

Using Generative AI Responsibly in Assessment Content Creation: Cambridge Principles and Practice



Ron Zeronis, Andrew Mullooly
and Andrew Somers



Contents

	Page
Introduction and context	4
Generative AI in assessment	5
Principles guiding Cambridge use of GenAI	7
Implementing the principles in practice	10
Illustrative case studies	16
Conclusion	19
References	21
Author biographies	23

Introduction and context

Generative artificial intelligence (GenAI) is now a familiar presence in everyday life. It is often used to generate content that is fluent and quickly produced, but also variable in quality and reliability. From automated news summaries to click-driven text, public experience has shown that AI-generated content can be efficient and persuasive but also misleading or not always fit for purpose. These tensions matter acutely in educational assessment, where content carries real consequences for learners' progression, certification and life opportunities. When AI is used to support assessment content creation, the question is therefore not simply whether it improves efficiency, but whether it can be **used in ways that preserve validity, fairness, trust and public confidence**.

This concern is reflected in emerging regulatory and ethical frameworks (e.g., the EU AI Act, 2024; Galaczi & Pastorino Campos, 2025) and is particularly pressing in high-stakes assessment contexts.

In this paper, assessment content refers to the texts, tasks, items, prompts, audio and visual materials that are designed to elicit evidence of learners' language ability for defined assessment purposes.

The use of GenAI to support the creation of this type of content presents both opportunities and risks that are now the subject of active debate within the language assessment field. GenAI offers the potential to support innovation, increase efficiency, and broaden the range of topics and voices represented in assessments. At the same time, modern GenAI systems do not simply assemble outputs at random or on a word by word basis. They generate content by drawing on complex patterns learned from large volumes of training data, producing material that reflects common structures, conventions and associations present in those data. While this

enables fluency and coherence in generated assessment content, it also raises real concerns that outputs may reproduce the assumptions, regularities and limitations of their training sources. This is particularly true in assessment contexts where errors or distortions cannot be easily tolerated.

Cambridge takes a clear position within that debate. We argue that GenAI can play a legitimate and valuable role in assessment content creation only when its use is guided by explicit ethical and assessment principles, robust human oversight, and empirically-grounded quality assurance processes. Our aim is therefore twofold: first, to identify principles of good practice for AI-mediated assessment content creation, grounded in established work in ethical AI and language assessment; and second, to illustrate how these principles are operationalised in practice through the Cambridge approach to designing, governing and monitoring GenAI-assisted content creation. Specifically, we:

- outline why GenAI raises distinctive opportunities and risks for assessment content creation
- articulate key ethical and assessment principles that should guide the responsible use of AI in this domain
- describe how these principles are embedded in Cambridge's GenAI-enabled content creation tools, workflows and governance processes
- explain how human oversight, quality assurance and psychometric validation are maintained at all stages
- present empirical case studies illustrating both the potential and the limits of GenAI in assessment contexts.

This paper sits alongside related Cambridge research on AI-supported scoring and automated marking (Xu et al., 2024), and forms part of our broader, integrated approach to the responsible use of AI across assessment processes.

Generative AI in assessment

Recent advances in GenAI have prompted renewed scrutiny of how assessment content is designed, generated and validated within the field of language assessment. A growing body of research now charts both the rapid uptake of large language models (LLMs) and the tensions they introduce for long-established assessment principles. A recent systematic review by [Chuang and Yan \(2025\)](#), synthesising 77 studies on GenAI in language testing and assessment, highlights a clear pattern: while generative tools are being increasingly experimented with across different stages of the assessment cycle, their implications for validity, fairness and governance remain patchily theorised across the literature and empirically limited. Much existing work has focused on classroom-based and low-stakes contexts, with far less publicly available evidence addressing large-scale assessment, where the consequences of error, bias or construct distortion are greatest.

From a validity perspective, GenAI presents challenges for assessment at precisely the point where validity theory places greatest emphasis: the principled representation of constructs (that is, the intentional and theory-based translation of defined language abilities into assessment tasks). Contemporary validity frameworks align on the view that test quality is grounded not in any single procedure or technology, but in the strength of an integrated validity argument linking test design, content, performance, interpretation and consequences ([Messick, 1989](#); [Kane, 2006](#); [Weir, 2005](#); [Chalhoub Deville & O'Sullivan, 2020](#)). Within this tradition, **a persistent risk in AI-generated**

content for language assessment is construct under-representation, or the failure to adequately sample the domain of language ability intended to be measured. Because generative models produce output based on statistical regularities in training data rather than on an understanding of language constructs or test purpose, there is a danger that AI-generated content emphasises frequent, generic or highly common, statistically prominent patterns at the expense of systematic construct coverage. Left unchecked, this can result in content that appears coherent and fluent, yet misaligns with intended testing focuses, level expectations or cognitive demands intended by the test producer.

These concerns are not unique to GenAI but are already recognised within professional standards for technology-based assessment. The recent [ATP/ITC Guidelines for technology-based assessment \(2025\)](#) explicitly frame risks such as construct under-representation and construct-irrelevant variance (variation caused by factors unrelated to the language skills being measured) as central validity threats in digital testing contexts, emphasising that technology should be adopted only where it enhances construct representation without changing what test scores meaningfully indicate about language ability or disadvantaging subgroups. The Guidelines further position human oversight, transparency and ongoing validation as necessary conditions for maintaining validity and fairness when automation or AI-based systems are introduced into assessment design, content creation or scoring processes.

A persistent risk in AI-generated content for language assessment is *construct under-representation*: the failure to adequately sample the domain of language ability intended to be measured.

At the same time, recent research also demonstrates why GenAI has attracted such intense interest from assessment organisations. Studies reviewed by Chuang and Yan (2025), as well as related work in AI-supported language assessment and item generation (e.g., Pastorino-Campos & Saville, 2025; UNESCO, 2023), indicate that, when carefully constrained, generative models can support rapid content drafting, the creation of parallel task variants (where an alternative version of a task is designed to be equivalent in difficulty and construct coverage), and the reduction of repetitive work for subject matter experts. These efficiencies are particularly evident in item generation and text-editing tasks, provided that human experts remain centrally involved in specifying constructs, reviewing outputs and validating performance. **The literature therefore increasingly converges on the insight that GenAI is neither inherently incompatible with sound assessment practice, but nor is it inherently safe. Its impact depends on how, where and under whose control it is deployed.**

Ethical and policy-oriented analyses reinforce this conclusion. Reviews of ethical AI frameworks in education consistently foreground principles such as human oversight, fairness, transparency, accountability and proportionality (e.g. UNESCO, 2023; Fjeld et al., 2020). In language assessment specifically, recent work by Pastorino-Campos and Galaczi (2025) articulates how these ethical considerations intersect with core assessment values, arguing that validity and reliability are not just technical qualities but ethical obligations in testing contexts. From this perspective, the use of GenAI in assessment content creation cannot be evaluated solely in terms of efficiency or innovation but must be judged in terms of its consequences for test takers, learners, score users and the level of public trust in assessment systems.

For large-scale assessment providers, these issues take on heightened significance. The scale at which content is produced and decisions are made means

that even small, systematic distortions introduced through AI-mediated processes can have wide-ranging effects. Assessment organisations therefore carry both a professional and social responsibility to ensure that emerging technologies do not erode established standards of quality, fairness and accountability. As validity theorists have long argued, responsible assessment practice requires anticipating consequences by design rather than treating them as an afterthought (Messick, 1989; Saville, 2009; Chalhoub Deville & O'Sullivan, 2020). In the context of GenAI, this implies the need for principled frameworks that explicitly integrate ethical AI considerations with assessment theory, and that embed human judgement, transparency and empirical validation throughout the content creation process.

Taken together, the literature points to a clear implication for the field.

Responsible use of GenAI in assessment cannot be achieved through isolated safeguards, informal experimentation or post-hoc review alone. It requires a coherent set of principles that align ethical AI frameworks with core assessment standards, and that embed human oversight alongside operational approaches which leverage assessment expertise to translate those principles into concrete design and governance decisions.

In the remainder of this paper, we articulate the principles that should govern the use of GenAI in assessment content creation and show how these are operationalised through human-centred tooling, workflows, quality assurance and validation practices.

Principles guiding Cambridge use of GenAI

Cambridge University Press & Assessment approaches the use of GenAI through the same principled lens that underpins all our learning, teaching and assessment activities. **Our use of GenAI is therefore guided not by the novelty or efficiency of the technology, but by our long-established commitments to fitness for purpose, fairness, quality, people and expertise, and sustainability**, as set out in our *Principles of Good Practice* (Caudwell & Somers, 2026).

We don't treat GenAI as a distinct or exceptional category of practice. Rather, in our view, it is one of a number of technologies that may be used where appropriate to support human expertise and judgement. Decisions about its use are therefore grounded first and foremost in educational, assessment and ethical considerations, rather than in technological capability alone.

The principles below articulate how Cambridge applies our existing values and standards to the specific opportunities and risks presented by GenAI in assessment contexts.

These principles are informed by, and build on, recent Cambridge-led research and thought leadership on ethical AI in language assessment. They draw on the ethical framework articulated by Galaczi and Pastorino-Campos (2025), which lays out a domain-specific set of ethical principles and considerations for the use of AI in high-stakes language assessment. They also reflect applied work by Pastorino-Campos and Saville (2025), which translates ethical and assessment theoretic concerns into guiding principles for the practical use of AI in language learning and assessment, with particular emphasis on validity, fairness, human control, transparency and sustainability.

The principles below are distilled from these strands of work and operationalised for the specific context of AI-assisted assessment content creation.

Principle 1: Validity and Fitness for Purpose

The centrality of validity when introducing AI into assessment is well established in recent assessment-focused ethical frameworks (e.g., ATP/ITC, 2025; Galaczi & Pastorino-Campos, 2025; Pastorino-Campos & Saville, 2025).

Any use of GenAI in learning, teaching or assessment, including assessment content creation, must support the intended purpose of the product or process and must not compromise the validity of the inferences drawn from it.

In assessment contexts, this means that GenAI may only be used where there is clear evidence that it supports accurate measurement of the intended construct and does not introduce construct-irrelevant variance or systematic bias, i.e., consistent patterns that unfairly advantage or disadvantage particular groups or responses. AI systems must meet the same standards of performance, consistency and appropriateness that Cambridge requires of human-designed and human-mediated processes.

Crucially, the introduction of GenAI must not drive changes to constructs, task design or scoring practices simply to accommodate technological constraints. Technology must adapt to assessment principles, not the reverse. However, in contexts where accessibility, affordability, or scalability are explicit design priorities, GenAI may inform the development of assessment approaches that use different but still credible content types, provided the resulting constructs are clearly defined, theoretically defensible, and demonstrably aligned to their intended purposes.

Principle 2: Human Centred Design and Oversight

There is a consistent emphasis in both ethical and professional guidance of the notion that AI systems should augment, not replace, human judgement in assessment decision-making (Pastorino-Campos & Saville, 2025).

Language learning and assessment are fundamentally human endeavours. GenAI therefore should be used to support, not replace, human expertise in the design, generation and review of assessment content, professional judgement and responsibility. This includes embedding assessment domain expertise into the design of GenAI-enabled tools and workflows, so that systems reflect established assessment principles, constraints and quality requirements from the outset.

Human oversight should be maintained at all stages of the AI system lifecycle, including design, development, deployment, monitoring and review. In assessment contexts, meaningful human involvement in decision-making is essential, particularly where outcomes may affect educational progression, employment or migration.

This principle explicitly recognises that accountability for decisions remains with people, not systems. GenAI outputs should be treated as inputs to human decision-making, rather than as autonomous or unquestionable outcomes.

Principle 3: Fairness, Equity and Inclusion

Cambridge is committed to ensuring that the use of GenAI does not disadvantage individuals or groups based on language background, culture, disability, socio-economic status or access to technology.

GenAI systems must be designed, trained, evaluated and monitored with explicit attention to fairness and bias, including the representativeness and appropriateness of training data and the differential performance of systems across learner groups. Where risks to fairness are identified, these must be mitigated or the use of the technology reconsidered.

For assessment content, this requires systematic attention to how topics, contexts, language features and representations accumulate across item banks, so that AI-supported generation does not introduce recurring patterns that privilege or disadvantage particular groups or experiences.

Equity considerations also apply to the ways in which GenAI is used in assessment content creation. Decisions about prompts, training sources, and human review processes need to take account of the risk that generated content may systematically privilege particular cultural norms, linguistic registers, or educational assumptions. Without careful oversight, such patterns can introduce bias or unnecessary barriers, affecting who can validly demonstrate the intended assessment outcomes.

Principle 4: Transparency and Explainability

Cambridge is committed to being open and clear about where and how GenAI is used in its products and services. Therefore, the use of GenAI in assessment content creation should be transparent and explainable at each stage of the content lifecycle. This includes clarity about where and how GenAI has been used in the development of assessment materials, the extent of human oversight and editorial judgement applied, and the rationale for decisions affecting task design, difficulty, and alignment to intended learning outcomes.

For assessment content, explainability is particularly important where GenAI contributes to item generation, adaptation, or modification. Assessment designers and governance bodies should be able to account for how generated content meets construct definitions, avoids bias, and adheres to established standards of validity, fairness, and reliability. Transparency in these processes supports confidence in the assessment, enables meaningful quality assurance and review, and ensures that responsibility for assessment judgements remains with human decision-makers.

Principle 5: Accountability and Governance

Ethical use should be an ongoing organisational responsibility. AI-enabled processes should be subject to regular monitoring, evaluation and audit, and be reviewed considering emerging evidence, regulation and societal expectations. Where risks outweigh benefits, or where acceptable standards cannot be met, GenAI will not be deployed.

For assessment content, this means that decisions about the creation, modification, deployment or withdrawal of AI-supported material should be made by formally accountable roles, supported by traceable evidence and subject to documented review and escalation processes.

The use of GenAI sits within Cambridge's existing governance and quality management and assurance systems. Responsibilities for decisions, outcomes and error remediation are clearly defined and documented.

Principle 6: Proportionality and Sustainability

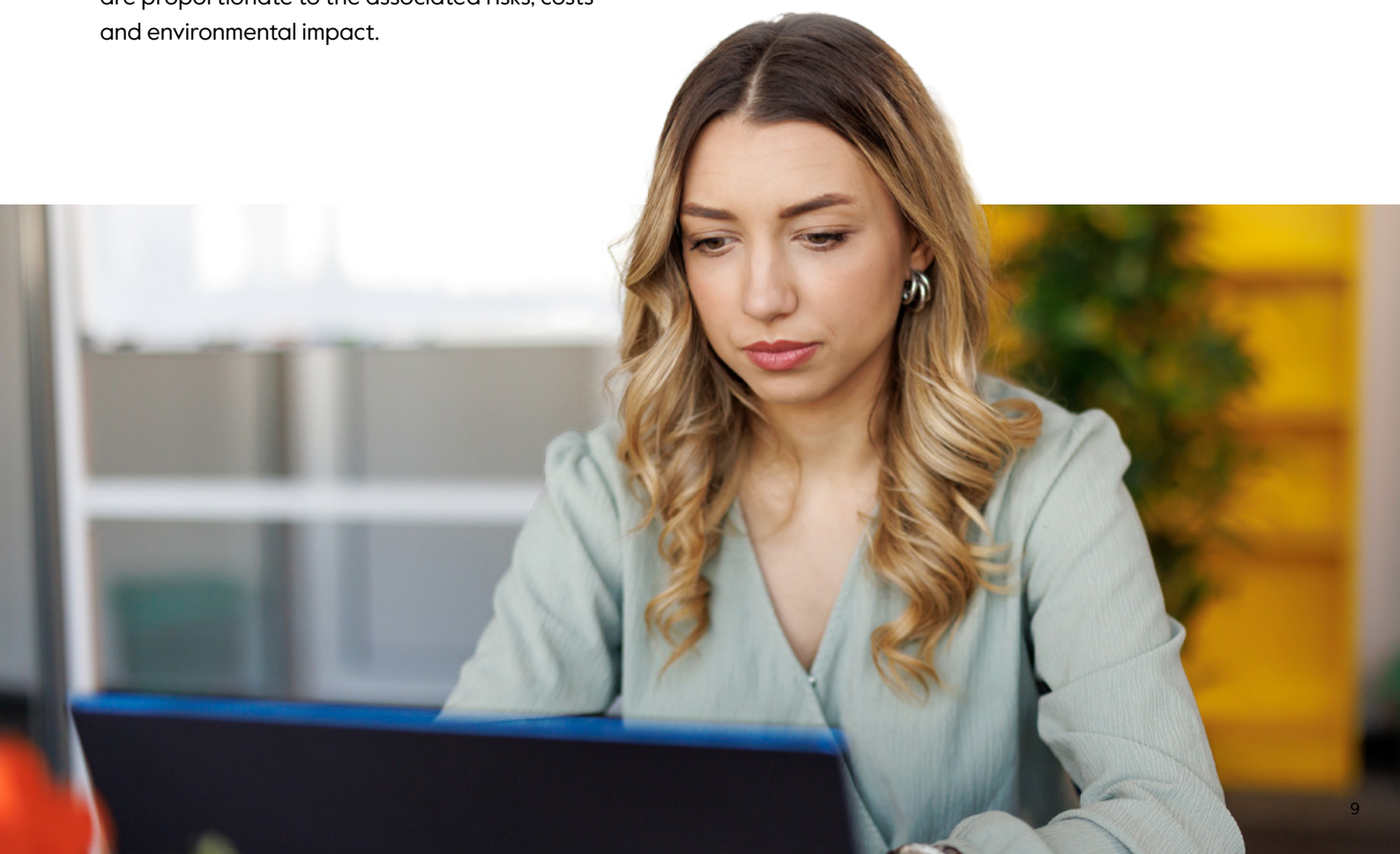
GenAI is used only where it offers clear educational or assessment benefit and where those benefits are proportionate to the associated risks, costs and environmental impact.

Cambridge recognises that different forms of AI vary significantly in their computational, ecological and social footprint. Decisions about adopting GenAI therefore involve consideration of sustainability alongside effectiveness, including whether simpler or more established approaches could achieve the same purpose with lower impact.

Where GenAI is used, tools and approaches must be designed and selected to minimise environmental impact, taking into account factors such as computational efficiency, frequency of use and the availability of lower impact alternatives. This includes incorporating clear assessment informed constraints, guidance and rules that support high-quality 'right first time' generation, reducing the need for repeated or exploratory model runs to achieve acceptable assessment content.

For assessment content creation, proportional use of GenAI means applying it where it adds clear value to the development and maintenance of assessment materials, in ways that are appropriate to the assessment's purpose and stakes and that support sustainable item development without compromising quality or governance.

This principle aligns the use of emerging technologies with Cambridge's wider commitments to sustainability and responsible innovation.



Implementing the principles in practice

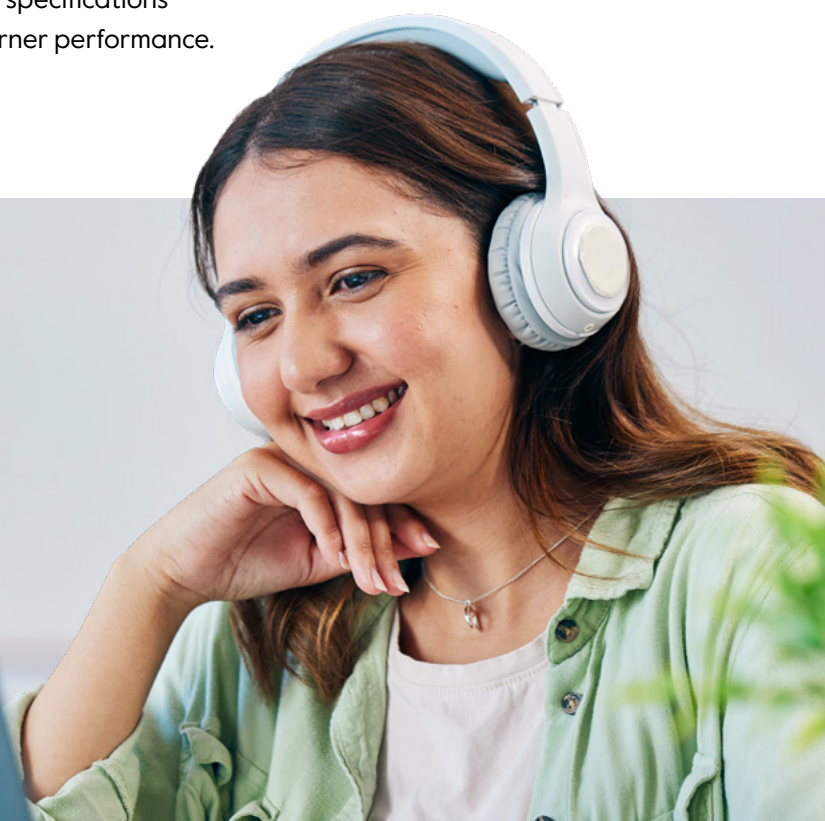
Together, the principles above demonstrate how Cambridge's long-standing assessment and educational values provide a robust foundation for the responsible use of GenAI to support content production. Our aim is to ensure continuity with our existing practice while explicitly addressing the new ethical, technical and societal considerations associated with generative technologies.

This section explains how Cambridge operationalises its principles for the responsible use of GenAI through tooling design and a structured assessment content creation workflow. Unlike generic AI content generation tools and pipelines, Cambridge's bespoke GenAI platform and workflow have been developed specifically for high-stakes assessment contexts and are grounded in, and leverage, Cambridge's longstanding expertise in language assessment, teaching and learning. This expertise is deliberately embedded into the design of the prompting engine (i.e., the system used to structure and guide AI generation) and supporting tooling, so that Cambridge standards of quality are present at the point of generation as well as downstream at the point of decision-making.

These standards of quality and assessment design draw on Cambridge's established assessment frameworks, test designs, task specifications and empirical evidence of learner performance.

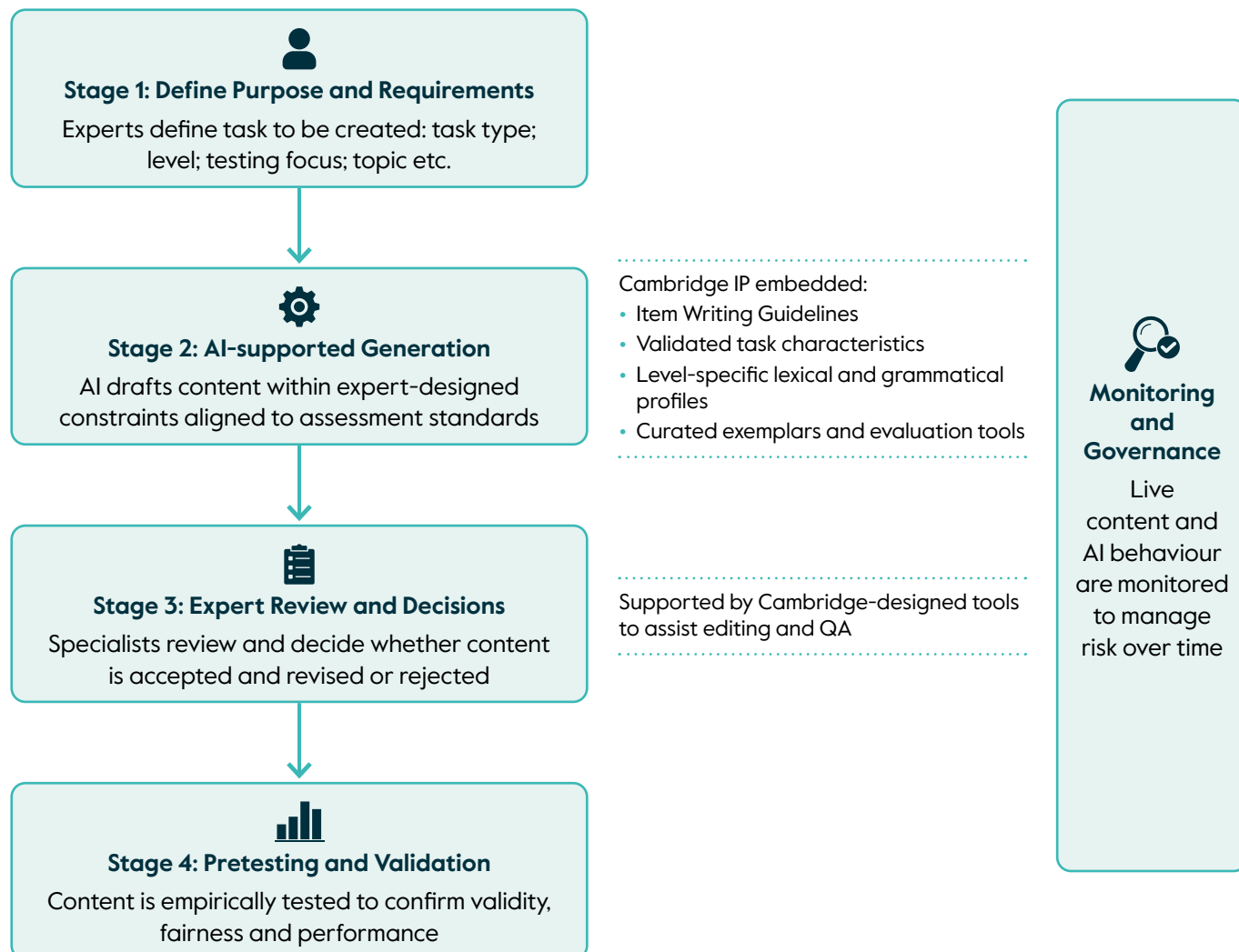
They are informed by expertise in language assessment, teaching and learning, and by fairness considerations, reflecting the principle of fairness, equity and inclusion through explicit attention to what aspects of language ability are represented and how.

A defining feature of this approach is that it does not treat AI capability as sufficient. **Effective and responsible use of generative technologies depends on the coordinated contribution of multiple forms of professional expertise, including language assessment methodology and test design, psychometrics, software engineering, and AI and machine learning.** The workflow described below shows how these forms of expertise interact at different stages to ensure that GenAI is used in ways that remain aligned with Cambridge's principles and with assessment purpose and validity.



A principle-led content creation workflow

At a high level, Cambridge's AI-supported content creation operates through a structured workflow as shown below.



Although represented here as a linear pipeline, this process is not intended to operate as a fully automated, end to end workflow. Instead, it reflects an ordered set of stages within a broader, iterative process.

At a micro level, generated assessment content may be refined through cycles of review, adjustment and regeneration. At a macro level, assessment experts may iteratively refine the constraints, prompts, and design parameters that shape generation over time. Human judgement therefore operates throughout the workflow, informing both the evolution of individual

assessment tasks and the ongoing design of the content generation system itself. At every stage, progression is gated by human decisions, institutional rules and quality thresholds.

The workflow exists precisely to ensure that the six principles set out previously are enforced through design and are present from the point of generation through every stage of quality assurance, rather than relying solely on correction at later stages. The sections that follow examine how these principles are enacted in concrete operational terms.

Stage 1:

Human Definition of Content Generation Requirements

All AI-supported assessment content creation begins with people rather than technology, through the definition of the specific content generation requirement within already-established assessment designs. In line with the principle of validity and fitness for purpose previously outlined, assessment specialists define the purpose of the content to be created, the target proficiency level, testing focuses and the intended use of the resulting material.

While these decisions are grounded in Cambridge's established test designs and validity frameworks, the focus at this stage is on specifying the requirements for a particular instance of content generation,

not on defining or redefining the design of the assessment itself, and GenAI plays no role.

Consistent with the principle of human-centred design and oversight, decisions about what is to be assessed, and why, remain fully the responsibility of assessment professionals. **This ensures that constructs are not implicitly shaped by the capabilities or limitations of generative models but remain grounded in assessment theory and educational purpose.**

The focus here is on translating established designs into clear, bounded requirements for a specific instance of content generation.

Stage 2:

AI-supported Generation Under Expert Defined Constraints

Once content generation requirements have been defined, GenAI is used to support content creation within carefully bounded conditions. In keeping with the principle of proportionality and sustainability, **AI is deployed selectively and only where it offers clear assessment benefit.**

AI-supported generation takes place within structured contexts defined by assessment experts. These constraints reflect decisions about task type, proficiency level, topic scope and permissible variation, or what can vary safely (for example, topic, context, names or wording) while still keeping the task equivalent in purpose, construct coverage and proficiency level. In doing so, these constraints operationalise the principle of validity and fitness for purpose by preventing unconstrained, open-ended generation that could undermine construct representation or level control.

Crucially, these constraints are not generic or model-driven. Their implementation relies on close collaboration between assessment specialists and technical experts, including software developers and machine learning engineers. This collaboration ensures that pedagogical and assessment intent is translated into system behaviour, that is, into how the AI actually generates content in practice.

The constraints themselves are derived from, and continuously informed by, Cambridge's proprietary assessment knowledge and intellectual property, developed over decades of research, item writing and operational delivery. This includes established item writer guidelines, empirically validated task characteristics, and carefully curated exemplars drawn from Cambridge's extensive item bank. Additionally, the incorporation of proprietary evaluation tools in the form of longstanding Cambridge IP such as level-specific vocabulary lists and level-sensitive lexical and grammatical profiles provide additional constraints. By incorporating this expertise into the design of prompts and evaluation mechanisms, Cambridge's conception of high-quality assessment content is embedded at the point of generation, rather than applied only through later review.

In this way, Stage 2 illustrates how the **principle of human-centred design is enacted not only through oversight, but through the deliberate encoding of Cambridge expertise into the tools themselves.**

Stage 3:

Expert Review, Editing and Decision-Making

AI-generated outputs are treated as provisional and subject to human judgement. Consistent with the principle of human-centred design and oversight, subject matter experts review all generated content to determine whether it aligns with assessment purpose, construct intent and fairness requirements.

At this stage, experts may edit content, request limited regeneration under revised constraints, or reject outputs entirely. To aid this process, the same Cambridge proprietary resources embedded in the content generation process also act as evaluative lenses and reference points to support expert judgement by providing structured evidence about the relationship between linguistic features, proficiency level and construct representation, i.e., how effectively the content reflects the language abilities the assessment is designed to measure. They inform evaluation and refinement without replacing human decision-making. Decisions are informed by professional knowledge of how linguistic and task features interact with difficulty,

interpretation and test taker experience. This reflects the principle of validity as something enacted through expert judgement in context, rather than enforced solely through technical thresholds.

This stage also operationalises the principle of accountability and governance. Acceptance and rejection decisions are made by individuals in defined roles and documented as part of routine quality assurance. **Responsibility for assessment content therefore remains clearly attributable to people, not systems, even where GenAI has been used to create initial drafts.** Crucially, AI-supported content is held to the same standards and acceptance criteria as human-produced content. At the same time, the use of GenAI introduces additional considerations, such as the risk of factual inconsistencies or ‘hallucinations,’ which are designed out where possible through constrained prompting, controlled sources and validation checks, but are explicitly checked during review.

Stage 4:

Pretesting and Psychometric Validation

In line with the principle of fitness for purpose, content that meets qualitative review standards proceeds to empirical validation. The use of GenAI does not alter the evidential requirements applied to assessment material. Items and tasks must demonstrate appropriate difficulty, discrimination (i.e. their ability to differentiate between stronger and weaker test takers) and absence of unintended bias through pretesting and psychometric analysis. **As with human produced content, AI-supported material is held to the same measurement standards and must demonstrate comparable empirical performance before it can be accepted for operational use.**

This stage draws on specialist psychometric expertise and reflects the principle of fairness, equity and inclusion by requiring evidence that content functions appropriately across relevant groups of test takers. Where empirical evidence indicates misalignment, content may be revised or withdrawn regardless of efficiency gains associated with AI-supported generation.

Stage 5:

Monitoring, Governance and Risk Management in Operational Use

The responsible use of AI-supported assessment content does not end at deployment. In line with the principles of accountability and governance, fairness, equity and inclusion, and fitness for purpose set out previously, Cambridge applies ongoing monitoring and structured oversight to both the performance of live assessment content and the behaviour of the AI systems that support its creation.

This stage recognises that risks to validity, fairness and reliability can emerge over time, even where earlier stages of design, review and validation have been rigorously applied. Continuous monitoring therefore functions as a safeguard against drift, that is, gradual changes in model behaviour or output characteristics that occur over repeated use or following system updates. It enables detection of instability, bias or unintended consequences arising from content accumulation effects, where individually acceptable items, taken together at scale, create imbalances or patterns that were not evident at the point of creation. Monitoring also accounts for evolving operational contexts, such as the expansion of an item bank over multiple test administrations, changes in the population of test takers, or shifts in delivery mode or usage patterns, any of which may alter how assessment content functions in practice.

Monitoring of Live Assessment Content

Once AI-supported content enters operational use, Cambridge monitors its performance using a combination of human expertise, empirical evidence and targeted analytical checks. From a validity and fitness for purpose perspective, monitoring examines whether content continues to function as intended, including whether it maintains appropriate difficulty, discrimination and construct representation over time.

From a fairness, equity and inclusion perspective, monitoring focuses on patterns that may only become visible at scale. This includes examining whether topics cluster too narrowly within the item bank, whether particular voices, groups or perspectives become over- or under-represented, and whether unintended stereotypes or imbalances emerge as content accumulates. Such patterns may not be visible at individual item level but can have material impact at system level.

Monitoring also considers the quality signals generated through human interaction with content, such as the typical extent and nature of edits specialists apply to AI-generated material. These signals provide evidence about where AI support is effective, where it introduces avoidable burden, and where continued use may no longer be proportionate, reflecting the principle of proportionality and sustainability.

Monitoring of AI Model Behaviour and System Performance

In addition to content-level monitoring, Cambridge applies oversight to the behaviour of the AI models and systems used to support content creation. This reflects the principle that responsibility resides with the organisation, not with the technology, and that systems must remain transparent and explainable to those governing their use.

Monitoring in this area focuses on detecting drift, instability or anomalous behaviour over time – that is, unexpected changes in how the model generates content, including changes in output characteristics following model updates, revisions by commercial providers to underlying models, or shifts in training data and system configuration. Specific attention is paid to indicators such as increased repetition of phrasing or stylistic patterns, reduced linguistic variety, variation in level control, or the emergence of content that conflicts with Cambridge's suitability rules.

Bias and representational issues are also a core focus. Monitoring seeks to identify whether model behaviour begins to favour or disadvantage particular topics, contexts, identities or forms of expression, even where such effects are subtle or indirect. The aim is to apply active scrutiny to model outputs in recognition of well-documented risks associated with large-scale language models.

Factual accuracy remains a further monitoring priority. Even where generated content has passed earlier review stages, ongoing checks help surface potential model 'hallucinations' or factual inconsistencies, supporting the principle of validity and protecting the integrity of assessment material.

Governance, Feedback and Remedial Action

All monitoring activity feeds into established governance structures with clearly defined roles, responsibilities and escalation routes. In keeping with the principle of accountability and governance, detected issues are assessed by appropriate experts, and decisions about remedial action are made explicitly and transparently within Cambridge's quality management processes.

Remedial actions may include revising or rejecting affected content prior to operational use, refining prompts or evaluative constraints, adjusting the scope of AI-supported generation, changing the LLM in use to a better aligned model or, where appropriate, suspending the use of specific AI capabilities altogether. Importantly, monitoring does not function as a passive reporting layer but as an active mechanism for maintaining alignment between AI-supported processes and Cambridge's assessment standards, with issues identified and addressed within controlled pre-operational processes wherever possible.

By combining monitoring of live content with oversight of AI model behaviour, this stage ensures that AI-supported assessment content remains trustworthy, fair and fit for purpose across its operational lifespan. It also reinforces the principle that responsible AI use is not a one-time design achievement, but an ongoing institutional commitment grounded in evidence, human judgement and governance.

How the workflow embeds assessment principles

Taken together, the stages described above show how the six principles previously articulated are embedded across the full lifecycle of AI-supported assessment content creation, from the definition of generation requirements through to operational use and ongoing monitoring. Principles such as validity, human-centred design, fairness, transparency, accountability and proportionality are enacted through deliberate decisions about who is involved, what evidence is required, how risk is managed, and under what conditions content may be generated, adapted, deployed or withdrawn.

A central feature of this workflow is the deliberate integration of diverse forms of professional expertise over time. Assessment specialists, item writers and reviewers, psychometricians and technical experts each contribute distinct knowledge at different stages, supported by structured governance and continuous oversight. Viewed as a whole, the workflow demonstrates that the quality and trustworthiness of AI-supported outputs are inseparable from the quality of the underlying assessment knowledge, evidence and judgement embedded in the system.

Cambridge-quality content emerges not from the generative model alone, but from the disciplined integration of proprietary expertise, curated evidence, monitoring and governance into the design, operation and ongoing use of AI-supported processes.

A further implication of this approach is its potential to support greater consistency in assessment content creation. While consistency is already maintained through expert review and quality assurance, AI-supported workflows enable more systematic encoding of assessment principles and constraints at the point of generation, reducing reliance on correction at later stages of inevitable human variation.

Illustrative case studies

The case studies presented in this section illustrate how the six principles and the workflow play out in specific assessment contexts. They are not intended as comprehensive evaluations of GenAI, nor as evidence that GenAI should be adopted wholesale across assessment practice. Rather, they demonstrate how GenAI can support assessment content creation in bounded and principled ways, and where its use remains limited or contingent on human expertise and empirical validation.

Each case study highlights both the potential benefits of GenAI when deployed within a principle-led framework and the constraints and risks that continue to require human judgement, governance and oversight.

Case Study 1:

AI-generated audio to support accent diversity in listening assessment

(Based on Karatay & Glasson, 2025)

Principles illustrated:

Validity and Fitness for Purpose; Fairness, Equity and Inclusion; Human-Centred Design and Oversight

Purpose and rationale

Accent diversity is a longstanding challenge in second language listening assessment. Assessments that aim to reflect real-world language use must balance exposure to a range of accents with concerns about intelligibility, comparability and unintended impacts on task difficulty. Generative speech technologies have been proposed as a way to increase accent diversity efficiently, but their suitability for high-stakes assessment contexts remains contested.

This case study explores whether AI-generated speech can complement human-recorded audio in listening tasks without undermining validity or fairness, and how expert judgement is required to interpret and constrain their use.

Approach

Parallel versions of listening task recordings were produced using both human speakers and AI-generated voices representing different accents of English. Panels of expert reviewers with diverse linguistic backgrounds evaluated the recordings in terms of intelligibility, comprehensibility and perceived accent strength. The focus was not on optimising AI output, but on assessing whether AI-generated audio met existing qualitative standards applied to human-recorded material.

Findings

The analysis suggested that, in many cases, expert reviewers could not reliably distinguish between human-recorded and AI-generated audio when accents were matched at a broad level. This indicates that generative speech technologies may have a role in supporting accent diversity in listening assessments, particularly where access to suitable human recordings is constrained.

Implications

At the same time, the case highlights important limitations. Perceptual similarity alone is not sufficient evidence of equivalence in assessment contexts. Even small differences in prosody, rhythm or consistency may interact with task demands in ways that affect difficulty or interpretive validity. For this reason, expert judgement and empirical monitoring remain essential, and AI-generated audio cannot be adopted on the assumption that perceptual indistinguishability guarantees equivalence.

This case illustrates the potential strength of GenAI in supporting greater representational breadth, aligned with the principle of fairness and inclusion. It also illustrates a clear limit in that decisions about suitability cannot be delegated to the technology. Human expertise is required both to evaluate whether AI-generated audio meets qualitative standards and to interpret empirical evidence about its effects on test performance.

Case Study 2:

Item cloning through structured manipulation of validated content

(Based on Mullooly, Machinska & Elliott, 2025)

Principles illustrated:

Validity and Fitness for Purpose; Accountability and Governance; Proportionality and Sustainability

Purpose and rationale

Large-scale assessment systems face ongoing challenges related to item security, renewal and resource efficiency. GenAI has been proposed as a means of rapidly producing large volumes of new items, but naïve generation risks construct drift and loss of psychometric control.

This case study examines a more constrained use of GenAI: the structured manipulation of previously validated cloze items to produce new variants, with the aim of extending item lifecycles while preserving construct integrity and comparability.

Approach

Previously validated cloze items were used as empirical anchors, serving as reference points with known performance characteristics. Generative models were employed to manipulate selected surface features, such as topical content or lexical choices, under expert defined constraints, while keeping the target answer unchanged. The following is an example of how this was done. In both, the key (soon) is the same:

Original

As as I pointed my skis towards the track, I began gaining speed

Clone (same topic)

As as I got off the lift, I carefully made my way to the edge of the slope

The resulting items were then subject to readability analysis and comparative psychometric evaluation, using established pretesting processes.

Findings

Initial evidence suggested that, under tightly controlled conditions, AI-supported item variants could display comparable readability profiles and score distributions to their source items.

Implications

This points to a potential role for GenAI in supporting efficient item development when constrained by expert judgement and empirical validation.

However, this approach does not eliminate the need for pretesting or governance. Comparability cannot be assumed on the basis of structural similarity alone, and responsibility for deciding whether cloned items are fit for operational use remains with assessment professionals.

This case illustrates how GenAI can support proportionate efficiency gains without redefining constructs or bypassing validation standards. It also underscores the notion that AI-supported item cloning is viable only when anchored to validated material and embedded within existing accountability frameworks. Used outside these conditions, the same techniques could introduce unacceptable risk.



What the case studies demonstrate

Taken together, these case studies illustrate that the value of GenAI in assessment content creation does not lie in automation or scale alone, but in targeted, constrained applications where GenAI supports human expertise, and where its outputs are subject to the same standards of evidence, review and accountability as any other assessment material.

Equally importantly, the case studies make visible the limits of GenAI. In both contexts, responsible use depends on human judgement, empirical validation and governance. These examples therefore reinforce the central argument of the paper: generative technologies can enhance assessment practice only when embedded within a principled, human-centred framework that foregrounds validity, fairness and trust.



Conclusion

GenAI is already reshaping the landscape of educational content creation, and assessment is no exception.

The central argument of this paper is not that GenAI should be adopted wholesale in assessment, nor that it should be resisted categorically. Rather, it is that the risks and opportunities associated with GenAI can only be managed responsibly when its use is grounded in explicit assessment principles, human expertise, and evidence-based governance.

By articulating a set of principles for AI-mediated assessment content creation and illustrating how these principles are operationalised through a bespoke toolkit and a structured workflow characterised by the embedding of assessment expertise and IP, this paper seeks to contribute to a growing field-level conversation about

what good practice looks like in this domain.

The case studies demonstrate that GenAI can add value in tightly bounded contexts, such as supporting representational diversity or extending item lifecycles, while reinforcing that human judgement, empirical validation and institutional accountability remain indispensable.

Implications for assessment practice

For readers involved in assessment design, governance or policy, this paper suggests three key implications.

First, **decisions about GenAI should be principle-led, not tool-led**. Questions about efficiency or innovation must be secondary to questions such as: *What construct is being measured? How might this use of AI affect validity or fairness? Where does accountability sit if something goes wrong?* Without explicit answers to these questions, the use of GenAI risks eroding, rather than enhancing, assessment quality.

Second, **people and professional expertise matter as much as technology**. Effective and ethical use of

GenAI in assessment depends on the coordinated contribution of assessment specialists, educators, psychometricians, technologists and governance leaders. Claims about AI capability that underplay this expertise risk obscuring the real work required to make AI-supported assessment both trustworthy and defensible.

Third, **validity, evidence and governance must keep pace with innovation**. GenAI-supported content should be subject to the same standards of empirical validation, monitoring and review as any other assessment material. Where those standards cannot be met, restraint is not a failure of innovation but a marker of responsible practice.

Responsibilities for assessment organisations

In practical terms, this paper invites assessment organisations, regulators and researchers to:

- Articulate explicit principles governing any use of GenAI in assessment content creation, grounded in validity, fairness and accountability rather than technology capability alone.
- Design workflows and systems that enforce those principles, embedding human oversight, decision points and validation requirements into practice rather than relying on informal checks.
- Invest in blended expertise, recognising that responsible use of GenAI requires deep assessment knowledge alongside technical capability.
- Share emerging evidence and practices transparently, contributing to a collective understanding of where GenAI strengthens assessment and where its limits remain.

As generative technologies continue to evolve, the challenge for the assessment field is not simply to keep up technologically, but to ensure that innovation remains aligned with longstanding commitments to quality, fairness and public trust.



References

- Association of Test Publishers, & International Test Commission. (2025). *Guidelines for technology-based assessment* (Version 1.1). Association of Test Publishers. <https://www.intestcom.org/upload/media-library/guidelines-for-technology-based-assessment-v20221108-16684036687NAG8.pdf>
- Caudwell, G., & Somers, A. (2026). *Principles of Good Practice* (in press).
- Chalhoub Deville, M., & O'Sullivan, B. (2020). Cognitive validity in language assessment. In G. Fulcher & L. Harding (Eds.), *The Routledge handbook of language testing* (pp. 41–55). Routledge.
- Chuang, Y., & Yan, X. (2025). A systematic review of generative artificial intelligence in language teaching, learning, and assessment. *System*, 115, 103026. <https://doi.org/10.1016/j.system.2025.103846>
- European Union. (2024). *The EU AI Act*. <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689>
- Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., & Srikumar, M. (2020). *Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI*. Berkman Klein Center for Internet & Society. <http://dx.doi.org/10.2139/ssrn.3518482>
- Galaczi E. D., & Pastorino-Campos, C. (2025). Ethical AI for language assessment: Principles, considerations, and emerging tensions. *Annual Review of Applied Linguistics*, 45, 294–314. doi:10.1017/S0267190525100081
- Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Praeger.
- Karatay, L., & Glasson, N. (2025). *Evaluating AI-generated speech for accent diversity in listening assessment* [Internal research presentation]. Cambridge University Press & Assessment.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). American Council on Education/Macmillan.
- Mullooly, A., Machinska, S., & Elliott, M. (2026). *Item cloning through structured manipulation of validated assessment content using generative AI* [Conference paper]. LTRC 2026.
- Pastorino Campos, L., & Saville, N. (2025). *Ethical AI for language learning and assessment*. Cambridge Papers in English Language Education. https://www.cambridge.org/sites/default/files/media/documents/Ethics_in_AI_CPELE_web.pdf
- Saville, N. (2009). *Developing a model for investigating the impact of language assessment within educational contexts by a public examination provider*. PhD thesis, University of Bedfordshire.
- UNESCO. (2023). *Guidance for generative AI in education and research*. UNESCO Publishing. <https://www.unesco.org/en/articles/guidance-generative-ai-education-and-research>
- Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. Palgrave Macmillan.
- Xu, J., Schmidt, E., Galaczi, E., & Somers, A. (2024). *Automarking in language assessment: Key considerations for best practice*. Cambridge Papers in English Language Education. <https://www.cambridge.org/sites/default/files/media/documents/cambridge-research-papers-in-english-language-education.pdf>



Author biographies



Ron Zeronis is Head of Assessment Content Quality at Cambridge University Press & Assessment. He has over 25 years' experience in English language education and assessment.

He holds a Master's degree in Applied Linguistics from the University of Leicester, UK. His current work explores how Generative AI and related technologies can be applied responsibly to support scalable, quality-driven approaches to assessment content development. He leads a team of specialists working on the practical application of assessment standards across AI-enabled content generation, accessibility, equity, diversity and inclusion and sustainability. The team help to ensure that innovation in assessment content enhances quality while maintaining validity, accessibility, inclusion and trust.



Andrew Mullooly has a master's degree in Applied Linguistics & TESOL. As part of his role as a Senior Assessment Manager at Cambridge University Press & Assessment, he has

co-authored several research papers. These include an exploration of interaction in video-call paired speaking tests, a review into the application of Large Language Models for language teaching and assessment purposes and publication of the Cambridge Multiple-Choice Questions Reading dataset.



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 to 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.



Find out more at
cambridge.org

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