Cognitive validity in listen-to-write summary tasks: A mixed-methods analysis of notetaking data

Rebecca Yeager
University of Illinois Urbana-Champaign, USA

GoMee Park
University of Texas Rio Grande Valley, USA

Junhee Park
University of Iowa, USA

This study examines the cognitive validity of a listen-to-summarize task on a university placement test. Cognitive validity is underinvestigated in listening assessment, especially for higher-order listening processes (Holzknecht et al., 2017; Rukthong, 2021). However, student notes from listening tests can record evidence of these processes (Yeager et al., 2024; Cushing, 1991). This study employed an exploratory sequential mixed-methods design to investigate the higher-order listening processes elicited by an integrated listen-to-summarize task. First, notes from two administrations of the same test form – one with notetaking scaffolding (n = 25) and one without (n = 25) – were analyzed qualitatively. Second, qualitative codes informed the development of a rubric to quantify higher-order listening processes across administrations, enabling comparison with summary and multiple-choice scores on another test form (n = 118). Qualitative analysis revealed substantial evidence of higher-order discourse construction processes in student notes, including Selecting, Integrating, Monitoring, and Structure-Building (Field, 2013), especially when scaffolding was provided. Linear regression models indicated that Selecting processes significantly predicted test scores for both the integrated task and the multiple-choice task, while Structure-Building processes were significant only

Email address for correspondence: ryeager3@illinois.edu

for the integrated task. Results support the cognitive validity of the listen-to-summarize task and inform improved test design and score interpretation.

Keywords: Integrated assessment; listening response formats; English for Academic Purposes; test validation; local testing

Introduction

Integrated assessment incorporates multiple language skills in one assessment task (Plakans & Gebril, 2013). It has been found to be more predictive of future academic performance than independent skills tasks (Llosa & Malone, 2019), and may promote positive washback on teaching and learning (Rukthong, 2021). For these reasons, integrated tasks have gained popularity for writing assessment, but are still underused for listening assessment (Rukthong & Brunfaut, 2020).

This study reports on the cognitive validity of an integrated listen-to-summarize task on a local placement test. We can say that an assessment task has cognitive validity when it elicits the same cognitive processes that are elicited in the target language use (TLU) domain (Weir, 2005). Our study builds on Weir (2005) and Taylor and Geranpayeh's (2013) sociocognitive framework for listening test validation, in which listening processes are impacted by characteristics of the test-taker, input, and task, and in turn impact features of the test-taker response, which is then scored and used as the basis for some decision with real world consequences. Figure 1 depicts an adapted version of this model for integrated listening tasks. In this figure, the bidirectional arrow between cognitive validity and context validity shows the close relationship between context variables, such as response format, and cognitive variables, including higher-order listening processes, which are elicited by features of the task. Cognitive validity in listening assessment is threatened when features of the listening task differ from features of listening in the TLU domain (Field, 2013). Therefore, in order to ensure proper interpretation of listening test scores, it is first necessary to confirm that tests elicit the intended cognitive processes.

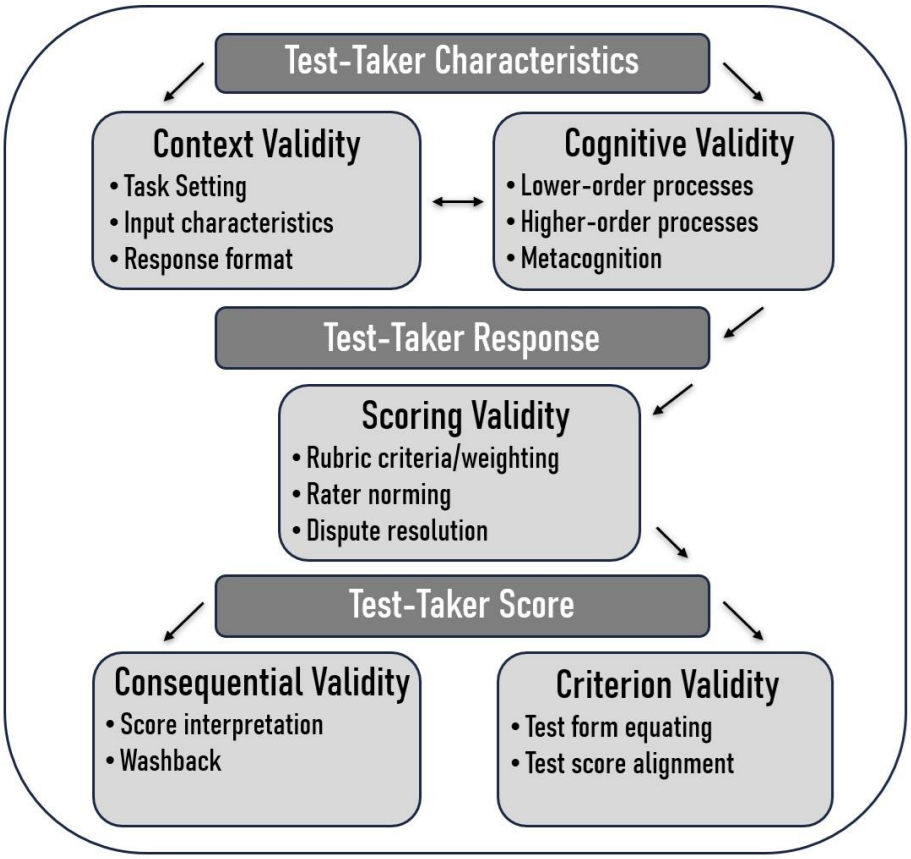


Figure 1. A sociocognitive framework for integrated listening task validation (adapted from Weir, 2005, and Taylor & Geranpayeh, 2013)

In this study, we adopt Field's (2013) taxonomy of lower-order and higher-order listening skills to represent the cognitive processes we want to elicit on our test. In his framework, lower-level processes draw on phonological knowledge, lexical knowledge, and syntactic knowledge, which contribute to input decoding, lexical search, and parsing. These processes help the listener interpret the acoustic cues in the signal as a propositional element. Field (2013) then divides higher-level processes into meaning construction and discourse construction processes. Meaning construction processes draw on pragmatic knowledge, world knowledge, and speaker knowledge to create a representation of the text which includes speaker intentions, context, inference, and reference. Finally, the listener combines all of these sources of knowledge in the last stage of listening: discourse construction. Discourse construction is comprised of four skills or processes: Selecting (choosing which ideas to focus on), Integrating (establishing links between adjacent

ideas), Monitoring (ensuring a coherent representation of the text), and Structure-Building (representing the relative importance of ideas).

Multiple studies of cognitive processes in second language (L2) listening assessment have observed fairly consistent evidence of lower-order processes, but marginal evidence of higher-order listening processes, even on tests that claim to assess them (Holzknecht et al., 2017, Rukthong, 2021). For that reason, although we recognize the value of lower-order listening skills, we focus our analysis on higher-order processes in this study. We attempt to do this by analyzing student notes taken during three administrations of a local placement test, using mixed methods to look for evidence of the four higher-order discourse construction processes identified in Field (2013)'s taxonomy of listening skills.

Cognitive validity in listening assessment

Listening assessment researchers have attempted to observe evidence of cognitive validity in a number of ways. However, this evidence is difficult to obtain because listening is both invisible and highly contextualized. Traditionally, cognitive validity in listening assessment has been investigated through self-report measures, including surveys, interviews, and stimulated recalls. Much of what we know today about cognitive validity in listening assessment comes from these measures; however, this data must be interpreted cautiously due to possible inaccuracies introduced by poor memory, poor self-awareness, or the desire to save face (Craig et al., 2020).

Recently, new forms of evidence are emerging which provide more direct evidence of listening behaviors, such as eye-tracking and functional near-infrared spectroscopy (fNIRS). These measures are less susceptible to self-report bias, but introduce their own challenges in interpretation. For instance, Holzknecht (2019) notes that focusing and zoning behaviors are indistinguishable in eye-tracking studies unless triangulated with stimulated recall data. Further, fNIRS data, which records blood flow near the surface of the brain, is only reliable at shallow depths (Klein et al., 2022) and may be imprecise unless adjusted for subject-specific anatomy (Bonilauri et al., 2023). In addition, both of these methods are limited in that they can only be used in laboratory studies, not in live

test administration contexts. The degree of generalizability between findings from laboratory studies and live testing contexts is still uncertain, but it may reasonably be expected that cognitive processes in particular may vary across these contexts due to the impact of affective factors such as anxiety and motivation. Finally, not all testing programs are able to afford the technology necessary to run eye-tracking or fNIRS experiments.

Each of the methods above provide valuable information but should be triangulated with multiple sources of evidence to account for known limitations. Further, we propose that it is desirable to attempt to collect at least some evidence of cognitive processes during live testing in a non-obtrusive way in order to ensure that the test is functioning as intended. Fortunately, many academic listening assessments enable the investigation of another data source which can be tapped for this type of evidence: notetaking. If notes are allowed on a test, they can be collected and analyzed for evidence of listening processes.

We are not the first to suggest the value of notetaking data as a source of evidence of cognitive validity; Cushing (1991), Faraco et al. (2002), and Rukthong (2021) have employed this source of evidence and serve as inspiration for our work. However, the majority of L2 notetaking studies have focused on the product or outcome of taking notes rather than the processes embodied in notes, often in service of determining whether or not to allow notetaking for a given task. The results of these studies are complex and will be discussed below, but here we want to reverse the focus of our inquiry. We are not investigating the value of notetaking but the value of the listening task itself. Rather than asking whether notetaking is useful for a task, we want to ask whether the task is cognitively valid; that is, whether it elicits and rewards the intended cognitive processes. If we find that notes do not show evidence of such processes, or if we discover that those processes in notes do not correlate with exam performance, we judge that as evidence of a threat to cognitive validity and an indicator that the task itself is in need of revision.

We recognize that we tread on delicate ground here. Just as other data sources which bear on cognitive validity are subject to limitations, so too is notetaking. Student notes are notoriously difficult to interpret precisely because they are sensitive to so many variables:

the amount, type, and effect of academic notes depends on characteristics of the notetaker, the listening context, and the assessment task. We attempt to summarize some of these variables below, and recognize that this complexity is likely the primary reason why notetaking has been overlooked for so long as a source of evidence about cognitive validity. However, we also believe this complexity is an asset which allows for more nuanced analysis of listening processes, especially when investigated qualitatively. This study documents an attempt to develop a mixed-methods protocol to analyze notes as evidence of cognitive validity in listening assessment.

Notetaking and listening assessment

A robust history of research on academic notetaking points to its ongoing importance for university students (Morehead et al., 2019; Potvin et al., 2023). Academic notetaking enables two functions which are beneficial to learning: deeper encoding and external storage, which refer to the process and product of notetaking, respectively (Kobayashi, 2005; 2006). However, the relative impact of notetaking depends on characteristics of the student, the lecture context, and the assessment task. Among students in university contexts, L2 listeners may have different challenges, such as identifying main ideas (Olsen & Huckin, 1990) and discourse markers (Clerehan, 1995), and advantages, such as the ability to encode notes in multiple languages (Canham et al., 2024). Among different contexts, the difficulty of the content appears to have a U-shaped relationship to note quantity: students take fewer notes on familiar material and more notes on challenging material (Witherby & Tauber, 2019), but sometimes abandon notetaking entirely if they feel too overwhelmed (Loughlin, 2015). Also, among assessment types, the impact of taking one's own notes is higher for productive tasks than recognition tasks (Kobayashi, 2005) and for delayed tasks than immediate tasks (Kim, 2018).

Surprisingly, the relationship between notetaking and L2 assessment performance does not always reflect the patterns established in academic notetaking research more broadly, which has led some L2 assessment researchers to conclude that notetaking is not useful for L2 test-takers (Clark et al., 2014). However, we submit that this disconnect can be better explained by a combination of the characteristics mentioned above. First, L2

students may have differing levels of experience and training in taking notes in lecture contexts (Siegel & Kusumoto, 2022). Secondly, L2 test-takers who are at ceiling or at floor (that is, who find the test extremely difficult or extremely easy) may cease taking notes entirely, and this could skew the relationship between notes and performance. Finally, the tasks typically used in L2 listening assessment often differ from academic listening tasks. A review of the literature on notetaking in L2 assessment points to a developing pattern, with stronger relationships between notetaking and performance on more authentic tasks. On L2 tests, notes generally do not improve performance on MCQ questions which are previewed before the lecture (Clark et al., 2014; Hale & Courtney, 1994; Sadeghi & Zeinali, 2015), but generally do boost performance on MCQ tests without preview, although the effects are relatively small (Borkovska & Irgin, 2026; Carrell, 2007; Hayati & Jalilifar, 2009; Irgin, 2025; Kim, 2023). Finally, relationships between notetaking and L2 test scores are most marked for productive or integrated tasks (Jin & Webb, 2023; Liu & Hu, 2012; Rukthong, 2021). We are concerned that simplified L2 listening tasks, especially in high-stakes contexts which incentivize washback, could promote test-taking strategies which compete with the development of academic notetaking skills (Yeager et al., 2024). This might explain why L2 notetaking studies do not always dovetail with academic notetaking studies, and raises concerns about the generalizability of performance on L2 listening tests to academic listening contexts as well.

For this reason, we theorize that higher-order listening processes in L2 notes will show stronger associations with L2 listening test performance on integrated tasks which are well-calibrated to students' ability level, especially where students are provided with scaffolding for notetaking and are allowed to use their full linguistic repertoire. We undertake an analysis of student notes collected across three placement test administrations in the hopes of discerning whether our test is eliciting the intended cognitive behaviors. Again, we undertake this analysis as a means of validating our listening test, not as a means of assessing the value of notetaking or a means of assessing the proficiency of individual test-takers. We begin from the premise that notetaking is a valuable academic skill, but do not assume that notes provide an exhaustive snapshot of an individual test-taker's listening comprehension, since notes may be influenced by

other factors beyond proficiency (e.g., notetaking experience and topic familiarity). Rather, we observe notetaking patterns at the corpus level across test administrations to discern whether our task format is eliciting and rewarding the cognitive processes targeted by our construct. To address these gaps, we designed an exploratory sequential mixed-methods study using notetaking data to test three validity questions.

Methods

In a previous study, we used mixed methods to analyze notes from two MCQ listening tasks: one with preview and one without (Yeager et al, 2024). We observed more higher-order listening processes in the format without preview, and more testwise strategies were rewarded in the preview format. We now want to apply this method to analyze notes from a placement test which assesses comprehension of one lecture with two response tasks: an integrated listen-to-summarize task, and a MCQ task without preview. The local Institutional Review Board granted an approval waiver for this study, as it relies exclusively on data collected during normal test administration processes and is intended to improve the quality of the local test.

Students who are held for the placement test have been accepted to a large public university in the United States with TOEFL scores between 80-99, IELTS scores between 6.5-7.5, and DET scores at or above 110 (University of Iowa Admissions, n.d.). The placement test is paper-based and determines placement into EAP Listening, Reading, Oral Skills, and Writing classes through a combination of four integrated writing tasks, 20 selected-response items, and an oral interview. Integrated writing tasks are rated by two human raters using a 5-point analytic scale, with disagreements resolved by a third human rater. The placement test rubric and a full practice test are available on the university website (University of Iowa ESL Programs, n.d.). Further information on the design, development, rater reliability, and predictive validity of the placement test is reported in Yeager and Martinez (2025a) and supplemental materials hosted by the Open Science Foundation (Yeager & Martinez, 2025b).

The listening portion of the test includes one listening input with two response tasks. The listening input is a 10-minute monologic lecture, played once, and semi-scripted with natural oracy features (Wagner & Wagner, 2016). Notetaking is allowed but not required, and students may take notes in any language. After the lecture, students must write a paragraph summarizing five key ideas from the lecture, and answer five MCQs targeting aspects of the construct which cannot be captured in a summary, including vocabulary, detail, and inference items. MCQ items are not previewed.

The study asked three validation questions:

1. To what extent does the integrated listening task ELICIT the higher-order processes targeted by the construct?

2. To what extent do higher-order processes observed in notes DIFFER when notetaking scaffolding is provided?

3. To what extent do higher-order processes observed in notes PREDICT integrated summary scores and MCQ scores?

To explore these questions, we analyzed test scores and notes from all test-takers who took the placement test in Spring 2023 (n = 25), Fall 2023 (n = 118), and Spring 2024 (n = 25). The two Spring administrations used the same test form, and differed from one another only by the introduction of notetaking scaffolding in Spring 2024 (described below). The Fall 2023 administration used a different test form and did not include notetaking scaffolding. Questions 1 and 2 were addressed through mixed methods analysis of the Spring 2023 and 2024 data, and Question 3 was addressed through quantitative analysis of data from Fall 2023, applying the rubric that was developed for Question 2. We restricted the analysis for Question 3 to Fall 2023, due to insufficient sample size from Spring 2023 and Spring 2024 for the type of model we wanted to explore. Our study protocol thus represents an exploratory sequential mixed-methods design (Creswell & Plano Clark, 2011).

To answer Question 1, notes from Spring 2023 and 2024 were analyzed qualitatively by the first two authors in ATLAS.ti (Mac Version 25.01). First round provisional codes were

adopted from our previous study (Yeager et al., 2024). These codes target the four higher-order discourse construction processes outlined by Field (2013): Selecting (selecting which ideas are noteworthy), Integrating (showing connections between neighboring ideas), Monitoring (processing uncertainty and conflicting evidence), and Structure-Building (representing the relative importance of ideas). In addition to these provisional codes, we used open coding (Saldaña, 2015) to identify unexpected patterns in our dataset.

After initial coding and discussion of ten random samples, the first two authors engaged in multiple rounds of coding over the entire dataset, leading to revision of the subcodes for the four processes and the addition of a new major code: Translanguaging (Canham et al., 2024). Figure 2 presents the finalized codes and subcodes.

Selecting	Integrating	Monitoring	Structure Building	Other Strategies
➢ Content: ○ Focused ○ Unfocused ○ Random ➢ Emphasis: ○ Text ○ Self	➢ Numbering ➢ Arrows	➢ Uncertainty Strategies: ○ Question marks ○ Blanks ○ Ellipses ○ Phonetic Dictation ➢ Revision Strategies: ○ Cross-outs ○ Inserts ○ Tagging ○ Delineation	➢ Unstructured: ○ Random/Messy ○ Linear ➢ Framing: ○ Introduction ○ Conclusion ➢ Subordination Strategies: ○ Indentation ○ Word Cloud ○ Brackets ○ Parentheses ○ Cohesive Devices	➢ Translanguaging

Figure 2. Qualitative codes and subcodes

Question 2 was concerned with the impact of providing students with a brief text box containing simple instructions on how to take good notes, in accordance with principles of Learning-Oriented Assessment (LOA; Carless, 2007; Thao & Trang, 2022). This scaffolding, which was introduced in Spring 2024, is presented in Figure 3.

Source 3: Solution 2

Listen to the lecture one time. You may take notes while you listen. Do not turn the page until the lecture is finished. After the lecture is finished, turn the page to answer the five multiple-choice questions and write a summary about Source 3. You may use your notes to help you answer the questions and write the summary.

How to Take Good Notes

1. Focus on main ideas.
2. Try to show the organization of the lecture in your notes (how smaller ideas connect to bigger ideas).
3. Don't worry about writing down every word; use short words and drawings to save time.

Figure 3. Notetaking scaffolding Spring 2024

To answer Question 2, the first two authors used the codes from the qualitative analysis to create an analytic rubric with five subscales to quantify the four discourse construction processes (plus translanguaging) in notes. The Selecting subscale was described in five bands to enable direct comparison of the number of key ideas present in notes and in the summary task. The remaining discourse construction subscales (Integrating, Monitoring, and Structure-Building) were described in four bands representing the subcodes from the qualitative analysis, focusing on nonverbal elements in notes to simplify the rating process. Finally, Translanguaging was indicated on a binary subscale, indicating presence or absence of any L1 use in notes.

Notes from Spring 2023 and Spring 2024 were rated by the first two authors in two rounds. After the first round, the rubric was revised, and the entire dataset was re-rated by both authors with satisfactory inter-rater reliability, as measured by Pearson correlation coefficients: Selecting (1.00), Integrating (0.86), Monitoring (0.96), Structure-Building (0.93), and Translanguaging (1.00). Translanguaged notes from Spring 2023 (Farsi n = 3, Japanese n = 1) and 2024 (Arabic n = 1, Farsi n = 3, Korean n = 5) were checked by fluent speakers for evidence of the five key ideas to confirm the Selecting subscores. The final rubric is displayed in Figure 4.

Higher-Order Discourse Construction Processes Rubric

	Selecting	Integrating	Monitoring	Structure-Building	Translanguaging
1	0-1 key idea[1] represented in notes	No evidence of linking strategies[2]	No evidence of uncertainty[3] and revision[4] strategies	No evidence of subordination strategies[5]	No evidence of L1 use
2	2 key ideas represented in notes	Minimal evidence of linking strategies	Minimal evidence of uncertainty and revision strategies	Minimal evidence of subordination strategies	Evidence of L1 use
3	3 key ideas represented in notes	Moderate evidence of linking strategies	Moderate evidence of uncertainty and revision strategies	Moderate evidence of subordination strategies	
4	4 key ideas represented in notes	Extensive evidence of linking strategies	Extensive evidence of uncertainty and revision strategies	Extensive evidence of subordination strategies	
5	5 key ideas represented in notes				
Notes	[1]See test Answer Key	[2]arrows and numbering	[3]questions, blanks, and ellipses [4]cross-outs, inserts, tagging, and delineation	[5]indentation, word cloud, brackets, and parentheses	

Figure 4. Analytic rubric for higher-order discourse construction processes

To compare the evidence of listening processes across semesters, five general linear models were constructed, one predicting each subscale on the rubric, with Semester as the sole binary predictor. Analysis was performed in R Statistical Software (v4.4.1; R Core Team, 2024). All code for this project is available at the Open Science Foundation (OSF; Yeager et al., 2026).

Question 3 was concerned with the relationship between evidence of higher-order processes in notes and test scores on the summary and multiple-choice tasks. To answer Question 3, notes from Fall 2023 were rated using the rubric developed under Question 2 (n = 118). First, fifteen percent of the notes from Fall 2023 were rated by the first and third authors. After an initial discussion, the Selecting band was re-rated for one minor point, after which all subscales exhibited satisfactory inter-rater reliability, as measured by Pearson correlation coefficients: Selecting (0.93), Integrating (0.93), Monitoring (0.87), Structure-Building (0.96), and Translanguaging (1.00). Afterwards, the first author rated the remaining notes from Fall 2023. The content of translanguaged notes

from Fall 2023 (Chinese n = 6, Farsi n = 1, Korean n = 3, Spanish n = 1) was checked by fluent speakers of each language to confirm the Selecting subscores.

To observe the effect of discourse construction processes on test performance, two linear main effects models were constructed, with the four discourse construction subscores as independent variables. Model 1 predicted summary task scores as the outcome variable, and Model 2 predicted MCQ task scores. Interaction effects were explored for both models but were not significant. This analysis was performed in R Statistical Software (v4.4.2; R Core Team, 2024), with code available at the Open Science Foundation (Yeager et al., 2026).

Results

With regard to Question 1, a qualitative analysis of notes from Spring 2023 and Spring 2024 revealed that students demonstrated evidence of multiple strategies exemplifying the four higher-order discourse construction processes from Field (2013). Across both semesters, Structure-Building was the most common category, making up nearly half of all codes, followed by Integrating, Monitoring, and Selecting. This overall pattern remained consistent across both datasets, regardless of how much note-taking occurred.

Building on Field's (2013) taxonomy, we subdivided Selecting codes into Content and Emphasis strategies. Students with a Focused Content selection strategy zeroed in on main ideas (with varying attention to detail for less important ideas). Students who adopted an Unfocused Content selection strategy recorded some sections of the lecture in some detail but left other major sections entirely blank, and students who adopted a Random Content strategy did not distinguish between main ideas and minor details, recording words seemingly at random. We further observed that some students emphasized certain words through underlining or circling. Text Emphasis marked key words in the instructions or scaffolding, while Self Emphasis marked student's own notes from the lecture. Figure 5 displays two samples with missing and extensive Emphasis codes (marked in red boxes). Due to test security constraints, we cannot provide examples of Content codes. Note that we have also chosen to blur connected text in these samples

to protect student anonymity, especially as handwriting identification tools become more widespread (Semma et al., 2021). We present notetaking samples to illustrate nonverbal signifiers such as numbers, arrows, underlining, and spatial organization.

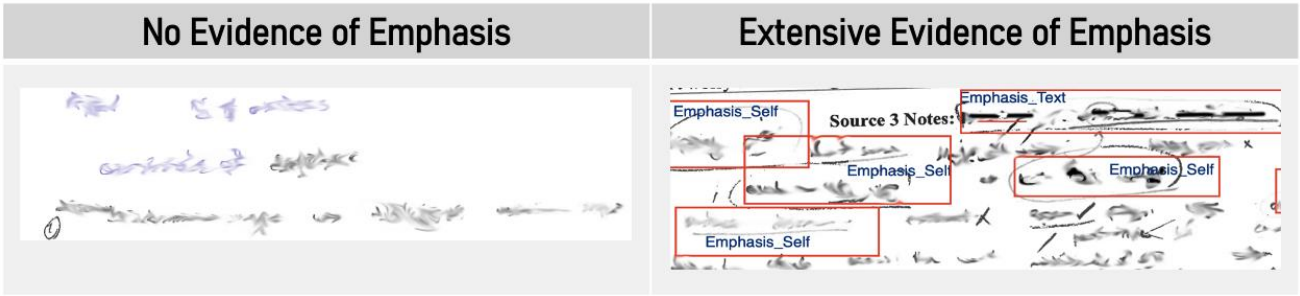


Figure 5. Examples of missing and extensive Selecting (emphasis) codes.

Integration codes included Numbers and Arrows to link ideas linearly or horizontally. Although we observed evidence of only two strategies within this category, both were employed liberally; these subcodes were the most common in our study, present even in low-scoring test samples. We distinguished Integrating arrows from Monitoring arrows (e.g., Insertion and Tagging arrows, see below) and Structure-Building arrows (e.g., Indentation or Word Cloud arrows, see below) by looking at the context. Arrows which inserted information above or below the line, or which continued an idea in a new location on the page, were classified as evidence of Monitoring strategies. Arrows which indicated a hierarchical or subordinate relationship between ideas were filed under Structure-Building. We initially marked but ultimately excluded from our coding protocol any arrows which appeared to be symbolic representations of concepts (such as ↑ increasing or ↓ decreasing), as the use of non-verbal symbols did not seem to clearly align with any of the discourse construction processes in our study. Only arrows which indicated a linear relationship between ideas (adjacency, chronology, or process) were filed under Integrating. Figure 6 depicts samples of missing and extensive Integrating codes (again marked in red boxes).

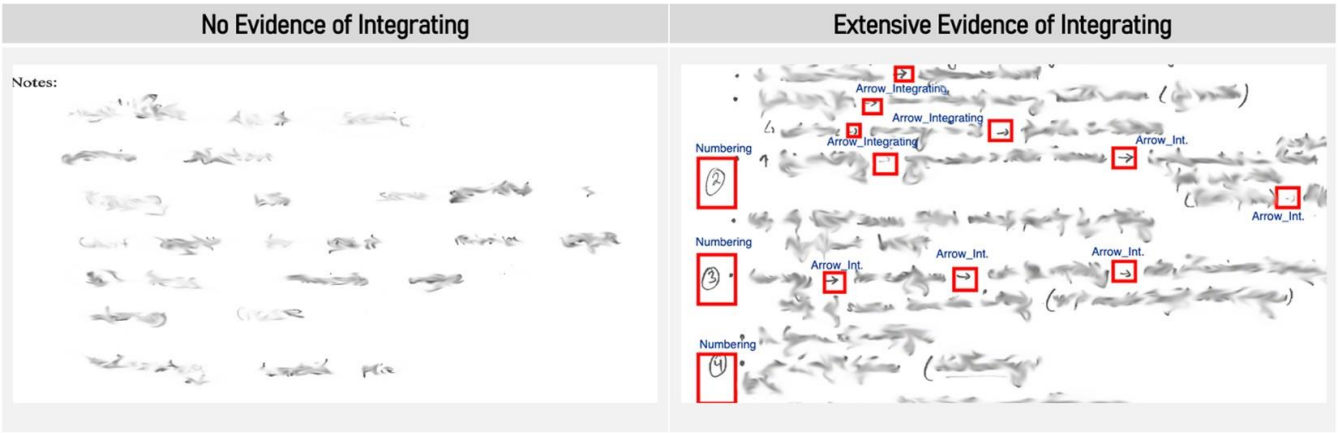


Figure 6. Examples of missing and extensive Integrating codes

Monitoring codes were subdivided into two categories. Uncertainty strategies included any indication that the test-taker was in doubt about whether they had heard something correctly. These strategies included Question Marks, leaving a space Blank in a list or a sentence, or trailing off with Ellipses. We also included Phonetic Dictation of unknown words. Revision strategies included Cross-outs and Insertions as well as strategies to connect or separate ideas on the physical page, such as drawing an arrow to another section (Tagging), or cramming two ideas closely together and drawing a line to separate them (Delineation). Figure 7 illustrates samples with missing and extensive Monitoring codes.

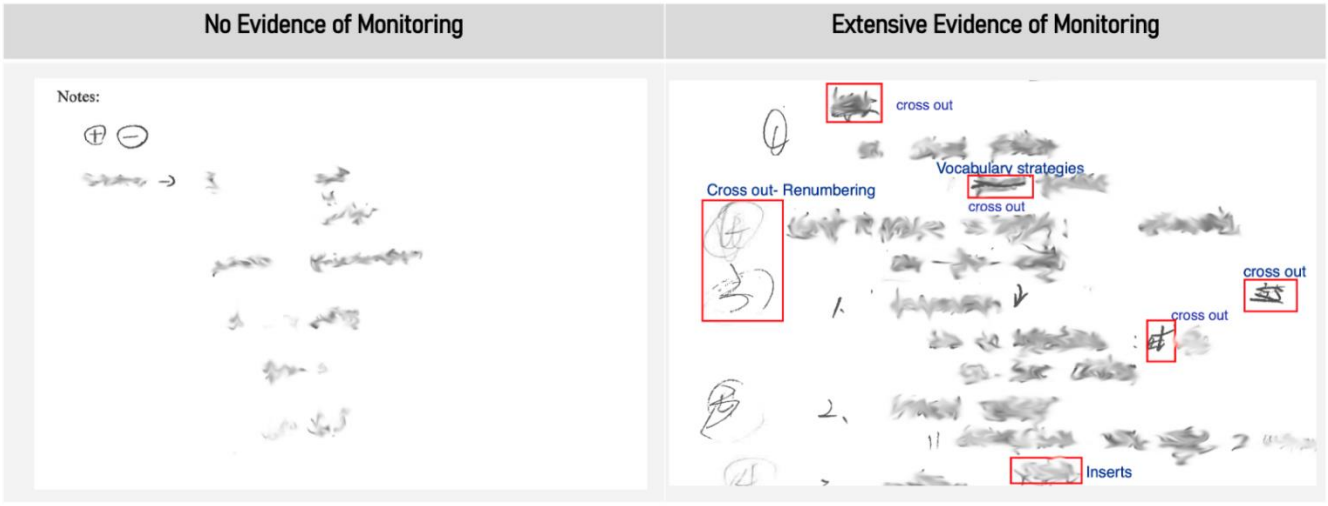


Figure 7. Examples of missing and extensive Monitoring codes

Structure-Building codes were subdivided into three categories: Unstructured, Framing, and Subordination. Unstructured notes were labelled either as Random/Messy (if no order was visible) or Linear (if points were listed with no substructure). We also looked for records of Framing devices from the Introduction or Conclusion. Under Subordination, we looked for evidence of a hierarchical structure to represent lecture propositions. Indentation was the most common strategy in this category, with up to four levels of indentation observed in some samples. We also saw substantial use of Word Clouds and Brackets, which allowed for more flexible representation of ideas. Finally, several students employed Parentheses to indicate trivial or deprioritized information. Figure 8 displays samples of missing and extensive Structure-Building codes.

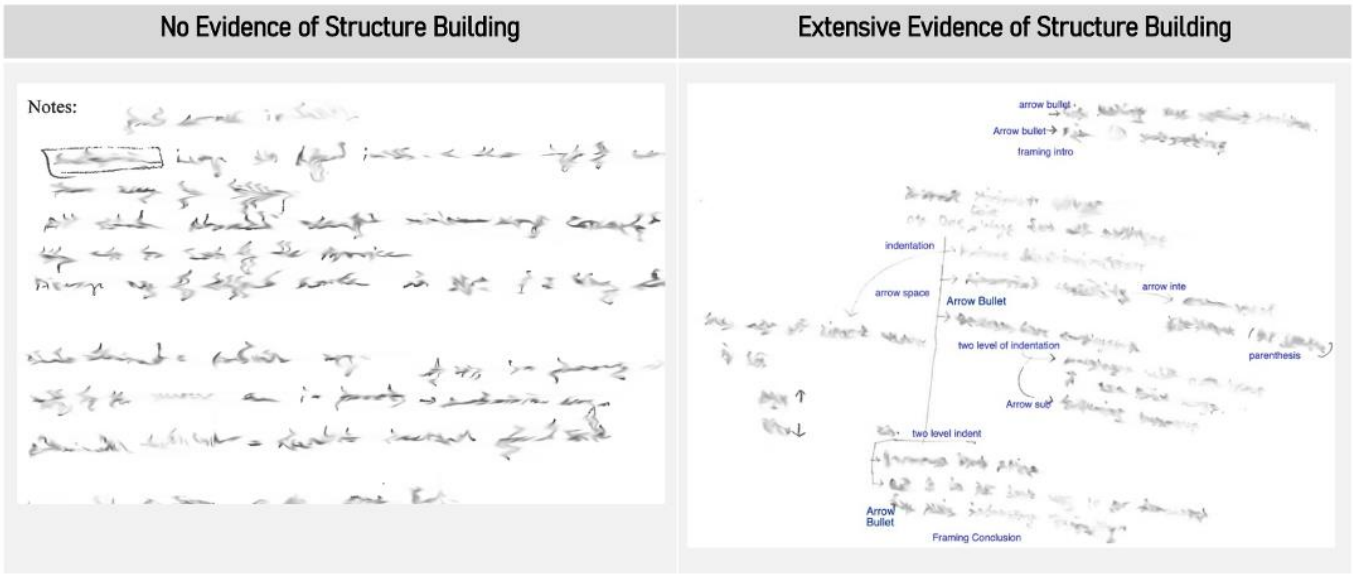


Figure 8. Examples of missing and extensive Structure-Building strategies

Although most students chose to record notes exclusively in English, we did observe several samples of Translanguaging. Most were primarily in English with a few words or phrases in L1. One student took notes primarily in L1 with numbers and key words in English. No samples were written exclusively in L1. Translanguaging allowed for fluid shifts between languages, especially in Spring 2024, frequently representing names, definitions, transitions, and exclamations. Some students inserted L1 in the middle of a phrase to connect two English words, while others used L1 phrases which were more concise than their L2 counterparts. Figure 9 displays a fairly typical example of

Translanguaging. Again, content words in L1 (in red boxes) and their translations (in blue font) are blurred for test security reasons.

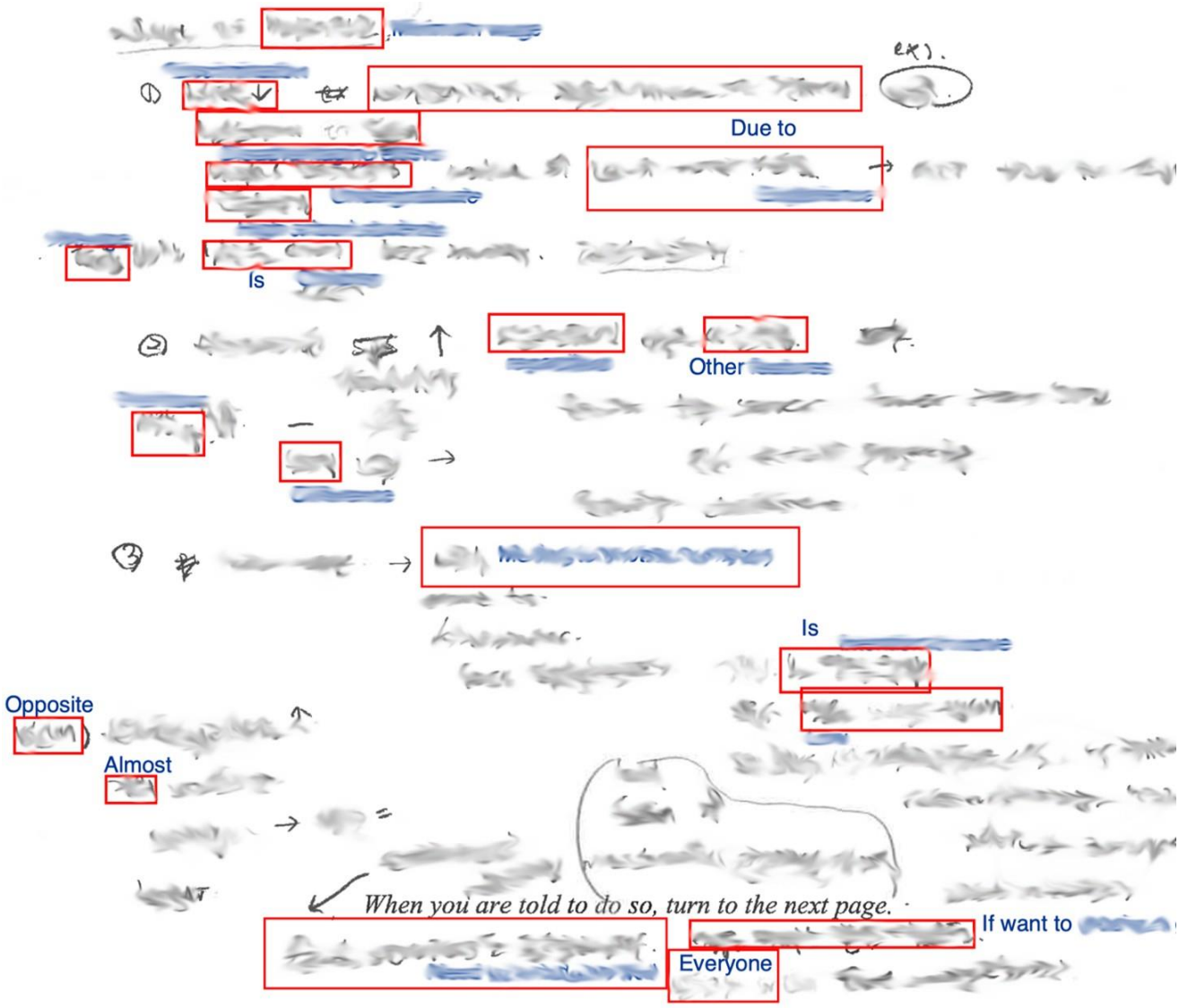
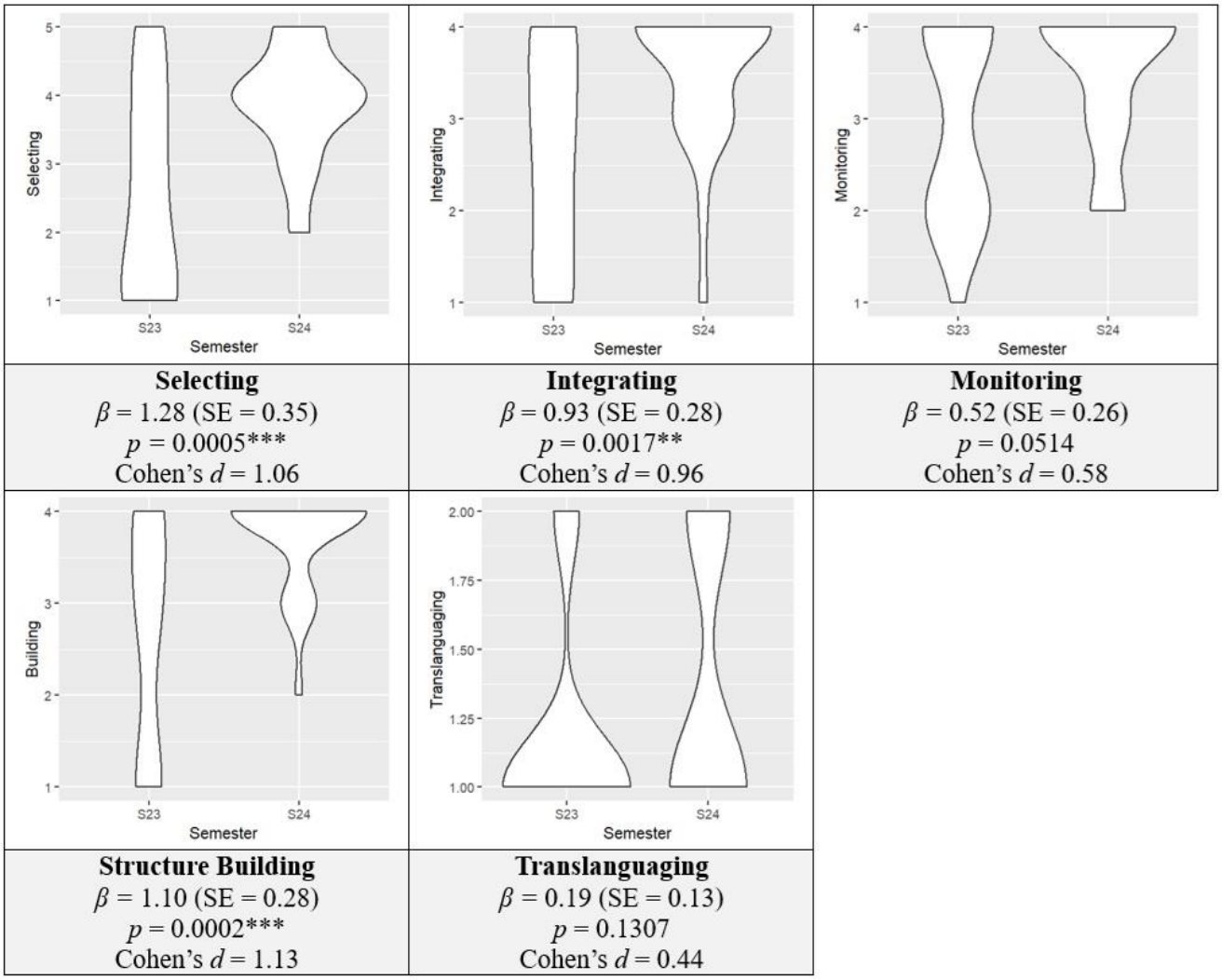


Figure 9. Translanguaging strategies

With regard to Question 2, we wanted to ascertain whether the provision of notetaking scaffolding elicited more higher-order listening processes. We examined the data across Spring 2023 and Spring 2024 both qualitatively and quantitatively. Our qualitative analysis indicated that although all strategies were observed across both semesters, students took substantially more notes in Spring 2024. In Spring 2023, several students took minimal notes, mostly consisting of random words, and one student took no notes

at all. In Spring 2024, all students took notes, and none could be classified as minimal. Specifically, we observed that in Spring 2024, students were more attentive to the five key ideas, used more integration arrows, and employed more subordination strategies such as Indentation and Word Clouds. Further, there were more examples of translanguaging in Spring 2024. However, monitoring strategies did not appear to differ systematically across semesters.

Figure 10 depicts the results of five general linear models predicting each subscore on the notetaking rubric, with Semester as the sole main effect in each model. The estimate for each model represents the difference between means across semesters. In each image, the violin plot on the left represents the distribution of subscores for Spring 2023, and the plot on the right represents the distribution for Spring 2024. Throughout the remainder of this paper, asterisks are used to represent significance levels: $p < .05^{*}$, $p < .01^{**}$, $p < .001^{***}$. The effect of Semester was significant for Selecting ($\beta = 1.28$, SE = 0.35, $p < .0005$, Cohen's $d = 1.06$), Integrating ($\beta = 0.93$, SE = 0.28, $p < .0017$, Cohen's $d = 0.96$), and Structure-Building ($\beta = 1.10$, SE = 0.28, $p < .0002$, Cohen's $d = 1.13$). However, Monitoring and Translanguaging strategies did not differ significantly across semesters. These findings suggest that students who received notetaking scaffolding in Spring 2024 showed significantly more Selecting, Integrating, and Structure-Building processes in their notes, but not Monitoring or Translanguaging.



Selecting
$\beta = 1.28$ (SE = 0.35)
$p = 0.0005$***
Cohen's $d = 1.06$

Integrating
$\beta = 0.93$ (SE = 0.28)
$p = 0.0017$**
Cohen's $d = 0.96$

Monitoring
$\beta = 0.52$ (SE = 0.26)
$p = 0.0514$
Cohen's $d = 0.58$

Structure Building
$\beta = 1.10$ (SE = 0.28)
$p = 0.0002$***
Cohen's $d = 1.13$

Translanguaging
$\beta = 0.19$ (SE = 0.13)
$p = 0.1307$
Cohen's $d = 0.44$

Figure 10. Results of five linear regression models predicting listening processes by Semester

Our third validation question asked whether the higher-order discourse construction processes in student notes from Fall 2023 (N = 118) would predict higher scores for the summary and MCQ task. Table 1 provides the descriptive statistics for the two listening task scores and the five categories on the notetaking rubric. The full range of scores was observed across all variables. Due to low representation of translanguaged samples in Fall 2023, Translanguaging was excluded from further analysis.

Table 1. Descriptive statistics for listening test scores and notetaking rubric from Fall 2023

Variables		Mean	SD	Min	Max
Test Scores	Summary	4	1.07	1	5
	MCQ	4.07	1.05	1	5
Notetaking Rubric Scores	Selecting	4.07	1.27	1	5
	Integrating	2.9	1.07	1	4
	Monitoring	2.74	0.98	1	4
	Structure-Building	3.28	0.9	1	4
	Translanguaging	1.08	0.28	1	2

Table 2 displays the results from a correlation matrix analysis for Fall 2023 for the two listening task scores and the four discourse construction process scores from the notetaking rubric. No evidence of multicollinearity was observed. Moderate correlations were observed between the two listening tasks and between Selecting, Integrating, and Structure-Building. Monitoring processes did not appear to correlate with the other variables.

Table 2. Correlation matrix for listening test scores and discourse construction processes from Fall 2023

Variables		Summary	MCQ	Selecting	Integrate	Monitor	Structure-building
Test Scores	Summary	1					
	MCQ	0.39	1				
Note-taking Rubric Scores	Selecting	0.49	0.41	1			
	Integrating	0.24	0.24	0.33	1		
	Monitoring	0.09	0.08	0.08	0.1	1	
	Structure-Building	0.38	0.29	0.43	0.39	0.14	1

Two separate linear models were then constructed to predict scores for each listening task. Table 3 shows the results of a linear main effects model predicting summary scores from Fall 2023 with the four discourse construction processes as main effects. The overall model was statistically significant, $F(4, 113) = 10.85$, $p < .0001$, explaining 27.8% of the variability in the summary score ($R^2 = .278$). Selecting ($\beta = .337$, SE = .077, $t = 4.40$, $p < .001$) and Structure-Building ($\beta = .233$, SE = .111, $t = 2.091$, $p < .001$) emerged as significant predictors of summary test scores. However, no significant relationships were observed for Integrating and Monitoring. This suggests that students who recorded more of the five key ideas from the lecture and who visually represented the structure of the

lecture in their notes were also more likely to receive higher scores from raters on the summary task.

Table 3. Results of a linear model to predict summary scores

Predictors	Estimate	SE	t-value	Pr(>|t|)
(Intercept)	1.699	0.418	4.066	0.000
Selecting	0.337	0.077	4.400	0.000***
Integrating	0.024	0.089	0.270	0.787
Monitoring	0.036	0.088	0.408	0.684
Structure-Building	0.233	0.111	2.091	0.039*

Table 4 depicts the results of a linear main effects model predicting MCQ task scores from Fall 2023 with the four discourse construction processes as main effects. The overall model was statistically significant, $F(4, 113) = 6.73$, $p < .0001$, explaining 19.2% of the variability in the MCQ score ($R^2 = .192$). The only significant predictor of MCQ scores was Selecting ($\beta = .282$, SE = .079, $t = 3.551$, $p < .001$). No significant relationships were observed for the other three processes. This suggests that students who showed more evidence of Selecting the five key ideas in their notes were also more likely to receive higher scores on the MCQs targeting vocabulary, inference, and detail items.

Table 4. Results of a linear model to predict MCQ scores

Predictors	Estimate	SE	*t*-value	Pr(>|t|)
(Intercept)	2.183	0.433	5.037	0.000
Selecting	0.282	0.079	3.551	0.001***
Integrating	0.075	0.093	0.810	0.420
Monitoring	0.037	0.092	0.408	0.684
Structure-Building	0.128	0.115	1.106	0.271

These results confirm that the listen-to-summarize task is eliciting a range of higher-order listening processes and several of these processes are associated with higher test performance.

Discussion

This study represented an attempt to discern whether two listening response tasks on an integrated placement test were functioning as intended to elicit and reward higher-order cognitive processes known to be important in the TLU domain. With regard to the first validation question, qualitative analysis revealed that the tasks did in fact elicit evidence

of the discourse construction processes identified by Field (2013), including Selecting, Integrating, Monitoring, and Structure-Building. Qualitative analysis also revealed that students employed Translanguaging strategies for differing purposes, indicating flexibility in bringing their full linguistic repertoire to bear on the task (Irgin, 2025). Regarding the second validation question, qualitative analysis suggested that the provision of minimal notetaking scaffolding resulted in greater evidence of Selecting, Integrating, and Structure-Building processes, with a slight increase in translanguaging strategies. These findings were generally confirmed by linear regression models which indicated that notes from Spring 2024 (which included notetaking scaffolding) contained significantly more evidence of Selecting, Integrating, and Structure-Building processes. However, the notetaking scaffolding provided in Spring 2024 did not significantly impact Monitoring processes or Translanguaging. Finally, regarding the third validation question, findings from linear regression models indicated that in Fall 2023 Selecting and Structure-Building processes were significantly predictive of summary task scores, while only Selecting processes were predictive of MCQ task scores.

These results suggest that not all higher-order listening processes are equally predictive of all listening tasks. In our study, Selecting and Structure-Building emerged as the most powerful predictors of test performance. The ability to identify and record main ideas was significant for both the summary task and the MCQ task, even though the MCQ items did not directly target main ideas. This suggests that students who attend holistically to the meaning of a lecture may be better prepared to process and integrate details later, even if they were not specifically focusing on those details at the time. This accords with evidence from Oakhill and Davies (1991) that students who take notes for a productive task tend to get better scores on recognition tasks to boot: “subjects expecting a recall task selected the main themes of the text for inclusion in their notes, perhaps assuming that, if they recalled those, they would be able to retrieve other low-level ideas” (p. 188). Further, the ability to represent the structure of the lecture visually in notes was significant for the summary task, though not the MCQ task. It is possible that structured notes taken during the lecture served as a kind of pre-writing task for the integrated writing (Brunfaut & Kormos, 2025).

However, Integrating and Monitoring strategies did not significantly impact test scores. This is worthy of discussion because previous research is divided on the relationship between nonverbal elements in notes and L2 listening test scores, sometimes finding significant positive effects (Carrell, 2007; Chaudron et al., 1988; Faraco et al., 2002) and sometimes finding nonsignificant effects (Chaudron et al., 1988; Kim, 2025). There is some difficulty in comparing effects across these studies because each study categorized nonverbal elements somewhat differently. For instance, Faraco et al. (2002) and Kim (2025) included arrows in a general measure of nonverbal elements. Carrell (2007) isolated the impact of arrows, diagrams, and emphasis strategies, but did not distinguish between subtypes of arrows. Chaudron et al. (1988) did distinguish between arrows used as symbols and arrows used in outlines; symbolic arrows were significant predictors of a MCQ test, and outlining arrows were significant for a cloze test. In our study, we classified arrows as evidence of either Integrating, Monitoring, or Structure-Building processes depending on whether they indicated a linear relationship between ideas, an insertion of a word above or below the line, or a subordinate relationship between ideas. Arrows which contributed to Structure-Building processes were predictive of summary task performance; arrows which contributed to Integrating and Monitoring processes were not. Overall, we anticipate that classifying nonverbal elements by their function, rather than by their form, may prove to be a more fruitful predictor of test performance. It has long been recognized that nonverbal elements are quite difficult to classify and interpret (Cushing, 1991), and might best be investigated in tandem with stimulated recalls which allow participants to describe the thought processes behind their nonverbal strategy choices.

We hypothesize that the non-significance of Integrating and Monitoring strategies in our dataset may be due to two factors. First, Integrating codes were the most frequent in our sample, and were visible across the range of scores. Numbering in notes was not always accurate (i.e., did not always match the points in the listening texts), and arrows were employed for a spectrum of purposes, some of which were opaque to the reader (seemingly indicating “here is another idea”). It is possible that some students may incorporate Integrating arrows in their notes when they are unsure how to otherwise represent the nature of the relationship between two ideas. Overall, Integrating Arrows

and Numbering strategies can be employed with greater or lesser skill, and warrant additional investigation. We would hypothesize that a measure which assessed accuracy of Integrating strategies would be more predictive of test scores than a measure which assesses only quantity of Integrating strategies.

Monitoring strategies represent another skillset in need of further investigation. As with Integrating strategies, we observed a range of Monitoring strategies across skill levels. We hypothesize that Monitoring strategies may pattern differently because they emerge in the face of difficulty. A listener who hears accurately and confidently may not record any Monitoring strategies, and therefore may be indistinguishable on paper from a listener who hears incorrectly but lacks the ability to monitor at all. Monitoring strategies may therefore only appear in contexts where a listener both struggles with the listening material AND possesses the strategic resources to cope with those challenges. In our study, we explored the possibility of a non-linear relationship between Monitoring strategies and test scores, but the data did not support such an analysis. However, the relationship between Monitoring skills and listening test performance invites future attention.

Finally, these results confirm our expectation that notetaking would be more valuable for the summary task than the MCQ task. Notetaking processes were predictive of scores on both tasks, but overall explained more of the variance for the integrated task, as indicated by the *F*-tests for the combined effect of higher-order processes in each model. This is in alignment with previous studies indicating that notetaking has closer relationships with productive tasks over recognition tasks (Jin & Webb, 2023; Kobayashi, 2005; Liu & Hu, 2012; Rukthong, 2021)

Implications of results for our context

The fundamental purpose of this study was to confirm that our placement test is eliciting the listening processes targeted by our test construct. Drawing on an established model of listening skills, we were able to construct a methodology to investigate these processes through both qualitative and quantitative measures. The results of this analysis give us

confidence in continuing to use both the summary task and the MCQ task without preview. In this study, we observed evidence of the desired cognitive processes across all test administrations. Further, we observed heightened levels of these processes when we provided brief notetaking scaffolding in accordance with LOA principles (Thao & Trang, 2022). This study confirms our intention to continue providing this scaffolding in future test administrations. Finally, we observed that several of the higher-order cognitive processes elicited by the test were rewarded by the scoring process, with Selecting and Structure-Building especially contributing to higher scores on the integrated task, and Selecting contributing to higher scores on the MCQ task. None of the processes were penalized by the scoring process or had a negative relationship with scores. This study therefore provides the confirmation needed to continue moving forward with our response formats, scaffolding, and scoring mechanisms.

Qualitative analysis also enabled us to identify unexpected areas for possible test revision. For example, in this study, we noted many examples of Tagging and Delineation strategies to connect or distance ideas on the page. These strategies were comparatively absent from our previous study (Yeager et al., 2024), which provided two pages for notetaking, while the current study provided only one. We note that the pressure to find space to continue notetaking may have led to additional demands on test-takers' working memory. Therefore, we recommend providing additional pages for notetaking in future versions of the placement test.

Implications of methods for other contexts

The *results* of our analysis should not be automatically assumed to generalize to other testing contexts, since validation studies, by definition, cannot employ a controlled experimental design due to variations in the test-taking population across semesters. However, we believe that the *methods* we developed in this study may be adopted by researchers in other contexts to explore similar questions about cognitive validity. As noted in our literature review, research on cognitive validity in listening assessment faces many challenges. We propose that notetaking provides an overlooked source of evidence about thought processes during a test event. In their notes, test-takers literally leave a

record of what they were thinking about during the test. Further, these notes can be collected during live testing at no extra cost, a substantial advantage especially for small local testing programs which may lack the resources to collect other forms of evidence. We recommend regular collection and analysis of notes whenever they are allowed on tests of listening comprehension. This analysis is especially valuable after the introduction of a new task type to ensure that it is performing as intended. Specifically, notes should be analyzed for evidence of higher-order listening processes, which are often under-elicited on listening assessments (Holzknecht et al., 2017; Rukthong, 2021).

Our study further confirms that notetaking is heavily personalized and contextualized, resisting overly simplistic analysis. However, this very complexity is an asset which can be mined through qualitative analysis. Building on previous work by Cushing (1991) and Rukthong (2021), we echo calls for more qualitative analysis of notetaking in listening assessment. We further recommend starting with provisional codes developed from a review of the literature on the test construct, but then allowing for open coding to acknowledge emerging patterns in the dataset, such as the Translanguaging, Tagging, and Delineation strategies we observed here.

Finally, our study pioneered a method for quantifying higher-order listening processes in notes, which allowed us to measure the effects of those processes on test scores. Quantifying qualitative data is always tricky, but the development of the rubric was assisted by our deep familiarity with the dataset after several rounds of qualitative coding. We were additionally encouraged to find that the third author, who had not participated in the qualitative stage of the study, was able to interpret and apply the rubric to a new dataset with relatively high reliability. We suspect the usability of this rubric was due in part to our decision to limit our Selecting scale to identifying five key ideas on a checklist, and to limit our Integrating, Monitoring, and Structure-Building scales to identifying nonverbal elements such as numbers, symbols, and spatial organization. This enabled the quantitative coding to actually go fairly quickly. For this reason, we tentatively offer our quantitative rubric as a quick-and-dirty metric for use by local testing programs short on resources to develop their own rubric for cognitive processes in listening assessment. However, we would urge any programs adopting our rubric to create their own subscale

for Selecting. Selecting is a process which is necessarily tied to content for a specific task, and therefore rubrics designed to measure Selecting for one task may not apply to another. We also note that different testing programs may target different listening constructs, in which case it would be necessary to develop a rubric from scratch anyway. The key idea which we offer here is not the rubric itself but the affirmation that it is possible to develop such a tool to measure the relationship between cognitive processes and listening performance.

Limitations and future research

Again, we limit our claims about the results to our testing context. Specifically, we would expect that the relationship between notetaking and listening performance might look different if the same task were administered to a population of much higher or lower listening proficiency. Our test targets a relatively narrow proficiency range: students with test scores sufficient to earn acceptance into the university but not sufficient to merit a placement test waiver. This enabled us to pinpoint the difficulty of our tasks, avoiding both a ceiling and floor effect. In other contexts, students with lower proficiency may struggle to take notes at all, while some highly proficient listeners may find it unnecessary to take notes, especially for simplified L2 listening tasks. The single test-taker in Spring 2023 who did not take any notes, for instance, did achieve a passing score (though not a perfect score) on the listening test. We cite this example as a reminder that notetaking is a highly personalized endeavor, and therefore we do not recommend the assessment of academic notes as a test task in itself, primarily because assessment of notes is uncommon in academia and therefore unrepresentative of the TLU domain. Further, where notes are assessed (especially for high-stakes purposes), students must attend to legibility in ways that may interfere with comprehension. However, we do support the use of notetaking as an assessment task in professional contexts where the TLU domain may include tasks such as taking notes for a colleague to review later (c.f. notetaking for medical professionals). In such contexts, legibility is relevant to the construct. It bears repeating that the context and purpose of a task impact both the processes and product of that task. We seek to understand and elicit the cognitive *processes* involved in academic listening, while assessing only the *products* that would be assessed in the TLU domain.

It is worth noting here also that in our study both the integrated task and the MCQ task pertained to the same lecture. Results may therefore differ in contexts where students listen to separate texts for each response task. Task expectancy has been shown to impact notetaking in university contexts, with students taking more notes when they expect to be given a writing task instead of an MCQ task (Oakhill & Davies, 1991). It remains to be seen whether Selecting skills are predictive of MCQs related to a separate listening text.

We further note that our placement test is paper-based and that the processes involved in notetaking may differ across modalities, especially for L2 notetakers (Kim, 2025). We do not see this as an inherent disadvantage, but simply a limitation of our work. Paper-based notetaking remains a common modality for notetaking in university classrooms, even post-pandemic (Potvin et al., 2023), and is often preferred by L2 students, in part due to its flexibility in accommodating translanguaging and non-verbal elements (Cubilo, 2017; Kim, 2025; Siegel, 2022). Further, several meta-analyses have established a slight advantage for paper-based notes over typed notes in classroom contexts (Lau, 2022; Voyer et al., 2022; Flanigan et al., 2024). There are therefore good reasons to believe that paper-based notetaking will remain part of the academic TLU domain for the foreseeable future. However, digital notetaking is also well-represented in academia (Potvin et al., 2023), and is a superior form of notetaking for some topics, such as computer science (Voyer et al., 2022). At the risk of oversimplifying, most digital notetaking tools maximize the external storage function of notetaking, while paper-based notes maximize the encoding function. Both functions are beneficial for learning, and 21st century students may need to draw on both notetaking modalities in different contexts. We call for more research on cognitive validity for both paper-based and digital notetaking tasks.

Finally, we call for further research on notetaking in combination with other data sources about cognitive processes in listening assessment, such as recall interviews (Brunfaut & Kormos, 2025) or eye-tracking (Kim, 2023). We value and rely on these sources of data in our own work, but were unable to collect such data for this study due to the limitations of working with live test data. Triangulation of notes with other data sources could lead to very productive insights about listening processes. For example, stimulated recalls could help determine what signals lead students to select information to include in their

notes, while eye-tracking data could help to trace the process of Monitoring strategies such as cross-outs and insertions. Drawing on all available data sources is critical if we want to support our inferences about cognitive validity in listening assessment.

Acknowledgements

The authors wish to thank ESL Programs director Melissa Meisterheim for granting approval for this test validation project, and the language instructors who rated the integrated summary tasks. I-Chun Vera Hsiao, Claire Jacobson, Alfonso Martinez, Ryan Lidster, and Sharry Vahed assisted with translation of the translanguaged notetaking samples. Melissa Meisterheim provided helpful commentary on an earlier version of this manuscript. Any remaining errors are our own. The authors acknowledge that this study was conducted on land taken from Native American communities, and we advocate for reparations to address that debt.

Funding

This research received no grant from any funding agency in the public, commercial, or not-for-profit sectors.

Conflict of interest disclosure

The authors declare the following conflicts of interest: The first author of this study was a language instructor and placement test coordinator for ESL Programs at the University of Iowa at the time when the study was conducted. The second and third authors were graduate students in the College of Education at the University of Iowa at the time of data collection. None of the authors received any funding or release time for their work on this validation study.

Ethics approval and informed consent statement

This project was reviewed and granted an approval waiver by the University of Iowa Institutional Review Board as it evaluates test performance as part of normal program

operations and is not intended to answer generalizable research questions about human subjects.

Data availability statement

For test security reasons, we are unable to provide access to the test forms or student data discussed in the study. However, a practice form of the placement test is available on the ESL Programs website (https://lllc.uiowa.edu/academic-areas/english-second-language/credit-program), and the R code used in the analysis is available on OSF (https://osf.io/7k53w/?view_only=7d8be3d9a9a34413bfc45f0cae222780)

Authorship statement

Author contributions:

- Rebecca Yeager: Conceptualization, Data Curation, Formal Analysis, Investigation, Metholodology, Project Administration, Resources, Software, Validation, Visualization, Writing Original Draft, Writing Review and Editing
- GoMee Park: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Resources, Software, Validation, Visualization, Writing Original Draft, Writing Review and Editing
- Junhee Park: Formal Analysis, Investigation, Methodology, Software, Validation, Writing Original Draft, Writing Review and Editing

Generative AI was not used at any point in the process of this study. All work is original to the authors.

ORCID iDs

Rebecca Yeager https://orcid.org/0000-0001-8017-5879

GoMee Park https://orcid.org/0000-0001-5107-0568

Junhee Park https://orcid.org/0009-0007-8085-6595

References

ATLAS.ti scientific software development GmbH. (2025). ATLAS.ti Mac (version 25.01). [Qualitative data analysis software]. https://atlasti.com.

Bonilauri, A., Intra, F., Baglio, F., & Baselli, G. (2023). Impact of anatomical variability on sensitivity profile in fNIRS–MRI integration. *Sensors, 23,* 2089. https://doi.org/10.3390/s23042089

Borkovska, N., & Irgin, I. (2026). The effect of structured and conventional note-taking on lecture listening comprehension and specialized vocabulary recall in EMI academic contexts. *English for Specific Purposes, 83,* 115—128. https://doi.org/10.1016/j.esp.2026.03.001

Brunfaut, T., & Kormos, J. (2025). Assessing multimodal viewing-to-write constructs: Task design, performance, processing, and rating. *Language Assessment Quarterly, 22*(4–5), 429–459. https://doi.org/10.1080/15434303.2025.2596374

Canham, A., Coumel, M., Manalova, J., & de Bruin, A. (2024). Note taking in bilingual students: Does using a first or second language influence note-taking quality and memory of newly learnt content? *International Journal of Bilingual Education and Bilingualism*. https://doi.org/10.1080/13670050.2024.2390862

Carless, D. (2007). Learning-oriented assessment: Conceptual bases and practical implications. *Innovations in Education and Teaching International, 44*(1), 57–66. https://doi.org/10.1080/14703290601081332

Carrell, P. (2007). Notetaking strategies and their relationship to performance on listening comprehension and communicative assessment tasks. *TOEFL monograph series No. RS 35*. ETS. https://files.eric.ed.gov/fulltext/EJ1111620.pdf.

Chaudron, C., Cook, J., & Loschky, L. (1988). Quality of lecture notes and second language listening comprehension. Center for Second Language Classroom Research. *Social Science Research Group University of Hawai'i at Manoi Technical Report #7.*

Clark, M., Wayland, S., Osthus, P., Brown, K., Castle, S., & Ralph, A. (2014). The effects of notetaking on foreign language listening comprehension. University of

Maryland Center for Advanced Study of Language. https://www.govtilr.org/Publications/Notetaking.pdf.

Clerehan, R. (1995). Taking it down: Notetaking practices of L1 and L2 students. *English for Specific Purposes, 14*(2), 137-155. https://doi.org/10.1016/0889-4906(95)00003-A

Craig, K., Hale, D., Grainger, C., & Stewart, M. (2020). Evaluating metacognitive self-reports: Systematic reviews of the value of self-report in metacognitive research. *Metacognition and Learning, 15,* 155–213. https://doi.org/10.1007/s11409-020-09222-y

Creswell, J.W. & Plano Clark, V.L. (2011). *Designing and Conducting Mixed Methods Research*. Sage Publications.

Cubilo, J. (2017). Video-mediated listening passages and typed note-taking: Examining their effects on examinee listening test performance and item characteristics. [Doctoral Dissertation. University of Hawai‘i at Mānoa].

Cushing, S. T. (1991). A qualitative approach to the study of notetaking in UCLA's English as a second language placement examination. Unpublished manuscript, University of California, Los Angeles.

Faraco, M., Barbier, M., & Piolat, A. (2002). A comparison between notetaking in L1 and L2 by undergraduate students. In S. Ransdell, & M. Barbier (Eds.), *Studies in Writing, Volume 11: New Directions for Research in L2 Writing* (pp. 145–167). Kluwer Academic Publishers.

Field, J. (2013). Cognitive validity. In L. Taylor & A. Geranpayeh (Eds.), *Examining Listening* (pp. 77–151). Cambridge University Press.

Flanigan, A., Wheeler, J., Colliot, T., Lu, J., and Kiewra, K. (2024). Typed versus handwritten lecture notes and college student achievement: A meta-analysis. *Educational Psychology Review, 36*(78). https://doi.org/10.1007/s10648-024-09914-w

Hale, G., & Courtney, R. (1994). The effects of note-taking on listening comprehension in the Test of English as a Foreign Language. *Language Testing, 11*(1), 29–47. https://doi.org/10.1177/026553229401100104

Hayati, A., & Jalilifar, A. (2009). The impact of note-taking strategies on listening comprehension of EFL learners. *English Language Teaching, 2*(1), 101–111. https://files.eric.ed.gov/fulltext/EJ1082250.pdf.

Holzknecht, F. (2019). Double play in listening assessment. (Publication No. 28278058) [Doctoral dissertation, Lancaster University]. ProQuest Dissertations & Theses Global.

Holzknecht, F., Eberharter, K., Kremmel, B., Zehentner, M., McCray, G., Konrad, E., & Spöttl, C. (2017). Looking into listening: Using eye-tracking to establish the cognitive validity of the Aptis listening test. ARAGs Research Reports Online AR-G/2017/3. British Council. https://www.britishcouncil.org/exam/aptis/research/publications/arags/looking-listening-using-eye-tracking.

Irgin, P. (2025). Note-taking in academic listening: A translanguaging perspective. *RELC Journal,* 1-13. https://doi.org/10.1177/00336882251322253

Jin, Z., & Webb, S. (2023). The effectiveness of note taking through exposure to L2 input: A meta-analysis. *Studies in Second Language Acquisition*. Advance online publication. https://doi.org/10.1017/S0272263123000529.

Kim, H. (2018). Impact of slide-based lectures on undergraduate students' learning: Mixed effects of accessibility to slides, differences in note-taking, and memory term. *Computers and Education, 123,* 13-25. https://doi.org/10.1016/j.compedu.2018.04.004

Kim, J. (2023). Test takers' interaction with context videos in a video-based listening test: A conceptual replication and extension of Suvorov (2015). https://doi.org/10.31219/osf.io/r83by

Kim, J. (2025). Handwritten versus typed notes: The impact of note-taking modes in second language listening tests. [Doctoral Dissertation. University of Hawai‘i at Mānoa]. https://scholarspace.manoa.hawaii.edu/items/6b359275-c3a8-40fe-ba52-e19a6b5415bd

Klein, F., Debener, S., Witt, K., & Kranczioch, C. (2022). fMRI-based validation of continuous-wave fNIRS of supplementary motor area activation during motor execution and motor imagery. *Scientific Reports, 12,* 3570. https://doi.org/10.1038/s41598-022-06519-7

Kobayashi, K. (2005). What limits the encoding effect of note-taking? A meta-analytic examination. *Contemporary Educational Psychology, 30,* 242–262. https://doi.org/10.1016/j.cedpsych.2004.10.001

Kobayashi, K. (2006). Combined effects of note-taking/reviewing on learning and the enhancement through interventions: A meta-analytic review. *Educational Psychology, 26,* 459–477. https://doi.org/10.1080/01443410500342070

Lau, T. (2022). The effect of typewriting vs. handwriting lecture notes on learning: A systematic review and meta-analysis (3982). [Doctoral dissertation, University of Louisville]. https://doi.org/10.18297/etd/3982

Liu, B., & Hu, Y. (2012). The effect of note-taking on listening comprehension for lower-intermediate level EFL learners in China. *Chinese Journal of Applied Linguistics, 35*(4), 506-518. https://doi.org/10.1515/cjal-2012-0036

Llosa, L., & Malone, M. (2019). Comparability of students' writing performance on TOEFL iBT and in required university writing courses. *Language Testing, 36*(2), 235-263. https://doi.org/10.1177/0265532218763456

Loughlin, C. (2015). Digitally mediated note-taking practices of students in higher education. [Master's Thesis. Kingston University.]

Morehead, K., Rhodes, M. G., & DeLozier, S. (2019). Note-taking habits of 21st Century college students: Implications for student learning, memory, and achievement. *Memory*. https://doi.org/10.1080/09658211.2019.1569694

Oakhill, J., & Davies, A. (1991). The effects of test expectancy on quality of notetaking and recall of text at different times of day. *British Journal of Psychology, 82*(2), 179-189. https://doi.org/10.1111/j.2044-8295.1991.tb02392.x

Olsen, L., & Huckin, T. (1990). Point-driven understanding in engineering lecture comprehension. *English for Specific Purposes, 9,* 33-47.

https://deepblue.lib.umich.edu/bitstream/handle/2027.42/28773/0000605.pdf;sequence=1

Plakans, L., & Gebril, A. (2013). Using multiple texts in an integrated writing assessment: Source text use as a predictor of score. *Journal of Second Language Writing, 22*(3), 217-230. http://dx.doi.org/10.1016/j.jslw.2013.02.003

Potvin, M., Chabot, M., Garrity, A., Hass, R., Zane, C., & Bower, A. (2023). Are iGen freshman different? Notetaking habits of STEM students: A descriptive study. *International Journal of Progressive Education, 19*(2). https://doi.org/10.29329/ijpe.2023.534.1

R Core Team (2024). R: A language and environment for statistical computing. R Foundation for Statistical Computing. https://www.R-project.org/

Rukthong, A. (2021). MC listening questions vs. integrated listening-to-summarize tasks: What listening abilities do they assess? *System, 97*. https://doi:10.1016/j.system.2020.102439

Rukthong, A., & Brunfaut, T. (2020). Is anybody listening? The nature of second language listening in integrated listening-to-summarize tasks. *Language Testing, 37*(1), 31-53. https://doi.org/10.1177/0265532219871470

Sadeghi, K., & Zeinali, M. (2015). The effect of item modality and note-taking on EFL learners' performance on a listening test. *Issues in Language Teaching, 4*(2), 81–101. https://doi.org/10.22054/ILT.2015.7227

Saldaña, J. (2015). *The coding manual for qualitative researchers*. Sage.

Semma, A., Hannad, Y., Siddiqi, I., Djeddi, C., & El-Kettani, M. (2021). Writer identification using deep learning with FAST keypoints and Harris corner detector. *Expert Systems With Applications, 184*, 115473. https://doi.org/10.1016/j.eswa.2021.115473

Siegel, J. (2022). Pen and paper or computerized notetaking? L2 English students' views and habits. *Computers and Education Open, 4*. Advance online publication. https://doi.org/10.1016/j.caeo.2022.100120

Siegel, J., & Kusumoto, Y. (2022). A cross-cultural investigation of L2 notetaking: Student habits and perspectives. *Journal of Multilingual and Multicultural Development*. https://doi.org/10.1080/01434632.2022.2036168

Taylor, L., & Geranpayeh, A. (Eds.). (2013). *Examining listening*. Cambridge University Press.

Thao, T., & Trang, H. (2022). High school EFL students' perceptions of scaffolding learning activities in learning listening skills. *TNU Journal of Science and Technology, 227*(13), 17-24. https://doi.org/10.34238/tnu-jst.6224

University of Iowa Admissions. (n.d.). English proficiency requirements. https://admissions.uiowa.edu/english-proficiency-requirements

University of Iowa ESL Programs. (n.d.). What is the English Placement Evaluation? https://lllc.uiowa.edu/academic-areas/english-second-language/credit-program

Voyer, D., Ronis, S., & Byers, N. (2022). The effect of notetaking method on academic performance: A systematic review and meta-analysis. *Contemporary Educational Psychology, 68*. https://doi.org/10.1016/j.cedpsych.2021.102025

Wagner, E., & Wagner, S. (2016). Scripted and unscripted spoken texts used in listening tasks on high-stakes tests in China, Japan, and Taiwan. In V. Aryadoust & T. Fox (Eds.), *Trends in language assessment practice and research* (pp. 438–463). Cambridge Scholars Publishing.

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

Witherby, A., & Tauber, S. (2019). The current status of students' note-taking: Why and how do students take notes? *Journal of Applied Research in Memory and Cognition, 8,* 139–153. https://osf.io/2mzts/

Yeager, R., Park, G., & Liao, R. (2024). Notetaking as validity evidence: A mixed-methods investigation of question preview in EAP listening assessment. *Journal of English for Academic Purposes, 68,* 101346. https://doi.org/10.1016/j.jeap.2024.101346

Yeager, R., & Martinez, A. (2025a). A sow's ear purse: Developing an integrated placement test in an under-resourced English for Academic Purposes program.

Language Testing. Advance online publication. https://doi.org/10.1177/02655322251346840

Yeager, R., & Martinez, A. J. (2025b, May 5). A sow's ear purse: Developing an integrated placement test in an under-resourced English for Academic Purposes program. Open Science Foundation. https://osf.io/52hmr

Yeager, R., Park, G., & Park, J. (2026, April 20). Cognitive validity in listen-to-write summary tasks: A mixed-methods analysis of notetaking data. Open Science Foundation. https://osf.io/7k53w/overview?view_only=10da637a967a471d9798f5c3d6247c5d