identifying the main features of speaking performances that help raters separate the adjacent levels on the scale. These features are represented in a binary (yes/no) decision chart, where raters make decisions about the performance level based on whether the feature is present in the performance or not. In other words, the speaking ability is represented into different levels of branching binary choices related to different questions about the performance characteristics. The questions tend to start from a holistic element (e.g., Does the candidate display fluency? No -- Level 1; Yes – Level 2 or 3) and gradually branch into more specific elements (e.g., Does the candidate display grammatical accuracy? No – Level 2; Yes — Level 3). EBB scales are intended to guide the raters' cognitive process when they judge performance data in particular context of language use.

Based on the EBB scale type, Fulcher et al. (2011) presented the performance decision tree (PDT) approach to scoring of L2 speaking ability. PDT scale design is based on analysis of performance data in real-world, domain-specific communicative situations or from test tasks designed to reflect these real-life situations. This approach is intended to add descriptive richness to the scale while maintaining the streamlined decision-making by the EBB. Arguably, the advantage of PDT over EBB is the increased relevance of the language use descriptions for the particular communicative contexts in which interaction takes place.

Raters and Rater Training

Although it has been argued that the rating scale characteristics contribute a great deal to the proficiency level differentiation and scoring reliability, an even more important factor in the scoring process is rater behavior. In order to achieve rater reliability, or rater agreement and consistency, a common practice is to involve raters into a rater-training program, in which they familiarize themselves with the scalar levels and descriptors. Fulcher (2003) warned, though, that scale validation based on performance data analysis should precede rater training because:

> If raters are trained, 'socialised' or 'cloned' before the validity argument is constructed, the training itself becomes a facet of the test that cannot be separated from the construct. (p. 301)

Regardless of how thick the scale descriptors are, it is the raters' interpretation of these descriptors that eventually leads to score assignment. Raters' linguistic backgrounds, accent familiarity, educational and instructional values, traditions, and philosophies, as well as previous assessment experiences and understanding of the context, influence raters' interpretations and interaction with the rating scale (Kim, 2009; Wei & Llosa, 2015; Winke et al., 2013; Yan, 2014; Zhang & Elder, 2011). The differences in scoring among raters are based on the degree of rater severity (or leniency) in their score assignment. In order to minimize the consequences of rater severity variation, L2 speaking performances are rated independently by at least two raters. An acceptable inter-rater reliability, i.e. consistency in score assignment, is at 0.8 or higher (Ginther, 2012). However, given that rating scales represent a continuum, rater agreement on borderline performances remains difficult, especially when it comes to assignment of holistic scores.

Structured and continuous rater training is essential in order to obtain and maintain higher inter-rater agreement and soring quality. Rater training programs typically follow several stages. First, novice raters are familiarized with the rating scale levels and descriptors. They listen to, or watch, benchmark performances for each level in order for raters to begin recognizing performances at different scalar levels and to link the scale descriptors to their perceptions of the speaking performance. In the last stage of the rater-training program, novice raters are invited to practice rating various speaking performance samples. In some cases, raters may be certified after they reach a certain level of agreement (70-80%) with the scores that the performance samples were originally assigned. The rater-training process may be individual (e.g., online training program [Elder et al., 2007; Knoch et al., 2016] in which raters access the training and practice materials) or in group (e.g., online or onsite meetings to discuss the scale levels, descriptors, and rating samples). After the initial rater training program, raters often engage in periodical training, or norming, sessions, especially before a large test admin or if there is an indication of falling inter-rater reliability.

Research regarding the effectiveness of rater-training programs indicates diverging findings. Some studies have shown that rater training is effective in that inter-rater reliability and agreement tends to increase and excessive severity or lenience tend to diminish after completion of a rater-training program for novice raters (Davis, 2016). Other studies suggest that excessive cases of severity may not be eliminated with rater training (Lumley & McNamara, 1995) or that rater-training enhances more the intra-rater than the inter-rater reliability (Weigle, 1998). Although individual feedback may not always make a difference in rating performance (Knoch, 2011), in some cases it may be effective (Elder et al., 2005), and it may promote self-reflection (Sundqvist et al., 2020).

Fulcher (2003) suggested that rater training helps raters to establish a common understanding of the scale descriptors -- this allows raters to align their perceptions of scale descriptors and approaches to the rating process but also to raise their awareness of their own biases in relation to specific aspects of L2 speaking performances (Sandlund & Sundqvist, 2021). Therefore, instead of periodical rater training sessions, opportunities for discussion and alignment could be established through consensus moderation or social moderation (Linn, 1993; Sadler, 2013). In addition to equity and accountability, moderation fosters collaboration and community building

(Bloxham et al., 2016). Community building allows raters to continuously verify their assessment judgements against standards and therefore build a shared understanding of the construct, the assessment tasks, and the rating criteria. It also promotes development of shared descriptor interpretations and assessment values (Sadler, 2013).

Moderation and community building have originally been discussed in relation to assessment of student achievement in a specific course in higher education. However, the same principles can be applicable to local language testing, especially when the local test is embedded within a specific language program. Given that language teachers, and in some cases other stakeholders, tend to also serve as a rater in the local context, community building provides opportunities to improve the assessment approaches and inform teaching through an ongoing discussion. Community building and consensus moderation can start with stakeholder involvement in scale design and verbalization of scale descriptors.

Local Language Testing

Fulcher draws on the philosophy of Pragmatism to suggest that the purpose of language testing should be viewed in terms of its practical uses, inferences, and successes, as well as its contribution to advancement of individuals (Fulcher, 2015). He emphasizes the role of context, the need for clear understanding of the linguistic characteristics of domain-specific task performances in test and scale design, and the importance of community-based rater training. The type of precise description of the sub-domains of language use that he proposes, grounded within particular communicative settings and local norms, values, and communicative expectations, seems most viable in local language testing. In the local context, Fulcher's data-driven approach to development of speaking assessment allows for a detailed definition of the speaking construct and its alignment with task and rating scale design, as well as for improved rater cognition of the rating processes and scale use.

In a specific local context, the pragmatism of language tests can be enhanced as the decisions about selection of tasks in test design can be guided by concrete needs analyses, in which local stakeholders' input is gathered in order to understand the test purpose and uses, as well as the local instructional and testing traditions and values (Dimova et al., 2020). For instance, the Oral English Proficiency Test (OEPT) at Purdue University is a semi-direct test for screening of international teaching assistants (ITAs) for oral English proficiency. The computer-delivery format was selected in this local context due to the need to administer the test efficiently to a large volume of students and rate their responses within a short time period before the beginning of the academic year (Ginther et al., 2012). The test format provides opportunities for simultaneous administration to a group of students and obtaining immediate access to their responses through an online interface. On the other hand, the Test of Oral English Proficiency for Academic Staff (TOEPAS), which is used for certification of L2 English speaking lecturers who teach English-medium instruction (EMI) courses, is a direct test based on a teaching simulation. The TOEPAS testing method was decided upon based on interviews with various stakeholders (e.g., lecturers, department heads, study program directors, union representatives) who expressed the need and value of obtaining feedback relevant for their English language use in the classroom in addition to a score (Kling &

Stæhr, 2011). These values could not be reinforced with the application of semi-direct tests, even though they could have been more efficient and cost-effective.

Any type of rating scale (holistic, analytic, EBB, PDT) could be developed and implemented in a local language testing setting. Regardless of which scale type is selected, an important possibility that the local context offers is involvement of local stakeholders in the design, validation, and implementation of the rating scale. Stakeholders' engagement with the rating scale from the point of its inception allows to represent the local values and to enhance stakeholders' understanding of the construct(s) the scale represents, and, therefore, the test results become more interpretable and meaningful for them. An enhanced understanding of the construct could lead to a more apparent link between instructional goals and tasks with assessment for teachers, and a lower incidence of score misinterpretation and misuse for decision-making purposes.

A mixed, or a hybrid, scale development method is easily implemented in local language test settings as the test developers and researchers get access to both speaking performance data and local expertise. Local test developers can involve local expert judgments in the initial wording of scale descriptors, while measurement-based approaches would allow for ensuring raters' consistency in the scale use, and the data-driven approaches help to link the scale with speaking performance characteristics (Dimova et al., 2020; Kuiken & Vedder, 2020).

For example, in the first phase of the development of the TOEPAS scale, the main constructs were identified through 1) reviewing theoretical models of speaking ability and existing speaking scales (e.g., ILR, CEFR, ACTFL) 2) observing university lecturer language use in the classroom, and 3) interviewing lecturers, students, department heads, and heads of study boards regarding the type of language functions needed for teaching and the expected communicative practices in the classroom (Kling & Stæhr, 2011). In the second phase, workshops with groups of stakeholders were held in order to obtain feedback on both the overall construction of the scale descriptors. The raters (English instructors, applied linguistics) and a group of 19 graduate students of English at the University of Copenhagen participated in a jigsaw exercise to rank the descriptors (lowest to highest proficiency) in order to confirm their natural progression from the lowest to the highest scalar levels. In the third phase, a field test trial was conducted under operational conditions, followed by a pilot test with 19 lecturers who volunteered to take the TOEPAS. Performance data from the pilot test were used to refine the scale and improve the precision of the descriptors. In the fourth phase, multi-facet Rasch (MFRM) analyses were conducted to determine the scale's ability to distinguish performances across the different levels, a procedure commonly applied in measurement-driven approaches (Dimova & Kling, 2015). Based on the results from the MFRM, the scale needed further refinement as some of the scale bands allowed for much more variation than other. In the fifth phase, speaking performance data were ranked based on proficiency level and the language use characteristics were analyzed in order to improve the precision of the scale descriptors. In other words, a data-driven approach was applied for scale revision purposes.

Local language tests offer a possibility for hybridity not only in the scale development method, but also in the characteristics of the rating scale, thus benefiting from both holistic and analytic scale features. For instance, the two local tests mentioned, OEPT and TOEPAS, use hybrid scales. The OEPT raters assign holistic scores for each item and overall holistic score for the test-taker's

performance across all items, but these overall score assignments are informed by scale descriptors across several categories, namely, pronunciation and intelligibility, fluency, lexical sophistication and accuracy, grammatical complexity and accuracy, and coherence (Dimova et al., 2020). Similarly, the TOEPAS raters assign one holistic score for the test-taker's performance, but the raters refer to different categories (addressing audience, fluency and cohesion, pronunciation, and grammar and vocabulary) to justify the score and point out the strengths and the weaknesses of the performance in the extensive formative feedback that accompanies the score (Dimova, 2020b).

Regarding rater training, local language testing provides opportunities for building a rater community of practice in which rater cognition of the construct and the scale can be developed through an ongoing dialog. Dimova et al. (2020) discuss the TOEPAS rater-training procedures as an example of a rater-training program that is grounded in a community building process. The TOEPAS rater team has grown into a community of practice with shared knowledge, rater identity, and rater cognition, due to the collaborative nature of the TOEPAS assessment procedure. The raters have been involved in collaboration at different stages of TOEPAS development and administration, i.e. they contributed to the establishment of scale descriptors, participated in group training sessions, engaged in post-score discussions, and collaboratively prepared written feedback reports for test-takers (p. 154). The post-score discussions and the provision of feedback reports after the test administration and the submission of independent scores force raters to interact actively with the rating scale levels and descriptors in their endeavor to justify their score assignment and describe the L2 speaking performance to raise test-takers' awareness about their strengths and weaknesses in their English use in the EMI classroom. These processes help raters to align their scale and descriptor interpretations and exemplify the descriptors with concrete examples from the speaking performance, which supports rating consistency. MFRM analyses suggest that despite variation in rater severity levels, inter-rater reliability and rater agreement remain relatively high (Dimova & Kling, 2018; Kling & Dimova, 2015).

Conclusion

The field needs domain-specific theoretical models for assessment of speaking (especially in the academic domains) that would allow for an improved understanding of the ability components and a more precise operationalization of the construct. The role of content and topical knowledge, the context, and the cognitive and metacognitive processes also requires further examination (Xi et al., 2021). Fulcher's pragmatic approach to language testing that highlights the usefulness of performance data embedded in a particular context in relation to specific testing needs and social and educational values provides us with tools to define the relevant construct for the testing purpose. However, given the contextual variation with regard to language use characteristics, language norms, communicative situations, and populations, identifying a single model may be a difficult feat.

Local language testing offers many possibilities regarding specific domain representation because of immediate availability of test data, accessibility of test-takers, and continuous communication with various stakeholders (e.g., students, instructors, policy-makers). Given that local language tests are embedded within the local context, be it a specific TLU domain or

curriculum, tasks can be developed to reflect the communicative characteristics of language use (Dimova, 2020a, 2021). For instance, relevant language varieties, norms, and pragmatic features of language use can be introduced in tasks because they are easier to capture when the context of language (and test result) use is delimited (e.g., tasks could include instances of codeswitching if that is a local norm and can be identified in the transcription of the TLU domain). From a pragmatic perspective, tests can be developed and adjusted over time to keep up with the evolving context and the changing needs of test-takers and score users.

The development, implementation, use, and validation of a rating scale could be conceptualized as an ongoing process that also offers various opportunities. Instructors, who may be potential raters, can be involved in scale design from its inception. This involvement allows for establishment of performance descriptors that are meaningful for the raters, and, therefore, contribute to a more reliable rating process once the scale is operational. Instructors' engagement with task and scale development and rating strengthens their language assessment literacy. Moreover, scale development in a local context has the possibility to incorporate hybrid methods, making use of intuitive, measurement-driven, and data-driven support.

Finally, involvement of potential raters (e.g., instructors) in scale development leads to establishment of a rater community of practice. Through iterative discussions during rater-training sessions and post-admin deliberations, instructors and testers align the conceptualization and operationalization of language ability and standardize their rating procedures and approaches. Therefore, local contexts can represent a fertile soil for experimentation and innovation through collaborative endeavor.

References

Al-Gahtani, S., & Roever, C. (2018). Proficiency and preference organization in second language refusals. *Journal of Pragmatics, 129*, 140-153. https://doi.org/10.1016/j.pragma.2018.01.014

Al-Gahtani, S., & Roever, C. (2012). Proficiency and sequential organization of L2 requests. *Applied Linguistics*, *33*(1), 42-65. https://doi.org/10.1093/applin/amr031

Bachman, L. F., & Palmer, A. S. (1996). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford University Press.

Bloxham, S., Hughes, C., & Adie, L. (2016) What's the point of moderation? A discussion of the purposes achieved through contemporary moderation practices. *Assessment & Evaluation in Higher Education, 41(*4), 638-653. https://doi.org/10.1080/026029.38.2015.1039932

Clark, J. L. D. (1979). Direct vs. semi-direct tests of speaking ability. In E. J. Briere & F. B. Hinofotis (Eds.), *Concepts in language testing: Some recent studies* (pp. 3-49). TESOL.

Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge University Press. https://rm.coe.int/16802fc1bf

Council of Europe. (2020). *Common European Framework of Reference for Languages: Learning, teaching, assessment – Companion volume.* Council of Europe Publishing. www.coe.int/lang-cefr

Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T., & McNamara, T. (1999). *Dictionary of language testing*. Cambridge University Press.

Davis, L. (2016). The influence of training and experience on rater performance in scoring spoken language. *Language Testing, 33*(1), 117-135. https://doi.org/10.1177/0265532215582282

Dimova, S. (2020a). Language Assessment of EMI Content Teachers: What Norms. In M. Kuteeva, K. Kaufhold & N. Hynninen (Eds.), *Language perceptions and practices in multilingual universities* (pp. 351-378). Springer. https://doi.org/10.1007/978-3-030-38755-6_14

Dimova, S. (2020b). The role of feedback in the design of a testing model for social justice. *Journal of Contemporary Philology, 3*(1), 21-34.

Dimova, S. (2020c). English language requirements for enrolment in EMI programs in higher education: A European case. *Journal of English for Academic Purposes, 47,* 100896. https://doi.org/10.1016/j.jeap.2020.100896

Dimova, S. (2021). Certifying lecturers' English language skills for teaching in English-medium programs in higher education. *ASp, 79*(1), 29-47. https://doi.org/10.4000/asp.7056

Dimova, S., & Kling, J. (2018). Assessing EMI lecturer language proficiency across disciplines. *TESOL Quarterly, 52*(3), 634-656. https://doi.org/10.1002/tesq.454

Dimova, S., & Kling, J. (2020). Current Considerations on Integrating Content and Language in Multilingual Universities. In S. Dimova & J. Kling (Eds.), *Integrating Content and Language in Multilingual Universities* (pp. 1-12). Springer. https://doi.org/10.1007/978-3-030-46947-4_1

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

Ducasse, A. M., & Brown, A. (2009). Assessing paired orals: Raters' orientation to interaction. *Language Testing*, *26*(3), 423-443. https://doi.org/10.1177/0265532209104669

Elder, C., Barkhuizen, G., Knoch, U., & von Randow, J. (2007). Evaluating rater responses to an online rater training program. *Language Testing, 24*(1), 37-64. https://doi.org/10.1177/0265532207071511

Elder, C., Knoch, U., Barkhuizen, G., & von 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

Fulcher, G. (2018). Assessing spoken production. In J. I. Liontas & M. DelliCarpini (Eds.), *The TESOL encyclopedia of English language teaching* (pp. 1-6). Wiley. https://doi.org/10.1002/9781118784235

Fulcher, G. (2015). *Re-examining language testing: A philosophical and social inquiry*. Routledge. https://doi.org/10.4324/9781315695518

Fulcher, G. (2012). Scoring performance tests. In G. Fulcher &F. Davidson (Eds.), *The Routledge handbook of language testing* (pp. 392-406). Routledge. https://doi.org/10.4324/9780203181287

Fulcher, G. (2003). *Testing second language speaking.* London: Longman/Pearson. https://doi.org/10.4324/9781315837376

Fulcher, G. (2000). The 'communicative'legacy in language testing. *System, 28*(4), 483-497. https://doi.org/10.1016/S0346-251X(00)00033-6

Fulcher, G. (1997). An English language placement test: issues in reliability and validity. *Language testing, 14*(2), 113-139. https://doi.org/10.1177/026553229701400201

Fulcher, G. (1996). Does thick description lead to smart tests? A data-based approach to rating scale construction. *Language Testing, 13*(2), 208-238. https://doi.org/10.1177/026553229601300205

Fulcher, G. (1993). *The construction and validation of rating scales for oral tests in English as a Foreign Language.* Unpublished PhD thesis, University of Lancaster, UK.

Fulcher, G. (1987). Tests of oral performance: The need for data-based criteria. *English Language Teaching Journal, 41*(4), 287-291. https://doi.org/10.1093/elt/41.4.287

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

Fulcher, G., & Reiter, R. M. (2003). Task difficulty in speaking tests. *Language testing, 20*(3), 321-344. https://doi.org/10.1191/0265532203lt259oa

Ginther, A. (2012). Assessment of speaking. In C. Chapelle (Ed.), *The Encyclopedia of applied linguistics* (pp. 1-8). Wiley-Blackwell. https://doi.org/10.1002/9781405198431.wbeal0052.pub2

Ginther, A., Redden, J., & Kauper, N. (2012). *OEPP 2012 Yearbok.* Purdue University. Retrieved from https://www.purdue.edu/oepp/documents/OEPP-Yearbook-2012.pdf

Hughes, A. (2003). *Testing for language teachers*. Cambridge University Press.

Ohta, R., Plakans, L. M., & Gebril, A. (2018). Integrated writing scores based on holistic and multi-trait scales: A generalizability analysis. *Assessing Writing, 38,* 21-36. https://doi.org/10.1016/j.asw.2018.08.001

Qian, D. D. (2009) Comparing Direct and Semi-Direct Modes for Speaking Assessment: Affective Effects on Test Takers. *Language Assessment Quarterly, 6*(2), 113-125. https://doi.org/10.1080/15434300902800059

Kasper, G., & Rose, K. R. (2002). Pragmatic development in a second language. *Language Learning, 52*(1), 1-352.

Kim, Y. H. (2009). An investigation into native and non-native teachers' judgments of oral English performance: a mixed methods approach. *Language Testing, 26*, 187-217. https://doi.org/10.1177/0265532208101010

Kling, J., & Dimova, S. (2015). The Test of Oral English for Academic Staff (TOEPAS): Validation of standards and scoring procedures. In A. Knapp & K. Aguado (Eds.), *Fremdsprachen in Studium und Lehre—Chancen und Herausforderungen für den Wissenserwerb* (pp. 247-268). Peter Lang. https://www.peterlang.com/document/1048011

Kling, J., & Stæhr, L. S. (2011). Certificering af universitetsunderviseres engelsksproglige kompetencer. [Certification of university lecturers for English proficiency]. *Sprogforum. Tidsskrift for sprog-og kulturpædagogik*, *17*(52). https://tidsskrift.dk/spr/article/view/102829

Knoch, U. (2011). Investigating the effectiveness of individualized feedback to rating behavior—a longitudinal study. *Language Testing, 28*(2), 179-200. https://doi.org/10.1177/0265532210384252

Knoch, U. (2009). *Diagnostic writing assessment: The development and validation of a rating scale (Vol. 17)*. Peter Lang. https://www.peterlang.com/document/1054114

Knoch, U., Fairbairn, J., & Huisman, A. (2016). An evaluation of an online rater training program for the speaking and writing sub-tests of the Aptis test. *Papers in language testing and assessment, 5*(1,2), 90-106. https://minerva-access.unimelb.edu.au/bitstream/handle/11343/115276/5_knoch_et_al.pdf

Kuiken, F., & Vedder, I. (2020). Scoring Approaches: Scales/Rubrics. In P. Winke & T. Brunfaut (Eds.), *The Routledge handbook of second language acquisition and language testing* (pp. 125-134). Routledge. https://doi.org/10.4324/9781351034784

Linn, R. L. (1993) Linking results of distinct assessments. *Applied Measurement in Education 6*(1), 83-102. https://doi.org/10.1207/s15324818ame0601_5

Lumley, T., & McNamara, T. (1995) Rater characteristics and rater bias: Implications for training. *Language Testing, 12*(1), 54-71. https://doi.org/10.1177/026553229501200104

Luoma, S. (2004). *Assessing speaking*. Cambridge University Press. https://doi.org/10.1017/CBO9780511733017

May, L. (2011). Interactional competence in a paired speaking test: Features salient to raters. *Language Assessment Quarterly*, *8*(2), 127-145. https://doi.org/10.1080/15434303.2011.565845

May, L. (2009). Co-constructed interaction in a paired speaking test: The rater's perspective. *Language Testing*, *26*(3), 397-421. https://doi.org/10.1177/0265532209104668

Nakatsuhara, F., Inoue, C., & Khabbazbashi, N. (2020). Measuring L2 speaking. In P. Winke & T. Brunfaut (Eds.), *The Routledge handbook of second language acquisition and language testing* (pp. 285-294). Routledge. https://doi.org/10.4324/9781351034784

Okada, Y. (2010). Role-play in oral proficiency interviews: Interactive footing and interactional competencies. *Journal of Pragmatics*, *42*(6), 1647-1668. https://doi.org/10.1016/j.pragma.2009.11.002

Okada, Y., & Greer, T. (2013). Pursuing a relevant response in oral proficiency interview role plays. In S.J. Ross & G. Kasper (Eds.), *Assessing second language pragmatics* (pp. 288-310). Palgrave Macmillan. https://doi.org/10.1057/9781137003522

O'Sullivan, B. (2000). Exploring gender and oral proficiency interview performance. *System, 28*(3), 373-386. https://doi.org/10.1016/S0346-251X(00)00018-X

Roever, C. (2011). Testing of second language pragmatics: Past and future. *Language Testing, 28*(4), 463-481. https://doi.org/10.1177/0265532210394633

Sadler, D. R. (2013). Assuring Academic Achievement Standards: From Moderation to Calibration. *Assessment in Education: Principles, Policy and Practice, 20*(1), 5-19. https://doi.org/10.1080/0969594X.2012.714742

Sandlund, E., & Sundqvist, P. (2021). Rating and reflecting: Displaying rater identities in collegial L2 English oral assessment. In R. Salaberry & A. R. Burch (Eds.), *Assessing speaking in context. Expanding the construct and its applications* (pp. 132-161). Multilingual Matters. https://doi.org/10.21832/9781788923828

Seedhouse, P., & Nakatsuhara, F. (2018). *The discourse of the IELTS speaking test: Interactional design and practice*. Cambridge University Press.

Stokoe, E. (2013). The (in) authenticity of simulated talk: Comparing role-played and actual interaction and the implications for communication training. *Research on Language & Social Interaction*, *46*(2), 165-185. https://doi.org/10.1080/08351813.2013.780341

Sundqvist, P., Sandlund, E., Skar, G. B., & Tengberg, M. (2020). Effects of rater training on the assessment of L2 English oral proficiency. *Nordic Journal of Modern Language Methodology, 8*(1), 3-29. https://doi.org/10.46364/njmlm.v8i1.605

Taguchi, N. (2018). Data collection and analysis in developmental L2 pragmatics research: Discourse completion test, role-play, and naturalistic recording. In A. Gudmestad&A. Edmonds (Eds.), *Critical reflections on data in second language acquisition* (pp. 7-32). John Benjamins. https://doi.org/10.1075/lllt.51.02tag

Taguchi, N., & Roever, C. (2017). *Second language pragmatics*. Oxford University Press.

Taylor, L., & Wigglesworth, G. (2009). Are two heads better than one? Pair work in L2 assessment contexts. *Language Testing, 26*, 325-339. https://doi.org/10.1177/0265532209104665

Turner, C. E., & Upshur, J. (2002). Rating scales derived from student samples: Effects of the scale marker and the student sample on scale content and student scores. *TESOL Quarterly, 36*(1), 49-70. https://doi.org/10.2307/3588360

Turner, C. E. (2000). Listening to the voices of rating scale developers: Identifying salient features for second language performance assessment. *Canadian Modern Language Review, 56*(4), 555-584. https://doi.org/10.3138/cmlr.56.4.555

Upshur, J., & Turner, C. E. (1999). Systematic effects in the rating of second-language speaking ability: Test method and learner discourse.' *Language Testing, 16*(1), 82-111. https://doi.org/10.1177/026553229901600105

Upshur, J., & Turner, C.E. (1995). Constructing rating scales for second language tests. *English Language Teaching Journal, 49*(1), 3-12. https://doi.org/10.1093/elt/49.1.3

Wei, J., & Llosa, L. (2015). Investigating differences between American and Indian raters in assessing TOEFL iBT speaking tasks. *Language Assessment Quarterly, 12,* 283-304. https://doi.org/10.1080/15434303.2015.1037446

Weigle, S. C. (1998) Using FACETS to model rater training effects. *Language Testing, 15*(2), 263-287. https://doi.org/10.1177/026553229801500205

Winke, P., Gass, S., & Myford, C. (2013). Raters' L2 background as a potential source of bias in rating oral performance. *Language Testing, 30*, 231-252. https://doi.org/10.1177/0265532212456968

Xi, X., Norris, J. M., Ockey, G. J., Fulcher, G., & Purpura, J. E. (2021). Assessing Academic Speaking. In X. Xi&J.M. Norris (Eds.), *Assessing academic English for higher education admissions* (pp. 152-199). Routledge. https://doi.org/10.1111/ijal.12405

Yan, X. (2014). An examination of rater performance on a local oral English proficiency test: a mixed-methods approach. *Language Testing, 31,* 501-527. https://doi.org/10.1177/0265532214536171

Youn, S. J. (2020). Managing proposal sequences in role-play assessment: Validity evidence of interactional competence across levels. *Language Testing*, *37*(1), 76-106. https://doi.org/10.1177/0265532219860077

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,* 31-50. https://doi.org/10.1177/0265532209360671

Acknowledgements

Not applicable.

Funding

Not applicable.

Ethics Declarations

Competing Interests

No, there are no conflicting interests.