Developing Programmable Listening Assessments via Generative AI: An Attribute-Driven Framework

Vahid Aryadoust[1,*] and Joann Wong[1]

[1] National Institute of Education, Nanyang Technological University, Singapore
*vahid.aryadoust@nie.edu.sg

Abstract. This chapter examines how generative artificial intelligence (GenAI) can be integrated into listening assessment development through an attribute-based framework. Listening tests require the creation of spoken texts, recordings, and task materials, yet traditional production processes are constrained by time, cost, and limited control over stimulus features. GenAI technologies, including large language models and neural text-to-speech systems, introduce new capabilities for specifying and manipulating task-related attributes such as speech rate, lexical density, accent variation, and mode. However, without theoretical grounding, generative tools risk automating existing paradigms rather than enhancing construct fidelity. The attribute-based approach to test development addresses this gap by identifying which features of listening input are construct-relevant and treating them as programmable design parameters. The chapter outlines how GenAI can operationalise task-related attributes (e.g., linguistic complexity, phonological variation, item format) and consider listener-related attributes (e.g., proficiency, prior exposure), enabling more precise calibration of difficulty and cognitive demand. Validity implications are discussed, including risks of construct underrepresentation, over-standardisation, and new sources of bias introduced by generative systems. The chapter concludes by proposing a framework for programmable listening assessment in which generative capabilities are guided by explicit, theory-driven attribute specifications to support valid and reliable test development.

Keywords: GenAI, Listening assessment, Attribute-based listening assessment, Programmable listening assessment.

1 The Generative Turn in Listening Assessment

The rapid ascent of generative artificial intelligence (GenAI), typified by large language models (LLMs) and multimodal generation systems such as GPT-4o, is fundamentally altering the landscape of educational assessment. Since the arrival of ChatGPT in late 2022, GenAI has catalysed the reevaluation of how knowledge and skills can be measured, prompting scholars and practitioners alike to reconsider long-standing assumptions about assessment design, authenticity, and fairness (Zhao et al., 2024). In mainstream educational contexts, much of the early research on GenAI has focused on automated feedback and scoring (e.g., AlGhamdi et al., 2026; Atkinson & Palma, 2025; Huang, 2014; Hussein et al., 2019; Li & Ng, 2024; Shadiev & Feng, 2024; Sun et al.,

© The Author(s) 2026
C. Y. Lye et al. (eds.), *Proceedings of the Pacific Rim Objective Measurement Symposium (PROMS 2025)*, Atlantis Highlights in Social Sciences, Education and Humanities 56,
https://doi.org/10.2991/978-94-6239-834-4_6

2025) as well as on the perceived risks and opportunities it creates for academic integrity and learning (Khlaif et al., 2024; Li & Ng, 2024; Yavariabdi et al., 2025). GenAI tools have been shown to support personalised feedback generation, scaling of assessment logistics, and consistency in scoring support, though expert oversight remains critical to avoid biases and ensure validity (Yavariabdi et al., 2025).

Listening assessment, however, represents a distinct and compelling frontier for generative technologies. In these tests, the core assessment stimuli (i.e., spoken texts) are typically custom-created, and although the use of authentic materials is recommended, it remains challenging to identify listening input that is appropriate for specific groups of learners and test takers. In listening assessments, therefore, test developers typically generate their own scripts, produce recordings, calibrate accent and delivery speed, and ensure that listening tasks reflect target communicative domains. These production requirements are key constraints for test developers due to time, cost, and logistical complexities. GenAI has the potential to overcome many of these constraints. Advanced language models can produce coherent, contextually appropriate texts on demand (Aryadoust et al., 2024); neural text-to-speech systems can synthesise voice input with controlled accents and prosody (Jia & Aryadoust, in press); and multimodal AI systems can even generate video and gestural cues (Zhang et al., 2025). Recent experimental work in language assessment demonstrates how generative methods can be used to produce listening assessment tasks at scale, using generative models to create conversational content and item variants for large-scale proficiency tests (Aryadoust et al., 2024; Runge et al., 2024).

For listening assessment specifically, the generative turn raises both promise and challenge: GenAI offers the capacity to programmatically vary and control listening input attributes, enabling highly systematic design of stimulus characteristics that were previously constrained by manual production processes (Aryadoust & Jia, 2025). But without a principled theoretical basis to guide this generative capacity, there is a risk that AI simply automates old paradigms without enhancing construct specificity or interpretability. This chapter takes the position that GenAI's most meaningful contribution to listening assessment lies in its ability to operationalise listening attributes in systematic, scalable, and theoretically aligned ways (Aryadoust, 2024; Aryadoust & Jia, 2025). To realise this potential responsibly, generative capabilities must be anchored in a robust understanding of the factors that influence listening comprehension — the *attributes* that shape performance under assessment conditions. In the following sections, we articulate how these two domains, AI generation and listening assessment attributes, intersect to create a new paradigm in test development.

2 Listening Assessment as an Attribute-Driven Design Problem

Listening assessment is often discussed in terms of subskills, cognitive processes, or attributes (Aryadoust, forthcoming; Aryadoust & Luo, 2023). While the former two perspectives illuminate how listeners process input (Field, 2008), they do not fully

address how listening performance is elicited and shaped in assessment contexts. From a test development standpoint, listening is less a static ability than a performance that emerges under specific task conditions. These conditions are defined by attributes. An attribute-based perspective conceptualises listening performance as the product of interacting variables that influence comprehension (Aryadoust & Luo, 2023; Buck & Tatsuoka, 1998). In assessment theory, an attribute refers to any feature that affects task performance, whether it is embedded in the task itself or located within the listener (Buck & Tatsuoka, 1998). In second-language listening research, this framework has been extended to encompass task-related characteristics (e.g., speech rate, accent, lexical density, modality / mode, item format) and listener-related characteristics (e.g., proficiency, exposure, anxiety, background knowledge) (Aryadoust & Luo, 2023).

This perspective shifts the focus from "what listening is" to "what conditions shape listening performance" (Aryadoust, forthcoming). For example, faster speech rates increase processing demands and reduce segmentation time, disproportionately affecting lower-proficiency listeners (Chiu & Chen, 2023; Griffiths, 1990, 1992). Accent unfamiliarity can introduce additional phonological processing load, even when the underlying comprehension construct remains constant (Field, 2008). Lexical density and syntactic complexity increase integration demands and working memory load (Buck, 2001). Response format also matters: selected-response items typically impose different cognitive demands than constructed-response formats (In'nami & Koizumi, 2009; Liu & Aryadoust, 2026). Seen from this angle, listening assessment becomes a design problem: test performance reflects how carefully these attributes are calibrated. For example, if speech is unnaturally slow, the construct may be underrepresented. If accents are artificially homogenised (see Aryadoust, 2024), situational authenticity may be reduced. If lexical load is uncontrolled, difficulty may fluctuate unpredictably across parallel forms.

Importantly, many of these task-related attributes are not incidental; they are central to construct representation. Authentic listening involves processing variable speech rates, encountering diverse accents, interpreting reduced forms, and integrating multimodal cues (Aryadoust, forthcoming). The attribute-based model therefore provides a structured way to identify which variables are construct-relevant, which are construct-irrelevant, and which require deliberate control. Thus, in this sense, listening assessment is inherently attribute-driven, and test validity depends not only on abstract construct definitions but also importantly on how systematically these attributes are operationalised in stimulus design.

2.1 Why the Attribute-Based Approach

The rise of GenAI dramatically expands the technical capacity to produce listening materials. LLMs can generate scripts at scale; neural text-to-speech systems can synthesise speech with adjustable accents and delivery rates; multimodal systems can introduce visual cues and gesture (Aryadoust, forthcoming). However, increased generative power does not automatically translate into improved assessment quality:

without a guiding framework, AI systems may produce fluent but construct-misaligned materials (Tan et al., 2025; Kowal et al., 2025).

The attribute-based perspective provides the theoretical control that generative systems require. First, it identifies which aspects of listening input are meaningful design variables. For example, speech rate, lexical sophistication, and accent are attributes that shape cognitive load and construct representation. When these variables are made explicit, they can be treated as programmable parameters rather than by-products of script writing. Second, the attribute framework guards against construct underrepresentation. Generative models tend toward highly coherent, standardised output. Neural text-to-speech systems often produce speech with reduced disfluency and controlled prosody. While this may enhance clarity, it can inadvertently remove features of authentic spoken language, such as reduced forms, hesitations, and accent variability, that listeners must navigate in real-world contexts (Buck, 2001; Field, 2008). An attribute-based approach ensures that generative efficiency does not oversimplify the listening construct. Third, the model clarifies sources of construct-irrelevant variance. For example, if lexical complexity fluctuates unpredictably across generated forms, difficulty and the resultant cognitive load may vary in unintended ways. By defining attributes in advance, developers can monitor and calibrate these dimensions systematically.

In this way, the attribute-based model functions as a control system for generative design. It translates validity concerns—such as contextual parameters and test takers' characteristics in Weir's (2005) validity framework—into operational variables that can be specified and adjusted. GenAI increases the scalability and flexibility of attribute manipulation, whereas the attribute-based model supplies the conceptual criteria for determining which attributes should be varied or controlled. Their integration supports a more systematic and transparent approach to listening test development, in which stimulus features are deliberately specified and evaluated rather than passively inherited from manual production constraints.

3 Operationalising Task-Related Attributes through GenAI

Listening assessment is fundamentally shaped by task-related attributes embedded in the stimulus and task design. These include contextual features (e.g., genre and communicative setting), linguistic properties (e.g., lexical density, syntactic complexity, speech rate), phonological variation (e.g., accent and reduced forms), mode (e.g., audio versus video), and item or task format. Traditionally, many of these attributes were determined through manual scripting and recording processes, limiting the degree to which they could be systematically controlled or varied. Generative AI introduces new possibilities for explicitly specifying and manipulating these variables during test development.

3.1 Control of Script Generation and Linguistic Attributes

LLMs enable the rapid generation of listening scripts across a wide range of genres, including academic lectures, interviews, dialogues, and narrative monologues. More importantly, they allow developers to iteratively and quickly adjust linguistic features of those scripts (Aryadoust et al., 2024). Lexical sophistication, for example, can be calibrated by prompting models to target particular proficiency levels or to avoid low-frequency vocabulary. Information load can be modulated by controlling the number of propositions introduced per unit of discourse (Field, 2012). These forms of control align directly with research demonstrating that lexical characteristics and discourse density of listening texts are significant predictors of comprehension difficulty (Révész & Brunfaut, 2013). By using various prompting techniques, such as few-shot exemplar prompting (Brown et al., 2020), progressive-hint prompting (PHP) method (Aryadoust et al., 2024; Zheng et al., 2023), and iterative refinement (White et al., 2023), developers can compare alternative versions of the same script while holding specific attributes constant. This enables more systematic exploration of how particular linguistic features contribute to task difficulty, which has traditionally required substantial expert effort, rewriting, and piloting cycles.

However, generative systems also introduce new risks. LLMs tend to produce highly coherent and standardised language. Without explicit prompting, they may underrepresent features characteristic of spontaneous speech, such as hesitation markers, false starts, or reduced forms. Because such features are common in authentic listening contexts (Field, 2008), their absence may narrow construct representation. Attribute specification is therefore necessary to ensure that generated scripts reflect the intended level of authenticity.

3.2 Speech Rate, Accent, and Phonological Variation through Neural Text-to-Speech

Neural text-to-speech (TTS) systems extend attribute control from textual features to auditory realisation. Contemporary TTS models allow developers to adjust speech rate, pitch, and accent characteristics, and to select from multiple synthetic voices. Speech rate is a well-documented determinant of listening difficulty (Chiu & Chen, 2023; Griffiths, 1990, 1992). With neural TTS, delivery speed can be calibrated to approximate empirical norms for particular discourse types (e.g., lectures versus conversations) rather than relying on the natural pacing of a single human speaker. This allows more consistent alignment between intended proficiency level and auditory processing demands. Accent representation can likewise be varied systematically. Instead of relying on the limited accent range available among available speakers, developers can generate parallel forms featuring different English varieties. Given longstanding discussions about accent diversity in listening assessment (Harding, 2012; Kang et al., 2020), this capacity enables more deliberate inclusion of diverse phonological varieties while monitoring for differential item functioning.

At the same time, synthetic speech often exhibits smoother prosody and fewer disfluencies than spontaneous human speech. Overly uniform pronunciation may reduce exposure to "authentic" variability. Consequently, developers must determine whether to introduce controlled disfluency or prosodic variation to maintain construct fidelity.

3.3 Modality and Multimodal Input

Generative systems are increasingly capable of producing video-based stimuli that integrate speech with visual cues such as facial expression and gesture. Research on video-based listening assessment indicates that multimodal input can have small but measurable effects on performance, moderated by proficiency level and task design (Liu & Aryadoust, 2024). GenAI enables developers to experiment with multimodal conditions more efficiently than traditional video production methods allow. For example, identical scripts can be rendered as audio-only, speaker-present video, or visually enriched formats. This creates opportunities to examine how modality influences comprehension and to determine whether visual cues are construct-relevant within a given target language use domain.

As with other attributes, modality must be justified within the assessment argument. The inclusion of visual support may enhance ecological validity in contexts where real-world listening is audio-visual, but it may also alter the construct by introducing visual literacy components. Generative tools expand the range of design possibilities; they do not resolve theoretical questions about construct boundaries.

3.4 Item and Format Generation

Task-related attributes extend beyond stimulus design to include response format and item characteristics. LLMs can generate multiple-choice questions, short-answer prompts, and inference-based items aligned with a given script. They can also produce alternative distractor sets, allowing developers to compare plausibility and cognitive demand across versions. Because item format affects processing demands and task difficulty (In'nami & Koizumi, 2009), generative systems can be used to systematically vary format while holding stimulus content constant. For example, a single listening passage may be paired with both selected-response and constructed-response items to examine differential effects on test performance. This capacity supports more precise alignment between format and intended construct representation.

Nonetheless, current generative systems may produce distractors that rely on superficial lexical overlap rather than deeper inferencing (see Aryadoust et al., 2024). Human review remains necessary to ensure that generated items tap intended comprehension processes rather than surface-level recognition.

4 Aligning GenAI with Listener-Related Attributes

While generative AI provides extensive control over task-related attributes, listening performance is also shaped by characteristics of the listener. These include proficiency, prior exposure to the target language, sociolinguistic background, and affective factors such as anxiety and motivation (Aryadoust & Luo, 2023). Unlike task attributes, these variables are not directly manipulable by test developers. However, GenAI introduces new possibilities for aligning listening input with listener profiles in more flexible and adaptive ways.

4.1 Proficiency and Experience as Design Constraints

Listening proficiency and L2 experience, rather than source of construct-irrelevant variance, should remain primary determinants of listening comprehension test performance. Higher-proficiency listeners demonstrate more efficient lexical access and better segmentation of continuous speech (Field, 2008; Vandergrift, 2007), and greater tolerance for variability in speech rate and accent (Griffiths, 1990, 1992; Harding, 2012). Less experienced listeners, by contrast, are more sensitive to increases in processing demands introduced by rapid delivery, dense information load, or unfamiliar phonological features. In traditional test development, particularly in assessment-for-learning contexts, accommodating such differences requires designing separate test forms for different proficiency bands. GenAI enables a more fine-grained approach. Because linguistic and auditory attributes can be adjusted programmatically, listening input can theoretically be calibrated more precisely to target proficiency levels (Aryadoust et al., 2024). For example, scripts can be generated with controlled lexical frequency and syntactic complexity, while speech rate and accent familiarity can be varied systematically. This allows developers to create graded input that reflects increasing levels of difficulty without fundamentally altering the underlying construct.

4.2 Toward Adaptive Listening Input

GenAI opens the possibility of *dynamic* adaptation, in which task attributes can be adjusted in (almost) real time in response to listener performance during the assessment itself. This represents a conceptually distinct contribution from the attribute manipulation discussed in Section 3: rather than pre-specifying a fixed attribute profile, the system responds to observed performance patterns and modifies input conditions accordingly. Unlike traditional adaptive testing, which typically adjusts overall item difficulty along a unidimensional ability estimate, attribute-based adaptation allows for targeted adjustment of specific input features. For example, attribute-based adaptation can incrementally reintroduce faster speech rates as a listener stabilises, or scaffold exposure to unfamiliar accents progressively rather than abruptly. This granularity supports more diagnostic interpretations of listening performance, since outcomes can be linked to particular input conditions rather than a single difficulty parameter. Such an approach also aligns with broader goals of learning-oriented assessment, where measurement and the support of learning are treated as complementary rather than competing aims.

4.3 Affective Factors and Test Engagement

Listener-related attributes also include affective factors such as anxiety and motivation. These variables influence attention, processing efficiency, and persistence during listening tasks (Elkhafaifi, 2005; Vandergrift, 2005; Vandergrift & Goh, 2012). While GenAI cannot directly modify these internal states, it may indirectly influence them through task design. For example, adaptive systems that begin with manageable input and gradually increase difficulty may reduce anxiety and support sustained engagement. Conversely, poorly calibrated tasks, for instance, those with excessive speech rate or unfamiliar accents introduced too early, may increase cognitive overload and negatively affect performance. This suggests that alignment between task attributes and listener characteristics is a matter of both difficulty calibration and maintaining appropriate levels of challenge and engagement.

4.4 Constraints and Risks

Despite these possibilities, several limitations must be acknowledged. First, listener attributes are often not directly observable. Proficiency, exposure, and affective states must be inferred from performance or external data, which introduces uncertainty. Second, adaptive and personalised systems require robust calibration to ensure that variations in input do not introduce unintended sources of construct-irrelevant variance. Third, there is a risk of construct dilution or underrepresentation. If input is excessively simplified to accommodate individual listeners, assessments may no longer reflect the demands of real-world listening in target language use (TLU) domains. This is particularly relevant in high-stakes contexts, where standardisation and comparability are essential.

4.5 Summary

GenAI does not alter listener-related attributes, but it changes how assessment systems can respond to them. Through enabling flexible adjustment of task attributes, GenAI allows for more precise alignment between input conditions and listener characteristics. This creates opportunities for more adaptive and diagnostic listening assessment, while also introducing new challenges related to validity and fairness.

5 Validity Implications in AI-Generated Listening Assessment

The integration of generative AI into listening assessment development introduces new possibilities for task design, but it also raises fundamental questions about validity. In language assessment, validity concerns the extent to which evidence and theory support the interpretations of test scores for their intended uses (Messick, 1989). From this perspective, the use of GenAI must be evaluated not in terms of technological capability alone, but in terms of how it affects construct representation and sources of construct-irrelevant variance.

5.1 Construct Representation

A central validity concern in AI-generated listening assessment is whether the resulting tasks adequately represent the intended construct. While earlier sections established that listening performance is shaped by a range of task-related attributes, the issue at this stage is not simply whether these attributes can be manipulated, but whether they are configured in ways that adequately reflect the TLU domain[1]. GenAI increases the degree of control over input features, but this control introduces new design decisions. Construct representation, then, depends on how developers select, prioritise, and combine attributes, and to what extent these attributes represent features of language in the TLU domain (Zhao & Aryadoust, 2025). For example, a listening task may be generated with carefully calibrated lexical and syntactic complexity, but if it lacks features such as reduced forms, variable speech rate, or discourse-level unpredictability, it may fail to represent the processing demands of authentic spoken interaction. In this sense, construct underrepresentation can arise not from insufficient control, but from overly simplified or overly regularised input. Zhao and Aryadoust (2025) and Tao and Aryadoust (2024) present methods for exploring construct representation in listening assessment from corpus-based perspectives.

A related issue concerns the distribution of attributes across tasks. Over-standardisation of listening input is not unique to AI-generated systems, and traditional test materials have similarly been criticised for overly polished, scripted language that underrepresents authentic speech features (Buck, 2001). However, the systematic and scalable nature of AI generation may intensify this tendency, since generative models optimise for coherence and fluency by default, potentially narrowing the range of conditions under which listening is assessed. For instance, consistently presenting moderately paced, clearly articulated speech may reduce variability in difficulty but fail to capture listeners' ability to cope with more demanding input conditions, such as rapid conversational speech or unfamiliar accents. While delivery speed can be adjusted relatively easily, the issue of "naturalness" in speech is not as readily addressed in either traditional or AI-generated materials. Naturalness, here, refers to features such as authentic intonation patterns and the presence of disfluencies (e.g., hesitations, repetitions, false starts, and self-repairs) and warrants deliberate attention in both cases. Construct representation, therefore, depends not only on the consideration of individual attributes, but also on their range and distribution across the assessment.

Generative systems also introduce a tendency toward structural and linguistic regularity: LLM-generated texts often exhibit clear organisation, explicit coherence markers, and relatively balanced information flow, with research demonstrating that LLMs reuse discourse structures more consistently than human authors and produce more homogenised output regardless of intended genre or register (Muñoz-Ortiz et al., 2024; Namuduri et al., 2025). Similarly, neural text-to-speech systems tend to produce

[1] This approach is also referred to as a task-based approach and can be distinguished from a competency-based approach to test development. The latter primarily focuses on the cognitive domain targeted by the test.

stable prosodic patterns with limited disfluency, as synthesised speech typically represents the average prosodic style of training data rather than the natural variability characteristic of spontaneous human speech (Raitio et al., 2020; Vallés-Pérez et al., 2021). While these characteristics may facilitate comprehension, they may also alter the nature of the listening construct by reducing exposure to features that require real-time inferencing and repair strategies. Listening research has long emphasised the role of such features, including hesitation, self-correction, and reduced forms, in shaping comprehension processes (Buck, 2001; Field, 2008). Their systematic absence may therefore lead to a construct that is narrower and less representative of real-world listening.

In sum, the attribute-based framework is suitable for evaluating these issues by requiring developers to make explicit decisions about which attributes are construct-relevant and how they should be instantiated. This shifts the focus from evaluating individual test items to evaluating attribute configurations at the test and listening comprehension levels. Rather than asking whether a given test item is "appropriate," developers must consider whether the overall set of tasks reflects the range of listening conditions implied by the construct definition. In this context, construct representation becomes an issue of relevance and coverage. GenAI enables fine-grained control over input features, but it also places greater responsibility on developers to ensure that this control is used to approximate the complexity and variability of real-world listening, rather than to produce optimised but simplified input (Aryadoust & Jia, 2025). The attribute-based approach provides a structured basis for making these decisions explicit and for evaluating whether AI-generated materials support valid score interpretations.

5.2 Sources of Construct-Irrelevant Variance

In addition to construct representation, validity depends on minimising construct-irrelevant variance, which refers to variation in test performance attributable to factors unrelated to the intended construct. GenAI introduces both opportunities and risks in this regard. On one hand, the ability to control attributes such as speech rate and lexical complexity can reduce unintended variability across test forms, supporting more consistent difficulty calibration. On the other hand, generative systems may introduce new sources of construct-irrelevant variance, and in the worst-case scenario, lead to bias. For example, when topic familiarity advantages one group of listening test takers while disadvantaging another group of comparable ability, the assessment risks (partially) reflecting differences in topic familiarity rather than listening comprehension. In such cases, the construct is confounded in favor of those with greater familiarity with specific topics covered in the test. With a sufficiently large sample, techniques such as differential item functioning (DIF; see Chen et al., 2025) can help identify these sources of construct-irrelevant variance and bias. DIF can lead to systematic bias in the test, which would jeopardize the validity of the interpretations and uses of test scores (Chen et al., 2025).

Similarly, generated scripts may reflect cultural or contextual biases embedded in model training corpora (Bender et al., 2021), introducing culturally specific content or discourse conventions that systematically advantage certain groups of test takers, which can become a further source of construct-irrelevant variance. From a validity perspective, both uncontrolled variability and excessive uniformity can distort score interpretation.

6 Reliability and Consistency

Reliability concerns the consistency of measurement across tasks, forms, and administrations (Shang et al., 2024). In listening assessment, this includes the stability of difficulty across parallel forms and the extent to which variations in input features introduce unintended differences in performance. GenAI has the potential to support reliability by enabling the generation of multiple parallel forms with controlled attribute profiles. Because task-related attributes can be specified explicitly, developers can aim to maintain consistency across versions in terms of speech rate, lexical difficulty, syntactic complexity, accent, and discourse type. This can also help improve the consistency of test scores. At the same time, generative processes introduce stochastic variation. Even when similar prompts or constraints are used, outputs may differ in subtle but consequential ways, such as shifts in information density, discourse organisation, or inferential demands, areas that have to be considered carefully in listening test development. These differences may not be immediately apparent but can affect task difficulty and, consequently, score comparability across forms.

Ensuring reliability in AI-generated listening assessment therefore requires systematic monitoring and calibration. This includes human review of generated materials, the use of computational measures (e.g., lexical and syntactic indices) to check consistency, and empirical validation through piloting and statistical analysis. GenAI increases the efficiency of producing candidate materials, but it does not remove the need for evidence-based calibration to ensure that parallel forms function equivalently.

7 Summary

Across these dimensions, a consistent theme is the continued importance of human expertise. While GenAI enables large-scale generation and manipulation of listening materials, it does not replace the need for theoretical grounding, empirical validation, and ethical judgement. The attribute-based framework can support this process by providing a structured framework for specifying and evaluating task characteristics. However, decisions about which attributes to prioritise, how to balance authenticity and standardisation, and how to interpret test scores remain fundamentally human responsibilities.

GenAI introduces new capabilities for listening assessment development, but it also reshapes the validity landscape. Its impact depends on how it is integrated into assessment design processes. When guided by an attribute-based framework and supported by systematic validation, generative technologies have the potential to enhance construct alignment and design precision. Without such guidance, however, they risk introducing new forms of bias, distortion, and construct misalignment.

8 Toward Programmable and Adaptive Listening Assessment

The integration of generative AI into listening assessment development introduces the possibility of moving beyond static test forms toward more flexible and responsive assessment systems. As discussed in previous sections, GenAI enables explicit specification and manipulation of task-related attributes, while also allowing alignment with listener-related characteristics. These capabilities point toward a model of listening assessment that is increasingly programmable. Programmable listening assessment refers to the ability to define, adjust, and combine listening task attributes in a systematic and transparent manner. Rather than relying on fixed sets of pre-constructed stimuli, developers can specify target attribute configurations, such as speech rate, lexical density, accent, and modality, and generate tasks that meet those specifications. This shifts the role of the test developer from producing individual items to designing attribute profiles that define the conditions under which listening is assessed. One implication of this shift is the potential development of attribute-calibrated test item banks, in which tasks are indexed not only by overall difficulty but also by their underlying attribute composition. Such item banks would allow developers to assemble test forms that systematically vary or control specific dimensions of listening input. This, in turn, supports more transparent alignment between construct definitions and task design, as well as more precise control over test form equivalence.

A related development is the possibility of adaptive listening assessment systems that adjust task attributes dynamically in response to test taker performance. As discussed in Section 4.2, attribute-based adaptation differs fundamentally from traditional adaptive testing by allowing targeted adjustment of specific input features rather than operating on a unidimensional estimate of ability, and supports more diagnostic interpretations of listening performance by linking outcomes to particular input conditions.

At the same time, the move toward programmable and adaptive systems raises important questions. First, increased flexibility must be balanced with the need for comparability across test takers, particularly in high-stakes contexts. If task attributes vary substantially between individuals, ensuring fairness and score equivalence becomes more complex. Second, the design of attribute configurations requires careful theoretical justification; not all forms of variation are construct-relevant. Third, the implementation of such systems depends on robust infrastructure for tracking, calibrating, and validating generated materials. Finally, as emphasised throughout this

chapter, programmable listening assessment does not eliminate the role of human expertise. Decisions about which attributes to include, how they should be combined, and how performance should be interpreted remain grounded in theoretical and empirical considerations. GenAI expands the range of possible designs, but it also increases the importance of principled decision-making. In this sense, the future of listening assessment is not defined by automation alone, but by the integration of generative capabilities with explicit, theory-driven models of listening. The attribute-based approach provides a foundation for this integration by linking construct definitions to the design and calibration of listening input in a systematic way.

9 Conclusion

This chapter has argued that the primary impact of generative AI on listening assessment lies in the transformation of listening test development into a more explicit and controllable design process. Because listening assessment is fundamentally shaped by input, the ability to generate and adjust that input at scale represents a significant shift in how tests can be constructed. However, increased generative capacity does not, in itself, ensure valid assessment. The attribute-based approach provides the necessary theoretical framework for identifying which features of listening input are construct-relevant and how they should be configured. Making task-related attributes explicit enables developers to move from implicit, experience-based practices toward more systematic and transparent design decisions. The integration of GenAI and the attribute-based approach therefore supports a more principled form of listening assessment development. It allows for finer control over stimulus characteristics, more consistent calibration across forms, and the possibility of adaptive and diagnostic applications. At the same time, it introduces new responsibilities related to validity, fairness, and ethics. Ensuring that generated materials reflect the intended construct, minimise construct-irrelevant variance, and support meaningful score interpretations remains essential. Ultimately, the value of GenAI depends on how effectively its capabilities are aligned with theoretically grounded models of listening comprehension.

Acknowledgments. This study was supported by an Academic Research Fund (AcRF) grant from the Ministry of Education, Singapore (grant number ARC 2/25 VSA (MOE-T2EP40224-0003)).

Disclosure of Interests. Nil

References

1. AlGhamdi, E. M., Li, Y., Gašević, D., & Chen, G. (2026). Leveraging prompt-based LLMs for automated scoring and feedback generation in higher education. *Computers & Education, 243*, Article 105511. https://doi.org/10.1016/j.compedu.2025.105511

2. Aryadoust, V. (2024). Topic and accent coverage in a commercialized L2 listening test: Implications for test-takers' identity. *Applied Linguistics, 45*(5), 765–785. https://doi.org/10.1093/applin/amad062
3. Aryadoust, V. (forthcoming). *Assessing listening in the age of generative artificial intelligence*. Routledge.
4. Aryadoust, V., & Jia, Y. (2025). Generative artificial intelligence in listening assessment. In L. McCallum & D. Tafazoli (Eds.), *The Palgrave encyclopedia of computer-assisted language learning*. Palgrave Macmillan. https://doi.org/10.1007/978-3-031-51447-0_202-1
5. Aryadoust, V., & Luo, L. (2023). The typology of second language listening constructs: A systematic review. *Language Testing, 40*(2), 375–409.
6. Aryadoust, V., Zakaria, A., & Jia, Y. (2024). Investigating the affordances of OpenAI's large language model in developing listening assessments. *Computers and Education: Artificial Intelligence, 6*, Article 100204. https://doi.org/10.1016/j.caeai.2024.100204
7. Atkinson, J., & Palma, D. (2025). An LLM-based hybrid approach for enhanced automated essay scoring. *Scientific Reports, 15*, Article 14551. https://doi.org/10.1038/s41598-025-87862-3
8. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610-623). Association for Computing Machinery. https://doi.org/10.1145/3442188.3445922
9. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., & Amodei, D. (2020). *Language models are few-shot learners*. arXiv. https://arxiv.org/abs/2005.14165
10. Buck, G. (2001). *Assessing listening*. Cambridge University Press.
11. Buck, G., & Tatsuoka, K. (1998). Application of the rule-space procedure to language testing: Examining attributes of a free response listening test. *Language Testing, 15*(2), 119–157. https://doi.org/10.1177/026553229801500201
12. Chen, X., Aryadoust, V., & Zhang, W. (2025). A systematic review of differential item functioning in second language assessment. *Language Testing*, *42*(2), 193-222. https://doi.org/10.1177/02655322241290188
13. Chiu, C.-W., & Chen, T.-P. (2023). Speech rate and young EFL learners' listening comprehension. *English Language Teaching, 16*(7), 74–74. https://doi.org/10.5539/elt.v16n7p74
14. Elkhafaifi, H. (2005). Listening comprehension and anxiety in the Arabic language classroom. *Modern Language Journal, 89*(2), 206–220. https://doi.org/10.1111/j.1540-4781.2005.00275.x
15. Field, J. (2008). *Listening in the language classroom*. Cambridge University Press.
16. Field, J. (2012). The cognitive validity of the lecture-based question in the IELTS listening paper. In L. Taylor & C. Weir (Eds.), *IELTS collected papers 2: Research in reading and listening assessment* (pp. 391–453). Cambridge University Press.
17. Griffiths, R. (1990). Speech rate and NNS comprehension: A preliminary study in time-benefit analysis. *Language Learning, 40*(3), 311–336. https://doi.org/10.1111/j.1467-1770.1990.tb00666.x
18. Griffiths, R. (1992). Speech rate and listening comprehension: Further evidence of the relationship. *TESOL Quarterly, 26*(2), 385–390. https://doi.org/10.2307/3587015

19. Harding, L. (2012). Accent, listening assessment and the potential for a shared-L1 advantage: A DIF perspective. *Language Testing, 29*(2), 163–180. https://doi.org/10.1177/0265532211421161
20. Huang, S.-J. (2014). Automated versus human scoring: A case study in an EFL context. *Electronic Journal of Foreign Language Teaching, 11*(Suppl. 1), 149–164.
21. Hussein, M. A., Hassan, H., & Nassef, M. (2019). Automated language essay scoring systems: A literature review. *PeerJ Computer Science, 5*, e208. https://doi.org/10.7717/peerj-cs.208
22. In'nami, Y., & Koizumi, R. (2009). A meta-analysis of test format effects on reading and listening test performance: Focus on multiple-choice and open-ended formats. *Language Testing, 26*(2), 219–244. https://doi.org/10.1177/0265532208101008
23. Kang, O., Thomson, R. I., & Moran, M. (2020). Which features of accent affect understanding? Exploring the intelligibility threshold of diverse accent varieties. *Applied Linguistics, 41*(4), 453–480. https://doi.org/10.1093/applin/amy053
24. Khlaif, Z. N., Ayyoub, A., Hamamra, B., Bensalem, E., Mitwally, M. A. A., Ayyoub, A., Hattab, M. K., & Shadid, F. (2024). University teachers' views on the adoption and integration of generative AI tools for student assessment in higher education. *Education Sciences, 14*(10), Article 1090. https://doi.org/10.3390/educsci14101090
25. Kowal, J. M., Bryant, K. H., Segall, D., & Kantrowitz, T. (2025). Harnessing generative AI for assessment item development: Comparing AI-generated and human-authored items. *International Journal of Selection and Assessment, 33*(3), e70021. https://doi.org/10.1111/ijsa.70021
26. Li, S., & Ng, V. (2024). Automated essay scoring: A reflection on the state of the art. In Y. Al-Onaizan, M. Bansal, & Y.-N. Chen (Eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (pp. 17876–17888). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-main.991
27. Liu, T., & Aryadoust, V. (2024). Does modality matter? A meta-analysis of the effect of video input in L2 listening assessment. *System, 120*, Article 103191. https://doi.org/10.1016/j.system.2023.103191
28. Liu, T., & Aryadoust, V. (2026). An eye-tracking study of the impact of item presentation and item format on cognitive processing in L2 listening assessment. *Studies in Second Language Acquisition*, 1–28. doi:10.1017/S0272263126101648
29. Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). Macmillan Publishing; American Council on Education.
30. Muñoz-Ortiz, A., Gómez-Rodríguez, C., & Vilares, D. (2024). Contrasting linguistic patterns in human and LLM-generated news text. *Artificial Intelligence Review, 57*, Article 265. https://doi.org/10.1007/s10462-024-10903-2
31. Namuduri, R., Wu, Y., Zheng, A. A., Wadhwa, M., Durrett, G., & Li, J. J. (2025). *QUDsim: Quantifying discourse similarities in LLM-generated text.* arXiv. https://arxiv.org/abs/2504.09373
32. Raitio, T., Rasipuram, R., & Castellani, D. (2020). *Controllable neural text-to-speech synthesis using intuitive prosodic features.* arXiv. https://arxiv.org/abs/2009.06775
33. Révész, A., & Brunfaut, T. (2013). Text characteristics of task input and difficulty in second language listening comprehension. *Studies in Second Language Acquisition, 35*(1), 31–65. https://doi.org/10.1017/S0272263112000678
34. Runge, A., Attali, Y., LaFlair, G. T., Park, Y., & Church, J. (2024). A generative AI-driven interactive listening assessment task. *Frontiers in Artificial Intelligence, 7*, Article 1474019. https://doi.org/10.3389/frai.2024.1474019

35. Shadiev, R., & Feng, Y. (2024). Using automated corrective feedback tools in language learning: A review study. *Interactive Learning Environments, 32*(6), 2538–2566. https://doi.org/10.1080/10494820.2022.2153145
36. Shang, Y., Aryadoust, V., & Hou, Z. (2024). A Meta-Analysis of the Reliability of Second Language Listening Tests (1991–2022). *Brain Sciences*, *14*(8), 746. https://doi.org/10.3390/brainsci14080746
37. Sun, J., Song, T., Peng, W., & Song, J. (2025). A survey of automated essay scoring: Challenges, advances, and future. *Neurocomputing, 650*, Article 130916. https://doi.org/10.1016/j.neucom.2025.130916
38. Tan, B., Armoush, N., Mazzullo, E., Bulut, O., & Gierl, M. (2025). A review of automatic item generation techniques leveraging large language models. *International Journal of Assessment Tools in Education, 12*(2), 317-340. https://doi.org/10.21449/ijate.1602294
39. Tao, X., & Aryadoust, V. (2024). A Multidimensional Analysis of a High-Stakes English Listening Test: A Corpus-Based Approach. *Education Sciences*, *14*(2), 137. https://doi.org/10.3390/educsci14020137
40. Vallés-Pérez, I., Roth, J., Beringer, G., Barra-Chicote, R., & Droppo, J. (2021). *Improving multi-speaker TTS prosody variance with a residual encoder and normalizing flows.* arXiv. https://arxiv.org/abs/2106.05762
41. Vandergrift, L. (2005). Relationships among motivation orientations, metacognitive awareness and proficiency in L2 listening. *Applied Linguistics, 26*(1), 70–89. https://doi.org/10.1093/applin/amh039
42. Vandergrift, L. (2007). Recent developments in second and foreign language listening comprehension research. *Language Teaching, 40*(3), 191–210. https://doi.org/10.1017/S0261444807004338
43. Vandergrift, L., & Goh, C. C. M. (2012). *Teaching and learning second language listening: Metacognition in action* (1st ed.). Routledge. https://doi.org/10.4324/9780203843376
44. Weir, C. J. (2005). *Language Testing and Validation: An Evidence-Based Approach*. Basingstoke: Palgrave Macmillan.
45. White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., & Schmidt, D. C. (2023). *A prompt pattern catalog to enhance prompt engineering with ChatGPT*. arXiv. https://arxiv.org/abs/2302.11382
46. Yavariabdi, A., Paudel, B., Carleton, T., & de Almeida, C. D. A. (2025). Generative AI in assessment and feedback generation in higher education: A systematic review. In *Proceedings of the 2025 International Conference on Education Technology and Computers (ICETC)* (pp. 361–371). https://doi.org/10.1109/ICETC66579.2025.11387416
47. Zhang, Y., Lucas, M., Bem-haja, P., & Pedro, L. (2025). AI versus human-generated voices and avatars: Rethinking user engagement and cognitive load. *Education and Information Technologies, 30*, 22547–22566. https://doi.org/10.1007/s10639-025-13654-x
48. Zhao, J., Chapman, E., & Sabet, P. G. P. (2024). Generative AI and educational assessments: A systematic review. *Education Research and Perspectives, 51*, 124–155. https://doi.org/10.70953/ERPv51.2412006
49. Zhao, Y., & Aryadoust, V. (2025). An automatized semantic analysis of two large-scale listening tests: A corpus-based study. *Language Testing*, *42*(3), 312-343. https://doi.org/10.1177/02655322241288598
50. Zheng, C., Liu, Z., Xie, E., Li, Z., & Li, Y. (2023). *Progressive-hint prompting improves reasoning in large language models.* arXiv. http://arxiv.org/abs/2304.09797

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (http://creativecommons.org/licenses/by-nc/4.0/), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

