Considerations in the development and validation of automated speaking assessment

Trevor Breakspear

This article builds on a TEASIG webinar given by the author on 24 November 2020.

Automated speech assessors (ASA) promise cheap, efficient and scalable scoring of student performance. Such benefits have led to the application of ASA in a range of contexts, including high-stakes testing as well as classroom and formative assessment. Yet there remains concern over how effectively ASA can measure certain features of speaking, such as discourse, coherence and task relevance (Weigle, 2010), and unease over a lack of transparency in disseminating the technologies powering them (Enright & Quinlan, 2010). In effect, we often just don't know how accurate or fit for purpose ASA are. To explore these issues, I will discuss a few key principles underpinning ASA, with reference to an autorated low-stakes test developed by the East Asia Assessment Solutions Team (EAAST) at the British Council. I aim to show how test developers and educational practitioners can add value in the development of ASA, whilst highlighting key questions to consider before we adopt such systems in our own learning or assessment contexts.

Automatic Speech Recognition (ASR)

The first challenge for ASA is to change the purely audial signal of speech into a form that machines can work with – written text. To achieve this, ASA use automatic speech recognition (ASR), the technology allowing virtual assistants and smart speakers to 'understand' our instructions, which converts the speech waveform into a text, or a transcript, of the spoken utterance (see Figure 1). The ASA system is only as accurate as its ASR; a poorly performing ASR will affect the reliability of the entire system, as incomplete or inaccurate text representations will bias the machine scoring of domains such as pronunciation, grammar and lexical expression.

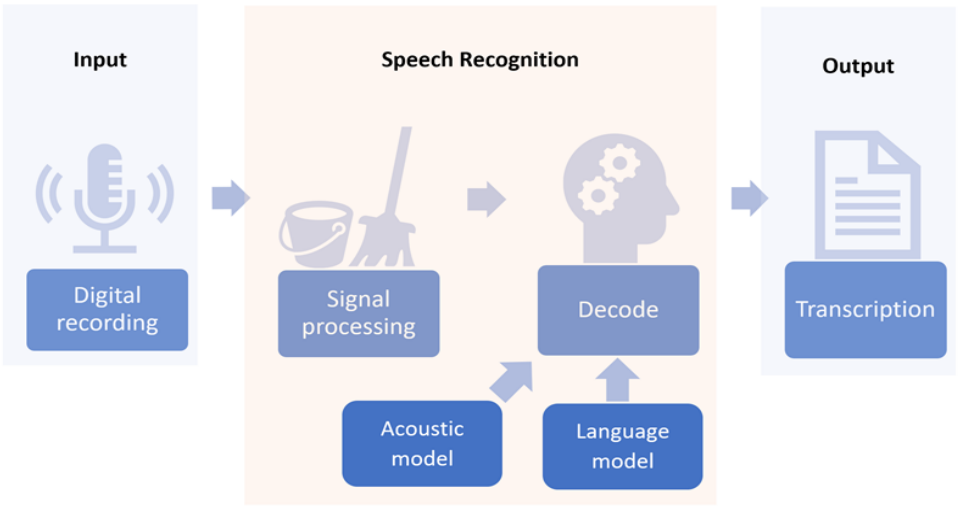


Fig 1. Overview of key ASR components.

The accuracy of the ASR depends primarily on the decoding stage, which combines output from two key processes, the acoustic and language models (see Figure 1). The acoustic model matches sounds from the 'cleaned' recording input with a suggested orthographic representation (the written text) using a huge database of digitally recorded speech segments tagged with their corresponding transcriptions (van Moere & Downey, 2016). The language model refines suggestions from the acoustic model by identifying the likely word sequences using a large database; rather like a predictive text system that 'corrects errors' or completes sentences on a mobile phone. Such optimization is shown in the following illustrative example whereby the language model could deduce from context that 'sea' is the most likely of the three orthographic representations of ['si:] suggested by the acoustic model.

The man walked by the [see/sea/c]. (acoustic model)
The man walked by the sea. (acoustic + language models)

As with predictive text, ASR systems are far from perfect. Accuracy can be affected by a variety of factors including (1) quality and relevance of the data used to train the models and (2) the predictability of the spoken content. Many argue that characteristics of the users (such as age, ethnicity, gender) should be represented in the speech data used to train the models (van Moere & Downey, 2016), to avoid variations in second language performance (such as L2 accent) affecting the accuracy of the ASR system. As a result, we should be asking whether the pronunciation characteristics of our learners or test takers are captured in the data used to train the acoustic models, and if not, whether the automated scoring of their performance will be affected.

As projected users of our placement test were Chinese learners aged 14-18, we commissioned a vendor with an ASR trained on a database of Chinese L2 (English) speaking samples representing the target age and covering a wide range of regional accents. In this way, we achieved a relatively close proximity between the speaking characteristics of the test user and the data on which the ASR was trained. However, should we launch the placement test in other countries with different L1 accents, we would need to validate the ASR using L2 speaking samples from the new target population.

Task types and ASA reliability

Generally speaking, the scoring reliability of ASA increases as the content of speech acts becomes more predictable, because the language models in ASR and algorithms used in scoring are more accurate when conditions are predictable. For these reasons, constrained task types, such as 'read aloud' and 'listen and repeat' are more commonly found in ASA tests than their human-scored counterparts. The types of tasks used in a particular ASA raise important questions for the teachers and learners who use them. An ASA solution comprising only constrained tasks is likely to underrepresent the speaking construct (unless we are only interested in pronunciation scoring, for example), whilst an ASA solution comprising open response tasks that elicit less predictable content will be more challenging for the technology and may lead to lower reliability.

The domain analysis for our test identified description, narration and argumentation as key speaking genres, and so we used open response tasks, such as asking

learners to describe personal experiences and give opinions on familiar topics, to elicit spoken performance (see Figure 2, step 1). We were aware that such a challenging choice of task type would put more stress on the ASA and therefore on us to demonstrate the validity of the human rating scales and the reliability of the machine scoring output. Our responses to these challenges are discussed later.

The human-machine rating divide

A key assumption underlying ASA is that human rater scores for a particular test purpose interpret the desired language traits, the things we want to measure (Messick, 1989a). As a result, ASA have been designed to identify and weigh features of spoken responses measurable through computational means in a way that best predicts the human score (van Moere & Downey, 2016). To measure fluency, for example, statistical models calculate the optimum weighting of a range of machine-measurable features (such as speech rate, length of spoken turn, duration of hesitation) in such a way as they correspond most closely with human ratings for the same sample. Therefore, developers use human-machine reliability (the correlation between human and machine scores) as key evidence in validating their ASA solutions.

A disadvantage of this assumption is that any errors or undesired variance in the design of task specifications and human rating scales will likely be amplified and operationalized in machine rating (Deane, 2013). To address these issues, and ensure we minimized variance in human-rated data, we spent significant time trialing and revising the rating scales using empirical evidence from learner responses, rater data, and rater perceptions (Figure 2). To ensure the reliability of the database used to train the machine, we conducted extensive human rater training and provided weekly feedback, ensured all responses were rated by 3 or more examiners, and used statistical tools, such as Many-Facet Rasch Measurement to reduce bias in the data.

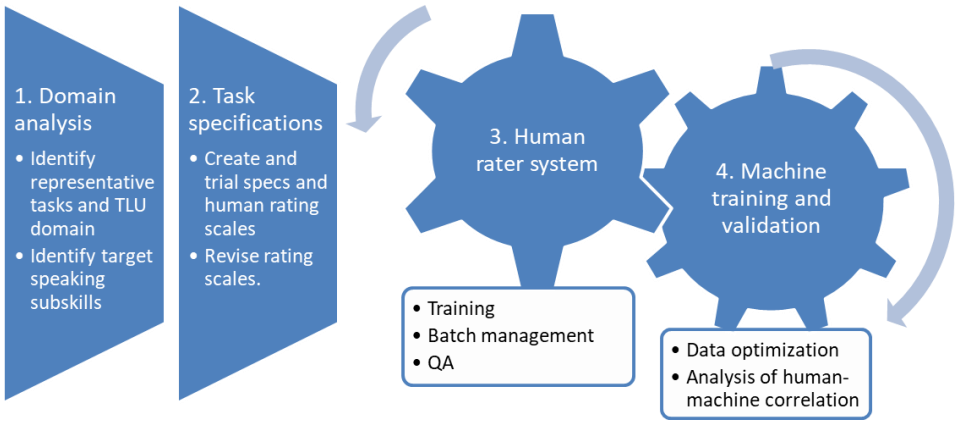


Fig 2. Simplified development and validation process for an autorated low-stakes test.

Significant differences remain, however, between what can be measured by machine raters and human raters. Machines can measure quantifiable features (such as speech rates, unique word counts etc.); however, they are less able to interpret or infer speaker meaning or intent, assess the efficacy of argumentation, or consider content-related accuracy (Deane, 2013). These features are often explicit or implicit

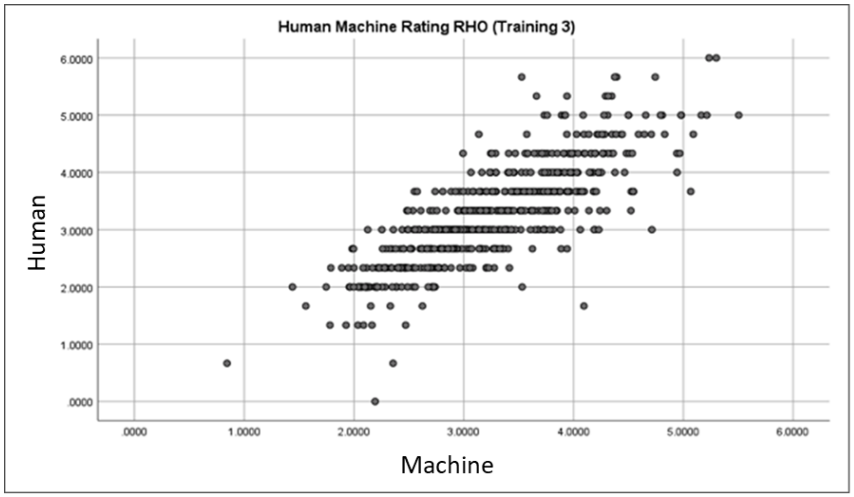
factors considered by human raters when scoring criteria such as topic relevance, task completion, discourse and argumentation. As a result, a generally acknowledged limitation of ASA is construct underrepresentation (Deane, 2013). Furthermore, even where human-machine features appear aligned, there are often clear differences in how they are observed. For example, a machine may be able to identify the number of phoneme errors in a spoken turn, but a human rater is arguably better placed to evaluate their impact on comprehensibility. So, whilst the same trait is being targeted by human and machine, interpretation, and thus impact on construct validity and washback (Messick, 1989), are arguably very different.

These challenges were particularly significant for us because the technologies used by our vendor were wrapped within a 'black box'. This lack of 'explainable' artificial intelligence (AI), caused by both the high level of technical expertise needed to understand the processes involved and a general reluctance of vendors to share their 'secrets', meant we were left somewhat in the dark.

An iterative approach to development and validation

There is, however, much that test developers can still contribute to the optimization of autorater systems even without a technical understanding of machine learning (ML) and automated scoring used by the tech vendor. One of the most common adages in ML is 'garbage in, garbage out', meaning the quality of speaking scores is only as good as the data on which the machine is trained. In developing our test, we adopted an iterative process of optimizing the ML inputs through an analysis of the outputs (Figure 2, steps 3 & 4) to ensure the data we provided to the vendor was as fit for purpose as possible.

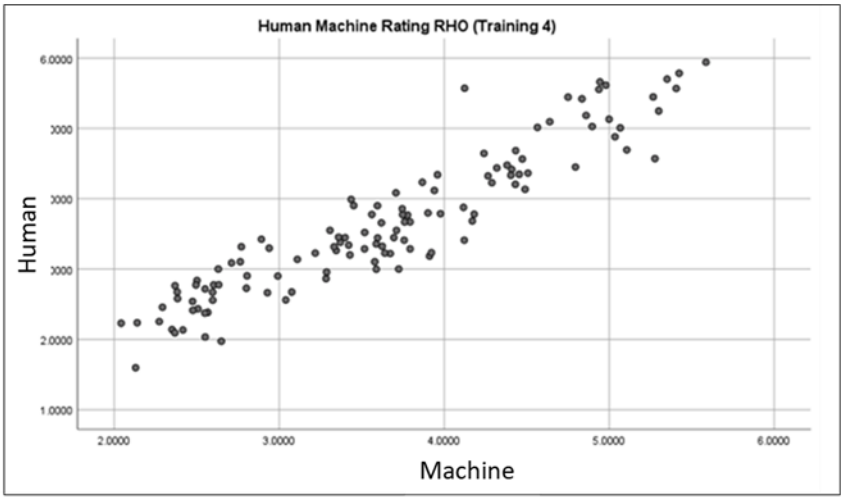
One of the key iterations we made during the ML process was to improve database representation. For our lower stakes purpose, Foltz et al. (2013) suggest using a training database of at least 200-300 samples evenly distributed across the proficiency levels for each task. Statistical methodologies used in ASA require the comparison of responses in the training database with those of test takers, meaning that high-scoring user responses will resemble (contain similar features present in) high-scoring samples from the training database (Dikli, 2006). For this reason, the training database should contain a variety of valid approaches to address the prompt to avoid disadvantaging speakers using an unconventional but appropriate response. As a result, we built a training database of 440 responses from a wide range of speaker backgrounds, proficiency levels, locales etc., to improve the generalizability of results to the general test taking population. Figures 3 and 4 show how human-machine reliability (the correlation of human and machine scores) improved (from rs 0.8 in iteration 3 to rs 0.93 in iteration 4), particularly at the tails (highest and lowest scores), as a result of an improved distribution of proficiency. With the literature suggesting that rs 0.7 is minimally adequate for low-stakes purposes (Green, 2013), the ASA human-machine reliability of rs 0.93 for our test appears more than adequate for purpose.



Training 3	
RHO Correlation Coefficient	.803**

**. Correlation is significant at the 0.01 level (2-tailed).

Fig 3. Human-machine reliability at iteration 3.



Training 4	
RHO Correlation Coefficient	.93**

**. Correlation is significant at the 0.01 level (2-tailed).

Fig 4. Human-machine reliability at iteration 4.

Conclusions and future directions

Whilst human-machine reliability is viewed as the key measure of whether ASA technology is working (Yan & Bridgeman, 2020), I would argue that more evidence is needed. A method commonly used is the concurrent validity study, which in this context measures the correlation between the scores a candidate receives on a machine-rated test (benchmarked for example to the CEFR) with the score that same candidate receives on a similarly benchmarked human-rated test. However, I would argue that such studies can be affected by a lack of equivalency, as even tests in the same language use domain (for example, tests of academic English) target different skills using different tasks. Therefore, the contribution of current validity studies towards the validity argument should be considered carefully. Another approach is to embed additional validation mechanisms within the machine learning process itself. For example, when we validated a pronunciation scorer linked to our test, outliers (responses that receive significantly different human and machine scores) were analysed individually by both our team and the vendor's data scientists to identify sources of error and review machine learning and modelling techniques to address them.

Overall, though, I believe the most significant way to improve the validation of ASA is to compel edutech vendors to open up about the technologies powering their systems by sharing the sources of data used to train the models, the features of speech the ASA can measure, and the scoring accuracy (reliability) of the system across different speaking domains and user groups (such as age and ethnicity). Edutech vendors should also collaborate more closely and transparently with test developers and teaching practitioners to ensure the best possible alignment of assessment and pedagogical practices with technical capabilities (Xi, 2010). The inappropriate use of ASA technologies could result in serious consequences for learners, such as a reliance on inaccurate scoring and evaluation as a replacement for the richer feedback and interaction traditionally offered by examiners and teachers. As Messick (1989b: 11) reminds us, social consequences "clearly have implications for both the science and ethics of assessment", and as such we should petition ASA vendors to play their part.

Bibliography

Deane, P. (2013). On the relation between automated essay scoring and modern views of the writing construct. *Assessing Writing 18*(1), 7–24. https://doi.org/10.1016/j.asw.2012.10.002

Dikli, S. (2006). An overview of automated scoring of essays. *Journal of Technology, Learning, and Assessment 5*(1), 1–35.

Enright, M. K. and Quinlan, T. (2010). Complementing human judgment of essays written by English language learners with e-rater® scoring. *Language Testing 27*(3), 317–334. https://doi.org/10.1177/0265532210363144

Foltz, P., Streeter, L., Lochbaum, K. and Landauer, T. (2013). Implementation and applications of the intelligent essay assessor. In Yan, D., Rupp, A. and Foltz, P. (Eds.). *Handbook of Automated Essay Evaluation: Current Applications and New Directions* (pp 68–88). Taylor & Francis. https://doi.org/10.4324/9780203122761

Green, R. (2013). *Statistical analyses for language testers. Statistical Analyses for Language Testers*. Springer Nature. https://doi.org/10.1057/9781137018298

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

Messick, S. (1989b). Meaning and Values in Test Validation: The Science and Ethics of Assessment. *Educational Researcher 18*(2), 5–11. https://doi.org/10.3102/0013189X018002005

Van Moere, A. and Downey, R. (2016). Technology and artificial intelligence in language assessment. In Tsagari, D. and Banerjee, J. (Eds.) *Handbook of Second Language Assessment*. Boston/Berlin: De Gruyter Mouton. https://doi.org/10.1515/9781614513827-023

Weigle, S. (2010). Validation of automated scoring of TOEFL iBT tasks against non-test indicators of writing. *Language Testing 27*(3): 335–353.

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

Yan, D. and Bridgeman, B. (2020). Validation of Automated Scoring Systems. In Yan, D., Rupp, A. and Foltz, P. W. (Eds.). *Handbook of Automated Scoring* (pp. 152-167). Taylor & Francis. https://doi.org/10.1201/9781351264808

Biodata

Trevor Breakspear works in an innovations role within the British Council's East Asia Assessment Solutions Team, leading on the academic development of new AI-rated assessment tools, as well as coordinating academic and stakeholder research into new test concepts leveraging both human and machine raters.

trevorjohn.breakspear@britishcouncil.org.cn