Exploring AI-Enabled Test Practice, Affect, and Test Outcomes in Language Assessment

Jill Burstein*
Ramsey Cardwell
Ping-Lin Chuang
Allison Michalowski
Steven Nydick
Duolingo
{jill, ramsey, pinglin, allison.michalowski, steven.nydick}@duolingo.com

Abstract

Practice tests for high-stakes assessment are intended to build test familiarity, and reduce construct-irrelevant variance which can interfere with valid score interpretation. Generative AI-driven, automated item generation (AIG) scales the creation of large item banks and multiple practice tests, enabling repeated practice opportunities. We conducted a large-scale observational study (N = 25,969) using the Duolingo English Test (DET)—a digital, high-stakes, computer-adaptive English language proficiency test to examine how increased access to repeated test practice relates to official DETscores, test-taker affect (e.g., confidence), and score-sharing for university admissions. To our knowledge, this is the first large-scale study exploring the use of AIG-enabled practice tests in high-stakes language assessment. Results showed that taking 1-3 practice tests was associated with better performance (scores), positive affect (e.g., confidence) toward the official DET, and increased likelihood of sharing scores for university admissions for those who also expressed positive affect. Taking more than 3 practice tests was related to lower performance, potentially reflecting *washback* – i.e., using the practice test for purposes other than test familiarity, such as language learning or developing test-taking strategies. Findings can inform best practices regarding AI-supported test readiness. Study findings also raise new questions about test-taker preparation behaviors and relationships to test-taker performance, affect, and behaviorial outcomes.

1 Introduction

For millions of international test takers, scores on high-stakes English language proficiency (ELP) assessments can profoundly impact their educational and professional goals. As a result, they engage in various test preparation strategies. For example, practice tests aim to build familiarity for a specific test; reading books and articles can improve English language reading skills; and, deliberate engagement in conversations with peers and instructors can strengthen English language speaking and listening skills.

This paper focuses on *practice tests*. Practice tests aim to build test familiarity to reduce test-design-related *construct-irrelevant variance* (CIV). CIV is associated with the introduction of factors unrelated to the skills a test is intended to measure (the *target construct*) (Messick, 1982; Powers, 1985). For instance, CIV can stem from unfamiliar technical features (e.g., *drag-and-drop*), lack of familiarity with the device required for taking a test (e.g., test requirements to use a laptop for test takers who have limited laptop experience (Koné et al. (2024)), or anxiety triggered by an unfamiliar format (Winke and Lim, 2017).

Conventional practice tests, often developed by testing organizations, aim to reduce CIV. However, they typically contain a limited number of fixed forms, restricting opportunities for repeated test practice. Modern generative AI-powered automated item generation (henceforth, AIG) alleviates this constraint by enabling the creation of large item pools for digital practice tests. As a result, practice test generation can be scaled to support repeated practice test opportunities for test takers.

The Duolingo English Test (DET)is a digital, AI-driven, high-stakes, computer-adaptive ELP assessment used for international student university admissions. The DET is taken by hundreds of thousands of test takers each year.

To help test takers become familiar with the test, the DET offers a *free* practice test that simulates the official DET. As such, the practice test provides exposure to the DET task types, mirroring the official test in both appearance and administration order. It also provides an estimated score range, giving test takers a sense of how they are likely to perform on

*Authors are listed alphabetically to reflect equal contributions.

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 1: Full Papers, pages 50–57

the official test. Like the official DET, the practice test is also computer-adaptive, but drawing from a separate item pool than the official test. The large practice-test item pool, enabled by AIG, is used to dynamically generate versions of the practice test with different item sets, offering test takers repeated opportunities for practice (Naismith et al., 2025).[1][2]

The study presented in this paper examines how access to repeated test practice (i.e., the number of tests taken)—enabled by AIG— relates to test-takers' official DET scores, test-taker affect (e.g., confidence), and test-takers' decision to share their official DET scores for university admissions.

2 Background

Language assessment research has examined various aspects of test preparation, including test-taker preparation preferences (O'Sullivan et al., 2021), the relationship between preparation and affect (such as anxiety) (Chang and Read, 2008; Powers and Alderman, 1983; Winke and Lim, 2017), and the link between preparation and test performance (Green, 2007; Knoch et al., 2020; Liu, 2014; Powers, 1985; Xie, 2013). These studies suggest that test preparation can reduce anxiety (Chang and Read, 2008; Powers and Alderman, 1983), increase confidence (Powers and Alderman, 1983), and improve test scores (Green, 2007; Knoch et al., 2020; Xie, 2013). Knoch et al. (2020) investigated repeat test takers, showing how they changed their test preparation strategies over time to try to improve their test score. Xie (2013) demonstrated how test takers use test preparation to develop strategies for score improvement. Green (2007) examined the comparative impact of test preparation courses for a high-stakes language assessment. These three studies highlight *washback effect* with regard to test preparation, whereby a test influences language teaching and learning (Messick, 1996).

Automated item generation research related to assessment and instruction is extensive, but much predates modern generative AI. For example, Mitkov et al. (2006) showed that NLP-assisted item generation with human review can be more time-efficient than manual creation. Heilman and Smith (2010) proposed a framework for automatically generating and evaluating questions from text, demonstrating the feasibility of transforming declarative sentences into fact-based questions. Similarly, Madnani et al. (2016) discussed the Language Muse system, which used NLP to generate reading comprehension exercises for U.S. middle school texts for English learners. More recent research has shifted toward evaluating item quality and comparing system performance using large language models. For instance, Laverghetta Jr and Licato (2023) investigated GPT-4 for test item generation, demonstrating its potential to create psychometrically valid items.[3]

AIG is now integrated into the development of digital, high-stakes language assessments. Specific to this paper, the official DET and its practice test are dynamically assembled using AIG-created item banks with human review (Attali et al., 2022). After generating items with prompts used to fine-tune the AIG, human experts conduct a review. To ensure item quality and appropriateness, a multistage process for human review is implemented. This process begins with automated checks for linguistic accuracy and social appropriateness, followed by human expert review focused on copyediting, fact-checking, and identifying potential fairness and bias issues that could disadvantage certain test-taker groups (Church et al., 2025).

An internally-developed review platform is used to coordinate item reviews, track reviewer performance, and ensure inter-rater consistency. The final items are used to automatically create the DET practice and *official* DET tests.

As mentioned earlier, prior research about test preparation for high-stakes assessment has studied test-taker preferences, and established links between test preparation, test-taker affect, and performance outcomes. However, we are unaware of research examining how test takers' access to repeated practice tests—now enabled by AIG—relates to these factors. This likely stems from the limited scalability of conventional practice tests, which rely on human test developers who cannot generate test items at the same scale as AIG. He et al. (2024) conducted an extensive literature review, including 66 studies about research for second language test preparation. No themes emerged demonstrating research that examined technology

[1] The practice test items are created using the same AIG methods as the official DET.

[2] Successive versions of OpenAI's GPT models were used to develop the practice test, reflecting generative AI advances.

[3] Also see Flor (2025) for a comprehension discussion of automated item generation.

or AI to enhance test preparation.

3 The Study

This *observational study* examined how access to repeated practice test opportunities—enabled by AIG—related to test takers' official DET performance, test-taker affect, and test-taker decisions to share their official DET scores for university admissions. The study addressed the research question: What are the *observed* relationships between the number of practice tests taken and test-takers' official DET performance, test-taker affect, and test-taker score sharing decisions?

3.1 Methods

3.1.1 Survey instrument

To measure test takers' affect, we developed a brief survey instrument (henceforth, *survey*) that elicited perceptions of *achievement, confidence, motivation, preparedness*, and *anxiety* in relation to the official DET. The survey items reflect affective factors commonly used in prior research on assessment (e.g., Winke and Lim, 2017) and instructional contexts (e.g., Ling et al., 2021). We acknowledge that typical affective surveys include more items per construct. However, because the DET is an operational, high-stakes assessment, there are required constraints: we had to limit the number of post-test, *offboarding*[4] questions to avoid overburdening test takers. Consequently, the survey consisted of five items, each rated on a six-point Likert-style scale. The survey was presented to all test takers as shown in **Figure 1**.

3.1.2 Data Collection

The survey was administered during September 2023. Upon completion of the DET, test takers were presented with the survey during the DET offboarding process.

Of the original 32,599 test-taker participants (henceforth, test takers) who took the survey, responses were retained from 25,969 test-takers for the analysis. Responses were retained only for participants who: (1) responded to all survey items; (2) were taking the official DET for the first time[5]; (3) received an official DET score that was validated by human proctors; and, (4) had taken the practice tests within 60 days prior to taking the official DET.

[4]Offboarding takes place once the test is completed. Test takers are asked questions related to, e.g., demographics and their target score.

[5]Prior testing may have provided additional practice, complicating the analysis.

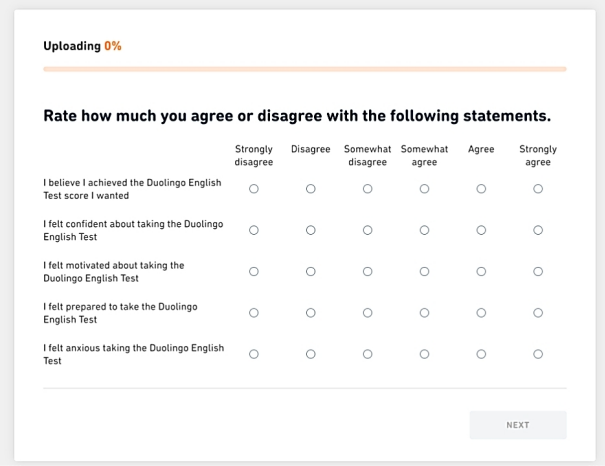

Figure 1: Post–DET Affective Perceptions Survey

	ACH	CON	MOT	PREP	ANX
ACH	1.00	0.74	0.59	0.67	0.04
CON	0.74	1.00	0.68	0.72	-0.03
MOT	0.59	0.68	1.00	0.64	0.07
PREP	0.67	0.72	0.64	1.00	0.06
ANX	0.04	-0.03	0.07	0.06	1.00

Table 1: Spearman Correlations Between Responses to Survey Items; ACH=Achieved; CON=Confident; MOT=Motivated; PREP=Prepared; ANX=Anxious

Table 1 shows the Spearman rank-order correlations between the survey items. The pairwise correlations between *I believed I achieved the DET score I wanted* (Achieved), *I felt confident about taking the DET* (Confident), *I felt motivated about taking the DET* (Motivated), and *I felt prepared to take the DET* (Prepared) are moderately high. This suggests that these positive affective statements may be related to a similar construct. By contrast, *I felt anxious taking the DET* (Anxious) is effectively uncorrelated with the other items.

3.1.3 Participant Demographics

Test taker demographic information is collected from test takers during the official DET's offboarding process. Offboarding items ask test takers about their *gender*, *age*, *testing intent* (i.e., obtaining an undergraduate or graduate degree), and *first language*.[6] **Table 2** shows the self-reported, test-taker demographics, also comparing the participant

[6]One hundred unique languages were reported by at least five participants.

Demographic	TTs (%)	DET(%)
Gender		
Female	44.0	47.6
Male	55.9	52.3
Age Group		
16–20 years	19.0	32.7
21–25 years	36.6	34.1
26–30 years	18.8	14.8
Testing Intent		
Undergraduate	43.0	47.1
Graduate	43.7	37.0
First Language		
English	13.7	9.5
Mandarin	10.8	17.8
Telugu	10.3	5.8
Spanish	8.8	10.0
Arabic	5.9	5.1

Table 2: Test-Taker Demographics; TTs=Test takers from this study; DET=DET population

sample to the DET test-taker population (Naismith et al., 2025). The sample includes all demographic subgroups from the DET population, though with some variation in proportions. This may be because the study included only first-time test takers, while the DET test-taker population includes both first-time and repeat test takers.

3.2 Analyses

This section discusses relationships that emerged between test takers' DET practice test engagement (i.e, *number of practice tests taken*), and their official DET scores, their affect (as self-reported in the survey), and their score-sharing decisions.[7]

Table 3 shows official DET scores by number of practice tests taken. Test takers were grouped into six bins (*count groups)* by number of practice tests completed (0, 1, 2–3, 4–6, 7+). We chose these categories to distinguish between 0, 1, and multiple practice test-taking sessions. Multiple practice test counts were grouped to balance the bin sample sizes.

Table 3 suggests a relationship between practice tests taken and official DET scores. For each practice test count group, we included 95% confidence intervals of the mean test score. The highest average scores were observed among those who took 1–3 practice tests (in **bold rows**). Confidence intervals of the mean test score for these rows did not overlap with those for 0, or 4 or more practice tests, showing significant differences. Those who took 0, or 4 or more practice tests scored slightly lower.[8]

# of PT	N	%	M	CI 95%
0	4,742	18.3	108.5	[107.8, 109.2]
1	**6,128**	**23.6**	**112.4**	**[111.8, 113.0]**
2–3	**6,469**	**24.9**	**112.3**	**[111.8, 112.8]**
4–6	4,142	16.0	111.1	[110.5, 111.7]
7+	4,488	17.3	108.6	[108.1, 109.1]
Total	25,969			

Table 3: Mean (M) Overall DET Score by Number of Practice Tests Taken (# of PT)

The finding that scores do not continue to increase with 4 or more practice tests aligns with expectations: practice tests are intended to build test familiarity, which on its own, should not facilitate large jumps in language proficiency.

Table 4 illustrates the relationship between number of practice tests taken, test-taker affect, and test takers' official DET score. As no clear differences emerged across the original Likert-scale categories (Figure 1), the six Likert-scale categories were collapsed into two. *Agree* contained: Strongly Agree, Agree, and Somewhat Agree. and *Disagree* contained: Strongly Disagree, Disagree, and Somewhat Disagree.

We included 95% confidence intervals of the difference between the mean scores for those who Agree and Disagree.[9] Rows in **bold** indicate that the confidence interval did not include 0, showing significant differences. Table 4 consistently shows that among test takers who took 0–3 practice tests, those who Agreed with positively-oriented items (Achieved, Confident, Motivated, Prepared) performed significantly better on the official DET than those who Disagreed. For those who Agreed they were Motivated and Prepared, better performance was also observed for 7+, and 4-6 and 7+ groupings, respectively.

Test takers who took 0 or 1 practice test showed a significant score difference between those who

[7]We used test takers' unique, official DET IDs to link to their practice test activity and score report sharing.

[8]Average scores across all groups hovered around the B2 CEFR level—a benchmark for independent language users and a common minimum for admission to English-medium universities (Council of Europe, 2020). However, it is important to note that where the test taker sits in the B2 CEFR range (lower vs. higher in the range) can impact their acceptance to a university.

[9]The Disagree mean score was subtracted from the Agree mean score.

Agreed and Disagreed across all positive statements. As well, test takers who practiced 2-3 times also showed significant differences between those who Agreed and Disagreed with the positive statements. This finding suggests that for some test takers, access to repeated test practice was related to positive affect and higher test scores.

Across the large proportion of test takers who indicated they felt Anxious (70.8%-75.3%), there was no signficant relationship found based on the number of practice tests taken. A possible explanation is the high-stakes nature of the DET. In recent work in classroom settings, Deho et al. (2025) found relationships between test anxiety and demographic factors. This is something that could be explored in future research.

Table 5 indicates a relationship between number of practice tests taken, likelihood of score sharing for university admissions, and test-taker affect.

We used 95% confidence intervals for the share rates (proportions) of those who Agreed or Disagreed with each of the statements. Rows in **bold** indicate that the corresponding Agree and Disagree confidence intervals did not overlap, which showed significant differences. Test takers who took 0, 1, or 2-3 practice tests and Agreed with the Achieved, Confident, and Prepared statements had non-overlapping confidence intervals with test takers who took 0, 1, or 2-3 practice tests and Disagreed with those statements. For those who Agreed with the Motivated statement, only those who took 2-3 practice tests had share rate confidence intervals that did not overlap with the corresponding confidence intervals with those who Disagreed. Note that test takers were always more likely to share their scores if they Agreed with positive statements.

As expected, further analysis showed that test takers who shared their scores tended to have higher mean scores. Scores typically aligned with a mid- to high B2 CEFR level. This is an expected outcome, as test takers are more likely to share scores that meet university requirements. Scores were highest among those who took 0–3 practice tests and Agreed with positive sentiment statements. For example, those who Agreed with the Achieved category had mean scores of 119.1, 120.9, and 119.0 for 0, 1, and 2–3 tests taken, respectively. This trend held across all positive sentiment categories. Scores declined slightly for those who took 4–6 tests (about 1 point lower) and more noticeably for those with 7+ tests (about 3 points lower). A similar pattern emerged for the Anxious category.

4 Discussion

Integrated into the DET pipeline, AIG generates large item pools. This scales the creation of DET practice tests, which increases test takers' access to repeated practice opportunities. To our knowledge, this is the first study to examine how AIG can contribute to increased practice opportunities and how, in turn, access to more practice is related to test-taker affect and outcomes. The study explored relationships between (1) practice test engagement and test score. (Table 3), (2) test-taker affect and official DET scores (Table 4), and (3) affect and score-sharing decisions for university admissions (Table 5).

Three key findings emerged from the analysis to address our research question: What are the *observed* relationships between the number of practice tests taken, and official DET performance, test-taker affect, and score-report sharing decisions?

First, repeated test practice was related to higher test scores to an extent. (Table 3). Those who took 1, or 2-3 practice tests had comparatively higher scores than those who took 0, or more than 3. As taking 2-3 practice tests was related to higher test scores, this suggests a potential benefit of access to repeated practice for some test takers. These test takers may have come to the practice test with higher proficiency and were using the practice test for its intended purpose—i.e., test familiarity.

By contrast, taking more than 2-3 practice tests was associated with lower performance. This may be related to washback effect (mentioned earlier). Specifically, test takers may have used the practice test for reasons beyond test familiarity, such as building English language skills (i.e., positive washback that supports language learning), or test-taking strategies, such as trying to *game* the test (i.e., negative washback that does not support language learning) (Knoch et al., 2020; Xie, 2013). In this scenario, test takers' repeated practice testing may be an example of *wheel spinning*, where learners repeated attempts to master a skill are unsuccessful (Beck and Gong, 2013; Mu et al., 2020).

Second, test takers who took more practice tests reported feeling more positively (Table 4). Based on the number of practice tests taken, higher proportions of test takers reported positive affect toward the official DET regarding their beliefs that

#	Agree		Disagree		CI 95%
	%	M	%	M	
Achieved					
0	**85.7**	**109.5**	**14.3**	**102.2**	**[5.1, 9.5]**
1	**82.8**	**113.8**	**17.2**	**105.6**	**[6.6, 9.9]**
2-3	**84.3**	**112.8**	**15.7**	**109.1**	**[2.3, 5.3]**
4-6	87.4	111.3	12.6	109.7	[-0.3, 3.5]
7+	91.0	108.6	9.0	108.3	[-1.6, 2.2]
Confident					
0	**85.4**	**109.9**	**14.6**	**100.5**	**[7.2, 11.6]**
1	**82.8**	**114.0**	**17.2**	**104.6**	**[7.8, 11.1]**
2-3	**84.4**	**113.2**	**15.6**	**107.4**	**[4.2, 7.2]**
4-6	86.6	111.3	13.4	109.6	[-0.2, 3.6]
7+	91.4	108.7	8.6	108.3	[-1.6, 2.4]
Motivated					
0	**90.9**	**109.1**	**9.1**	**102.1**	**[4.1, 9.9]**
1	**89.8**	**113.0**	**10.2**	**107.5**	**[3.2, 7.7]**
2-3	**91.7**	**112.6**	**8.3**	**108.8**	**[1.7, 5.8]**
4-6	93.0	111.2	7.0	110.1	[-1.6, 3.7]
7+	**95.2**	**108.8**	**4.8**	**105.8**	**[0.4, 5.5]**
Prepared					
0	**85.3**	**110.0**	**14.7**	**99.9**	**[7.8, 12.2]**
1	**82.5**	**114.3**	**17.5**	**103.6**	**[9.1, 12.3]**
2-3	**85.4**	**113.3**	**14.6**	**106.2**	**[5.5, 8.6]**
4-6	**88.3**	**111.6**	**11.7**	**107.3**	**[2.3, 6.2]**
7+	**92.7**	**108.9**	**7.3**	**105.4**	**[1.4, 5.5]**
Anxious					
0	70.8	107.6	29.2	110.6	[-4.6, -1.6]
1	72.9	112.2	27.1	113.1	[-2.2, 0.4]
2-3	73.9	112.3	26.1	112.1	[-0.9, 1.4]
4-6	75.3	111.2	24.7	110.6	[-0.7, 1.9]
7+	75.0	108.5	25.0	108.9	[-1.5, 0.8]

Table 4: Mean (M) Overall DET Score by Practice Tests Taken (#) and Affective Perceptions

#	Agree		Disagree	
	%	CI 95%	%	CI 95%
Achieved				
0	**41.9**	**[40.3, 43.4]**	**32.2**	**[28.6, 35.7]**
1	**43.5**	**[42.1, 44.8]**	**31.9**	**[29.1, 34.7]**
2-3	**43.4**	**[42.1, 44.7]**	**33.6**	**[30.7, 36.5]**
4-6	42.2	[40.6, 43.8]	38.0	[33.9, 42.2]
7+	44.3	[42.8, 45.8]	40.2	[35.5, 45.0]
Confident				
0	**42.2**	**[40.7, 43.7]**	**30.3**	**[26.8, 33.7]**
1	**43.6**	**[42.2, 44.9]**	**31.4**	**[28.6, 34.2]**
2-3	**43.5**	**[42.2, 44.8]**	**32.8**	**[29.9, 35.7]**
4-6	42.1	[40.5, 43.7]	38.8	[34.7, 42.8]
7+	44.0	[42.5, 45.5]	43.3	[38.4, 48.2]
Motivated				
0	41.0	[39.5, 42.4]	35.5	[31.0, 40.0]
1	41.9	[40.6, 43.2]	37.9	[34.1, 41.8]
2-3	**42.5**	**[41.2, 43.8]**	**34.7**	**[30.7, 38.7]**
4-6	41.7	[40.1, 43.2]	41.2	[35.6, 46.9]
7+	44.2	[42.7, 45.6]	39.6	[33.1, 46.1]
Prepared				
0	**42.1**	**[40.6, 43.6]**	**31.1**	**[27.7, 34.5]**
1	**43.5**	**[42.1, 44.9]**	**32.0**	**[29.2, 34.8]**
2-3	**43.5**	**[42.2, 44.8]**	**32.3**	**[29.3, 35.3]**
4-6	42.1	[40.5, 43.7]	38.0	[33.7, 42.3]
7+	44.3	[42.8, 45.8]	38.9	[33.6, 44.2]
Anxious				
0	39.7	[38.0, 41.3]	42.4	[39.8, 45.0]
1	41.1	[39.6, 42.5]	42.6	[40.3, 45.0]
2-3	42.0	[40.6, 43.4]	41.4	[39.1, 43.8]
4-6	41.4	[39.6, 43.1]	42.5	[39.5, 45.6]
7+	43.2	[41.5, 44.8]	46.3	[43.4, 49.2]

Table 5: Proportion of Test Takers who Shared Their DET Score by Number of Practice Tests Taken and Affective Perceptions

they achieved the score they wanted, and their confidence, motivation, and preparedness. As such, the 7+ group consistently had the highest proportion of test takers reporting positive affect. Reported feelings of anxiety were similar across the number of practice tests taken (Table 4). While not surprising in a high-stakes context, the finding is novel compared to prior work suggesting that test preparation could reduce anxiety (Chang and Read, 2008; Powers and Alderman, 1983; Winke and Lim, 2017). However, previous work was conducted in no- or low-stakes experimental settings.

Regarding DET performance, test takers who agreed with the positive statements had higher official DET scores, on average, than those who disagreed; this finding was significant (Table 4). Those who took 1-3 practice tests had the highest scores, on average. Test scores trended lower after taking more than 3 practice tests.

Third, test takers who reported positive perceptions were more likely to share their official DET score report for university admissions (Table 5). This finding was consistent across the number of practice tests taken with comparatively higher proportions for those who Agreed than Disagreed with the positive survey items. Share rates were significantly higher for those who took 0-3 practice tests and Agreed with the Achieved, Confident and Prepared statements, and for those who

took 2-3 practice tests and Agreed with the Motivation statement, as compared to those who Disagreed. Like other outcomes we investigated, Anxiety did not show significant differences in share rates by agreement status.

5 Limitations

This section notes *two* study limitations.

First, as an observational study, our findings are **not causal**. Independent of practice test use, higher English proficiency may underlie positive perceptions, higher scores, and share rates.

Second, the number of survey items was necessarily limited to reduce the burden test takers after taking a high-stakes test. Given this real-world constraint, we prioritized items related to test-taker affect, and did not include an item eliciting information about alternative test strategies. As a result, we lacked data on test takers' use of alternative preparation methods. Related, we do not have information about what motivated test takers' repeated practice. As we continue with this research, we are exploring ways to address this limitation.

6 Conclusions

The DET's practice test simulates the official DET. As a computer-adaptive test, the practice test aims to familiarize test takers with the official DET's item types, its adaptive administration, and the official DET score scale (by providing an estimated test score range). Integrating AIG into the test development pipeline enables scalable production of DET practice tests. This facilitates the creation of multiple practice test versions, offering test takers repeated opportunities to build test familiarity.

The study analysis showed that test takers who took 1-3 practice tests tended to have higher official DET scores. Higher test scores were also related to positive affect (i.e., agreeing with the positive survey items). Higher share rates were also linked to positive affect. This *may* be related to those test takers having higher underlying English proficiency. Therefore, test takers may have used the practice test for its intended purpose—test familiarization, whereby 1-3 practice test repetitions may have been sufficient. This also suggests that for some test takers—those who took 2-3 practice tests— that *limited* repeated practice may have provided extra needed support to sufficiently build their test familiarity.

By contrast, test takers who took more than 3 practice tests had lower performance, on average. It is possible that these test takers may have come to the test with lower proficiency. Their additional test practice may be related to washback, whereby test takers used the practice test for reasons besides building test familiarity (e.g., English language learning or building test-taking strategies). However, we lack data about test takers' preparation strategies, beyond the DET practice test, as well as test-taker goals for taking the practice test. Therefore, this limits interpretation. At the same time, it raises interesting questions with regard to appropriate guidance about test preparation, especially with regard to mitigating negative washback effects, such as using test practice to develop test gaming strategies.

AIG for high-stakes assessment is still in its early stages. The study examines how repeated practice—enabled by AIG—may relate to test-taker performance, affect, and behavioral outcomes (i.e., score sharing). It also raises important questions about test preparation practices when test-takers have access to repeated test practice. Our findings—and future research—could be useful in helping to inform best practices for AI-enhanced test readiness in high-stakes contexts.

Acknowledgements

The authors would like to thank our Duolingo colleagues for helpful reviews of earlier versions of this paper: Audrey Kittredge, Nitin Madnani, Ben Naismith, and Alina von Davier. Thanks to the anonymous reviewers for their feedback.

References

Yigal Attali, Andrew Runge, Geoffrey T LaFlair, Kevin Yancey, Sarah Goodwin, Yena Park, and Alina A Von Davier. 2022. The interactive reading task: Transformer-based automatic item generation. *Frontiers in Artificial Intelligence*, 5:1–13.

J. E. Beck and Y. Gong. 2013. Wheel-spinning: Students who fail to master a skill. In *Artificial Intelligence in Education: 16th International Conference, AIED 2013*, pages 431–440. Springer.

A. C. S. Chang and J. Read. 2008. Reducing listening test anxiety through various forms of listening support. *TESL-EJ*, 12(1). N1.

Jacqueline Church, Yena Park, and Jill Burstein. 2025. Guidelines for fair test content: The Duolingo English Test example. Duolingo Research Report DRR-25-02, Duolingo. 19 pages.

Council of Europe. 2020. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment – Companion Volume*. Council of Europe Publishing.

Oscar Blessed Deho, Srecko Joksimovic, Maria Vieira, and Ryan Baker. 2025. Beyond predictive accuracy: Fairness and bias in predicting test anxiety. In *Proceedings of the International Conference on Artificial Intelligence and Education*.

Michael Flor. 2025. Question generation with large language models and generative ai. In *Automatic Question Generation*, pages 137–147. Springer.

Anthony Green. 2007. Washback to learning outcomes: A comparative study of ielts preparation and university pre-sessional language courses. *Assessment in Education*, 14(1):75–97.

Shanshan He, Anne-Marie Sénécal, Laura Stansfield, and Ruslan Suvorov. 2024. A scoping review of research on second language test preparation. *Language Testing*, 42(1):11–47.

Michael Heilman and Noah A. Smith. 2010. Question generation via overgenerating transformations and ranking. Technical Report CMU-LTI-10-008, Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA.

Ute Knoch, Annemiek Huisman, Cathie Elder, Xiaoxiao Kong, and Angela McKenna. 2020. Drawing on repeat test takers to study test preparation practices and their links to score gains. *Language Testing*, 37(4):550–572.

Kadidja Koné, Paula Winke, and Matthew Gordon. 2024. "we would like to see ourselves in the test:" the experiences of francophone african english learners in high-stakes english proficiency testing.

Antonio Laverghetta Jr and John Licato. 2023. Generating better items for cognitive assessments using large language models. In *Proceedings of the 18th workshop on innovative use of NLP for building educational applications (BEA 2023)*, pages 414–428.

Guangming Ling, Norbert Elliot, Jill C Burstein, Daniel F McCaffrey, Charles A MacArthur, and Steven Holtzman. 2021. Writing motivation: A validation study of self-judgment and performance. *Assessing Writing*, 48:100509.

O. L. Liu. 2014. Investigating the relationship between test preparation and toefl ibt performance. *ETS Research Report Series*, (2):1–13.

Nitin Madnani, Jill Burstein, John Sabatini, Kristy Biggers, and Slava Andreyev. 2016. Language muse™: Automated linguistic activity generation for english language learners. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, Berlin, Germany.

S. Messick. 1982. Issues of effectiveness and equity in the coaching controversy: Implications for educational and testing practice. *Educational Psychologist*, 17:67–91.

Samuel Messick. 1996. Validity and washback in language testing. *Language testing*, 13(3):241–256.

Ruslan Mitkov, Le An Ha, and Nikiforos Karamanis. 2006. A computer-aided environment for generating multiple-choice test items. *Natural Language Engineering*, 12(2):177–194.

Tong Mu, Andrea Jetten, and Emma Brunskill. 2020. Towards suggesting actionable interventions for wheel-spinning students. *International Educational Data Mining Society*.

B. Naismith, R. Cardwell, G. LaFlair, S. Nydick, and M. Kostromitina. 2025. Duolingo English Test: Technical manual. Duolingo research report, Duolingo.

B. O'Sullivan, K. Dunn, and V. Berry. 2021. Test preparation: An international comparison of test takers' preferences. *Assessment in Education: Principles, Policy & Practice*, 28(1):13–36.

D. E. Powers. 1985. Effects of test preparation on the validity of a graduate admissions test. *Applied Psychological Measurement*, 9(2):179–190.

D. E. Powers and D. L. Alderman. 1983. Effects of test familiarization on sat performance. *Journal of Educational Measurement*, 20(1):71–79.

P. Winke and H. Lim. 2017. The effects of test preparation on second-language listening test performance. *Language Assessment Quarterly*, 14(4):380–397.

Q. Xie. 2013. Does test preparation work? implications for score validity. *Language Assessment Quarterly*, 10(2):196–218.