

Recent Empirical Studies of Paired Speaking Assessment: A General Survey

Yunhui Gou

Graduate School of Education, University of Pennsylvania, Philadelphia, Pennsylvania, 19104, USA

Abstract: This literature review surveys empirical research on paired speaking assessment published between April 2012 and April 2022, following a narrative review procedure. Seventeen empirical studies retrieved from the University of Pennsylvania Franklin Library were analyzed and organized into two thematic groups: sources of variability in paired assessment (interlocutor effects, rater effects, and planning time) and paired assessment in use (rating scales, practicality, and the interactional skills test-takers display). The review finds that interlocutor proficiency does not consistently predict overall test scores but is closely associated with asymmetric interactional patterns, that raters attend to both verbal and non-verbal co-constructed features when scoring, and that planning time has a limited effect on holistic scores while still shaping discourse features. The review also highlights an ongoing gap between the theoretical construct of interactional competence and its practical operationalization in classroom and large-scale testing contexts. Implications for future rating-scale development and rater training are discussed.

Keywords: Paired Speaking Assessment; Interactional Competence; Language Testing; Rater Cognition; Interlocutor Proficiency; Literature Review.

1. Introduction

Some language testers argue that speaking is the most challenging aspect to assess among all four language skills due to the requirements for eliciting speech, rating scales, and well-trained raters [1]. Given the interactive and context dependent nature, language test developers need to consider multiple perspectives beyond just the interlocutor's linguistic competences while designing assessments. While traditional individual speaking assessment still takes the majority of the speaking assessment market, paired speaking assessment has gained more attention as it allows test-takers to interact with peers and demonstrate competencies that are hard to observe in traditional individual speaking assessments. This literature review aims to provide a general idea of what topics researchers have focused on from April 2012 to April 2022 regarding paired speaking assessment.

2. Foundational Theoretical Background on Paired Speaking Assessment

2.1. Interactional Competence and Communicative Competence

Paired speaking assessment connects closely to the concept of interactional competence (IC). Interactional competence differs from communicative competence by its sociocultural dimension as it refers to the ability to jointly co-construct meaning in a shared interactional context [2]. Communicative competence was first brought up by Hymes [3], who defines communicative competence as “grammatical competence” and “sociolinguistic competence”. Through researchers' [4,5,6] discussion and evolution about what should be included in communicative competence, the definition expands and changes to organizational competence [4,5], pragmatic competence [4,5], and strategic competence [6]. Bachman and Palmer's [6] model of communicative competence has been widely influential in language

assessment.

While this model has been influential in language assessment, later researchers argued that it did not fully capture the sociocultural nature of language use, which should involve “spontaneous and vivid interactions” [7]. Many researchers also suggest that language proficiency is not a global measure of general competence but language use in a specific local context [8,9]. In this context, the idea of interactional competence [2] provides a framework to view language competence through interlocutor interactions where co-construction happens. Interactional competence refers to the ability to participate in socially situated interaction by drawing on linguistic and interactional resources with one or more interlocutors to accomplish a particular discursive practice. In this view, competence is not treated only as an individual possession; it is also demonstrated through the ways participants coordinate their contributions, establish mutual understanding, manage turns, develop topics, and respond to one another in context [8,9]. This concept builds the foundation for the growing attention to paired speaking assessments, as they assess test-takers based on co-constructed interaction rather than the isolated production of one test-taker.

Paired and group speaking assessment is an alternative arrangement that has been discussed and used in both high-stake and low-stake test contexts. The discussion has continued since the University of Cambridge Local Examinations Syndicate (UCLES) adopted this exam arrangement for the First Certificate in English (FCE) exam, in which two test-takers are grouped to perform speaking tasks while two examiners score each individual's language performance. Another high-stake test that adopts paired speaking assessment is the Examination for the Certificate of Proficiency in English (ECPE), which measures English proficiency in academic and professional contexts and uses a similar format to UCLES. Paired assessments therefore provide a way to examine speaking performance in interaction. McNamara [10] argues that paired speaking assessment filled the gap left by individual speaking

assessment, which “focuses too much on the individual candidate rather than the candidate in interaction” (p. 86).

However, since the format was first proposed by Folland and Robertson [11], paired speaking assessment has remained in a minor role in the speaking assessment market compared with the dominant arrangement of individual assessment. High-stake and even many low-stake tests use individual speaking assessments because the drawbacks of paired speaking assessment are as salient as its merits. An earlier researcher [12] was concerned about the lack of empirical data to support the benefits of paired speaking assessment. Although later empirical research filled this gap, other debates continued relating to paired speaking assessment. The central debate shifted from whether paired assessment should be used to more specific questions about validity, variability, and how interactional performance can be interpreted and scored.

The validity of paired speaking assessment has been a prolific area of research. Researchers have focused on two major aspects: the rater and the interlocutor. In paired speaking assessment, the rater is the person who evaluates the performance, whereas the interlocutor is the other participant with whom a test-taker must interact. Research on raters examined how they interpret co-constructed interaction and what they attend to when assigning scores [13]. May [14] also focused on raters and highlighted how raters perceive mutual achievements in paired assessment settings. Other researchers [15] examined variability, finding that new raters are inconsistent in rating group language assessments.

Another aspect contributing to variability is the interlocutor. Over the past two decades, researchers have demonstrated that the personal characteristics of interlocutors can influence how test-takers perform. More specifically, researchers found that gender [16,17], acquaintanceship [18,17], and personality [19] affect accuracy in paired speaking assessment. Beyond personal characteristics and relationships, other researchers have specifically focused on interlocutors' language proficiency levels and their impact on paired speaking assessment [20,21].

3. Empirical Research from 2012 to 2022

Discussion of paired speaking validity has continued in recent years as researchers have expanded their focus to specific events during paired speaking assessments. April 2012 to April 2022 was used as the review period because it represents the most recent ten-year period immediately preceding the completion of this paper. The 2012 cutoff is chronological rather than theoretical: earlier studies provide the conceptual and empirical background, while studies published during 2012–2022 constitute the primary evidence reviewed here. Following the narrative review procedure [22], the keywords “paired speaking assessment”, “paired oral assessment”, “paired speaking tests”, and “dyadic speaking test” were used under the disciplines of “education” and “languages & literature” to search the University of Pennsylvania Franklin Library. The initial search identified 88 records. After screening for topic and format, 21 records met the broad eligibility criteria. Within the 21 records, 17 empirical studies were retained for analysis and 4 articles were excluded because they did not match the topic or were not empirical studies. For example, Ahmadi [23] defined paired speaking assessment as the interaction between

examiners and test-takers, while this review focuses only on peer-peer interactional speaking assessment; Nyroos [24] and Hüttner [25] focus on a different topic within SLD but not on paired speaking assessment; and Galaczi and Taylor [26] provide a historical analysis rather than empirical research. The search strategy was necessarily limited to the databases available, so this review should be understood as a focused survey rather than an exhaustive systematic review.

Most of the selected work is from major, influential journals, including but not limited to *Language Assessment Quarterly*, *Applied Linguistics*, and *Language Testing*. This diversity of trusted, reputable journals provides readers with a general survey of what researchers have done in recent years. In this literature review, the following questions are addressed: (1) What topics are researchers focusing on in recent research? (2) What conclusions have been drawn?

3.1. Coding and Analysis

The 17 studies were coded into six topics and organized into two broad groups. The first group concerned sources of variability in paired assessment, including planning time ($n=3$), raters' impact ($n=3$), and interlocutor impact ($n=5$); the second group focused on paired assessment in use, including rating scale ($n=1$), implications for future practice ($n=3$), and interactional skills for co-construction (i.e., turn-taking, etc.) ($n=2$).

4. Paired Speaking Assessment Variability

4.1. Interlocutor-Related Issues

Interlocutors' impact on paired speaking assessment still attracts researchers' attention, especially interlocutors' proficiency level. Bennett [27] approached this question differently from many earlier studies, studying 43 Italian participants preparing for a Cambridge paired oral examination and collecting paired oral test data together with pre- and post-test questionnaires about participants' perceptions of partner proficiency. Rather than using a conventional pre-test/post-test design to measure changes in language ability, the study examined whether variation in the linguistic ability of the paired partner was associated with participants' performance and perceptions. Bennett reported that most candidates believed their scores could be affected by their partner, but the test evidence did not support a systematic effect of partner linguistic ability on scores. Bennett suggests that interlocutors' proficiency level has little or no impact on test scores, which contradicts earlier work [28].

Isaacs [29] reached a different conclusion when the focus shifted from composite scores to interactional patterns. In a study of 84 international engineering graduate students completing a paired speaking task, interaction was examined in relation to proficiency and other participant characteristics. The analysis showed that unevenly matched pairs were more likely to display asymmetric interactional patterns, in which one participant became more dominant and the other more passive. Thus, partner proficiency appeared to matter for how the interaction was jointly constructed even when the relationship between proficiency and an overall speaking score was less straightforward. Such an asymmetric dynamic could lead to a less positive “interactional outcome” (p. 23). Galaczi [30] similarly demonstrated that paired speaking interactions can be characterized by different patterns of co-

construction, making the structure of the interaction itself an important object of analysis.

The discrepancy between these findings may reflect differences in research questions, task types, and measurements rather than a simple contradiction. Bennett [27] examined whether partner ability produced a systematic effect on performance and scores, whereas Isaacs [29] examined the interactional consequences of proficiency differences. Moreover, Hu's [31] empirical research on role-play speaking assessment showed that matched-proficiency pairs are more likely to have overlapping interactional resources that "successfully demonstrate their interactional competence" (p. 484). Hu argued that interlocutor effects may become more visible in role-play tasks than in discussion tasks. In conclusion, these studies show that interlocutors' influence on paired speaking exams cannot be ignored, even though it does not show a consistent effect on global test scores.

This distinction is also shown in Artunç and Hart's [32] study: when comparing low-low, low-high, and high-high pairs, they found differences in turn-taking, topic management, repair, and task management across pairing conditions.

Moving away from the proficiency level of interlocutors, Ockey and Chukharev-Hudilainen [33] used quantitative methods to compare test-takers' scores between a human interlocutor and a computer interlocutor. Their results showed similar scores for fluency, pronunciation, and grammar/vocabulary, while interactional competence scores were higher in the human-partner condition. However, the computer partner produced more dependable ratings for interactional competence and was perceived by raters as more standardized. This study demonstrates that the design of the interlocutor can affect both the interaction and the dependability of assessment within a paired format.

In short, although there are mixed findings on the relationship between interlocutors' proficiency levels and test scores, recent studies have shifted focus to the interactional pattern between interlocutors. The more consistent finding is that proficiency differences can shape interactional patterns through turn-taking, support, topic development, and task management.

4.2. Rater-Related Issues

Rater-related research has focused on how raters' judgement contributes to variability in paired speaking assessment. Grubor [34] found inter-rater reliability on two test-takers' ratings of paired speaking assessment using Pearson's Product Moment correlation and Cronbach's alpha. This finding contributes to the discussion on paired assessment rating consistency. However, other work suggests that raters may attend to interactional features that do not always fully represent the grading scales.

Borger [35] found that topic development moves, turn-taking management, and interactive listening strategies are three primary "interactional resources" that raters pay attention to. Vo [36] suggests that raters attend to non-verbal interactional features such as hand gestures, nods, or eye contact. Although raters also report these features in individual speaking assessments, the reported frequency is higher in paired speaking assessments.

These studies show that during paired speaking assessment, raters focus not only on what test-takers say, but also on how test-takers respond to and co-construct the interaction, both

verbally and non-verbally. This finding suggests that raters of paired speaking assessment need more well-established training and a clearer grading rubric. Bonk and Ockey's [15] work found rater severity differences, especially among new raters, whereas returning raters were more consistent in severity. This work shows a gap in assessment practice regarding how to optimize the interactional competence rubric and criteria in paired speaking assessments.

4.3. Planning Time and Task Instruction

Another source of variability is the planning condition before a speaking task. Sandlund and Sundqvist [37] showed through conversation analysis how pre-task instruction can shape how test-takers understand the topic, and how instruction can shape the direction of the conversation even when the exam is standardized. Wang and Song [38], using four strategic-planning conditions with 72 Chinese EFL learners, found greater fluency and syntactic complexity under some planned conditions but no interaction between planning time and proficiency. Nitta and Nakatsuhara [39] also investigated planning in paired oral tests, but their findings were more nuanced: rating scores and discourse measures suggested limited effects of planning, whereas conversation analysis identified differences in the interactional mode under planned and unplanned conditions.

The apparent discrepancy between Wang and Song [38] and Nitta and Nakatsuhara [39] is therefore partly explained by what counts as "performance." Planning time may change particular discourse features, such as syntactic complexity or fluency, without producing a large change in overall ratings. It may also alter the way participants co-construct the interaction even when score differences are small.

To conclude, researchers remain interested in factors that affect test-takers' performance in paired speaking assessment, particularly from the interlocutor, rater, and planning-time perspectives. Partner proficiency can shape interactional patterns, raters can attend to interactional and non-verbal features differently, and planning or task instruction can influence the discourse that test-takers produce.

5. Paired Assessment in Use

While most researchers have reached agreement on the factors that affect testing scores, researchers remain in dispute over the use and implications of paired speaking assessment. Accordingly, researchers focus on a diverse set of topics relating to use and implication, including rating scale, test practicality, usefulness, and the interactional patterns test-takers show during paired speaking assessment.

Paired speaking assessment is often used to measure test-takers' interactional competence. However, in putting theory into practice, test developers question how interactional competence can be measured. Wang [40] proposed a modified CAP scale to measure interactional competence quality. Ramazani [41] used exploratory factor analysis to test construct validity and found that the CAP rating scale showed "a reasonable reliability" (p. 180), suggesting that certain functions measure interactional competence: "building argument, developing discussion, offering support and shaping connection" (p. 200). Similarly, May et al. [42] proposed a detailed checklist for classroom assessment of interactional competencies. These checklists and rating scales provide more opportunities for paired arrangements in future assessments, whether high-stake or low-stake.

However, contradicting prior empirical research on the

benefits of paired speaking assessment [43] and the consistency of inter-rater reliability [34], Becker et al. [44] found low practicality [45] and insufficient inter-rater reliability for paired speaking assessments in an Intensive English Program placement assessment. They suggest that paired speaking assessment may work better for formative classroom assessments than for large-scale summative assessments. East's [46] work reveals a similar concern about feasibility and fairness of scoring, while also finding that teachers perceive paired speaking assessment in classrooms relatively positively and consider it more "useful" (p. 116), indicating a good fit for the test purpose [6] in classroom contexts.

These studies show that while researchers are actively putting theory into practice and finding ways to measure interactional competence, fieldwork analysis reveals a gap between what paired speaking assessment intends to achieve and how it works in real-life situations, especially in high-stake summative assessments. Other researchers have shifted focus to the interactional skills test-takers use in paired speaking assessments, given the belief that such assessments can elicit interactional competence.

Galaczi [47] analyzed test-takers' interaction and found that test-takers use three categories of features to co-construct interaction in paired speaking assessments: "topical development, listener support, and turn-taking management" (p. 559). This typology expands Ducasse and Brown's [13] findings on the management skills test-takers use to co-construct interaction, and sheds light on directions for future study. Butler and Zeng [48] analyzed differences in interactional patterns between different age-group pairs, finding that lower-grade pairs showed "formulaic and fixed utterance patterns" (p. 67), implying that turn allocation is an advanced interactional skill. Both studies aim to expand the range of interactional functions and features considered.

6. Conclusion and Discussion

After reviewing 17 empirical studies on paired speaking assessment from various journals, it can be concluded that interlocutors' effect on paired speaking assessment remains a central focus. Findings are mixed on whether interlocutors' proficiency affects test-takers' scores; still, most researchers agree that a mismatch in proficiency level is more likely to cause an asymmetric interactional pattern. Researchers have also studied other factors contributing to variability in paired speaking assessment, including raters' perceptions and planning. Studies of rater cognition indicate that raters attend to verbal and non-verbal interactional features and may interpret some aspects of performance as mutual achievements. Other findings show that planning time has a limited impact on test-taker performance, but pre-task instruction and interaction may play an important role.

Research on assessment in use highlights a related challenge: the construct of interactional competence is easier to defend theoretically than to operationalize consistently across testing contexts. Researchers have found conflicting results on inter-rater reliability between laboratory studies and real-life situations, indicating a gap in the practicality of paired oral assessment from theory to practice, especially in formative assessment contexts. Other studies focus on what interactional skills should be included and measured, and what patterns test-takers demonstrate during paired speaking assessments.

This literature review has several limitations. The search

was limited to the University of Pennsylvania Franklin Library and a specific set of keywords, so some relevant studies may not have been identified. Future reviews could expand the dataset to include more empirical research and take a closer look at the research methodology of the studies reviewed.

References

- [1] Ginther, A. (2012). Assessment of speaking. *The Encyclopedia of Applied Linguistics*.
<https://doi.org/10.1002/9781405198431.wbeal0068>
- [2] Kramsch, C. (1986). From language proficiency to interactional competence. *The Modern Language Journal*, 70(4), 366–372.
<https://doi.org/10.1111/j.1540-4781.1986.tb05249.x>
- [3] Hymes, D. (1972). On communicative competence. In J. B. Pride & J. Holmes (Eds.), *Sociolinguistics* (pp. 269–293). Penguin.
- [4] Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, 1(1), 1–47.
<https://doi.org/10.1093/applin/1.1.1>
- [5] Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford University Press.
- [6] Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests* (Vol. 1). Oxford University Press.
- [7] Sun, D. (2014). From communicative competence to interactional competence: A new outlook on the teaching of spoken English. *Journal of Language Teaching & Research*, 5(5). <https://doi.org/10.4304/jltr.5.5.24>.
- [8] He, A. W., & Young, R. (1998). Language proficiency interviews: A discourse approach. In R. Young & A. W. He (Eds.), *Talking and testing: Discourse approaches to the assessment of oral proficiency* (Vol. 14, pp. 1–24). John Benjamins.
- [9] Young, R. F. (2011). Interactional competence in language learning, teaching, and testing. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 426–443). Routledge.
- [10] McNamara, T. F. (1996). *Measuring second language performance*. Addison-Wesley Longman.
- [11] Folland, D., & Robertson, D. (1976). Towards objectivity in group oral testing. *English Language Teaching Journal*, 30, 156–167. <https://doi.org/10.1093/elt/XXX.2.156>.
- [12] Foot, M. C. (1999). *Relaxing in pairs*. Cambridge ESOL.
- [13] Ducasse, A. M., & Brown, A. (2009). Assessing paired orals: Raters' orientation to interaction. *Language Testing*, 26(3), 423–443. <https://doi.org/10.1177/0265532209104668>.
- [14] May, L. (2009, 2011). [Reference details not present in the original manuscript — please supply the full citation for this source.].
- [15] 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>
- [16] Brown, A., & McNamara, T. (2004). The devil is in the detail: Researching gender issues in language assessment. *TESOL Quarterly*, 38(3), 524–538. <https://doi.org/10.2307/3588349>.

- [17] Norton, J. (2005). The paired format in the Cambridge speaking tests. *ELT Journal*, 59(4), 287–297.
<https://doi.org/10.1093/elt/cci054>
- [18] O'Sullivan, B. (2002). Learner acquaintanceship and oral proficiency test pair-task performance. *Language Testing*, 19, 277–295. <https://doi.org/10.1191/0265532202lt206oa>.
- [19] Berry, V. (2007). Personality differences and oral test performance. Peter Lang.
- [20] Davis, L. (2009). The influence of interlocutor proficiency in a paired oral assessment. *Language Testing*, 26(3), 367–396. <https://doi.org/10.1177/0265532209104667>.
- [21] Nakatsuhara, F. (2006). The impact of proficiency level on conversational styles in paired speaking tests. *Cambridge ESOL Research Notes*, 25, 15–20.
- [22] Cooper, H., Hedges, L. V., & Valentine, J. C. (2019). *The handbook of research synthesis and meta-analysis*. Russell Sage Foundation.
- [23] Ahmadi, A. (2016). Assessing English language learners' oral performance: A comparison of monologue, interview, and group oral test. *Language Assessment Quarterly*, 13(4), 341–358. <https://doi.org/10.1080/15434303.2016.1216472>.
- [24] Nyroos, L., Sandlund, E., & Sundqvist, P. (2017). Code-switched repair initiation: The case of Swedish *eller* in L2 English test interaction. *Journal of Pragmatics*, 120, 1–16. <https://doi.org/10.1016/j.pragma.2017.08.008>.
- [25] Hüttner, J. (2014). Agreeing to disagree: 'Doing disagreement' in assessed oral L2 interactions. *Classroom Discourse*, 5(2), 194–215. <https://doi.org/10.1080/19463014.2014.912174>.
- [26] Galaczi, E., & Taylor, L. (2018). Interactional competence: Conceptualisations, operationalisations, and outstanding questions. *Language Assessment Quarterly*, 15(3), 219–236. <https://doi.org/10.1080/15434303.2018.1472776>.
- [27] Bennett, R. (2012). [Reference details not present in the original manuscript — please supply the full citation for this source.]
- [28] Iwashita, N. (1999). The validity of the paired interview format in oral performance assessment. *Melbourne Papers in Language Testing*, 5(2), 51–66.
- [29] Isaacs, T. (2013). International engineering graduate students' interactional patterns on a paired speaking test: Interlocutors' perspectives. In K. McDonough & A. Mackey (Eds.), *Second language interaction in diverse educational settings* (pp. 227–246). John Benjamins.
- [30] Galaczi, E. D. (2008). [Reference details not present in the original manuscript — please supply the full citation for this source.]
- [31] Hu, J. (2015). Interaction in assessment-oriented role play: A conversation analytic approach. *Chinese Journal of Applied Linguistics*, 38(4), 472–489. <https://doi.org/10.1515/cjal-2015-0030>.
- [32] Artunç, E. K., & Hart, D. O. (2020). Interactional competence in paired speaking tests: A study on proficiency-based pairings. *System*, 89, 102194.
<https://doi.org/10.1016/j.system.2020.102194>
- [33] Ockey, G. J., & Chukharev-Hudilainen, E. (2021). Human versus computer partner in the paired oral discussion test. *Applied Linguistics*, 42(5), 924–944.
<https://doi.org/10.1093/applin/amab012>
- [34] Grubor, J. (2013). The construct of perceived L2 speaking proficiency in a paired testing format. *Papers and Studies in Contrastive Linguistics*, 49(2), 185–203.
<https://doi.org/10.2478/pscl-2013-0010>
- [35] Borger, L. (2019). Assessing interactional skills in a paired speaking test: Raters' interpretation of the construct. *Apples – Journal of Applied Language Studies*, 13(1), 151–174. <https://doi.org/10.17011/apples/urn.nbn:se:su:diva-167453>.
- [36] Vo, S. (2021). Evaluating interactional competence in interview and paired discussion tasks: A rater cognition study. *TESOL Journal*, 12(2). <https://doi.org/10.1002/tesj.573>.
- [37] Sandlund, E., & Sundqvist, P. (2013). [Reference details not present in the original manuscript — please supply the full citation for this source.]
- [38] Wang, H., & Song, F. (2015). Effects of strategic planning time on L2 paired oral test performance. *Chinese Journal of Applied Linguistics*, 38, 263–276. <https://doi.org/10.1515/cjal-2015-0017>.
- [39] Nitta, R., & Nakatsuhara, F. (2014). A multifaceted approach to investigating pre-task planning effects on paired oral test performance. *Language Testing*, 31(2), 147–175.
<https://doi.org/10.1177/0265532213514420>
- [40] Wang, L. (2015). Assessing interactional competence in second language paired speaking tasks (Unpublished doctoral dissertation). Northern Arizona University.
- [41] Ramazani, M., Behnam, B., & Ahangari, S. (2018). Psychometric characteristics of a rating scale for assessing interactional competence in paired-speaking tasks at micro-level. *Journal of English Language Pedagogy and Practice*, 11(23), 180–206.
- [42] May, L., Nakatsuhara, F., Lam, D., & Galaczi, E. (2020). Developing tools for learning oriented assessment of interactional competence: Bridging theory and practice. *Language Testing*, 37(2), 165–188.
<https://doi.org/10.1177/0265532219898092>
- [43] Ffrench, A. (2003). The change process at the paper level. Paper 5, speaking. In C. Weir & M. Milanovic (Eds.), *Continuity and innovation: Revising the Cambridge proficiency in English examination 1913-2002* (pp. 367–471). Cambridge University Press.
- [44] Becker, A., Matsugu, S., & Al-Surmi, M. (2017). Balancing practicality and construct representativeness for IEP speaking tests. *Asian-Pacific Journal of Second and Foreign Language Education*, 2(1), 1–16. <https://doi.org/10.1186/s40862-017-0026-1>.
- [45] 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.
- [46] East, M. (2015). Coming to terms with innovative high-stakes assessment practice: Teachers' viewpoints on assessment reform. *Language Testing*, 32(1), 101–120.
<https://doi.org/10.1177/0265532214551646>
- [47] Galaczi, E. D. (2014). Interactional competence across proficiency levels: How do learners manage interaction in paired speaking tests? *Applied Linguistics*, 35(5), 553–574. <https://doi.org/10.1093/applin/amu011>.
- [48] Butler, Y. G., & Zeng, W. (2014). Young foreign language learners' interactions during task-based paired assessments. *Language Assessment Quarterly*, 11(1), 45–75.
<https://doi.org/10.1080/15434303.2013.853060>