Designing Construct-Valid L2 Academic Listening Assessments: An Evidence-Informed Framework for Managing Cognitive Demands in High-Stakes EAP Testing

Raihan Md Ghazali[1], Mohd Adib Izzat Mohd Nor[2]

Academy of Language Studies, Universiti Teknologi MARA, Melaka Campus, Malaysia, Faculty of Hotel and Tourism Management, Universiti Teknologi MARA, Melaka Campus, Malaysia

DOI: https://doi.org/10.47772/IJRISS.2026.101000139

Received: 13 October 2026; Accepted: 19 October 2026; Published: 30 October 2026

ABSTRACT

Academic listening tests in English for Academic Purposes (EAP) settings should support inferences about learners' ability to interpret spoken arguments, distinguish central claims from supporting details, and recognize discourse relationships. Yet particular item and delivery choices may introduce reading, memory, or test-wiseness demands that are only partly relevant to the intended listening construct. This conceptual paper integrates an argument-based approach to assessment validity with cognitive load theory and L2 listening research to propose an operational design framework. It specifies complementary macro-level and micro-level listening targets, outlines a discourse-mapping and item-review workflow, illustrates how answer options can privilege meaning over superficial lexical overlap, and presents preview and repeat-listening conditions as testable design alternatives. An illustrative environmental-policy lecture is used to demonstrate item construction rather than to report empirical results. The framework also identifies threats and trade-offs, including construct underrepresentation, altered task authenticity, coaching effects, and the possibility that repeated playback changes what a score means. A staged validation agenda combines expert review, cognitive interviews, randomized format comparisons, psychometric analysis, and fairness checks. The contribution is a falsifiable design proposal, not evidence that any particular delivery protocol already improves validity or fairness.

Keywords: academic listening; English for Academic Purposes; construct validity; cognitive load; item writing; listening assessment; validation

INTRODUCTION

University study in English-medium contexts frequently requires learners to follow lectures, integrate propositions across discourse segments, interpret the speaker's stance, and select relevant details. Listening assessment therefore involves more than recognizing words or recalling numbers. The interpretation of scores depends on which listening operations the test samples and which additional processes its format elicits (Bachman & Palmer, 2022; Fulcher, 2025). Research on L2 listening identifies interactions among perceptual decoding, contextual interpretation, and metacognitive regulation (Buck, 2001; Field, 2008; Vandergrift, 2007; Goh & Vandergrift, 2022).

In a multiple-choice listening examination, a learner may identify an option because it repeats a salient phrase from the recording without integrating the argument. Conversely, a proficient listener may answer incorrectly after losing the referent of a densely written option while listening. Neither possibility establishes that conventional tests are necessarily invalid. Rather, each indicates a potential mismatch between the intended construct and the response process that must be investigated empirically. This is a plausible response-process threat rather than an established characteristic of all EAP tests (Buck, 2001; Field, 2008).

This paper asks how EAP listening assessments can represent discourse-level comprehension while limiting avoidable demands introduced by item wording and administration. It advances three design principles: articulate the construct before writing items, require evidence of semantic and rhetorical comprehension, and compare

delivery conditions rather than assuming that additional support is always beneficial. The framework is offered as a conceptual model and a validation agenda. No participants were recruited, no scores were collected, and no intervention effects are claimed. The framework treats validity as an evidence-based argument concerning proposed score interpretations and uses (Messick, 1989; Kane, 2013).

THEORETICAL AND VALIDITY FOUNDATIONS

2.1 An argument-based view of validity

Validity concerns the evidential justification for interpretations and uses of test scores, not an intrinsic property conferred on an item by its wording alone. An assessment-use argument should connect the proposed score interpretation, task specifications, observed response processes, scoring consistency, generalization, and consequences of decisions (Bachman & Palmer, 2022). A seemingly sophisticated main-idea question can still perform poorly if proficient listeners disagree about its key or if it measures topic familiarity more strongly than listening ability. This argument-based interpretation follows Messick (1989) and Kane (2013), who emphasize support for score-based inferences rather than merely desirable item features.

Construct underrepresentation occurs when the task sample omits relevant capabilities, such as tracking claims, qualifications, and relations across lecture segments. Construct-irrelevant variance may arise when visual reading load, audio defects, specialized background knowledge, or unintended option cues influence performance beyond the intended inference. Importantly, working memory is involved in real-world lecture comprehension. The defensible target is not to remove memory demands altogether but to identify which demands are integral to the target language-use domain and which arise from avoidable testing artifacts. These threats follow the distinction articulated by Messick (1989) and operationalized for listening assessment by Buck (2001).

2.2 Cognitive load as a design lens

Cognitive load theory distinguishes demands associated with the complexity of information processing from demands imposed by how information is presented (Sweller et al., 2019). The conventional labels intrinsic, extraneous, and germane load are useful heuristics, but germane load should not be treated as an independently measurable reserve of capacity generated by reducing extraneous load. More recent accounts integrate schema-building activity with processing intrinsic complexity. Accordingly, this framework makes no arithmetic claim that total load equals a fixed sum or that reducing an item's reading burden necessarily improves score validity. Working-memory limitations are pertinent to assessment design, but cognitive load theory alone does not prove that any particular delivery format is invalid (Sweller et al., 2019).

Simultaneously reading long options and monitoring ongoing speech may encourage divided attention. However, previewing questions can also direct attention toward isolated answers and change the cognitive operations elicited. Similarly, a second listening may reduce transient memory demands but depart from courses in which students normally hear lectures once. Such format decisions require evidence about the intended domain and their consequences for score interpretation (Fulcher, 2025; Goh & Vandergrift, 2022). Research comparing listening-test media likewise shows that changes in presentation mode can change what test takers attend to (Ockey, 2007).

2.3 L2 listening and discourse representation

L2 listening involves interaction between perceptual decoding and higher-level interpretive processes. Listeners must identify words and clause relations while building and revising a representation of speaker intentions and discourse organization (Goh & Vandergrift, 2022). Macro-level interpretation and micro-level verification are analytically distinct assessment targets, although they are interdependent rather than separate cognitive faculties. Failure to decode a critical detail can compromise inference about the main argument, and macro-level context can facilitate detail interpretation. Listening models therefore support assessing discourse comprehension alongside, rather than instead of, linguistic decoding (Buck, 2001; Field, 2008; Vandergrift, 2007; Goh & Vandergrift, 2022).

A CONSTRUCT SPECIFICATION FOR ACADEMIC LISTENING

Test specifications should derive from local academic language-use needs, for example actual first-year lectures and seminars. The proposed blueprint partitions score evidence into two complementary domains and adds discourse-function interpretation as a cross-cutting target. These are design categories, not an asserted two-factor latent structure. Such target-domain specification follows established principles of listening-test development (Buck, 2001; Fulcher, 2025).

Target	Illustrative operation	Evidence in a response
Macro-level synthesis	Identify the controlling argument, compare claims, interpret qualifications	Correct paraphrase preserves scope and argumentative priority
Micro-level verification	Identify a supporting reason, example, condition, or contrast	Response ties a detail to its intended function, not isolated recall
Discourse relations	Recognize shifts from problem to remedy or claim to rebuttal	Response identifies a rhetorical relationship across segments

Blueprint proportions should be justified by a domain analysis rather than prescribed universally. A course whose learning outcomes emphasize seminar interaction may require substantially different tasks from a lecture-comprehension placement examination. Item formats other than multiple choice, such as note completion, short constructed responses, and integrated summaries, should be considered when their scoring quality and accessibility can be established. Test specifications should be justified against intended interpretations and the educational setting (Kane, 2013; Fulcher, 2025).

ITEM-ENGINEERING PROCEDURE

Step 1. Define intended score use and target domain. Document the decisions supported by test scores and sample relevant lecture genres, linguistic demands, interactional conditions, and permitted note-taking practices. The score-use starting point is consistent with an argument-based approach to validation (Kane, 2013).

Step 2. Select and document audio. Obtain rights or permissions, specify speaker, register, length, speech rate, recording quality, and topic familiarity. Avoid treating accent variation as a defect. Assess intelligibility and fairness against the intended listening domain. Audio selection should follow the intended target-language-use domain rather than an abstract notion of authenticity (Buck, 2001).

Step 3. Map discourse structure. Independently annotate the main claim, supporting reasons, examples, counterarguments, qualifications, and transitions. Resolve disagreements before drafting keyed responses.

Step 4. Draft items from mapped propositions. For macro-level items, require selection of a proposition that captures both the claim and its limits. For detail items, test functional understanding where appropriate rather than numerical recall alone. Direct quotation is not automatically invalid, but a key should not be recoverable from lexical overlap without adequate comprehension.

Step 5. Audit distractors. Classify each distractor as a plausible misinterpretation, such as mistaking an example for the main claim, reversing causality, or overstating a qualified statement. Avoid misleading distractors that are implausible to all examinees or require information absent from the recording. The classification is a proposed item-review heuristic, not a taxonomy whose prevalence has been empirically established (Field, 2008; Buck, 2001).

Step 6. Conduct expert and learner review. Elicit judgments about domain relevance, language complexity, one-best-answer defensibility, and the source of distractor attractiveness. Revise after cognitive interviews and pilot statistics; do not rely on arbitrary zero-overlap targets or fixed discrimination cutoffs. An expert review alone cannot establish validity without response and score evidence (Kane, 2013).

WORKED EXAMPLE: AN ILLUSTRATIVE LECTURE EXCERPT

The following excerpt is constructed for item-writing demonstration. It is not a transcript of an authentic lecture and provides no verified national statistics or causal evidence about the impacts of carbon taxation.

"Today I will discuss carbon pricing as one approach to industrial emissions. When a factory releases greenhouse gases, the costs may be borne partly by people outside the transaction. A charge on emissions aims to make firms consider those wider costs in production decisions. In principle, this gives companies a financial reason to reduce emissions when doing so is less costly than paying the charge. Yet the policy can create concerns about firms shifting production to places with weaker climate rules. Policymakers sometimes propose border measures to address this risk. These measures, however, raise questions about administration and international trade. Thus, carbon pricing has a clear economic rationale, but its effectiveness depends partly on policy design and implementation."

5.1 Example assessment items

Item 1. Which statement best captures the lecturer's overall position?

A. Border measures are a complete solution to industrial emissions.
B. Charging for emissions can influence firms' decisions, although practical design affects the outcome. [Key]
C. Firms always relocate when governments introduce environmental charges.
D. Industrial emissions arise mainly from international trade regulations.

Rationale. The key integrates the mechanism and qualification. A misreads a supporting policy response as a complete solution. C overgeneralizes the stated risk. D introduces an unsupported causal priority. The item still requires empirical evidence that the keyed option is uniquely defensible.

Item 2. Why does the lecturer refer to possible relocation of production?

A. To explain a potential unintended effect of domestic policy. [Key]
B. To provide proof that carbon charges never reduce emissions.
C. To define the costs imposed on people outside a transaction.
D. To argue that all border measures are legally prohibited.

Rationale. This item targets the function of a supporting detail, not mere recognition of a repeated phrase. The remaining options misuse the relation between mechanism, risk, and response.

Item 3. How is the argument organized after the explanation of the emissions charge?

A. A risk is identified, followed by a possible response and its complications. [Key]
B. The initial argument is withdrawn after a contradictory experiment.
C. Several unrelated national case studies are presented chronologically.
D. A mathematical derivation replaces the policy discussion.

Rationale. This item targets discourse relations spanning multiple propositions. It should be piloted to establish that the item measures organization rather than recognition of overt signal words alone.

DELIVERY CONDITIONS AS COMPETING HYPOTHESES THESE CONDITIONS ARE HYPOTHESES BECAUSE ASSESSMENT DELIVERY CAN ALTER PROCESSING DEMANDS AND THE CONSTRUCT BEING ELICITED (SWELLER ET AL., 2019; OCKEY, 2007).

The original proposal of a 60–90-second preview and two audio exposures is retained here as an experimental design option, not a universal standard. Duration, audio length, response timing, and permission to take notes

should follow domain analysis and pilot testing. A two-pass format might improve comprehension estimates by reducing transience effects, but it can also change the target task from real-time comprehension to recognition after repetition.

Condition	Proposed administration	Principal validity question
A: Single pass	Brief standardized instructions; one playback; questions available according to existing practice	Does this approximate target academic listening, including note taking?
B: Preview plus single pass	Comparable items; calibrated stem/option preview; one playback	Does preview reduce irrelevant reading demand or induce answer hunting?
C: Preview plus two passes	Same preview; one global listening and one verification listening, with standardized interval	Does repetition improve relevant comprehension or alter the construct and score scale?

The conditions should be counterbalanced using equivalent audio forms to control passage difficulty, fatigue, order, and practice. Inferences from one condition should not automatically be transferred to another. Pre-item preview may disadvantage candidates with slower written reading if it is timed too tightly, while unlimited preview may shift the task toward anticipatory search. Accessibility accommodations should be assessed separately from changes to standard administration. Experimental comparisons should additionally consider the possibility that a second exposure legitimately measures a different listening task rather than simply removing irrelevant difficulty (Buck, 2001; Kane, 2013).

EVIDENCE REQUIRED BEFORE OPERATIONAL ADOPTION

Content evidence. Invite EAP specialists and disciplinary lecturers to rate representativeness of lectures and tasks. Record disagreements about key propositions, vocabulary demands, and the authenticity of each administration condition. This evidence addresses representativeness of the target domain and construct definition (Buck, 2001; Messick, 1989).

Response-process evidence. Use retrospective cognitive interviews, targeted verbal reports after listening, and selective eye tracking where feasible. Interpret eye fixation patterns cautiously: looking at options cannot by itself establish lexical matching or comprehension. Contrast the processes of higher- and lower-performing candidates. Response-process methods must be interpreted cautiously because reports and observable attention do not uniquely identify underlying comprehension (Ockey, 2007; Kane, 2013).

Internal-structure and reliability evidence. Examine item difficulty, distractor functioning, corrected item-total correlations, internal consistency, conditional standard errors, and dependence among items sharing a passage. Factor analysis may compare one-factor and related multidimensional representations if sample size and blueprint support these models. Neither a two-factor solution nor a universal point-biserial threshold should be presumed in advance. Such analyses support, but do not independently establish, intended score interpretations (Kane, 2013; Fulcher, 2025).

Relations with external variables. Test whether scores align with independently scored lecture summaries, course-relevant listening performances, and other suitable criteria while addressing topic familiarity and language proficiency. Estimate the incremental information provided by discourse-oriented items over simpler recall items.

Fairness and consequences. Examine differential item functioning and measurement comparability across justified groups such as proficiency levels, educational backgrounds, and familiarity with accents. Group score differences alone do not demonstrate bias. Evaluate score use, access arrangements, instructor consequences, and whether changes alter classification decisions. Fairness claims require evidence relevant to the proposed decision and population rather than assuming identical effects across learner groups (Messick, 1989; Kane, 2013).

Decision rules. Preregister the principal comparisons, sample-size rationale, exclusion rules, and criteria for revising or retaining items. Use effect sizes and uncertainty intervals rather than classifying results solely by statistical significance. Where material changes to administration are made, undertake linking or equating studies before treating scores as interchangeable.

LIMITATIONS AND IMPLICATIONS

The proposed model has several limitations. First, it does not establish that widespread EAP tests systematically reward lexical matching: the prevalence of that behavior must be demonstrated in specified populations. Second, the worked example is invented for pedagogical clarity, so its ecological validity remains untested. Third, the macro/micro distinction may be useful for blueprinting without corresponding to psychometrically separable traits. Fourth, reading load, domain knowledge, and working-memory demands cannot be declared irrelevant without defining the academic performance being assessed. Finally, double playback could increase accessibility in some contexts while decreasing representativeness in others. In particular, theoretical plausibility must not be confused with demonstrated item-level processing or test-score improvement (Kane, 2013; Ockey, 2007).

For practitioners, the defensible immediate implication is to document claims about the construct, create discourse maps, audit answer cues, and pilot alternative delivery protocols. A high-stakes program should adopt a proposed format only after evidence supports the interpretation and consequences of its scores in that local setting.

CONCLUSION

Construct-valid academic listening assessment requires an explicit account of the comprehension processes that an examination intends to elicit and the alternative processes its items might permit. The framework proposed here connects discourse-oriented construct specification to item design, delivery comparisons, and empirical validation. Its value lies in making test-development decisions transparent and testable. Claims that previewing, repeat listening, or paraphrase-based options improve fairness or accuracy should follow, rather than precede, evidence from appropriate validation studies. The next research step is therefore validation of interpretations and uses in the intended EAP context, rather than declaring the proposed format inherently valid (Kane, 2013; Buck, 2001).

REFERENCES

1. Buck,G.(2001). Assessing listening. Cambridge University Press. https://doi.org/10.1017/CBO9780511732959
2. Field, J. (2008). Listening in the language classroom. Cambridge University Press.
3. Fulcher, G. (2025). Practical language testing (2nd ed.). Routledge.
4. Goh, C. C. M., & Vandergrift, L. (2022). Teaching and learning second language listening: Metacognition in action (2nd ed.). Routledge.
5. Kane, M. T. (2013). Validating the interpretations and uses of test scores. Journal of Educational Measurement, 50(1), 1–73. https://doi.org/10.1111/jedm.12000
6. Messick, S. (1989). Validity. In R. L. Linn (Ed.), Educational measurement (3rd ed., pp. 13–103). American Council on Education/Macmillan.
7. Ockey, G. J. (2007). Construct implications of including still image or video in computer-based listening tests. Language Testing, 24(4), 517–537. https://doi.org/10.1177/0265532207080771
8. Sweller, J., van Merriënboer, J. J. G., & Paas, F. (2019). Cognitive architecture and instructional design: 20 years later. Educational Psychology Review, 31(2), 261–292. https://doi.org/10.1007/s10648-019-09465-5
9. 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