Editorial

Use of innovative technology in oral language assessment

Fumiyo Nakatsuhara
University of Bedfordshire
fumiyo.nakatsuhara@beds.ac.uk
https://orcid.org/0000-0002-0429-7417

Vivien Berry
British Council, London
https://orcid.org/0000-0002-5610-4787

The theme of the very first Special Issue of *Assessment in Education: Principles, Policy and Practice* (Volume 10, Issue 3, published in 2003) was 'Assessment for the Digital Age'. The editorial of that Special Issue notes that the aim of the volume was to "draw the attention of the international assessment community to a range of potential and actual relationships between digital technologies and assessment" (McFarlane, 2003, p.261). Since then, there is no doubt that the role of digital technologies in assessment has evolved even more dynamically than any assessment researchers and practitioners had expected. In particular, exponential advances in technology and the increased availability of high-speed internet in recent years have not only changed the way we communicate orally in social, professional, and educational contexts, but also the ways in which we assess oral language. Revisiting the same theme after almost two decades, but specifically from an oral language assessment perspective, this Special Issue presents conceptual and empirical papers that discuss the opportunities and challenges that the latest innovative affordances offer.

The current landscape of oral language assessment can be characterised by numerous examples of the development and use of digital technology (Sawaki, 2022; Xi, 2022). While these innovations have opened the door to types of speaking test tasks which were previously not possible and have provided language test practitioners with more efficient ways of delivering and scoring tests, it should be kept in mind that "each of the affordances offered by technology also raises a new set of issues to be tackled" (Chapelle, 2018). This does not mean that we should be excessively concerned or sceptical about technology-mediated assessments; it simply means that greater transparency is needed. Up-to-date information and appropriate

guidance about the use of innovative technology in language testing and, more importantly, what language skills are elicited from test-takers and how they are measured, should be available to test users so that they can both embrace and critically engage with the fast-moving developments in the field (see also Khabbazbashi et al., 2021; Litman et al., 2018). This current Special Issue therefore aims to contribute to and to encourage transparent dialogues by test researchers, practitioners, and users within the international testing community on recent research which investigates both methods of delivery and methods of scoring in technology-mediated oral language assessments.

Of the seven articles in this volume, the first three are on the application of technologies for speaking test delivery. In the opening article, Ockey and Neiriz offer a conceptual paper examining five models of technology-delivered assessments of oral communication that have been utilised over the past three decades. Drawing on Bachman and Palmer's (1996) qualities of test usefulness, Ockey and Hirch's (2020) assessment of English as a *lingua franca* (ELF) framework, and Harding and McNamara's (2018) work on ELF and its relationship to language assessment constructs, Ockey and Neiriz present a framework for evaluating technology-driven speaking test delivery models. Although earlier models of tape-mediated communication are briefly discussed, the authors' main focus is on more recent examples such as synchronous oral communication assessment with mediated visual presence (through video-conferencing and virtual reality technologies) and spoken dialogue systems (SDSs) as these are the most current models that are consistently being researched and updated. They conclude that synchronous assessments with mediated visual presence most closely represent the oral communication construct as defined in Ockey and Hirch's (2020) framework and may also have the most positive impact on language learning and instruction. However, they find them to be the least practical and also potentially the least reliable depending on a number of factors including interlocutor inconsistencies in delivery. They are also concerned about access and fairness in the global divide relating to access to technology. They suggest that SDSs, although still in need of further development, may reasonably well represent the oral communication construct and positively impact learning and instruction, while being generally reliable and practical. We believe that the contemporary conceptualisation of oral communication and its novel evaluation framework laid out in this paper will pave the way for further development of different technology-mediated delivery models, while carefully scrutinising the way in which new models are expanding the construct measured by them.

Our second article continues the theme introduced in Ockey and Neiriz's paper of delivery of speaking assessments and their relationship to the oral communication construct. Nakatsuhara et al. investigate the comparability of scores obtained and language functions produced by test-takers in face-to-face and video-conferencing versions of the IELTS Speaking test. Unlike many of the earlier studies that used paired t-tests or other inferential statistics as a means of comparing the scores awarded in the two modes of delivery, their study conducted a more robust analysis with Many-facet Rasch Measurement (MFRM) analysis. Results from the MRFM analysis on the test scores of 99 test-takers who took both face-to-face and video-conferencing tests show that the scores obtained on the two modes were fundamentally equivalent, thus supporting the findings of numerous researchers of mode comparability over the years. However, test scores are, of course, only one aspect of comparability studies of different modes of delivery. Another important aspect is to explore how test-takers' and examiners' discourse compares between the two modes, and this is what the authors do in their language functions analysis. In terms of the language output, one function *asking for clarification* was used significantly more often in the video-conferencing mode, with 36.6% more test-takers asking clarification questions in one of the test parts. The authors propose that this could possibly indicate a slight change of construct in communication under the video-conferencing mode. That is, explicit verbalisation of potential communication problems and explicit attempts to clarify possible misunderstandings seem to be key to successful interaction via video-conferencing, where supplemental information and subtle cues from the interlocutor, such as gestures and voice inflection, could be less salient.

The third article continues the topic of delivery mode of speaking assessments. However, in their study, Zhang and Jin are not concerned with the score comparability of the two modes but with similarities and differences in communication strategies used by test-takers taking both a face-to-face version of the Chinese College English Test – Spoken English Test (CET-SET) and its computer delivered version. The computer-based CET-SET uses an innovative paired discussion task to assess interactional competence with pairs randomly allocated by the computer. This paired format of testing, which allows for peer-peer interaction, is considered to elicit more natural and symmetrical conversations than are possible in the one-to-one oral interview. They adopted Galaczi and Taylor's (2018) definition of interactional competence: "the ability to co-construct interaction in a purposeful and meaningful way, taking into account sociocultural and pragmatic dimensions of the

speech situation and event" (p. 226). For their experimental study, a face-to-face test with equivalent task design was developed, and 30 pairs of students took the face-to-face test immediately after taking the computer-based CET-SET. Results showed that most communication strategies were used equally across the two modes with the exception of some cooperative strategies (asking a question, giving feedback) and meaning-negotiation strategies (making/responding to a repetition request, making/responding to a comprehension check) which test-takers used more frequently in computer-mediated discussions than in face-to-face discussions. The results also present some evidence that the audio-only computer-mediated discussion task is effective in facilitating test-takers' active participation, although this is likely to become even more effective when increased bandwidth allows test-takers to see each other virtually. Since computer delivery is almost certainly the future of testing oral ability, especially in a country like China with its vast population of test-takers, it is important that there is ongoing research into whether and how different modes of delivery affect the oral construct. This study provides useful insights that other researchers can take as a starting point.

The remaining four articles concern the use of technology for scoring oral performances. In recent years, a number of automarkers have been designed to score test-takers' spoken performances, and it is not uncommon that international examination boards offer high-stakes speaking tests that are partly or fully assessed by an automarker. Such increase in popularity of automated scoring entails a greater need for transparency in how test-taker speech is scored and how reliable the resulting scores are (e.g. Xi et al., 2016). In our fourth study, Xu et al. contribute to the endeavour to enhance transparency in automarker research and validation, by exploring the performance of the Custom Automated Speech Engine (CASE v1.9) that was developed for the Cambridge Assessment English Linguaskill General Speaking test. The study explored four specific aspects of the automarker guided by Xi's (2010) framework: (a) the accuracy of the automarker, (b) the consistency and severity of the automarker, (c) automarker performance at different confidence levels, and (d) the robustness of the automarker against abnormal test behaviours. They also introduce 'limits of agreement' (Bland & Altman, 1999), a method from clinical measurement, and argue for its advantages over traditional reliability measures, such as correlation and Kappa, that are commonly used in automarker validation. Using 'limits of agreement' and MFRM analyses on automarker scores and scores from three human raters, the study revealed that the automarker gave slightly higher scores than human raters' fair average scores, particularly on

lower-proficiency speech samples (CEFR A1 and A2), while its internal consistency was excellent. The results also show that *Language Quality scores* – an automarker's measure of confidence/uncertainty in its speech recognition element in transcribing speech input to text – was useful for predicting automarker reliability and for flagging speech with anomalies that should be passed on to human raters. In other words, considerably more accurate automarker performance was observed when high confidence in speech recognition was indicated, while reduced intelligibility and audio quality were likely to be detrimental to automarker reliability. This study is noteworthy in that it uses empirical data to investigate varying automarker reliability on speech at different levels of proficiency. Based on these findings, the authors conclude that there is insufficient evidence for the version of the automarker in the study to be used on its own for a high-stakes English language assessment. Their study is a welcome addition to the body of the automated speaking assessment literature, enhancing our understanding of the actual performance, potential, and limitations of the operationally-used automated scoring system.

Recent years have seen much discussion on the potential of human-machine hybrid marking, by making the best use of complementary strengths of human raters and automated scoring systems (e.g. Isaacs, 2018). It is generally suggested that specific micro-linguistic features can be measured by automated scoring systems, while more global evaluation of test-taker speech, such as discoursal and content-related features, can be marked by human raters. However, empirical studies on how best to combine human and machine scores are still scarce. The exploratory study presented in our fifth article, by Davis and Papageorgiou, is therefore a timely contribution to the ongoing debates. Using spoken samples on the TOEFL iBT Speaking test, trained raters, and their SpeechRater automated scoring system, they compared four different combinations of human and machine scores on three analytic criteria: Delivery (fluency and pronunciation), Language use (grammar and lexis), and Topic development (content and discourse-level features). The four combinations were:

i) Delivery (machine) + Language use (human) + Topic development (human)

ii) Delivery (human) + Language use (machine) + Topic development (human)

iii) Delivery (machine) + Language use (machine) + Topic development (human)

iv) Delivery (human) + Language use (human) + Topic development (human)

The results show that the highest reliability was observed from all human analytic scores (iv, Cronbach's alpha = .94); this was followed by those from two human and one machine analytic scores (i & ii, .92); finally, from one human and two machine scores (iii, .90). All of

these reliability figures of the composite scores were marginally higher than human holistic scores (.88), obtained from operational administrations of the TOEFL iBT Speaking test, although the modest increase in reliability is likely to be attributed to the multiple observations included in composite scores rather than the use of a hybrid approach. While a decrease in reliability was observed as composite scores incorporated more machine analytic scores, the authors themselves acknowledge that the sample size of test-taker speech used to train and evaluate the automated scoring system in this experimental study was smaller than would normally be used to build an operational automated scoring system. They note that there is considerable opportunity for language testers to investigate different hybrid models of human and machine scoring using larger datasets, more efficient machine-learning algorithms, and evolving natural language processing technology. There is no doubt that this study will encourage further research and discussion on the potential of human-machine hybrid marking to capture complementary aspects of the speaking construct.

The above two articles discussed automarkers that involve three main components typical of most automated scoring systems: a speech recogniser, a feature extraction module, and a scoring model. In contrast, in our penultimate article, De Jong et al. take a different approach, focusing solely on speech fluency. Using PRAAT – a computer program that allows for precise analyses of speech delivery (Boersma & Weenink, 2016), they updated and revised De Jong and Wempe's (2009) PRAAT script to create a user-friendly tool that can measure filled pauses (e.g. "uhm") as well as silent pauses and speed of speaking, automatically, without the need for automated/manual transcribing or manual annotations and measurements. A detection algorithm for filled pauses is the newly added feature in their script, and the development was informed by the body of literature on acoustic properties of filled pauses. They developed two versions of the tool - one for Dutch and the other for English. They then examined how accurately the script can detect filled pauses in two types of datasets (Dutch and English-speaking performances collected under language testing conditions), by comparing the outcomes of the script and manual annotations of filled pauses. The results show that most manually annotated filled pauses in both datasets can be correctly identified by the algorithm, although they detected a relatively high number of false positives and scope for further improvements are discussed. Gauging the validity of the automated measurement for potential use in language assessment, they found that if filled pauses are an explicit part of the fluency construct in a given rating scale, up to 20% of the variance of fluency ratings could be explained by the automatically measured filled pauses alone. While

the algorithm is promising, the authors cautiously conclude that the developed tool is not yet ready for automated measurement of fluency for the purpose of assessing learners' proficiency. The authors share the updated script to detect syllables and a new script to identify filled pauses in the online appendices. Those scripts themselves are a valuable contribution to the field of oral fluency and spoken assessment and will facilitate future work in this area.

In the final article of this volume, Hunte et al. provide detailed accounts of their step-by-step exploration of the extent to which computational linguistic and acoustic indices identified through natural language processing approaches could predict human scores of children's speaking performances on story retell tasks. Unlike the rest of the articles in this volume, Hunte et al.'s research context involved mostly L1 English speakers: children at a school in Canada. They address a research niche for the use of automated scoring on children's speech, with the aim of promoting assessment of oral proficiency at schools – a fundamental element in children's language development and educational success. A total of 184 children aged 9 to 11 performed two story retell tasks, one guided by a written stimulus and the other by an aural stimulus. The performances were rated by trained raters on five criteria: vocabulary, grammar, idea development, task-fulfilment, and speech delivery, and a composite score was derived. The authors explain how natural language processing techniques were applied to the speech data, detailing the procedures for linguistic and acoustic feature extraction, feature engineering, and supervised machine learning. An iterative model building process taken to determine which features and machine learning algorithms best predict each of the human analytic scores is then illustrated. While the authors acknowledge the limited generalisability of the composite score predictions, they demonstrate that their closely engineered features predicted children's composite scores from human raters with almost 90% accuracy and accounted for 70% of the total variance in the samples. However, when each analytic criterion was scrutinised, a challenge was also presented with the 'speech delivery' criterion that indicated the lowest prediction, namely, 63% with the aural stimulus and 53% with the written stimulus, showing human-machine correlations of .54 and .27 respectively. The authors speculate reasons for the considerable difference in the prediction accuracy between the two modes of elicitation stimuli and call for further research.

With the aim of contributing to enhanced transparency for technology-mediated spoken language assessment research and practice, this Special Issue offers a selected

collection of the latest research that addresses both the opportunities and challenges presented by various technological innovations in the field. Digital technology in language testing was already a booming trend prior to the pandemic and most research in this volume was conducted before the advent of COVID-19. However, the need for alternative, digital solutions in language testing over the last two years (Kremmel & Isbell, 2020; Ockey, 2021) has only increased the relevance of the focus of this volume. It is hoped that this Special Issue will serve as a useful resource to help test users make more informed decisions about technology-mediated speaking tests, and to provide the language testing community with timely insights for future oral test development.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

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

Bland, J. M., & Altman, D. G. (1999). Measuring agreement in method comparison studies. *Statistical Methods in Medical Research, 8*(2), 135–160. https://doi.org/10.1177/096228029900800204

Boersma, P., & Weenink, D. (2016). PRAAT: Doing phonetics by computer [Computer program]. Version 6.1.29. http://www.PRAAT.org/

Chapelle, C.A. (2018, May 25-27). *The contested role of technology in building better language tests* [Keynote paper presentation]. 15th EALTA conference, Bochum, Germany.

De Jong, N. H., & Wempe, T. (2009). PRAAT script to detect syllable nuclei and measure speech rate automatically. *Behavior Research Methods, 41*(2), 385-390. https://doi.org/10.3758/BRM.41.2.385

Galaczi, E., & Taylor, L. (2018). Interactional competence: Conceptualisations, operationalisations, and outstanding questions. *Language Assessment Quarterly, 15*(3) 219-236. https://doi:10.1080/15434303.2018.1453816

Harding, L., & McNamara, T.F. (2018). Language assessment: The challenge of ELF. In J. Jenkins, M. J. Dewey & W. Baker (Eds.), *Routledge Handbook of English as a Lingua Franca* (pp. 570-582). Routledge.

Isaacs, T. (2018) Shifting sands in second language pronunciation teaching and assessment research and practice. *Language Assessment Quarterly, 15*(3), 273-293. https://doi.org/10.1080/15434303.2018.1472264

Khabbazbashi N., Xu J., Galaczi E. (2021). Opening the black box: Exploring automated speaking evaluation. In B. Lanteigne, C. Coombe & J.D. Brown (Eds.), *Issues in language testing around the world: Insights for language test users* (pp.333-343). Springer.

Kremmel, B., & Isbell, D. (2020). Test review: Current options in at-home language proficiency tests for making high stakes decisions. *Language Testing*, *37*(4), 600–619. https://doi:10.1177/0265532220943483

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:10.1080/15434303.2018.1472265

McFarlane, A. (2003). Editorial. Assessment for the digital age, *Assessment in Education: Principles, Policy & Practice, 10*(3), 261-266. https://doi:10.1080/0969594032000148127

Ockey, G. J. (2021). An overview of COVID-19's impact on English language university admissions and placement tests. *Language Assessment Quarterly*, *18*(1), 1-5, https://doi: 10.1080/15434303.2020.1866576

Ockey, G. J., & Hirch, R. R. (2020). A step toward the assessment of English as a lingua franca. In G. J. Ockey & B. A. Green (Eds.), *Another generation of fundamental considerations in language testing* (pp. 9-28). Springer.

Sawaki, Y. (2022). Computer-based testing. In G. Fulcher & L. Harding (Eds.), *The Routledge handbook of language testing* (2nd ed.) (pp. 530-544). Routledge.

Xi, X. (2010). Automated scoring and feedback systems: Where are we and where are we heading? *Language Testing*, *27*(3), 291–300. https://doi.org/10.1177/0265532210364643

Xi, X. (2022). Validity and the automated scoring of performance tests. In G. Fulcher & L. Harding (Eds.), *The Routledge handbook of language testing* (2nd ed.) (pp.513-529). Routledge.

Xi, X., Schmidgall, J., & Wang, Y. (2016). Chinese users' perceptions of the use of automated scoring for a speaking practice test. In G. Yu & Y. Jin (Eds.), *Assessing Chinese learners of English: Language constructs, consequences and conundrums* (pp. 150–175). Palgrave MacMillan.