3-2024

Notetaking as validity evidence: A mixed-methods investigation of question preview in EAP listening assessment

Rebecca Yeager

GoMee Park

Ray J. T. Liao

Notetaking as Validity Evidence:

A Mixed-Methods Investigation of Question Preview in EAP Listening Assessment

Keywords: listening, test format, testwise strategies, washback, cognitive validity

Abstract

Recent scholarship has questioned the cognitive validity of listening tests with preview, in which test-takers can see test questions before listening. This study mined student notes for evidence of cognitive processes in listening tests with and without preview, using a mixed-methods design that explored the effect of test format on notetaking behaviors. Qualitative analysis indicated that students who previewed items were more likely to systematically omit information, highlight keywords, and engage in shallower structural representation. Conversely, Kruskal-Wallis tests revealed that students who listened without preview took more notes, especially of main ideas and details, and had better coverage of the lecture. However, correlation and hierarchical linear regression analyses found these notetaking achievements did not predict higher scores in the no-preview condition, while in the preview condition, only note quantity and focus on minor ideas predicted scores. Both strands of data suggest that students' cognitive processes were shaped by the format of the exam they experienced. These findings may bear on validity arguments for listening assessment and inform the way that language instructors prepare their students for academic listening.

Introduction

In university classrooms, students engage in many tasks which involve listening. Foremost among these is the academic lecture, in which students typically listen, take notes, ask clarification questions, and review those notes for study later (Lynch, 2011; Siegel, 2020). Listening instruction in English for Academic Purposes (EAP) contexts might be expected to focus on these key skills. However, when listening skills are assessed in these contexts, assessments are often stripped of notetaking, questioning, and review components in an attempt to get at a "pure" listening construct (O'Grady, 2021). Comprehension is often measured through multiple-choice questions (MCQs) answered during or immediately after the lecture (Rukthong, 2021). However, the inauthenticity of this task type raises questions about the validity of claims made on the basis of such assessments, and further about the consequences of preparing students to expect test formats they will not

encounter again outside the EAP classroom.

This study is concerned with the impact of test format on cognitive validity, in which the mental processes engaged by the test match the mental processes engaged by authentic listening tasks (Weir, 2005). Threats to cognitive validity arise from the introduction of "testwise strategies" enabled by features of the test which differ from the target language use environment (Cohen, 2007). Weir's (2005) socio-cognitive validity framework has been applied to listening assessment by Taylor and Geranpayeh (2013), in a model which combines test-taker characteristics and evidence of cognitive, context, scoring, consequential, and criterion-related validity to support claims about test-taker listening ability. We have selected Taylor and Geranpayeh's (2013) model to undergird our study because it recognizes the importance of social context in interpreting scores. In an academic context, learners frequently have access to reading materials, office hours, and other discussion opportunities before taking tests. In a language testing context, such resources are rarely available, which may reasonably be expected to impact the cognitive processes that test-takers employ.

This study explores the impact of one common listening task type, question preview, on cognitive validity. In question preview, multiple-choice questions are presented to test-takers in full (full-preview) or in part (stem-preview or option-preview) before listening. We are interested in this task type because it can be found in EAP assessments and textbooks around the world, but there is little guidance from the literature concerning its validity in EAP contexts. The existing research on preview mostly focuses on its impact on difficulty and affect. In terms of difficulty, full-preview and stem-preview appear to perform similarly (Iimura, 2010; Koyama et al., 2016; Li et al., 2017; O'Grady, 2021; Yanagawa & Green, 2008), but full-preview tends to be easier than option-preview (Koyama et al., 2016; Sadeghi & Zeinali, 2015; Yanagawa & Green, 2008) or no-preview (Iimura, 2010; Koyama et al., 2016; O'Grady, 2021). In other words, the advantage of preview primarily lies in access to the question stems, with access to the options providing only marginal benefit. In terms of affect, students generally express preferences for preview (Iimura, 2010; Li et al., 2017). However, the impact of this task type on cognitive processing is unclear. Therefore, our study explores the cognitive validity of listening tests with and without preview through mixed-methods analysis of student notes and test scores.

Literature Review

Preview and Cognitive Validity

Theoretically, preview could impact cognitive processing in one of two ways. First, it is possible that preview could simulate the processes activated in an academic context, serving as a replacement for other classroom resources which aid the listener in preparing to learn. Evidence for this possibility comes from preview's generally positive impact on scores, along with evidence comparing different forms of prelistening activities in which preview outperforms vocabulary activities (Chang & Read, 2006) and prereadings (Alavi & Janbaz, 2014). However, both vocabulary activities (Berne, 1995; Madani & Kheirzadeh, 2022) and prereadings (Chang & Read, 2006) sometimes do as well or better than preview. In general, then, it would appear that preview can function as a listening resource, but other activities may accomplish this same goal.

Second, it is also possible that preview could alter the processes employed by learners during the listening task to the extent that their cognitive processes do not match those of the listening construct. This effect would be undesirable, as it would make it difficult to generalize from a student's performance on a listening test to their future performance in a university classroom. It could also mislead students and language instructors to prioritize test preparation strategies which will not be transferable to a university context.

Several studies have investigated strategy use in tests with preview through surveys or stimulated recall. Many have uncovered troubling patterns, such as using preview to selectively attend only to points in the lectures that will be assessed (Field, 2011), guessing or eliminating options (Cheng, 2004; Field, 2012), and aural scanning for keywords without comprehending the structure of the text (Field, 2011; 2012). Badger and Yan (2012) in fact discovered that testwise strategies were used equally by L1 and L2 listeners when completing a listening test with preview, suggesting that test format may have more impact than test-taker characteristics on strategy use. On the other hand, In'nami and Koizumi (2022) compared metacognitive survey responses with performance on while-listening-performance (WLP) tests with preview and post-listening-performance (PLP) tests without preview, and found that only scores on the WLP test were related to planning and evaluation strategies. They interpret these results as evidence that students may have used the questions in the WLP format to help them plan for the listening task.

Cognitive processes during listening tests have also been explored via eye-tracking and Functional Near-

Infrared Spectroscopy (fNIRS; Aryadoust et al., 2022; Zhai & Aryadoust, 2022). These studies indicate that test-takers exhibit differences in eye gaze behavior, fixations, and neural activity when taking a WLP test with preview and a PLP test without preview. Although some of these results may be explained by the difference in response timing, many of the behavioral patterns during WLP tests would not be possible without access to the questions during the lecture. These results provide evidence that preview may enable and reward listening strategies which are not possible in EAP contexts. Aryadoust et al. (2022), for example, observed that "the gaze behavioral patterns exhibited during the WLP tests suggested that the test-takers adopted keyword matching and 'shallow listening,'" and further that "test-takers displayed lower activity levels across brain regions supporting comprehension during the WLP tests relative to the PLP tests" (p. 56). Together, these studies indicate reason for concern that some cognitive processes enabled by preview may not be transferable to academic listening tasks.

Notetaking as Validity Evidence

One data source for observing the impact of preview on cognitive processes has been underexplored: notetaking. Notetaking is notoriously difficult to analyze because it is known to vary widely across learners and contexts. Variables impacting note quantity include lecture topic and speed (Siegel, 2022), access to visuals (Cubilo & Winke, 2013), and task type (Oakhill & Davies, 1991). Confounding these factors, students may deploy a range of efficiency strategies which make later interpretation of notes difficult, including abbreviations, symbols, and translanguaging (Zhou et al., 2022). Considering these complications, it is perhaps not surprising that notetaking has been pushed to the side in the search for evidence of cognitive processing on listening exams.

However, this oversight is unfortunate. Notetaking is a valuable source of data about student comprehension and can be used as an assessment tool in its own right (Nakayama et al., 2017; Song, 2011). More importantly, evidence of notetaking behavior during a listening test should be systematically collected as part of test validation. Test formats which reward empirically-supported notetaking choices should be favored over test formats which reward testwise notetaking strategies. This evidence should be evaluated as part of an ongoing attempt to ensure positive washback.

The benefits of notetaking have been theorized to fall into two categories: encoding and review (Kim, 2018). Encoding refers to the advantages that arise when students are forced to selectively attend to the key points of a lecture, paraphrase, and visually represent its structure. Review refers to the external storage function of notes, allowing learners to revisit key points later. Meta-analyses have confirmed moderate effects for encoding and strong effects for review (Kobayashi, 2005; 2006). In L1 academic contexts, notetaking appears to aid comprehension particularly where the test is delayed (Chen et al., 2017; Kim, 2018), the task is productive (Kobayashi, 2005; Oakhill & Davies, 1991), or the content is unfamiliar (Brobst, 1996), with main ideas predicting success better than total notations (Northern et al., 2023). Notation of details appears to matter comparatively little on immediate tests, but becomes important on cumulative exams (Kiewra et al., 1987).

Failure to learn notetaking skills continues to impact student success. Even digital resources cannot compensate for a deficiency in this area: meta-analyses confirm a significant advantage for handwritten notes over typed ones in classroom contexts (Allen et al., 2020; Voyer et al., 2022). The provision of guided notes by the instructor may improve performance short-term but runs the risk of creating a dependence on resources which may not always be available (Chen et al., 2017; Konrad et al., 2009), while students who rely solely on slides from the instructor miss out on the encoding function of notetaking (Kim, 2018).

L2 notetaking studies have generally replicated these findings from L1 contexts. The preponderance of evidence suggests that L2 students perform better when allowed to take notes (Carrell, 2007; Hayati & Jalilifar, 2009; Kim, 2023), especially after instruction in notetaking strategies (Siegel, 2020; Yang & McAllister, 2023). As in L1 contexts, measures of content (Dunkel, 1988) and structure (Chaudron et al., 1994; Cushing, 1991) seem to be more meaningful than measures of length. Further, the benefits of notetaking are more pronounced for productive tasks (Cubilo & Winke, 2013; Liu & Hu, 2012; Song, 2011), and where review is allowed (Carrell, 2007; Hayati & Jalilifar, 2009).

Surprisingly, however, some L2 studies found no effects for notetaking (Clark et al., 2014; Sadeghi & Zeinali, 2015), and in one study students scored lower after being forced to take notes (Hale & Courtney, 1994). These findings may be partially explained by use of different comprehension tasks. The L2 studies above with positive associations for notetaking generally used productive tasks or MCQs without preview. In one study,

notetaking was associated with summary scores but not with MCQ scores (Liu & Hu, 2012). Among the L2 studies we identified which explicitly used preview tasks, two found no relationship between notetaking and score (Clark et al., 2014; Sadeghi & Zeinali, 2015), and one found no effect when notetaking was allowed and negative effects when it was forced (Hale & Courtney, 1994).

From this brief review, we can observe a few general principles for notetaking in academic contexts: notes that represent the structure of key ideas on paper appear to lead to higher scores, especially on tasks which are productive and allow for review. Unfortunately, L2 notetakers almost universally underperform L1 notetakers in these contexts, especially when it comes to capturing main ideas (Asaly-Zetowi & Lipka, 2019; Clerehan, 1995; Olsen & Huckin, 1990), organizing notes to replicate the macrostructure of the text (Faraco et al., 2002; Olsen & Huckin, 1990), and self-efficacy (Desselle & Shane, 2019; Dunkel & Davy, 1989). These studies underscore the lack of preparation that L2 students have for the demands of notetaking in university contexts, and motivate a closer look at the impact of test tasks on the way that learners conceive of and prepare for EAP listening.

Theoretical Framework

This study builds on Field's (2013) model of listening comprehension, which is referenced in other studies on cognitive validity in listening assessment (Holznecht et al., 2017; Rukthong, 2021), and was specifically developed through examination of the differences between listening processes in tasks with and without preview (Field, 2012). In his model, the final stage in listening comprehension includes four discourse-construction processes: *selecting* (determining which ideas are worthy of attention), *integrating* (relating points to each other), *monitoring* (deciding whether incoming information makes sense against what has been heard before), and *structure-building* (mapping propositional hierarchy). We focus specifically on these processes because they are sometimes critically under-represented in tests that purport to measure academic listening ability (Field, 2011; Holznecht et al., 2017).

Research Questions

Our study seeks to explore the impact of stem-preview on student notes in an attempt to establish the cognitive validity of this task type. We have chosen to focus on stem-preview out of all the preview types

because it is the type that has the most theoretical justification (Iimura, 2010; O'Grady, 2021; Yanagawa & Green, 2008). By allowing access to the question stems but not the response options, stem-preview may plausibly be compared to the use of guided notes or a study guide, both resources seen in university contexts.

Accordingly, this study investigates student notes for evidence of cognitive processes in listening tests with and without stem-preview. This research agenda is addressed through three nested lines of inquiry:

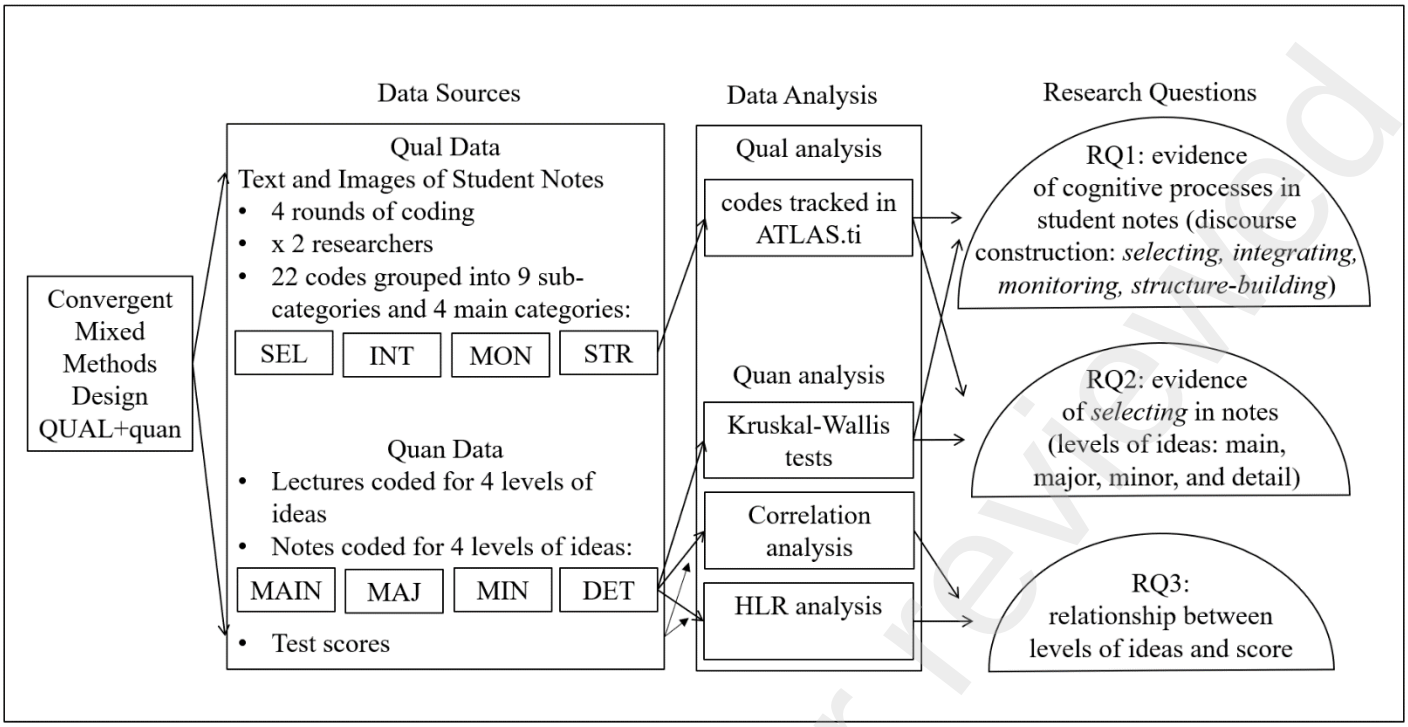
1. What evidence of discourse-construction processes is discernable in student notes with and without preview?
2. Which ideas (main, major, minor, and detail) are selected and recorded most frequently in student notes with and without preview?
3. What is the relationship between ideas in notes and test scores with and without preview?

Methods

Our study adopted a convergent mixed-methods design (Creswell & Plano Clark, 2018), depicted in Figure 1. Through qualitative analysis of text and images (Saldaña, 2015), we sought to establish an overarching view of student notetaking choices with and without access to preview. Qualitative analysis enabled us to observe the intersubjectivity of student notetaking strategies, focusing on the four discourse-construction processes identified by Field (2013). These observations were then supported by quantitative analysis of the first of those processes, *selecting*. Finally, we investigated the relationship between selection choices and test scores. Both strands of analysis converge, presenting an overall depiction of test-taking processes across conditions.

[Insert Figure 1.]

Figure 1. *Research design.*



178

179 ***Research Context***

180 This data was collected at a large public university in 2019 while exploring a possible revision to a local
181 EAP placement exam. At the time of data collection, the listening portion of the exam included two ten-minute
182 lectures followed by eight multiple-choice questions each. We wanted to investigate the impact of adding stem-
183 preview to determine which format would elicit the most ecologically valid test behaviors. An earlier study
184 relying on the same dataset focused on item difficulty, item type, and item discrimination (Author, Year).
185 Instruments and analysis from that study are available on the Open Science Framework (OSF) (anonymized):
186 https://osf.io/7x5yd/?view_only=13b96a9619214f49ae4320fb8d23a305.

187 ***Instruments***

188 Two ten-minute lectures with eight MCQs each were developed following specifications for the
189 placement exam. Both lectures were semi-scripted (Wagner & Wagner, 2016), included naturalistic oracy
190 features including repair, redundancy, and hesitation phenomena (Taylor & Geranpayeh, 2013), and were edited
191 in Audacity for sound quality and length (Audacity Team, 2019). Four MCQs were global items targeting main
192 ideas and inference, and four were local items targeting details and vocabulary. One additional item targeting a

trivial detail was designed to explore the impact of preview on item type, a key focus of the first study, and was excluded from the current analysis. All lectures and items underwent two rounds of piloting and revision. Following Koyama et al. (2016), we calculated reliability and dependability estimates separately for each combination of lecture and condition. All materials and reports are available on OSF.

Participants

Notetaking samples and test scores (n = 94) were collected from consenting undergraduate students in eight intact listening classes. Students in these classes had a TOEFL score between 80-99 or an IELTS score between 6.5-7.5. Following local Institutional Review Board protocol for intact classroom research, we did not collect identifying information about students. However, registrar data from 2019 indicates that the majority of international students enrolled in Fall 2019 were from China (50.8%), and that female-identifying students (52.9%) outnumbered male-identifying students (47.1%).

Data Collection

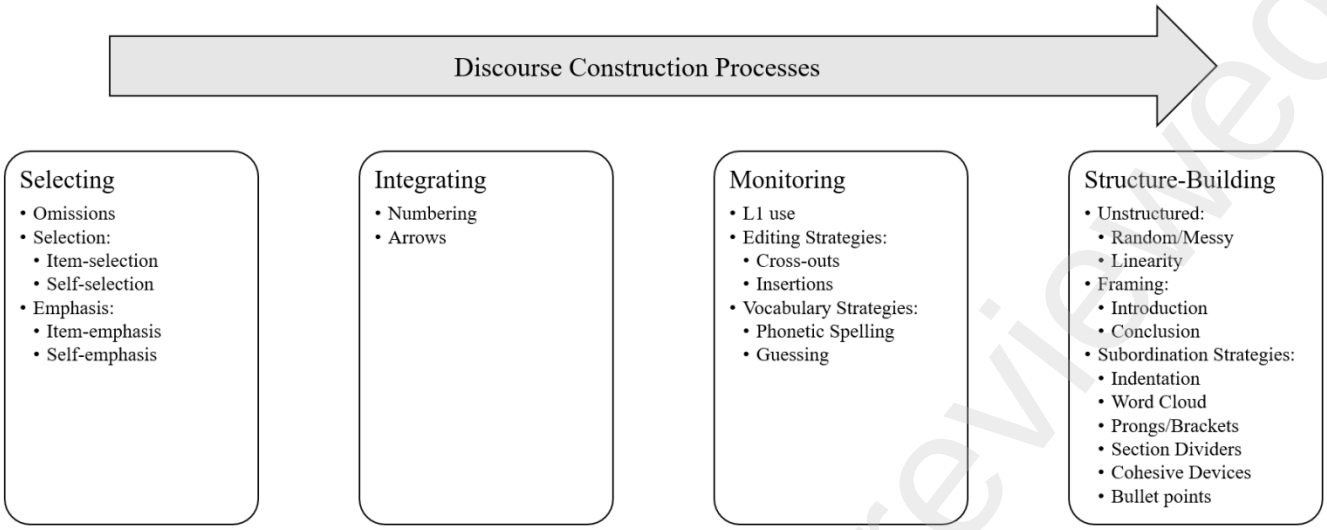
During the second week of the semester, Lecture A was administered to four classes with stem-preview, and to four classes without; one week later, Lecture B was administered with preview condition counterbalanced. All students were given two pages on which they were encouraged but not forced to take notes; the no-preview group had two blank sheets, while the preview group received one blank sheet and one preview sheet.

Qualitative Analysis

To address RQ1 and RQ2, we undertook a qualitative analysis of student notes using ATLAS.ti (Mac Version 22.0.6.0). Provisional codes (Saldaña, 2016) were developed based on a literature review of notetaking studies. Two researchers independently coded 100% of the data, and resulting codes were revised. In second-cycle focused coding, codes were grouped into categories which aligned with the four discourse-construction processes described in Field (2013). Five total rounds of coding were completed by two researchers, followed by discussion in which codes and categories were combined, condensed, and finalized, as summarized in Figure 2.

[Insert Figure 2.]

Figure 2. *Finalized qualitative codes.*



Quantitative Analysis

To address RQ2 and RQ3, it was first necessary to analyze both lectures for propositional structure. We adopted procedures from Kiewra et al. (1987) and Song (2011), identifying each proposition as either a main idea, major idea, minor idea, or detail. Two researchers analyzed both lectures independently; absolute agreement for Lecture A was .95 and for Lecture B .98. Disagreements were resolved through discussion.

Next, student notes were transcribed and analyzed for evidence of these ideas at each level. To assist with accuracy, words that were unique to each proposition were identified and highlighted in student notes. Two researchers then independently rated 10% of the data; inter-rater agreement was at 100%, and subsequently, one researcher rated the remainder of the data.

Following Nakayama et al. (2017), each notetaking sample was further scored in two ways. First, we wanted to control for differences in number of ideas across lectures. To accomplish this, we divided the number of ideas at each level in student notes by the number of ideas at that level in the lecture. This provided a measure of lecture coverage. Secondly, we wanted to control for differences in student writing fluency. We accomplished this by dividing the number of ideas students took at each level by the number of ideas they captured overall. This provided a measure of which level students focused on in their notes. We labelled these measures Coverage and Focus, respectively.

Quantitative analysis was conducted in SPSS (IBM Corp, 2022). We ran nonparametric Kruskal-Wallis tests to investigate whether preview affected notes students took at each level. Before running the analyses, we checked the assumptions based on Thorndike and Thorndike-Christ (2009). The results of Shapiro-Wilk test of normality showed that a few dependent variables (e.g., Main, Minor, and Detail Totals, and Minor Coverage) were not normally distributed. Thus, we decided to employ Kruskal-Wallis tests instead of multivariate analysis of variance. In the Kruskal-Wallis tests, the independent variables were preview and no-preview test conditions, while the dependent variables (summarized in Table 1) were Total Notations (TN) and Ideas Total (IT); Main, Major, Minor, and Detail Totals (T1, T2, T3, T4); Main, Major, Minor, and Detail Coverage (C1, C2, C3, C4); and Main, Major, Minor, and Detail Focus (F1, F2, F3, F4).

[Insert Table 1.]

Table 1. *Abbreviations, labels, and definitions for notetaking variables.*

Notetaking Variable Abbreviations	Notetaking Variable Labels	Notetaking Variable Definitions
TN	Total Notations	Total number of notations including words, abbreviations, and symbols
IT	Ideas Total	Total number of ideas referenced in student notes across all four levels; the sum of T1, T2, T3, and T4
T1	Main Idea Total	Total number of main ideas referenced in student notes
T2	Major Idea Total	Total number of major ideas referenced in student notes
T3	Minor Idea Total	Total number of minor ideas referenced in student notes
T4	Detail Idea Total	Total number of detail ideas referenced in student notes
C1	Main Idea Coverage	T1 divided by the number of main ideas in the lecture
C2	Major Idea Coverage	T2 divided by the number of major ideas in the lecture
C3	Minor Idea Coverage	T3 divided by the number of minor ideas in the lecture

C4	Detail Idea Focus	T4 divided by the number of detail ideas in the lecture
F1	Main Idea Focus	T1 divided by IT
F2	Major Idea Focus	T2 divided by IT
F3	Minor Idea Focus	T3 divided by IT
F4	Detail Idea Focus	T4 divided by IT

To understand the relationship between levels of notes and listening performance across conditions, we conducted correlation and multiple regression analyses. In terms of hierarchical linear regression (HLR) analyses, we did not find multicollinearity in the data, as the tolerance scores for the predictor variables were all well above .1 (ranging from .75-.99). The scatterplots of standardized predicted values by standardized residuals showed that our data were homoscedastic. The residuals were also normally distributed in the histogram and normal probability plots. Moreover, based on the results of Cook's distance, we did not detect any outliers. When conducting HLR analyses, the dependent variables from the Kruskal-Wallis tests above served as the independent variables in the HLR analysis, while the dependent variables were students' listening test scores. We conducted HLR analyses separately for each condition.

Results

Qualitative Results

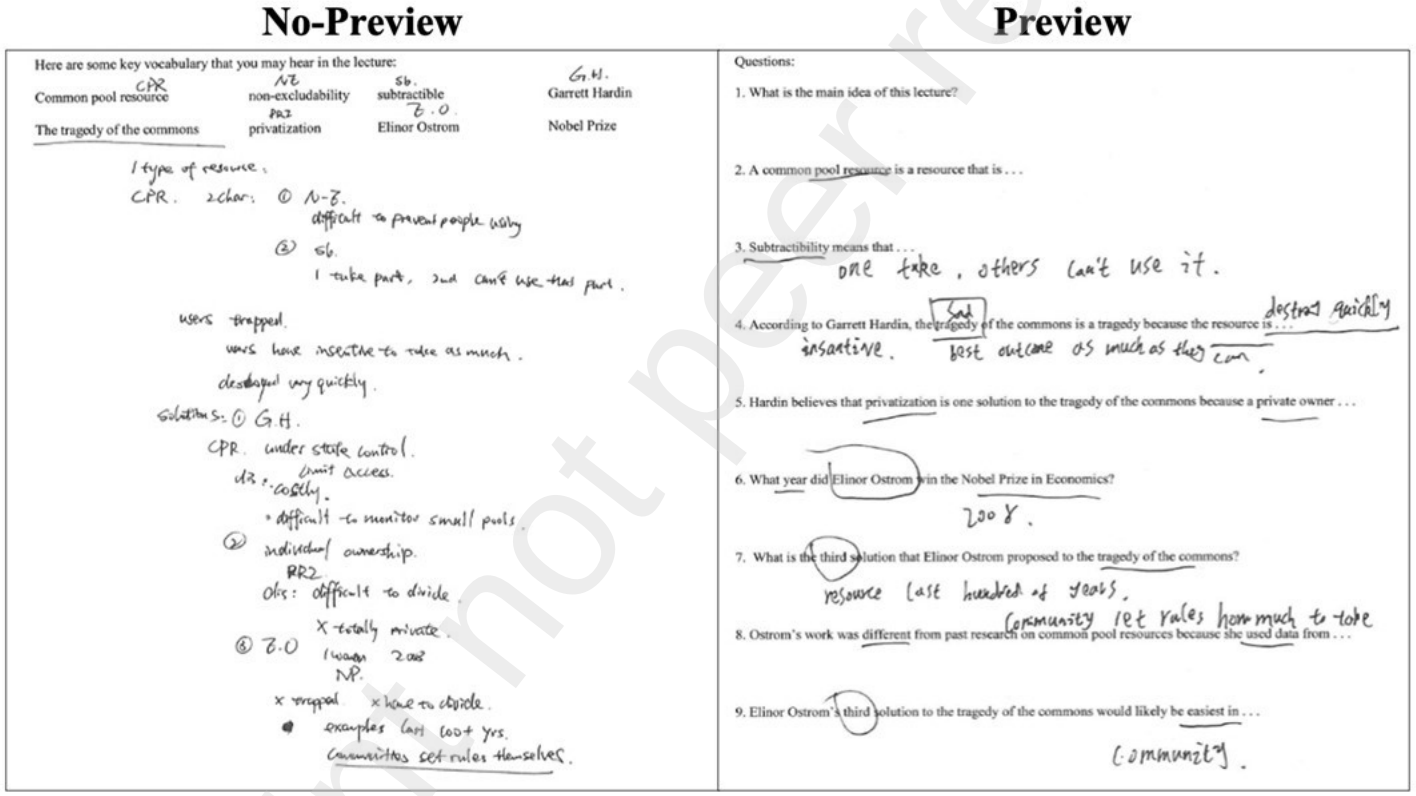
Qualitative analysis revealed several salient patterns in student notes across the four dimensions of discourse-construction identified by Field (2013). Within the category of *selecting*, two distinguishable patterns emerged. First, students in the preview condition tended to omit major sections of the lecture from their notes if no questions clearly addressed those sections (e.g., Lecture B Solution 2). These *omissions* were systematic and predictable, while in the no-preview condition, omissions were inconsistent and unpredictable. Second, differences emerged in student *emphasis* of notes (indicated visually by underlining, circling, and starring words). Students in both conditions engaged in self-emphasis of words they had written, though this was more common if notes were taken freestyle (in the no-preview condition or on the blank page in the preview condition). Marked differences appeared, however, in terms of item-emphasis: over half of the students in the

preview condition underlined or circled keywords in the question stems, indicating that they were relying on those words to help them make connections in the lecture. This behavior was less frequent in the no-preview condition. Figure 3 displays representative examples of emphasis across conditions.

In the second lecture administration, there was a notable uptick in students who attempted to employ item-emphasis in the no-preview condition by underlining or circling words in the heading or gloss (36% Lecture A; 56% Lecture B).

[Insert Figure 3].

Figure 3. *Representative examples of item-emphasis and self-emphasis in preview and no-preview conditions.*



Integrating strategies were the most frequently represented in student notes across both conditions. There were only two codes in this category, but each was heavily used, with *numbering* being the most common, followed by *arrows*. Both of these strategies were frequently used in both conditions, but especially in the no-preview condition or in freestyle preview notes.

Student use of *monitoring* strategies appeared to be fairly constant across conditions. Four students used *translanguaging* in each condition (likely the same four, judging from handwriting). *Editing* strategies, such as

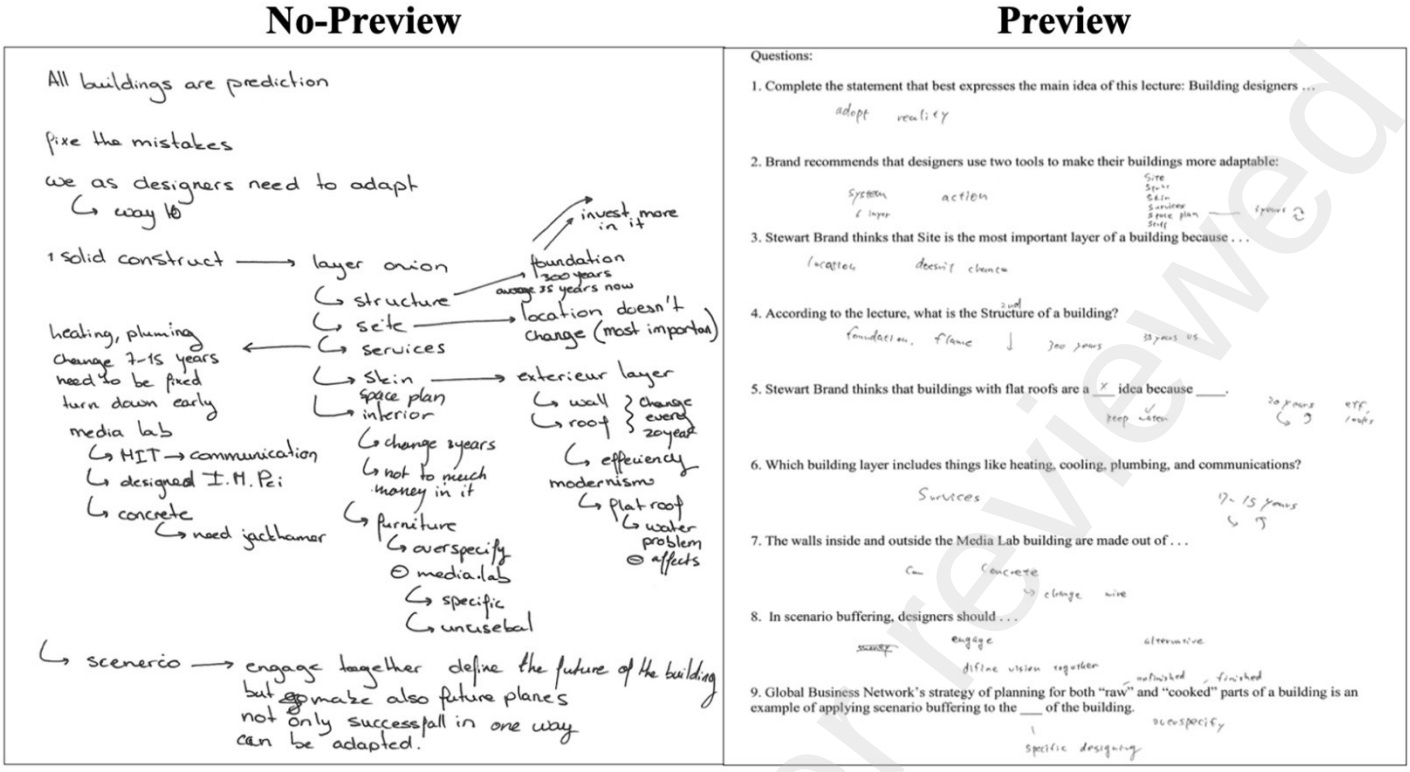
cross-outs and insertions, were frequent in both conditions, with slightly higher prevalence in the preview condition. *Vocabulary* strategies (such as guessing or phonetic spelling of unknown words) appeared to be consistent across conditions.

The greatest number of codes were clustered in the *structure-building* category, in which three salient patterns emerged. First, *unstructured* codes (random/messy or linear) were over-exemplified in the preview condition. Second, *framing* references (introductions and conclusions) were nearly absent in the preview condition. Finally, *subordination* strategies (indentation, word clouds, brackets, section dividers, and cohesive devices) were used extensively in the no-preview condition; indentation, for instance, sometimes reached up to five levels. Conversely, indentation in notes taken under the question stems in the preview condition was extremely rare (only six examples) and never exceeded two levels. Figure 4 illustrates indentation use across both conditions.

Notably, in Lecture B administration, there was a marked increase in the number of students in the preview condition who chose to take notes freestyle (23% Lecture A; 43% Lecture B). Multiple levels of indentation were sometimes observed in freestyle preview notes.

[Insert Figure 4.]

Figure 4. *Representative examples of subordination strategies in preview and no-preview conditions.*



Quantitative Results

Table 2 details descriptive statistics of students' test scores and notes at each level across conditions.

[Insert Table 2.]

Table 2. *Means and standard deviations of scores and notes in preview and no-preview conditions.*

Notetaking Variables	Condition			
	Preview (n = 55)		No-Preview (n = 40)	
	M	SD	M	SD
Test Scores	5.43	1.79	4.90	2.09
TN	105.96	47.40	140.58	55.21
IT	29.91	12.55	36.88	14.60
T1	1.69	1.25	2.40	1.39
T2	10.16	4.57	11.80	4.39
T3	9.69	4.92	11.47	6.12

T4	8.40	5.07	11.25	5.25
C1	.37	.29	.54	.30
C2	.45	.19	.55	.17
C3	.41	.20	.52	.19
C4	.21	.11	.27	.13
F1	.05	.04	.06	.03
F2	.34	.09	.33	.08
F3	.32	.08	.30	.07
F4	.28	.12	.30	.09

Kruskal-Wallis tests revealed that test condition had significant effects on students' notes at each level, as indicated in Table 3. In particular, test condition had statistically significant effects at the $p < 0.05$ level on IT ($\chi2(1) = 5.93$), T1 ($\chi2(1) = 6.10$), and C4 ($\chi2(1) = 5.89$), and at the $p < 0.01$ level on TN ($\chi2(1) = 9.23$), T4 ($\chi2(1) = 7.08$), C1 ($\chi2(1) = 8.24$), C2 ($\chi2(1) = 6.86$), and C3 ($\chi2(1) = 6.75$). Specifically, the results of Kruskal-Wallis tests showed that the mean ranks for Total Notations, Ideas Total, Main Total, Detail Total, and Coverage measures at all four levels were significantly higher in the no-preview condition. In other words, students without preview tended to take more notes overall, especially main ideas and details, and had better coverage of ideas from the lecture across all levels.

[Insert Table 3.]

Table 3. *Kruskal-Wallis test of statistical significance (degree of freedom = 1).*

Notetaking Variables	Mean rank		Kruskal-Wallis chi-square
	Preview (n = 55)	No-Preview (n = 40)	
TN	40.67	58.08	$\chi^2(1) = 9.23$, $p = 0.00$**
IT	42.13	56.08	$\chi^2(1) = 5.93$, $p = 0.02$*

T1	42.17	56.01	$\chi^2(1) = 6.10$, $p = 0.01$*
T2	43.55	54.13	$\chi^2(1) = 3.43$, $p = 0.06$
T3	44.47	52.85	$\chi^2(1) = 2.15$, $p = 0.14$
T4	41.60	56.80	$\chi^2(1) = 7.08$, $p = 0.00$**
C1	41.13	57.45	$\chi^2(1) = 8.24$, $p = 0.00$**
C2	41.69	56.68	$\chi^2(1) = 6.86$, $p = 0.01$*
C3	41.75	56.60	$\chi^2(1) = 6.74$, $p = 0.00$**
C4	42.15	56.04	$\chi^2(1) = 5.89$, $p = 0.02$*
F1	44.75	52.48	$\chi^2(1) = 1.83$, $p = 0.18$
F2	49.81	45.51	$\chi^2(1) = 0.53$, $p = 0.45$
F3	50.72	44.26	$\chi^2(1) = 1.27$, $p = 0.26$
F4	45.57	51.34	$\chi^2(1) = 1.01$, $p = 0.31$

**$p < .01$.

*$p < .05$.

To explore the relationship between levels of notes and listening scores, we adopted correlation and HLR regression for both conditions. Correlation analyses displayed in Table 4 show that in the preview condition, Total Notations, Minor Total, Minor Coverage, and Minor Focus in student's notes were statistically and positively correlated with test scores. In other words, in the preview condition, students scored higher if they took more notes overall, and especially if they focused on minor ideas.

[Insert Table 4.]

Table 4. *Correlations between notes and test scores: Preview*

	TN	IT	T1	T2	T3	T4	C1	C2	C3	C4	F1	F2	F3	F4
Score	.30*	.23	.09	.17	.39**	.02	.11	.12	.34**	.11	.02	-.20	.48**	-.20

**$p < .01$.

*$p < .05$.

In the no-preview condition, no significant correlations were observed between level of notes and test scores, as shown in Table 5.

[Insert Table 5.]

Table 5. *Correlations between levels of notes and test scores: No-Preview*

	TN	IT	T1	T2	T3	T4	C1	C2	C3	C4	F1	F2	F3	F4
Score	.01	.22	.11	.28	.28	.03	.07	.25	.30	.09	-.01	-.03	.25	-.14

* $p < .05$.

Finally, for both test conditions, all independent variables were entered into a stepwise regression. The regression analyses showed that Total Notations and Minor Idea Focus in the preview condition were the only significant predictors of score. The resulting models in Table 6 showed that these two variables explained 33% (r =.58, R^2 = .33) of the variance in test scores of the preview condition, with sole contributions from Minor Idea Focus (Model 1, ΔR^2 = .23) and Total Notations (Model 2, ΔR^2 = .10).

[Insert Table 6.]

Table 6. *Summary of stepwise regression model: Preview*

Entry	Predictors	*r*	Total R^2	R^2 change	*B*	*SE B*	*β*
1	Minor Idea Focus (F3)	.48	.23	.23	9.67	2.43	.48**
2	Total Notations (TN)	.58	.33	.10	.01	.00	.32**

** $p < .01$.

Discussion

This study explored the impact of stem-preview on cognitive validity in EAP listening assessment by examining student notes and test scores in two conditions. In response to RQ1, qualitative analysis revealed evidence of similar *integrating* and *monitoring* strategies across preview conditions, but distinctions emerged in *selecting* and *structure-building*. Students who previewed questions were much more likely to omit information if a question was not directly targeting it, even if that information was structurally important to the lecture. They were also more likely to highlight keywords in the stems than they were to highlight keywords in their notes. In

terms of *structure-building*, students who previewed questions were more likely to adopt a random or linear notetaking style, and to omit introductory and concluding material. Students who did not have access to preview were more likely to incorporate *subordination* strategies such as indentation, word clouds, and brackets, with up to five levels of indentation observed in some no-preview samples.

Quantitative analysis corroborates these findings for RQ1, specifically with regard to *selecting* ideas within notes. In response to RQ2, we found that students without preview were more likely to take more notes, capture more ideas overall (with more main ideas and details in particular), and have better coverage of ideas at all four levels. However, in answer to RQ3, none of these advantages predicted scores in the no-preview condition. In the preview condition, students scored higher if they took more notes overall, and specifically if they focused on minor ideas. This finding is especially noteworthy considering that the items in the test were designed to focus on global and local ideas equally (with four questions about each). Regardless, students in the preview condition who focused on minor ideas tended to perform better on the test.

Concerningly, our qualitative analysis revealed unexpected evidence of a washback effect across the two administrations. Students who experienced Lecture A with preview were more likely to employ *item-emphasis* when they took Lecture B without preview a week later. Without access to the question stems, some students reverted to circling keywords in the heading and gloss. The reverse was true for students who experienced Lecture A without preview; they were more likely to take notes for Lecture B freestyle rather than under the stems, even though they had access to the questions. Students who opted to take notes freestyle were much more likely to apply no-preview-style strategies, such as *self-emphasis*, *framing*, and *subordination*. In other words, students who had access to preview first tried to rely on keywords even when taking a no-preview test later, while students who had no-preview first were more likely to ignore the stems, even when they had access to them later. This suggests that strategy use on listening tests may be susceptible to washback from test format.

Overall, these findings corroborate concerns that preview promotes passive listening strategies. Regarding selection of material to include in notes, Field (2011) observes that in MCQ tests with preview, "much of the necessary decision-making is taken care of by item writers. They, not the listener, determine which points of information are relevant and which are not; and they reduce the information in the recording to

a string of discrete points, regardless of how each contributes to the line of argument" (p. 110). In contrast, when students are forced to create a mental representation of the text on their own, they must make these choices independently. *Selecting* and *structure-building* strategies can also facilitate crucial aspects of the writing process, which could explain why productive tasks consistently reward notetaking, even when MCQ tasks do not (Liu & Hu, 2012). Rukthong (2021) observed that students took disorganized or linear notes when readying for an MCQ task, but used indentation and arrows in their notes in preparation for a summary task. Our study contributes evidence that tasks without preview elicit more discourse-construction processes, while students with preview employed more passive strategies.

Specifically, these results confirm reports that preview facilitates the use of testwise strategies such as keyword matching, aural scanning, and guessing (Badger & Yan, 2012; Field, 2011; 2012). This evidence may also partially explain findings that WLP tests with preview elicit eye gaze behaviors and neural activation patterns consistent with shallow processing (Aryadoust et al., 2022; Zhai & Aryadoust, 2022). In other words, the differences observed in these studies could be potentially attributable to preview instead of response timing, in that the behaviors observed would not have been possible without access to the questions while listening.

Beyond influencing which strategies students use, task format appears to reward those strategies differentially. Our results suggest that preview actually rewarded the use of shallower processing strategies, while students in the no-preview condition who took more organized notes and focused on main ideas did not see gains in scores. These results may contextualize the findings from In'nami and Koizumi (2022), who observed a relationship between self-reports of planning and evaluation strategies and WLP test scores, but not PLP scores. In our study, test-takers used question stems to predict keywords and make selections about what to include in their notes, which can be interpreted as evidence of planning strategies, and these strategies were rewarded with higher scores. However, these planning strategies may not transfer outside the test context.

Implications

A number of implications can be drawn from this study for language pedagogy, test development, and assessment research. First, language instructors in EAP contexts are often expected to prepare students for standardized listening assessments that employ question preview. In this context, they may feel pressure to

prioritize test preparation over authentic listening tasks. This tension can be addressed through honest discussion with students about the limitations of listening tests with preview, and by varying classroom assessment types to include practice for standardized exams along with integrated listening tasks, while simultaneously providing instruction in notetaking strategies. Such instruction can be very effective (Siegel, 2020; Yang & McAllister, 2023), especially when measured by productive tasks (Cubilo & Winke, 2013; Song, 2011). In some cases, instructors may wish to provide notetaking scaffolding through the use of guided notes, which can range in specificity from basic headings to cloze tasks (Chen et al., 2017; Cushing, 1991; Song, 2011). Konrad et al. (2009) recommends a "systematic fading" of guided notes, with greater specificity provided at the beginning of the semester which is gradually withdrawn until students are able to take organized notes on their own (p. 440). Instructors can also find it valuable to collect notes periodically in order to provide students with feedback on things like omissions and subordination strategies.

Second, in terms of test development, MCQ presentation formats should not be assumed to be interchangeable. Access to preview may impact the strategies that are available to test-takers, resulting in potential threats to cognitive validity and the interpretation of test scores. In particular, EAP test developers should ensure that listening tasks facilitate and reward listening behaviors which will transfer to academic listening contexts. Tasks which are observed to foster and reward testwise strategies should be questioned.

Finally, the time is ripe for a Copernican revolution in L2 notetaking research. Some have questioned the value of notetaking for L2 learners after finding only weak associations between notetaking and L2 standardized test scores (e.g., Clark et al., 2014). However, our findings suggest that this interpretation should be reversed: rather than questioning of the value of notetaking, we ought to turn our critical gaze around and question the appropriacy of listening tasks which do not facilitate good notes. Notetaking has proven to be indicative of success on L2 integrated tasks (Field, 2012; Liu & Hu, 2012; Rukthong, 2021; Rukthong & Brunfaut, 2020), and critical to success in university contexts (Asaly-Zetowi & Lipka, 2019; Clerehan, 1995; Olsen & Huckin, 1990). Further, notes provide a visible record of a listener's cognitive processes during a test, drawing our attention to the *process* and not only the *product* of listening (Faraco et al., 2002). As such, notes can serve as a form of cognitive validity evidence in listening assessment, supplementing other measures used

for this purpose, including self-report, eye-tracking, or fNIRS. While self-report measures can reliably indicate test-takers' self-knowledge and self-regulation, they may not be not as reliable in indicating behavior (Craig et al., 2020). Conversely, if used in isolation, eye-tracking and fNIRS data may reveal behavioral patterns that are difficult to interpret; for example, Holznecht (2019) observes that behaviors such as focusing and zoning out may appear indistinguishable in eye-tracking data unless supplemented with stimulated recalls. Alongside these measures, notes can provide a more interpretable record of test-taker behavior, and thus should supplement these data sources in validity research. While notetaking data is less readily quantifiable, it yields itself readily to qualitative analysis. We should not shy away from analyzing notetaking because of its complexity, but mine those complexities for validity evidence.

Limitations, Future Research, and Conclusions

The nature of intact classroom research limits our confidence in making inferences about the effect of preview in other contexts. First, because the listening syllabus emphasized the importance of notetaking, we would expect that, regardless of preview condition, students were more motivated to take notes than might be expected in some other contexts. Second, we allowed students in the preview condition a choice about whether to take notes under the stems or on the blank page, a choice that is not permitted on most standardized tests with preview. This decision probably did minimize the differences we might otherwise expect to find between the two groups; however, it provided us with an unexpected opportunity to observe a washback effect across the two lecture administrations. Third, our sample was limited by class size, which limits the conclusions we can draw from our quantitative analysis. Although our data met test assumptions for correlation and regression analyses, the quantitative strand of our study ought to be interpreted as explorative and supportive of our qualitative findings, which constitute the main pillar of this study. The anonymous nature of our data collection additionally prevents us from making claims about the interaction between performance and individual characteristics. Finally, the absence of a follow-up interview means a loss of opportunity to hear test-takers explain their notetaking choices in their own words.

Despite these limitations, our findings motivate further exploration of preview and notetaking in EAP contexts. We hope that future studies will examine the impact of preview on notetaking in other contexts,

varying the type of preview (preview types which include options might conceivably have a greater impact on notetaking), the assessment task (students might employ different notetaking strategies when expecting a MCQ test or a summary), and the time of testing (notetaking may impact immediate and delayed post-tests differently). Beyond this, we hope to see investigation of even more innovative listening test formats which include opportunities for integrating sources, discussion, and review. Recent scholarship has established the importance of assessing writing in EAP contexts through authentic tasks which facilitate positive washback. We hope the time has come to put listening assessment under similar inspection (Lynch, 2011).

The present study found evidence of major omissions, shallower structural representation, and minor idea focus in student notes when stems were previewed before a listening test. In the absence of preview, notes were more comprehensive, represented more levels of structure, and focused more on main ideas. However, these behaviors were not rewarded when the task was scored. Given the importance of notetaking for student success in academic contexts, it is vital to ensure that listening assessment elicits and rewards cognitively valid notetaking behaviors. We hope to see more investigation of notetaking and a commitment to the development of test formats which better prepare students for success in academic contexts.

References

Alavi, S., and Janbaz, F. (2014). Comparing two pre-listening supports with Iranian EFL learners: Opportunity or obstacle. *RELC Journal, 45*(3), 253–267. https://doi.org/10.1177/0033688214546963

Allen, M., Lefebvre, L., Lefebvre, L, and Bourhis, J. (2020). Is the pencil mightier than the keyboard? A meta-analysis comparing the method of notetaking outcomes. *Southern Communication Journal, 85*(3), 143–154. https://doi.org/10.1080/1041794X.2020.1764613

Aryadoust, V., Foo, S., and Ng, L. (2022). What can gaze behaviors, neuroimaging data, and test scores tell us about test method effects and cognitive load in listening assessments? *Language Testing, 39*(1), 56-89. https://doi-org.proxy.lib.uiowa.edu/10.1177/02655322211026876

Asaly-Zetowi, M., and Lipka, O. (2019). Note-taking skill among bilingual students in academia: Literacy, language and cognitive examination. *Frontiers in Psychology, 10.* https://doi.org/10.3389/fpsyg.2019.00870

ATLAS.ti Scientific Software Development GmbH [ATLAS.ti 22 Windows]. (2022). https://atlasti.com

Audacity Team. (2019). *Audacity.*

Author. (Year).

Badger, R., and Yan, X. (2012). The use of tactics and strategies by Chinese students in the Listening component of IELTS. *IELTS Research Reports, 9,* 67-96. https://www.ielts.org/-/media/research-reports/ielts_rr_volume09_report2.ashx

Berne, J. (1995). How does varying pre-listening activities affect second language listening comprehension? *Hispania*, *78*(2), 316–329. https://doi.org/10.2307/345428

Brobst, K. (1996). The process of integrating information from two sources, lecture and text (9631667). [Doctoral dissertation, Teachers College, Columbia University]. ProQuest Dissertations Publishing.

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

Chang, A., and Read, J. (2006). The effects of listening support on the listening performance of EFL learners.

TESOL Quarterly, 40(2), 375-397. https://onlinelibrary.wiley.com/doi/pdf/10.2307/40264527

Chaudron, C., Loschky, L., and Cook, J. (1994). Second language listening comprehension and lecture note-taking. In J. Flowerdew (Ed.), *Academic listening: Research perspectives* (pp. 75-92). Cambridge University Press.

Chen, P., Teo, T., and Zhou, M. (2017). Effects of guided notes on enhancing college students' lecture note-taking quality and learning performance. *Current Psychology, 36*(4), 719-732. http://dx.doi.org/10.1007/s12144-016-9459-6

Cheng, H. (2004). A comparison of multiple-choice and open-ended response formats for the assessment of listening proficiency in English. *Foreign Language Annals, 27*(4), 544-553. https://doi.org/10.1111/j.1944-9720.2004.tb02421.x

Clark, M., Wayland, S., Osthus, P., Brown, K., Castle, S., and 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

Cohen, A. (2007). The coming of age for research on test-taking strategies. In J. Fox, M. Wesche & D. Bayliss (Eds.), *Language testing reconsidered* (pp. 89-111). University of Ottawa Press.

Craig, K., Hale, D., Grainger, C., and 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., and Plano Clark, V. (2018). *Designing and conducting mixed methods research, 3rd ed.* Sage.

Cubilo, J., and Winke, P. (2013). Redefining the L2 listening construct within an integrated writing task: Considering the impacts of visual-cue interpretation and note-taking. *Language Assessment Quarterly, 10*(4), 371-397. https://doi.org/10.1080/15434303.2013.824972

Cushing, S. (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.

518 Desselle, S., and Shane, P. (2019). Native English speakers and English as a Foreign Language (EFL) students'
519 performance and notetaking in a Doctor of Pharmacy health systems course. *Research in Social and*
520 *Administrative Pharmacy, 15*(9), 1154-1159. https://doi.org/10.1016/j.sapharm.2018.09.023
521 Dunkel, P. (1988). The content of L1 and L2 students' lecture notes and its relation to test performance. *TESOL*
522 *Quarterly, 2*(2), 259-281. https://doi.org/10.2307/3586936
523 Dunkel, P., and Davy, S. (1989). The heuristic of lecture notetaking: The American university perceptions of
524 American international students regarding the value & practice of notetaking. *English for Specific*
525 *Purposes, 8*(1), 33-50. https://www.sciencedirect.com/science/article/pii/0889490689900057
526 Faraco, M., Barbier, M., and Piolat, A. (2002). A comparison between notetaking in L1 and L2 by
527 undergraduate students. In S. Ransdell & M. Barbier (Eds.), *Studies in Writing, Volume 11: New*
528 *Directions for Research in L2 Writing* (pp. 145-167). Kluwer Academic Publishers.
529 Field, J. (2011). Into the mind of the academic listener. *Journal of English for Academic Purposes, 10*(2)*,* 102–
530 112. https://doi.org/10.1016/j.jeap.2011.04.002
531 Field, J. (2012). The cognitive validity of the lecture-based question in the IELTS Listening paper. *IELTS*
532 *Collected Papers 2,* 391-453. https://www.ielts.org/-/media/research-
533 reports/ielts_rr_volume09_report1.ashx
534 Field, J. (2013). Cognitive validity. In L. Taylor & A. Geranpayeh (Eds.), *Examining listening* (pp. 77–151).
535 Cambridge University Press.
536 Hale, G., and Courtney, R. (1994). The effects of note-taking on listening comprehension in the Test of English
537 as a Foreign Language. *Language Testing, 11*(1), 29-47. https://doi.org/10.1177/026553229401100104
538 Hayati, A., and Jalilifar, A. (2009). The impact of note-taking strategies on listening comprehension of EFL
539 learners. *English Language Teaching, 2*(1), 101-111. https://files.eric.ed.gov/fulltext/EJ1082250.pdf
540 Holznecht, F., Eberharter, K., Kremmel, B., Zehentner, M., McCray, G., Konrad, E., and Spöttl, C. (2017).
541 Looking into listening: Using eye-tracking to establish the cognitive validity of the Aptis Listening Test.
542 *ARAGs Research Reports Online AR-G/2017/3. British Council.*
543 https://www.britishcouncil.org/exam/aptis/research/publications/arags/looking-listening-using-eye-

tracking

Holznecht, F. (2019). Double play in listening assessment. [Doctoral Dissertation, Lancaster University.] Retrieved from https://eprints.lancs.ac.uk/id/eprint/139699/1/2019holzknechtphd.pdf#:~:text=Candidates%20displayed%20more%20higher%2D%20order,and%20were%20markedly%20less%20anxious.

IBM Corp. (2022). *IBM SPSS Statistics for Windows, Version 29.0.* IBM Corp.

Iimura, H. (2010). Factors affecting listening performance on multiple-choice tests: The effects of stem/option preview and test characteristics. *Language Education and Technology, 47*, 17-36. https://doi.org/10.24539/let.47.0_17

In'nami, Y., and Koizumi, R. (2022) The relationship between L2 listening and metacognitive awareness across listening tests and learner samples. *International Journal of Listening, 36*(2), 100-117. https://doi.org/10.1080/10904018.2021.1955683

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

Kiewra, K., Benton, S., and Lewis, L. (1987). Qualitative aspects of notetaking and their relationship with information processing ability and academic achievement. *Journal of Instructional Psychology, 14*(3), 110-117.

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

Konrad, M., Joseph, L, and Eveleigh, E. (2009). A meta-analytic review of guided notes. *Education and*

Treatment of Children, 32(3), 421-444. https://www.jstor.org/stable/42900031

Koyama, D., Sun, A., and Ockey, G. (2016). The effects of item preview on video-based multiple-choice listening assessments. *Language Learning & Technology*, *20*(1), 148–165. http://lib.dr.iastate.edu/engl_pubs/73

Li, C., Wu, M., Kuo, Y., Tseng, Y., Tsai, S., and Shih, H. (2017). The effects of cultural familiarity and question preview type on the listening comprehension of L2 learners at the secondary level. *The International Journal of Listening, 31*(2), 98-112. https://doi.org/10.1080/10904018.2015.1058165

Liu, B., and 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

Lynch, T. (2011). Academic listening in the 21st century: Reviewing a decade of research. *Journal of English for Academic Purposes, 10*, 79-88. https://doi.org/10.1016/j.jeap.2011.03.001

Madani, B., and Kheirzadeh, S. (2022). The impact of pre-listening activities on Efl learners' listening comprehension. *International Journal of Listening, 36*, 53–67. https://doi.org/10.1080/10904018.2018.1523679

Nakayama, M., Mutsuura, K., and Yamamoto, H. (2017). The possibility of predicting learning performance using features of note taking activities and instructions in a blended learning environment. *International Journal of Educational Technology in Higher Education, 14*(6). https://doi.org/10.1186/s41239-017-0048-z

Northern, P., Tauber, S., Hilaire, K., and Carpenter, S. (2023). Application of a two-phase model of note quality to explore the impact of instructor fluency on students' note-taking. *Journal of Applied Research in Memory and Cognition, 12*(1), 94-104. https://doi.org/10.1037/mac0000032

Oakhill, J., and 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

O'Grady, S. (2021). Adapting multiple-choice comprehension question formats in a test of second language

listening comprehension. *Language Teaching Research.* Advance online publication. https://doi.org/10.1177/1362168820985367

Olsen, L., and 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

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

Rukthong, A., and 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., and 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). *Thinking qualitatively: Methods of mind.* Sage.

Saldaña, J. (2016). *The coding manual for qualitative researchers.* Sage.

Siegel, J. (2020). Effects of notetaking instruction on intermediate and advanced L2 English learners: A quasi-experimental study. *Journal of English for Academic Purposes, 46,* 1-10. https://doi.org/10.1016/j.jeap.2020.100868

Siegel, J. (2022). Factors affecting notetaking performance. *International Journal of Listening.* Advance online publication. https://doi.org/10.1080/10904018.2022.2059484

Song, M. (2011). Notetaking quality and performance on an L2 academic listening test. *Language Testing, 29*(1), 67-89. https://doi.org/10.1177/0265532211415379

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

Thorndike, R., and Thorndike-Christ, T. (2009). *Measurement and evaluation in psychology and education, 8th ed.* Pearson.

Voyer, D., Ronis, S., and Byers, N. (2022). The effect of notetaking method on academic performance: A systematic review and meta-analysis. *Contemporary Educational Psychology, 68.* Advance online

publication. https://doi.org/10.1016/j.cedpsych.2021.102025

Wagner, E., and Wagner, S. (2016). Scripted and unscripted spoken texts used in listening tasks on high-stakes tests in China, Japan, and Taiwan. In V. Aryadoust & J. 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.

Yanagawa, K., and Green, A. (2008). To show or not to show: The effects of item stems and answer options on performance on a multiple-choice listening comprehension test. *System, 36*(1), 107-122. https://doi.org/10.1016/j.system.2007.12.003

Yang, M., and McAllister, G. (2023). 'Drawing out the whole picture': Positive and gestalt effects of taking sign-based notes on listening performance in Chinese ESL classrooms. *Behavioral Sciences, 13,* 395. https://doi.org/10.3390/bs13050395

Zhai, J., and Aryadoust, V. (2022). The metacognitive and neurocognitive signatures of test methods in academic listening. *Frontiers in Psychology.* Advance online publication. https://doi.org/10.3389/fpsyg.2022.930075

Zhou, X., Chen, X., and Wang, Z. (2022). The effect of linguistic choices in notetaking on academic listening performance: A pedagogical translanguaging perspective. *International Review of Applied Linguistics in Language Teaching.* Advance online publication. https://doi.org/10.1515/iral-2022-0127

Notetaking as Validity Evidence: A Mixed-Methods Investigation of Question Preview in EAP Listening Assessment

Authors
Rebecca Yeager
University of Iowa
https://orcid.org/0000-0001-8017-5879

GoMee Park
University of Texas Rio Grande Valley
https://orcid.org/0000-0001-5107-0568

Ray J. T. Liao
National Taiwan Ocean University
https://orcid.org/0000-0003-4246-1449

Correspondence concerning this article should be addressed to Rebecca Yeager
Rebecca-yeager@uiowa.edu
ESL Programs, University of Iowa, Iowa City, IA 52242, USA

Acknowledgements
The authors wish to thank the students and instructors at the University of Iowa who participated in data collection for this study. We are grateful to Zach Meyer and Ryan Lidster for their helpful contributions at the conceptual stage of this study, and to Stacy Sabraw, Melissa Meisterheim, and Jieun Kim for their comments on an earlier version of this manuscript. Any remaining errors are our own. Finally, we acknowledge that our research was conducted on land stolen from Native American tribes, and are eager to see these tribes recompensed for the use of their land.

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

Abstract
Recent scholarship has questioned the cognitive validity of listening tests with preview, in which test-takers can see test questions before listening. This study mined student notes for evidence of cognitive processes in listening tests with and without preview, using a mixed-methods design that explored the effect of test format on notetaking behaviors. Qualitative analysis indicated that students who previewed items were more likely to systematically omit information, highlight keywords, and engage in shallower structural representation. Conversely, Kruskal-Wallis tests revealed that students who listened without preview took more notes, especially of main ideas and details, and had better coverage of the lecture. However, correlation and hierarchical linear regression analyses found these notetaking achievements did not predict higher scores in the no-preview condition, while in the preview condition, only note quantity and focus on minor ideas predicted scores. Both strands of data suggest that students' cognitive processes were shaped by the format of the exam they experienced. These findings may bear on validity arguments for listening assessment and inform the way that language instructors prepare their students for academic listening.

Keywords: listening, test format, testwise strategies, washback, cognitive validity