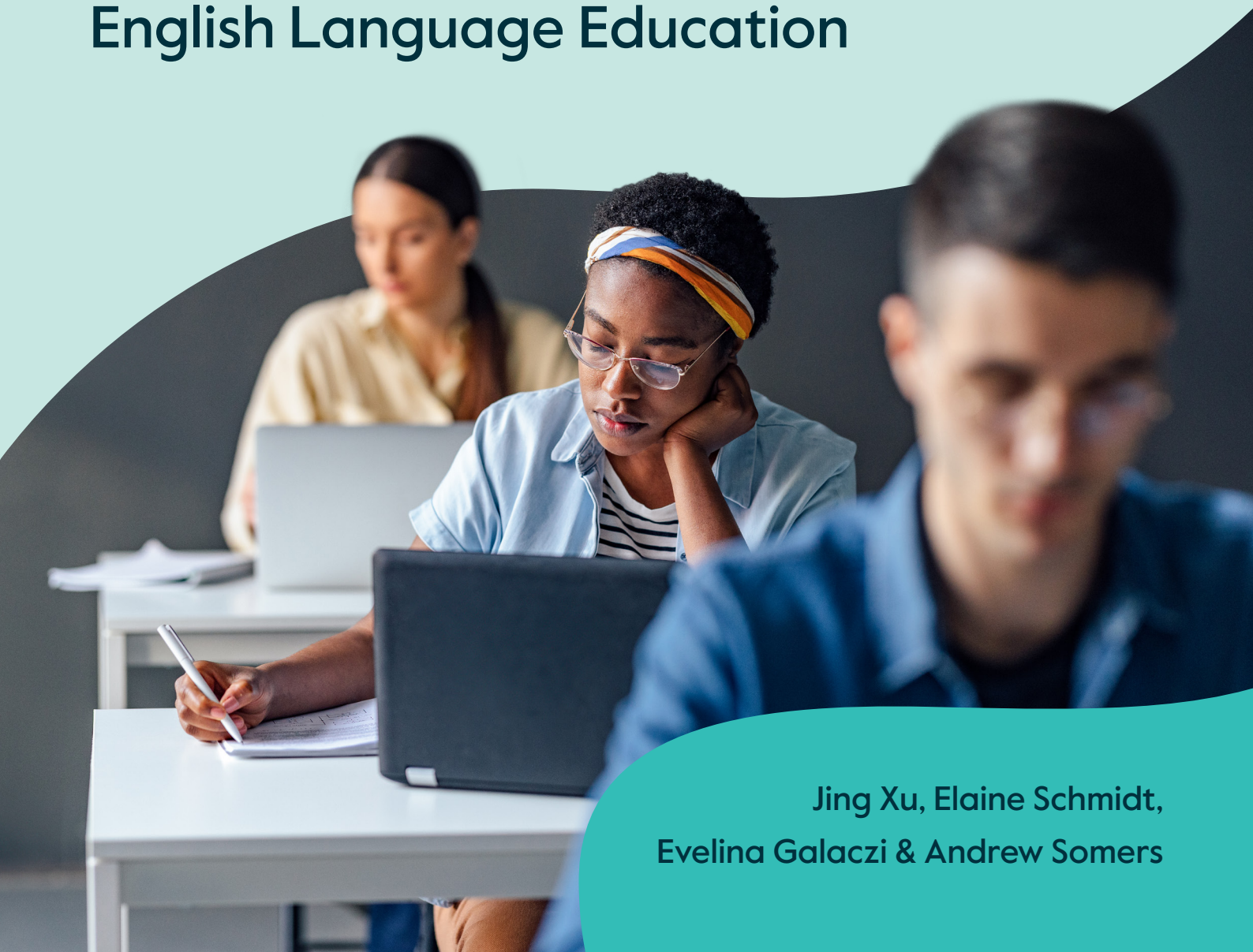
CAMBRIDGE

Automarking in language assessment: Key considerations for best practice

Cambridge Papers in English Language Education



Jing Xu, Elaine Schmidt, Evelina Galaczi & Andrew Somers

Author biographies

Jing Xu is Head of Propositions Research-English at Cambridge University Press & Assessment. His research interests are in the application of cutting-edge technologies, particularly AI, in language assessment and learning and the related validity issues. His latest research focuses on automarking of L2 English speaking and writing assessments and using spoken dialogue systems in computer-based speaking tasks. He was the winner of the 2017 Jacqueline Ross TOEFL Dissertation Award and a co-recipient of the 2012 International Language Testing Association (ILTA) Best Article Award. Jing is currently Co-Chair of the ILTA Automated Language Assessment Special Interest Group. He received his PhD in Applied Linguistics and Technology from Iowa State University with a focus on language assessment.

Evelina Galaczi is Director of Research-English at Cambridge University Press & Assessment. She has worked in English language education for over 30 years, and her current work focuses on the challenges – and exciting opportunities – of using AI in language learning, teaching and assessment. Evelina has explored issues in assessment and learning through her academic publications and international presentations. She currently leads a modern and dynamic team of experts in language learning, teaching and assessment, serves as a Trustee for the International Research Foundation for English Language Education and is co-editor of the journal *Language Assessment Quarterly*. She holds a Master's and a Doctorate degree in Applied Linguistics from Columbia University, USA.

Elaine Schmidt is a Senior Researcher at Cambridge University Press & Assessment. Her research focuses on cognitive aspects of language processing and learning using eye tracking and electroencephalography (EEG). She obtained her PhD in Linguistics and Language Acquisition from the University of Cambridge. After her PhD she worked on cognitive processes of L1 and L2 speech perception in Sydney, Australia, before she moved back to the Linguistics Department at the University of Cambridge. A few years later she then decided to combine her research with more practical applications and joined English-Research at Cambridge University Press & Assessment, where she brings her expertise in speech production and perception, eye tracking and EEG in second language learning to an assessment context. More recently she has also focused on research in other types of technology for assessment, including automarking or research centring around remote proctoring.

Andrew Somers is Director of Assessment Excellence-English at Cambridge University Press & Assessment. He has worked in English language assessment for over 20 years. His current focus includes the application of AI to various aspects of language assessment, particularly how it complements existing processes and offers additional benefits to learners and business, but also how we manage its many challenges. He has worked extensively in the application of new methods, processes and systems to enhance the security, validity and reliability of assessment processes. He currently leads a varied team of experts in language assessment, psychometrics, data science and AI engineering focusing on the development and implementation of new assessment capabilities across the Cambridge University Press & Assessment portfolio. He holds a PhD from the University of Cambridge.

Contents

Introduction

In the past decade, automarking technology has become increasingly widespread in large-scale second/foreign (L2) language assessment thanks to the progress made in Machine Learning (ML) and Artificial Intelligence (AI). Automarkers, also known as automated scoring systems or AI-based scoring systems, are computer algorithms which are trained to mark complex constructed responses such as written essays and extended speech. Currently, automarking of writing is more advanced than speaking, largely because of the challenges of capturing speech and working with spoken data.

Automarkers are algorithms (also called models) which are trained to predict the scores that human examiners would give. Typically, an automarker works alongside additional algorithms that could determine, for example, how confident an automarker is in its score prediction and whether a test taker's response is unusual in any way and thus needs flagged for review (Gao et al., 2024). All these components form part of an **automarking system**.

Automarking is a transformational application of AI in language education. The decisions we make based on assessments that are enabled by AI can significantly impact life opportunities such as education admission, employment or immigration decisions. **The aim of our paper is to present key principles underlying good practice in the use of automarking in L2 language assessment, and to offer an illustration of how these principles are applied in practice** through a look at an automarked Cambridge writing exam.

We believe that such a grounding in essential considerations will increase test users' understanding of automarking technology, enable informed decisions on test takers, and promote responsible use of automarking in language assessment that optimises the benefits and minimises the limitations of this valuable tech capability.

Benefits and limitations of automarking systems

Automarking has emerged as a promising solution to the demand for efficient and accessible assessment of writing and speaking skills (Xi, 2021). It offers a number of benefits, such as:

- **speed of marking** – writing and speaking can be marked much faster than when only human examiners are involved
- the ability to facilitate **on-demand testing** which is not constrained by examiner availability
- in the case of well-trained automarking systems, the ability to **perform consistently** to a high standard over time, in contrast to human examiners who must be regularly trained and standardised to provide reliable ratings over time
- the enabling of **adaptive testing** in computer-based speaking assessments where the difficulty of tasks is adjusted throughout the test to test taker performance
- the **integration of learning and assessment**, since instant automated scores can inform teaching and learning and allow for individualised learning paths based on test takers' strengths and areas that need improvement

Despite these advantages, expert opinion among language testing professionals towards automarking of writing and speaking performances is only cautiously optimistic, on account of certain risks and concerns. The common limitations associated with automarking centre around:

- how **authentic** and representative of real-life language use automated language tests are, since automarking systems perform best with constrained tasks such as a short written response to a question, reading sentences aloud or replying to a set of questions instead of engaging in free-flowing dialogues. This is especially the case with speaking, where the robust automarking of interaction is currently beyond the capabilities of AI models.
- the **validity** of scoring algorithms, due to the relatively narrow range of language features automarking algorithms can deal with compared to human marking. Currently automarking technology is unable to fully measure communicative language ability. That includes not only lexico-grammatical features (complexity, accuracy and fluency) but also discourse organisation, argument development, implied meaning, and the appropriateness of language use in a social context. In the case of speaking, communicative language ability also includes tone, turn-taking management, dealing with communication breakdowns and non-verbal behaviour.

All of these aspects of successful language use are currently extremely difficult to be marked automatically in a reliable manner due to:

- the difficulty in **giving learners and stakeholders visibility of the assessment criteria** used by the automarking systems in relation to the language abilities targeted by the assessment and evaluated by human markers
- the increased risk of **exam malpractice** and challenges to exam integrity by strategies to cheat on automarked tests (Xi et al., 2016)
- the potential **negative washback** in instructional contexts that prompts learners to practice constrained, non-communicative language skills
- inadequate **AI literacy** among test users leading to misuse of automated language assessment

In this paper, we will show how Cambridge University Press & Assessment is addressing these challenges, while adhering to the principles of good practice.

Principles of assessment underpinning automarking

We start our discussion with a brief overview of two principles of good practice in assessment – validity and fairness. These principles are the foundation of tests in general, including automarked tests.

Validity

Validity is the most fundamental consideration in developing and evaluating tests. Broadly speaking, **the concept of validity** refers to how well a test measures what it is intended to measure. In technical terms, validity refers to how well research evidence and theory support the use of test scores for their intended purposes. Test validation is then a process of gathering relevant evidence to provide "a sound scientific basis" for the proposed test uses (AERA et al., 2014, p. 11).

Validity is not an inherent test quality – no test is valid or not valid in an absolute sense. A test should always be judged as **valid for a specific purpose and target a test taker group**. Validity, therefore, is a judgement made in specific contexts where the test is used. And test validity should thus be evaluated based on clearly defined assessment purposes, e.g. is the purpose to make decisions on student admission, certificate proficiency, place learners into appropriate classes, screen job applicants, or diagnose learners' strengths and weaknesses?

We now briefly outline specific aspects of validity through the prism of an influential language assessment framework, **the socio-cognitive model of validity** (Weir, 2005).

Validity

Cognitive validity
What knowledge and skills am I testing?

Context validity
How am I testing the knowledge and skills of interest?

Reliability
How far can I trust the accuracy of the test scores?

Test impact
How will the test affect the way teachers teach, learners learn, and society more broadly?

Criterion-related validity
How does the test performance compare with other measures?

Fairness

How equitably are all test takers treated?

Automated assessments of writing and speaking need to be guided by the principles of validity and fairness.

1 Cognitive validity – What knowledge and skills am I testing?

Cognitive validity relates to how well the language knowledge and skills involved in completing test tasks reflect those underlying the real-world communication activities related to the assessment purpose. For example, a writing prompt may require test takers to organise points coherently, use appropriate syntax and employ relevant vocabulary. The cognitive nature of this writing process should not be altered by the assessment.

If an automarker is used in writing assessment, it must be able to evaluate tasks that simulate such authentic language use situations rather than just constrained, non-communicative responses such as gap filling, matching and rearranging words.

2 Context validity – How am I testing the knowledge and skills of interest?

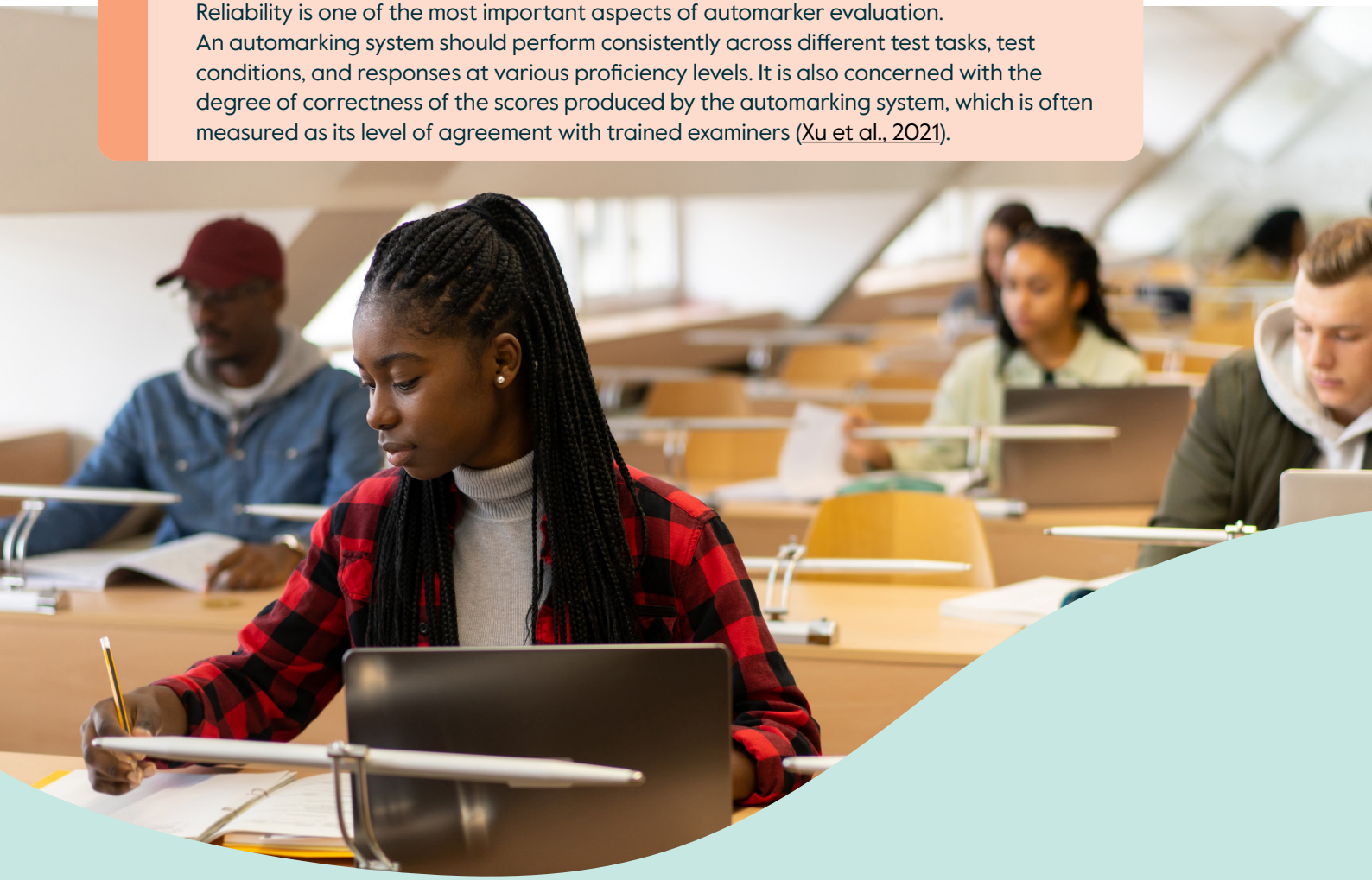
Context validity relates to how accurately the tasks in the test reflect the real-world language use of interest. In practice this means including diverse scenarios and tasks that are realistic and relevant to the assessment contexts.

The notion of context validity has implications for training and evaluating an automarker. Automarker training is always context-dependent as an automarker cannot be trained to be accurate for all task types or all test takers. The training data must be carefully selected to reflect the targeted test taker population and their various patterns of responses. An automarker that is well trained for one assessment context is not necessarily robust for another.

3 Reliability – To what extent can I trust the accuracy of the test scores?

Reliability (also called scoring validity) refers to the accuracy and consistency of the test results. For example, if a speaking or writing test is marked by two different examiners using the same assessment criteria, reliability is about the level of agreement between the two examiners.

Reliability is one of the most important aspects of automarker evaluation. An automarking system should perform consistently across different test tasks, test conditions, and responses at various proficiency levels. It is also concerned with the degree of correctness of the scores produced by the automarking system, which is often measured as its level of agreement with trained examiners (Xu et al., 2021).



4 Test impact – How will the test affect the way teachers teach, learners learn, and society more broadly?

Test impact (also called consequential validity) considers the broader impact of the test on stakeholders such as test takers, educators, and wider society. This includes both the intended and unintended consequences of test use. A valid assessment should promote positive educational outcomes, such as enhancing learning and guiding effective teaching practices, while avoiding negative educational outcomes such as disadvantaging certain groups of test takers (Cheng & Sultana, 2021).

To ensure fair opportunities for all test takers, automarkers need to be continuously evaluated and adjusted in accordance with the changing test taker population. In addition, the washback of using an automarker in a high-stakes language test on how test takers prepare for the test and how teachers teach language courses must be carefully investigated (Xi et al., 2016).

5 Criterion-related validity – How does the test performance compare with other measures?

Criterion-related validity examines the degree of alignment between test scores and other established measures of language proficiency, e.g. performance on other tests or in classroom settings.

If a language test is marked by an automarking system, its criterion-related validity may be evidenced by a high correlation between test takers' automated scores and their scores in another exam that assess the same test construct. Alternatively, criterion-related validity evidence could be gathered by examining the relationship between test takers' automated scores and their non-test performance in the real world such as academic achievement or teachers' observations.

Fairness

Fairness in language assessment is about making the entire testing process just, unbiased, and equitable for all (Walters, 2021; Xi, 2010). At its core, fairness is about minimising the influence of any factors irrelevant to the test construct. For example, if raters in a speaking test consistently assign higher or lower scores to speakers of a particular native language because of their familiarity/unfamiliarity with the accent, this would be an instance of unfairness. Such presence of systematic errors in test scores either in favour of or against certain subgroups of test takers is called **test bias** (AERA et al., 2014). Such biases need to be identified and removed so that test results accurately reflect test takers' true abilities. In speaking tests that involve examiners, rater bias can be reduced or eliminated through ongoing rigorous training (Davis, 2021).

Many stakeholders have responsibility in upholding fairness.

Test developers have a responsibility to create bias-free test content and marking procedures for all test takers, regardless of their native languages, gender, ethnic backgrounds, levels of education, or other personal characteristics.

Test developers are also responsible for providing clear guidelines on test administration and score interpretation. Based on these guidelines, test administrators are accountable for eliminating potential factors that could affect test takers and raters' performances, such as noise in the test environment, malpractice, and technical glitches.

Both test developers and administrators share a responsibility to accommodate test takers with special needs so that no test takers face barriers in demonstrating their true language abilities.

Fairness is also concerned with test takers having equal opportunities to learn and prepare for a test. Thus, test familiarisation materials such as descriptions of the test format, sample questions, and practice tests should be easily accessible.

Finally, test users involved with test-based decision-making assume the responsibility of interpreting test scores accurately and justifying the use of the test for specific contexts.

If you are interested in finding out more about validity and fairness, consult these seminal publications:

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

Chapelle, C. A. (2020). *Argument-based validation in testing and assessment*. Sage.

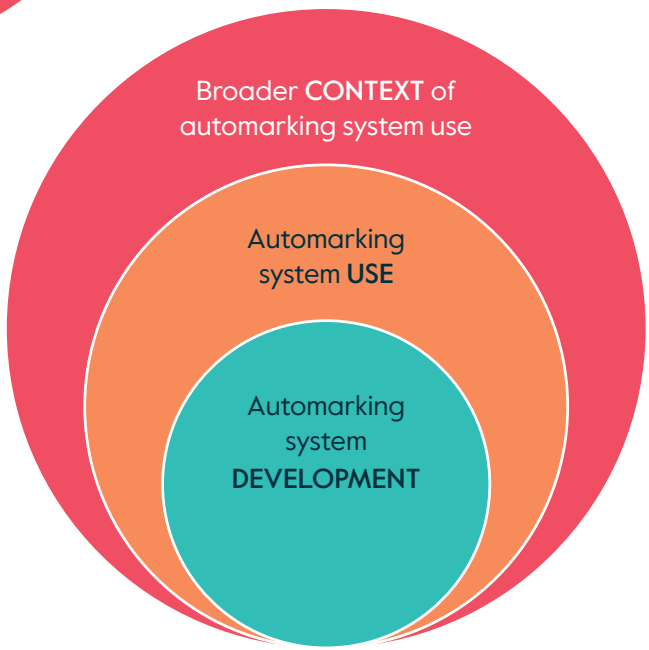
Walters, F. S. (2021). Ethics and fairness. In G. Fulcher & L. Harding (Eds.), *The Routledge handbook of language testing* (2nd ed.) (pp. 563–577). Routledge.

Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. Palgrave Macmillan.

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

Principles of good practice in automarking

The broad assessment principles we have discussed so far – validity and fairness – underpin three areas of importance for automarking in language tests: the **development** of the automarking system, its **use** in practice, and the **broader context** of its employment.



We propose **12 fundamental principles** that underpin the development, use, and broader contexts of automarked language tests.

Automarking system development

1 What data has the automarking system been trained on?	**Training data** The breadth of training data is critical for training an automarking system, since it determines how accurately the automarking system performs with a wide range of test takers. Carefully selected training data based on the test purpose ensures that the automarking system will display no positive/negative bias towards one demographic group over another.

- Training data needs to be gathered from a range of test taker performances, which include different ability levels, language backgrounds that are representative of the test taker population, a balanced sample of male/female test taker responses, and different age groups if relevant.
- Once trained, statistical analyses should be used to investigate potential biases in automarking on specific subgroups of test takers.
- Automarking system performance should be monitored continuously to ensure that it is reflective of changes in the test taker population, e.g. if the test taker population shifts to younger age ranges.

2 What scoring features are extracted to inform an automarked score?	**Scoring features** A range of features that represent the test construct (i.e. what the test aims to measure) adequately needs to be captured in an automarked test.

- The scoring features should extend beyond just grammar and vocabulary or pronunciation and fluency in the case of speaking. These linguistic features are foundational to language but not sufficient to capture the complex nature of language use.
- Discourse-level features related to organisation, relevance of content, and topic development need to be captured as well.

3 What is the accuracy level of the automarking system, and how is the accuracy evaluated?	**Accuracy evaluation** Ensuring that automarker results are accurate is one of the most important aspects of deploying automarking technology. A range of approaches to evaluating score accuracy should be employed to provide a full representation of the accuracy of the automarking system. This must be an ongoing process and not just a one-off activity because of the possible change in test taker behaviours over time.

- The accuracy of automarking should be evaluated by comparing computer marks with 'gold standard' marks provided by trained and certified human examiners.
- As a minimum, a 'gold standard' should be established through double marking of test responses by human examiners, i.e. each script/speaking response is independently marked by two examiners (ITC & ATP, 2022).
- Highly reliable and accurate 'gold standard' human marks should be used for automarking system evaluation. The 'gold standard' human marks can be obtained using statistical procedures, such as the Many-Facet Rasch Model. Such methodologies produce average scores adjusted for examiner severity and are therefore more accurate estimations of test takers' true abilities (Myford & Wolfe, 2003, 2004) than raw average scores. Alternatively, the 'gold standard' human marks may be obtained via adjudication in which a panel of expert markers discuss and resolve any discrepancies in marking.
- Measures showing human/machine agreement across the entire test scale used and not just on average should be implemented. This will address the fact that automarking system accuracy varies, sometimes substantially, across the proficiency levels of responses (Xu et al., 2021).
- Additional criteria for evaluating the performance of the automarking system should be included as well. Examples of these are: comparing the distribution of automarker scores to that of human scores (i.e. which values are common or uncommon); the likelihood of errors of automarker scores at different proficiency levels and across different types of test takers; how well the confidence scores provided by an automarking system are useful for indicating automarking errors; how well the findings of automarking system evaluation can be generalised to a much larger group of test takers not included in the evaluation.

Automarking system use in practice

4 What is the test design and what tasks are used in the automarked test?	**Test design** The design and tasks of an automarked language test should reflect the test purpose.

- For example, an automarked language test of communicative language ability should include a range of task types. The tasks need to go beyond highly controlled tasks (e.g. producing single words or short sentences, reading aloud, sentence repetition) to unrestricted tasks (e.g. essay-length writing, extended speech).
- The use of less controlled tasks ensures that a wider sample of language is elicited, leading to more robust interpretations about language proficiency.

5 What is the potential for malpractice on the automarked test?	**Malpractice detection** While the vast majority of test taker responses are legitimate attempts, measures are needed to identify any attempts to unfairly obtain a score.

- The malpractice detection methods need to target and flag a wide range of test-taking behaviours or strategies which are considered unusual and deviant significantly from expected or typical responses, e.g. unexpected context which is off-topic or irrelevant, repetitious language, prompt copying.
- The range of scoring features extracted also indirectly impacts the potential for cheating. If content relevance and topic development, for example, are included in the extracted features, then the potential for malpractice is reduced.

6 How is the automarked test administered in practice?	**Test administration** The accuracy of automarker output relies on appropriate test administration conditions which enable optimal test performance.

- Such conditions include high internet speed, provisions for taking the test offline, and clear test instructions.
- In the case of automarked speaking tests, the administration conditions need to additionally include minimal background noise

7 What is the degree of examiner involvement?	**Human involvement** To mitigate the limitations of automarking technology and uphold high standards of marking quality, a hybrid human/machine approach may be best suited for certain contexts.

- Robust metrics need to be employed, such as reliable automarking confidence measures which determine the level of trustworthiness of scores produced by automarking systems.
- Test responses which fall below confidence thresholds should be escalated to human marking to ensure the automarking system does not award or penalise test takers unjustly.

8 What is the degree of change of the automarking system in operational use?	**Automarking model change** Automarking systems need to apply rules consistently on each written/spoken response they encounter. This involves adherence to predetermined algorithms, thus maintaining human control and intent without allowing for adaptability or autonomy of the AI system when it is used.

- AI systems that learn and evolve on their own, as seen in Generative AI models, might be less suited to automarking test contexts, since there is less transparency in the process and the features underpinning the automarking decisions.
- Being in control of model change and also implementing change when needed (e.g. if the responses to be marked change during operational use) is also important.

Automarking system use in practice

9 How explainable are the automarker system decisions?	**Automarker system explainability** The 'black box' nature of automarking tools, i.e. the level of transparency into how an automarker reaches its decisions (also known as explainability and interpretability), is especially important when the automarking system is entrusted to make high-stakes decisions. Level of explainability needs to be a key consideration in decisions on which automarking model to deploy in which contexts.

- The interpretability of automarking systems is enhanced through visibility of the features in a written essay or spoken response which are the basis for automarking decisions.
- Automarking interpretability is not a dichotomous concept, and appropriate degrees of interpretability need to be explicitly defined, justified and adopted for different contexts. For example, in addition to scoring features, explicit consideration of the scoring rubrics, task types, and examiner role would be part of an approach which embraces explainability.

Broader context of using the automarking system

10 Is there a good fit between the purpose of the test and the automarking system used?	**Fitness-for-purpose** The use of an automarking system needs to be considered in the context of the test purpose and the stakes of the decisions made based on the test. There needs to be a suitably good fit between the test purpose and the use of automarking within that purpose.

- For example, using an automarked test in a low-stakes practice test context would involve different considerations compared to a high-stakes university entry test and the underlying automarking system used.

11 What is the impact of the automarked test on language learning, teaching and society?	**Impact** A test should have beneficial impact on classroom practices and learning (also known as washback), and positive impact on society more broadly.

- The broader the range of tasks and extracted features for scoring, the more representative the test will be of real-life use, and the higher the potential for positive impact.
- For example, a writing test which only focuses on narrow sentence-level tasks will have lower positive impact than a test which requires extended writing; a speaking test which includes mostly reading or repeating sentences, or short responses, will have lower positive impact than a test which includes extended speech.

12 How transparent is the information underpinning the automarked test?	**Transparency** How automarking systems work and what their scores mean must be easily understood by everyone, from regulatory bodies to other stakeholders such as examiners, teachers, learners, parents. This information needs to be publicly available.

- Commitment to transparency must be evident in the information provided for all stakeholders.
- The information should be available in open-access technical reports or research articles that are published in peer-reviewed journals, and should be aimed at a range of audiences, such as peer-reviewed academic publications aimed at technical experts, and papers or other content (e.g. videos such as this one) for non-technical stakeholders.

If you're interested in finding out more about automarking and related validity and fairness issues, consult these seminal publications:

Khabbazbashi, N., Xu, J., & Galaczi, E. (2021). Opening the black box: Exploring automated speaking evaluation. In B. Lanteigne, C. Coombe & J. D. Brown (Eds.), *Challenges in language testing around the world* (pp. 333–343). Springer. https://doi.org/10.1007/978-981-33-4232-3

Shermis, M. D., & Burstein, J. (2013). *Handbook of automated essay evaluation*. Routledge.

van Moere, A., & Downey, R. (2016). Technology and artificial intelligence in language assessment. In D. Tsagari & J. Banerjee (Eds.), *Handbook of second language assessment* (pp. 341–358). Walter de Gruyter. https://doi.org/10.1515/9781614513827-023

Xi, X. (2021). 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.

Yan, D., Rupp, A. A., & Foltz, P. W. (Eds.). (2020). *Handbook of Automated Scoring: Theory into Practice*. Taylor & Francis.

From principles to practice: A case study

We now turn to a concrete example of how the 12 principles of good practice outlined above are implemented in operational conditions: the use of automarking in Cambridge writing exams. Cambridge University Press & Assessment employs various automarking systems tailored to specific exams, each trained on unique datasets that reflect diverse test taker demographics and task types. In this section we will describe our general approach to the development of these systems, and in certain cases will use the automarker for the Cambridge B2 First exam as an illustration.

How does the Cambridge automarking system work?

The Cambridge automarking system is a consistent system that behaves exactly the same way every time it marks a script. It uses natural language processing (NLP) techniques to extract language features from essays and applies a complex set of scoring rules to these features to evaluate the quality of writing against a set of scoring criteria. The deployed model is trained offline and does not carry out any further learning or evolve by itself while being used in operational tests.

The Cambridge automarking system contains three different components (i.e. models):

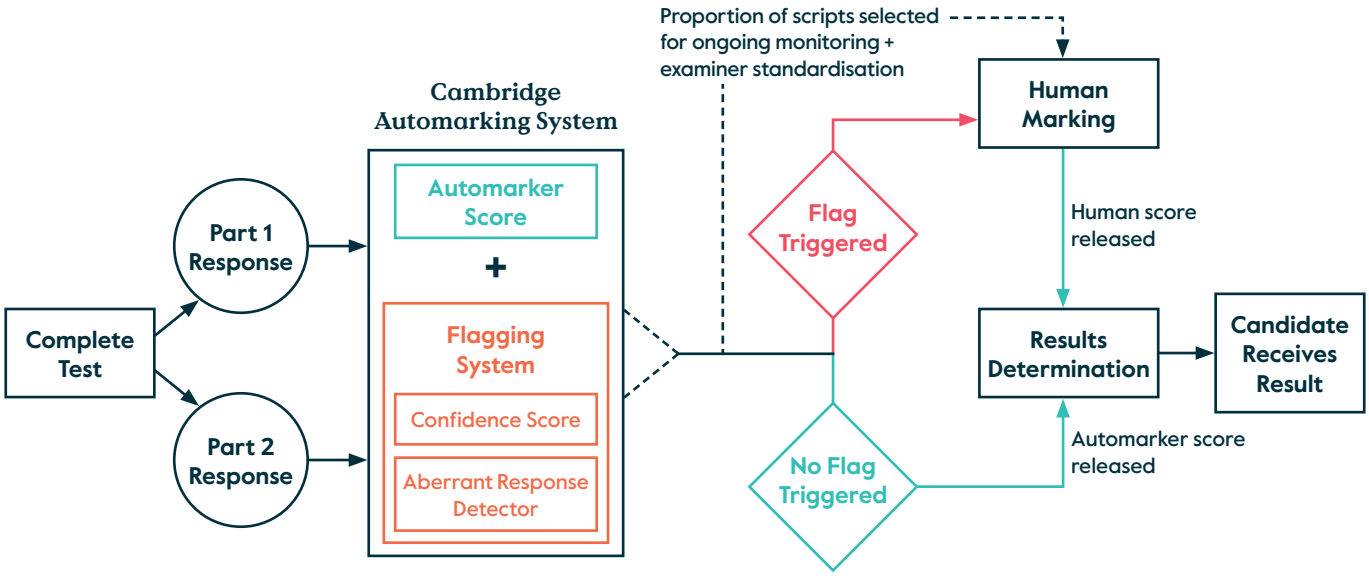
- a scoring model
- a confidence model
- a model to detect aberrant responses.

When deployed operationally, individual test taker responses are entered into the automarking system. The automarker tags the test taker's response according to a range of defined computational features that we have developed for the purpose. These features represent the writing construct and the aspects of language that Cambridge examiners would look for in assessing writing. In the case of Cambridge B2 First, this process includes identifying up to approximately 330,000 language features which focus on the language and organisation of the response (e.g. vocabulary, sentence length, grammatical relationships, cohesive devices, readability measures, relevance to the prompt) and errors (e.g. spellings, unknown words, grammatical errors). After the automarking system tags the response for features, the features are weighted and summarised to give a score.

Some of the computational features the automarker uses are directly related to the assessment criteria human examiners use, while others apply more broadly across multiple different criteria. For example, there is a unique link in use of cohesive devices as a feature for 'Organisation', or prompt relevance for 'Content'. Examples of features that span multiple different criteria are grammatical sequences which contribute to 'Communicative Achievement', 'Organisation', and 'Language'. Similar to human examiners, the automarker uses features that are easily quantifiable such as lexico-grammatical errors for the 'Language' criterion while 'Communicative Achievement' and 'Organisation' are assessed by features that are proxies for success, such as readability scores for 'Communicative Achievement'. In this approach, humans and automarkers are all using a range of specific and proxy examples to form a view based on their training, creating a degree of comparability across the two approaches used by machine and by humans.

In summary, our automarking process involves five stages:

1. All responses are marked by the trained automarking system, which includes a scoring model, a confidence model and a model to detect aberrant responses.
2. Those with a high confidence metric and not flagged as anomalous are released to test takers.
3. Marks which do not meet our confidence threshold are all sent to our pool of examiners for allocation and marking. Any scripts flagged as potentially anomalous are also sent to human markers for review. In cases of discrepancies between human and automarking system scores, the test taker receives the human score.
4. Additionally, a random selection of responses is sent for human marking regardless of the score confidence metric. This helps to ensure examiners are exposed to a full range of scripts, rather than just those flagged by the confidence system, to maintain their standard of marking.
5. Human marking is then subject to further statistical monitoring, and reviewed as with our current marking practices.



In addition to these steps, ongoing monitoring and validation of the automarking system performance is also implemented, to ensure it continues to behave robustly across different test settings.

We now turn to overviewing in more detail essential steps in our development and use of the automarking system.

How did we train the automarking system?

The training process includes carefully selected data which is representative of our targeted test taker population to ensure that our model is generalisable and unbiased.

To train the automarker model, we use a **comprehensive set of features related to L2 writing, along with operational examiner scores**. During the training phase, the model establishes which features are represented in the training data and how they combine to characterise test taker writing at specific proficiency levels. These levels are defined by the scores provided by certified examiners. In other words, during training the automarking system works to systematically learn which features are associated with the corresponding examiner scores. For this, the automarking system is trained on tens of thousands of scripts.

This large-scale data allows the model to determine the weightings of each feature. This means that it determines how and how much each feature contributes to an overall assessment of writing proficiency for the particular exam. The initial objective in this phase is to ensure that the automarking system can rank a set of scripts and thus accurately distinguish stronger performances from weaker ones in a robust way consistent with our examiners.

As well as training the model to predict the examiner scores, the Cambridge automarking system uses a **confidence score model to determine how well the automarker performs**. That involves a *confidence measure*, and an indication of how confident we can be in the scores the automarker generates. The automarking system is trained to identify which scripts/scores are unlikely to be a strong prediction of an examiner score. In such cases the automarking system recognises its own limitations and provides us with a low confidence score. Any scripts with a low confidence score are automatically sent for marking by human examiners. This confidence measure is used to ensure that we only release automarked scores where we are highly confident that the automarking system will give a reliable outcome, i.e. the same results that the test taker can expect to receive from a certified human examiner. The confidence measure threshold can be dialled up or down, depending on the purpose of the test.

The automarking system also includes a range of **aberrant response detectors**. Aberrant responses are those which may be unusual in some way, and hence more likely to lead to automarking inaccuracies. Examples of this include detection of repeated text, copying of the prompt, or non-English words. If a test taker response is flagged as aberrant, it is automatically sent to a human examiner for review. This helps to ensure that the system remains a valid assessment of writing and withstands atypical attempts on which the automarking system hasn't been trained sufficiently.

How did we evaluate performance of the automarking system?

The evaluation of automarking system performance involves examining the accuracy of the scoring and confidence models, and scrutinising the automarking system for bias across various demographic groups. The process involves several steps which seek to ensure we have a consistent, reliable automarking system that applies the correct standards.

- We evaluate the general performance of the **scoring** model – to ensure it is able to score responses accurately.
- We look at the performance of our **confidence** model – how well we can identify responses where the automarker is performing to the required standard, and where it is not. This allows us to incorporate human marking efficiently, to ensure that test takers get a fair outcome.
- We evaluate the overall system of marking to confirm that **combining human marking and automarking** gives consistent outcomes and preserves the overall standard of performance on a par with existing operational marking.
- We conduct analyses to **monitor the performance** of the automarking system against different demographic groups to guard against any bias.

As part of the evaluation stage, we generate multiple models which we evaluate and choose the model that best fits, i.e. the one that aligns most closely with human examiner scores. For the initial evaluation of automarking models we generate a 'gold standard' data set of scripts. For the Cambridge B2 First exam, this set comprises a sample of 1,700 test taker responses that represent the current distribution of test taker scores. It is also adjusted to ensure that we have adequate coverage of the more extreme scores to ensure it works across the full range of potential performance and is robust to potential future changes in demographics and/or ability. These scripts also cover a range of different topics and all the different task types a test taker will see, and are drawn from multiple separate administrations of the tests.

These scripts are then multi-marked by a representative team of examiners from our operational marking pool. Using the Many-Facet Rasch Model (which is an established method in the assessment field to arrive at fair average scores for test takers), we determine a 'gold standard' score for each script that best represents a test taker's true ability. We then use this gold standard set of marks to compare the performance of our automarking system and the performance of our current operational examiner pool. We aim to show that automarking is comparable with human marking in terms of score accuracy, and thus is appropriate and fair to use in a live test where a test taker's script may be marked by either the automarking system or a human or both.

When comparing the performance of the automarking system and the examiner gold standard we look at a range of measures:

- RMSE (Root-Mean-Square-Error, which calculates the average difference between a model's predicted score and its actual score) is used to measure the overall level of agreement between two sets of marks to give a general summary of model accuracy. The smaller the value of RMSE is, the more accurate the model is.
- A range of agreement indices on classifications (e.g. CEFR classifications, pass/fail classifications, A/B/C grade classifications) are also generated to ensure that we get consistent outcomes from the automarking system. For example, we analyse how many test takers would get the same outcome in terms of pass/fail decisions when marked by machine or human, or in terms of A/B/C grades or CEFR levels.

The above measures provide an *overall* assessment of the performance of the automarking system across the whole range. We also review how the automarking system performs at *different score points* within the overall distribution, since its performance at the lower or higher ends of the score distribution may be different – similar to how a human may vary when they have less information/experience with specific contexts.

Part of the evaluation involves the use of confidence metrics. Automarker confidence scores are invaluable in the use of the automarking system in this part of the process, as part of a hybrid model. In our hybrid approach, the automarking system will not be the sole evaluator of scripts in all instances; human examiners will get involved in cases where automarker confidence scores are low. With our confidence score indicators, we define an optimal threshold that ensures the best overall marking outcomes, without negatively affecting individual test takers. By adjusting the threshold, we seek to find the point at which we can release automarking system scores without human review. For any script below the threshold, we believe it is more likely that the human marking would result in a fairer score; for any script above the threshold, we are confident that it would receive the same result if it was marked by our examiners.

The final stage of our evaluation is to apply the hybrid marking system, automarker, confidence metrics and human scores to an operational setting to ensure that the overall outcomes are consistent with the current human marking practice, and that the same marking standards can be maintained after hybrid marking is introduced.

The models we have developed thus far perform well on these measures and are on a par with how our examiner pool performs on the same measures.

How is the Cambridge automarking system used live?

Whilst the automarker enables us to meet our customer needs in terms of speed of results and availability of tests, we need to ensure that the results remain fair and valid. Thus, we use a combination of automarking and human marking – our **hybrid marking model** – to optimise all of these requirements. This approach allows us to **use the best of both automarking and human marking to get the optimal outcome for our test takers**, in the most efficient way.

In operational conditions the Cambridge automarking system directly releases results to test takers in cases where it is confident in the awarded scores; in cases where it is uncertain, the marks get sent to an examiner; additionally, a randomly chosen sample is sent to examiners for human review.

How does the Cambridge automarking system compare with a human examiner?

Similar to a human examiner, the automarking system identifies features from the writing scripts to assess a test taker's language proficiency, such as (but not limited to) relevance to the prompt, readability measures, cohesive devices, vocabulary, grammatical relationships and grammatical errors. These features relate to the four main rating criteria human examiners use in B2 First. The automarking system attaches a weighting to the features for each criterion in order to produce the test taker's final score for that task.

This **automarking process is intrinsically similar to the marking process carried out by a human examiner, especially since both are marking based on the same mark scheme**. An examiner reads a response and establishes how much evidence of different features the text contains across the marking criteria. Specifically, examiners identify textual features in the responses that are related to each marking criterion and arrive at the final mark based on the descriptors in the mark scheme (Lumley, 2005). The use of a range of features and their corresponding weightings ensures that no one feature (e.g. text length) is solely responsible for a mark and that examiners cover the whole breadth of the writing construct. The fact that all responses, regardless of whether they are assessed by a human or an automarking system, are covered by a comprehensive mark scheme underscores the reliability and fairness of the assessment process.

The **key distinction between humans and the automarking system lies in the scale and efficiency enabled by the automarking system's extensive training on a vast dataset**. The automarking system is trained on a diverse range of responses, which encompass a broader spectrum of writing styles and nuances than a cohort of human examiners could feasibly encounter. As a result, the automarking system can apply these rules more consistently, and as accurately, as a group of human examiners across a multitude of responses. This advanced capability stems from the automarking system's ability to process and analyse large volumes of data, allowing for an in-depth analysis of language use and writing proficiency that contributes to a robust and reliable assessment mechanism. Additionally, the automarking system is only a single rater, applying the same approach time and time again, so has less inherent variation than a pool of examiners. As a result, we can achieve greater consistency more efficiently, with a reduced need for monitoring and quality checks to review and remark specific scripts.

What are the key standards underpinning the quality of the Cambridge automarking system?

The automarking of Cambridge tests is enabled by a number of principles of good practice, which ensure that our automarking system is suitable for high-stakes assessment contexts and produces scores which are as reliable as scores from trained and certified examiners because of the standards in its development and use:

- **comprehensive training** ensures the automarking system has been exposed to various writing styles and intricacies, contributing to a thorough training process
- the automarking system provides a **consistent and reliable** evaluation which is much quicker compared to human markers
- the extensive sampling in training ensures **fairness and lack of bias** in the automarking system
- the automarking system covers a **broad spectrum of language features** essential for our assessments, offering a comprehensive analysis that aligns with the evaluation criteria used by examiners
- while the performance of the automarking system is at least on par with humans, we maintain the **flexibility to incorporate human markers** in situations where their judgment is likely to yield a more reliable outcome. This hybrid approach underpins the overall quality of the assessment process, combining the efficiency of automation with the nuanced insights provided by human examiners.

What next for automarked language tests?

The need for efficiency and marking speed in L2 tests is here to stay. At the same time, principles of validity and fairness need to remain an important anchor to good practice in L2 assessment. Interesting new developments in Generative AI will inevitably impact assessment, with the promise to provide content generation at speed and at scale, automarking, personalisation in learning and assessment, and auto-generated feedback.

Generative AI risks and concerns need to be kept front of mind, since they raise the legal and ethical stakes of use in L2 education. Those risks involve the likelihood of digital hallucinations (i.e. inaccurate content), the risk of amplifying prejudice and biases, the risk to assessment integrity, the interpretability of AI models, the Intellectual Property and copyright considerations, and environmental impact (see the 'Generative AI and Language Education: Opportunities, Challenges and the Need for Critical Perspectives' paper of Cambridge Papers in English Language Education).

It is important, therefore, as we move forward into the exciting and uncharted terrain of AI models, to be mindful of the need to control the associated risks while leveraging the benefits. That needs to be done through human-centred AI which puts people first. And in the case of automarking systems, that entails ensuring the judicious involvement of human examiners alongside automarking models.

Acknowledgement

We must note that our work on this paper has been aided significantly by many people. We would like to express appreciation to Mark Brenchley, Yan Huang, Graham Seed, Andrea Vinkler, Nick Saville and Angela Wright for reviewing the earlier drafts of the paper and for their highly constructive feedback. We also humbly acknowledge that our summary of Cambridge's approach to automarking rests upon the diligent and excellent work of many frontline NLP engineers and language assessment researchers such as Ian Iewin, Shilin Gao, Saeid Mokaram, Edmund Jones and Trevor Breakspear. We are also grateful to John Savage for his helpful editorial support.

References

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

Cheng, L., & Sultana, N. (2021). Washback: Looking backward and forward. In G. Fulcher & L. Harding (Eds.), *The Routledge handbook of language testing* (2nd ed.) (pp. 136–152). Routledge.

Davis, L. (2021). Rater and interlocutor training. In G. Fulcher & L. Harding (Eds.), *The Routledge handbook of language testing* (2nd ed.) (pp. 322–338). Routledge.

Gao, S., Gales, M., & Xu, J. (2024). Detecting aberrant responses in automated L2 spoken English assessment. In C. Chapelle, G. H. Beckett, & J. Ranalli (Eds.), *Exploring artificial intelligence in applied linguistics* (pp. 96–117). Iowa State University Digital Press.

ITC, & ATP. (2022). *The ITC/ATP guidelines for technology-based assessment*. ATP. https://www.intestcom.org/upload/media-library/tba-guidelines-final-2-23-2023-v4-167785144642TgY.pdf

Lumley, T. (2005). *Assessing second language writing: The rater's perspective*. Volume 3. Peter Lang.

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet rasch measurement: Part I. *Journal of Applied Measurement, 4*(4), 386–422. https://pubmed.ncbi.nlm.nih.gov/14523257/

Myford, C. M., & Wolfe, E. W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *Journal of Applied Measurement*, *5*(2), 189–227. https://psycnet.apa.org/record/2004-13366-007

Walters, F. S. (2021). Ethics and fairness. In G. Fulcher & L. Harding (Eds.), *The Routledge handbook of language testing* (2nd ed.) (pp. 563–577). Routledge.

Weir, C. J. (2005). *Language testing and validation: An evidence-based approach.* Palgrave Macmillan.

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

Xi, X. (2021). 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.

Xu, J., Jones, E., Laxton, V., & Galaczi, E. (2021). Assessing L2 English speaking using automated scoring technology: Examining automarker reliability. *Assessment in education: Principles, policy & practice*, *28*(4), 411–436. https://doi.org/https://doi.org/10.1080/0969594X.2021.1979467

Find out more at
cambridge.org/english

We believe that English can unlock a lifetime of experiences and, together with teachers and our partners, we help people to learn and confidently prove their skills to the world.

Where your world grows

CAMBRIDGE
UNIVERSITY PRESS & ASSESSMENT

All details are correct at the time of going to print in December 2024.

ENG_07213_V1_Dec24_TemplateCPELE