Performance of High School Students in Multiple-Choice Questions in English

Abdulrahman bin Mustafa Olwi[(1)*]

(1) Associate Professor, Taibah University, Medina, Saudi Arabia.

Received: 01/11/2024
Accepted: 24/03/2025
Published: 30/06/2024

* ***Corresponding Author:***
aolwi@taibahu.edu.sa

DOI:https://doi.org/10.59759/educational.v4i2.740

Abstract

This study investigated how 301 high school students performed differently on English language tests based on their gender and grade. Descriptive statistics were used to summarize performance trends, and a Multivariate Analysis of Variance (MANOVA) was conducted to examine the effects of grade and gender. The results revealed significant grade-level effects, with grade 12 students' outperforming their peers in grades 10 and 11 across all metrics. Modest gender differences were observed, with female students' achieving higher scores in writing skills. However, no significant interaction effects between grade level and gender were found. The findings encourage schools to integrate test preparation into the curriculum earlier, particularly for students in grades 10 and 11. Future research should explore longitudinal trends, socioeconomic influences, and alternative assessment formats to provide a deep understanding of students' performance in multiple-choice questions.

أداء طلاب المرحلة الثانوية في أسئلة الاختيار من متعدد في اللغة الإنجليزية

عبد الرحمن بن مصطفى علوي[١]

(١) أستاذ مشارك، جامعة طيبة، المدينة المنورة، المملكة العربية السعودية.

ملخص

بحثت هذه الدراسة أداء ٣٠١ طالبا وطالبة في المرحلة الثانوية في اختبارات اللغة الإنجليزية بناءً على جنسهم وصفهم الدراسي. استُخدمت الإحصاءات الوصفية لتلخيص اتجاهات الأداء، وأُجري تحليل التباين المتعدد المتغيرات (مانوفا) لفحص آثار الصف والجنس. كشفت النتائج عن تأثيرات كبيرة على مستوى الصف، حيث تفوق طلاب الصف الثاني عشر على أقرانهم في الصفين العاشر والحادي عشر في جميع المقاييس. لوحظت فروق ليست ذات دلالة إحصائية بين الجنسين، حيث حققت الطالبات درجات أعلى في مهارات الكتابة. ومع ذلك، لم يتم العثور على تأثيرات تفاعلية كبيرة بين مستوى الصف والجنس. تشجع النتائج المدارس على دمج التحضير للاختبارات في المناهج الدراسية في وقت مبكر، وخاصة للطلاب في الصفين العاشر والحادي عشر. كذلك ينبغي للبحوث المستقبلية أن تستكشف الاتجاهات الطولية والتأثيرات الاجتماعية والاقتصادية والتقييمات البديلة لتوفير فهم عميق لأداء الطلاب في أسئلة الاختيار من متعدد.

Introduction

Standardized tests are essential tools for assessing the academic readiness of high school students. These tests evaluate a wide range of skills, including reading comprehension, writing proficiency, and critical thinking, and they are seen as key predictors of success in higher education. Performance on these tests often varies based on demographic factors, including grade level and gender. Investigating these variations is important to enhance educational practices and address performance disparities.

Gender differences in standardized testing outcomes have also been widely documented, with research consistently showing that female students tend to excel in language-focused tasks whereas male students often perform better in quantitative domains (Reardon et al., 2018; Taylor & Lee, 2021). These differences could stem from several factors, including cognitive development and sociocultural influences. For instance, female students generally outperform male students on open-ended

or writing-intensive questions. This aligns with their stronger verbal reasoning and communication skills. On the other hand, multiple-choice formats often favour male students (Williams & Richards, 2020; Kwon & Park, 2020). Although these are established patterns, the extent to which gender differences interact with grade level remains relatively unexamined.

Academic progression plays a significant role in shaping students' performance on standardized tests. Students in higher grades benefit from increased exposure to curriculum content, enhanced cognitive development, and improved test-taking strategies. These factors contribute to the consistently superior performance observed among students in Grade 12 compared to their peers in Grades 10 and 11. For example, research has demonstrated that cumulative academic experiences significantly impact test outcomes, with structured learning environments providing older students with a competitive edge (Matheny et al., 2022). However, the relationship between grade level and performance in standardized testing, especially in critical sub-skills like reading and writing, remains underexplored, which brings a gap that this study aims to address.

This study investigated performance across high school grade levels 10, 11, and 12 and examined how gender influences performance within these grades. The primary aim of the study was to better understand how demographic factors, specifically grade level and gender, contribute to variations in standardized test outcomes.

In the context of this study, several key terms are defined to provide clarity. Performance refers to the scores derived from standardized tests designed to assess high school students' readiness for college within the context of English language learning. Grade level is defined as academic progression categorized into Grades 10, 11, and 12. Last, gender represents the demographic variable examined in terms of male and female performance differences.

While this study aims to provide valuable insights, it is important to acknowledge potential limitations. The reliance on multiple-choice assessments, although effective for evaluating foundational skills, might not fully capture higher-order cognitive

abilities such as critical thinking (Schuwirth & Van Der Vleuten, 2020). Also, disparities in sample size across grade levels, particularly the smaller representation of Grade 10 students, could affect the generalizability of the findings.

Multiple-choice questions are a usual assessment instrument, but their effectiveness can vary depending on several factors. While multiple-choice questions present many advantages, they may not be the best tool for all assessment purposes. In education, the use of multiple-choice questions is a complex concern with continuing debate and research being conducted on the subject. Nevertheless, it is critical to use different assessment tools to get a comprehensive understanding of student learning. Since multiple-choice questions have often been used in standardized testing like language proficiency exams, research studies have been conducted to compare the usefulness of multiple-choice questions to other assessment tools like open-ended questions.

Still, multiple-choice questions, called objective questions, are often criticized. However, many universities and schools around the world use multiple-choice questions extensively as a basic measurement tool, and item format and its effect on assessment quality have been a topic for researchers for years. Moreover, nearly all international language tests, such as the TOEFL and IELTS exams, use multiple-choice questions in their separate sections that assess grammar, reading, and listening most objectively. Indeed, multiple-choice questions are considered the most dependable, reasonable, and affordable type of questioning, especially in mass testing that deals with hundreds of thousands of students taking the same test at the same time (Rauch & Hartig, 2010).

Klufa (2015) found that multiple-choice questions have several advantages in testing: they easily provide scoring reliability in crowded groups; they accommodate large item numbers; they cover much more critical content in the subject area; and they increase high content validity, and so on. The research showed significant differences between multiple-choice questions and open-ended questions in terms of the levels of item difficulty and item discrimination in both grammar and reading testing. Moreover, it was found that multiple-choice questions were easier than open-ended questions.

Purpose and Question of the Research Study

The purpose of this research study was to check high school students' performance in English language multiple-choice questions and the difference between Grade 12 students and other students who have not yet reached Grade 12. It was noticed that high school students need to start preparing themselves for their university exams earlier than Grade 12, but, in fact, they do not start preparing themselves for university until they actually reach Grade 12. The impact of being a Grade 12 student only a few months away from entering college and the importance of the readiness of undergraduate students needs to be examined as a result. Furthermore, there is a need to look at what other pre-Grade 12 high school students are doing concerning their preparation for university entrance exams. After all, these exams are usually multiple-choice questions, especially in the section that measures English language skills.

Based on the purpose of the research study, the following research question was addressed: Is there a significant difference between Grade 12 and pre-Grade 12 students in their performance regarding English language multiple-choice questions?

Literature Review

Testing has been a research topic in the educational field since time immemorial, and multiple-choice questions and open-ended questions are among the testing topics that have been investigated within the language assessment context, mainly because these two types of questions measure different language aspects. More recently, for example, Ozuru et al. (2007) reported moderate to high correlations between performance on multiple-choice questions and open-ended questions of the same questions when administered as memory-based questions. They additionally found a limitation that both multiple-choice questions and open-ended questions were moderately correlated with prior knowledge.

Additionally, the difficulty of each test type has also been examined. It was found that multiple-choice questions were easier than open-ended questions.

Educators use multiple-choice questions because they are more reliable, cost-effective, and time-saving. Furthermore, their scoring is quite objective. Multiple-choice questions allow for an accurate clarification of content validity, the form of student learning outcomes, the level of difficulty, and the reliability of the questions. Although conducting multiple-choice questions is quite simple, these measures bring to the surface information on certain specific skills, whereas other skills like critical thinking and synthesis might not be evaluated through multiple-choice questions. In contrast, open-ended questions reduce errors in assessment by eliminating chance success. They are suitable for partial scoring, and preparation is easier than multiple-choice questions (Allan & Driscoll 2014).

However, open-ended questions take a much longer time to implement and score. Moreover, they can make it difficult for the evaluator to provide content validity. Open-ended questions also usually appear less frequently than multiple-choice questions on exams simply because of the time constraint. Most importantly, their scoring is more objective (Palmer & Devitt, 2007). High-quality skills like generating hypotheses, evaluating ideas, establishing cause-and-effect relationships, organizing problems, problem-solving, generating original ideas, applying information in different situations, generalizing, and making comparisons between options are better assessed by open-ended questions (Kwon et al., 2006).

Both multiple-choice and open-ended format questions are two of the most prevalent types of questions measuring language skills and text comprehension in educational settings. However, some research studies have suggested that these two forms of questions assess comprehension differently. Nonetheless, such a difference is not well understood. Although the processes in using the two forms of questions are very different, the distinction might not be as clear as it seems due to many affecting factors (Graesser et al., 2010). Furthermore, when talking about reading comprehension measurement, short answer questions should be more effective than multiple-choice questions.

Shohamy (1984) found that, on reading comprehension tests, multiple-choice questions were easier to answer than open-ended questions, especially among those students whose reading proficiency is low. Cahill and Leonard (1999) found out skilled students perform well just because of their increased ability to analyze the test

itself after getting over-tested by multiple-choice questions. Martinez (1999) also underlined the same problem, believing that multiple-choice questions are weak at eliciting upper-level cognition. Other educators found that the simplest testing forms are multiple-choice questions, as they rely on memorization and thinking, which are necessary but insufficient, and enhancing the power of expression and interpretation is still needed.

In fact, multiple-choice questions are not conducive to students' abilities to express their ideas as they wish. Instead, when responding, students answer depending on the given options. Thus, it is pointed out that multiple-choice questions are only capable of examining the basics of information about facts (Walsh & Seldomridge, 2006). However, well-constructed multiple-choice questions could yield scores as reliable as those given by open-ended questions (Bacon, 2003). Moreover, thoughtfully written multiple-choice questions could serve to assess higher-level cognitive processes, although creating such items requires more skill (Palmer & Devitt, 2007). Even more, multiple-choice questions could accurately discriminate between those students whose achievement is high and those whose achievement is low (Schuwirth & Van Der Vleuten, 2003).

Multiple-choice questions could be different from yes–no questions. According to research on memory, it is unclear if yes–no and multiple-choice questions are different in terms of the recollection process and the familiarity that contributes during the process of answering. On testing reading comprehension, for example, research studies showed a significant difference between open-ended questions and multiple-choice questions in terms of difficulties and the level of discrimination. Both types of questions differ in their strengths and weaknesses, none of which could be considered perfect for all language testing purposes (Cheryl et al., 2017).

Like multiple-choice questions, open-ended questions also have advantages and disadvantages. Magliano et al. (2007) stated that the most basic advantage of open-ended questions is that they provide detailed responses from students, thereby increasing their validity. Open-ended questions work better in assessing productive language skills of speaking or writing because these two skills require

higher-order thinking. Furthermore, open-ended questions include higher levels of cognitive processing, such as analysis and synthesis, with some disadvantages such as being less reliable and more time-consuming. Supporters of open-ended questions in language testing claim that open-ended questions allow for the measuring of high levels of foreign language skills. Nonetheless, students find open-ended questions more difficult than multiple-choice questions (Oksuz & Guven Demir, 2019). This might be because open-ended questions require remembering information and expressing ideas in both written and spoken forms. The factors influencing this idea could be the lack of options for answers and the absence of luck-based success (Turgut & Baykul, 2012).

Open-ended questions cannot be answered by a simple "yes" or "no" response, so they are phrased as statements that require written or spoken answers. These open-ended questions eliminate the possibility of guessing because no answer choices are given with the questions. Useful insights into the theoretical understanding levels are required. If carefully crafted, open-ended questions could reflect students' generating abilities rather than memorizing skills. Not only do they require complex thinking from students, but they also require teachers and evaluators to use multiple criteria during evaluating responses. They require full answers based on students' knowledge or ideas (Cooney et al., 2004).

Lee et al. (2011) supported the same fact that open-ended questions do not have the chance factor that exists in multiple-choice questions and the possibility of guessing the right answer through using the process of elimination. In addition to their superiority in measuring higher skills, open-ended questions make students provide answers that are more useful in determining and diagnosing the quality of teaching (Cooney et al., 2004). If including open-ended questions, the assessment process focuses on perception and justification. Also, it measures the ability to utilize information, allowing students to reflect on their differences (Melovitz Vasan et al., 2017).

Black et al. (2003) said the attempts to test pre-planned educational outcomes using more valid and reliable methods have led to the development of various test techniques in foreign language assessment. While these assessment tools in education differ according to cognitive and affective factors, they are also categorized according to their test preparation techniques, such as multiple

choice, open-ended, matching, true/false, and completion items. Some studies have shown that questions in multiple-choice tests can be answered without fully understanding the passage, and therefore, it would not be healthy to measure the level of reading comprehension with multiple-choice questions. On the other hand, multiple-choice questions have been asserted by some researchers to be reliable and effective in assessing the level of understanding of students when used with other appropriately prepared questions.

When multiple-choice questions, open-ended questions, and summary questions were used, it was found that in reading comprehension there was no significant difference among these three types of questions that had the same trait of the subjects. However, she still found that multiple-choice questions were the easiest and summary questions were the most difficult. Also, in reading comprehension, using multiple-choice questions, true-false questions, and short-answer questions, it was found that the type of questions influenced students' performance. It was found that students with high proficiency levels were more easily affected than students with low proficiency levels. Significant differences among the scores were found in the three different types of questions. Finally, research conducted on different types of questions revealed that different measurement tools, such as true-false questions, matching questions, homework assignments, projects, portfolios, graduation theses, and so on, are preferred for better reliability (Karadağ, 2014).

Methodology

The design adopted in this research study was a comparative study, which involves examining both Grade 12 and pre-Grade 12 students to identify similarities and differences. To minimize bias, participants were randomly assigned to collect the data for the research study. No intervention was done during this research study. Rather, a dataset was gathered and analyzed.

Participants

A dataset collected from different international high schools was used as a sample for this research study. The dataset consisted of a total of 301 high school students preparing themselves for university entrance exams. As shown in Table 1, 51.2% of the participants were female, whereas 48.8% were male. Most participants were Grade 11 students at 59.8%, followed by Grade 12 students, who made up 32.9% of the participants, and Grade 10 students, who made up 7.3% of the participants.

Table 1. Participants

Characteristic	N
Gender	
Female	154 (51.2%)
Male	147 (48.8%)
Educational Grade	
Pre-G12	202 (67.1%)
G12	99 (32.9%)

Results

This study employed a retrospective observational design to examine the performance of high school students, focusing on differences across three grade levels of high school. The primary objective was to analyze how grade and gender influence performance in multiple-choice assessments designed to evaluate math-related language skills.

The dataset comprised 301 students from various international high schools, including 147 male students (48.8%) and 154 female students (51.2%). Most participants were in Grade 11 (59.8%), followed by Grade 12 (32.9%), and finally, Grade 10 (7.3%). Students were included in the study if they had completed a standardized assessment and had complete demographic and performance data. Participants with missing demographic information or performance scores were excluded to ensure the accuracy and reliability of the analysis.

The assessment tool used in this study was modeled after standardized university entrance exams and aimed to evaluate language skills through multiple-choice questions. The test covered seven distinct domains, including general reading, evidence search, contextual usage, sharing perspectives, norms of structure, facts and ideas, and design and organization. The tool's design aligned with established principles of standardized testing, ensuring its validity and reliability.

Data for this study were collected from anonymized academic records. The dataset was cleaned and prepared to ensure its integrity and suitability for analysis. Any records with missing demographic data were excluded, while missing performance scores were replaced using column-wise mean values. This imputation approach minimized data loss while maintaining the overall structure and reliability of the dataset.

Statistical analysis was conducted using SPSS (version 27) to evaluate the relationships between grade, gender, and performance. Multivariate Analysis of Variance (MANOVA) was employed to examine the collective and individual effects of grade and gender on multiple performance metrics, including Total Score, Reading Test Score, and Writing Test Score. Post hoc comparisons were performed to identify specific grade or gender differences. Descriptive statistics, including group-level means and standard deviations, were calculated to summarize performance trends across grades and genders. To ensure the robustness of the analysis, assumptions of homogeneity of variances and multivariate normality were tested prior to conducting MANOVA.

All data used in this study were anonymized to protect participant privacy. Ethical guidelines for the use of secondary academic records were followed, and no direct interventions or interactions with participants were conducted.

Performance metrics across grades and genders revealed clear trends that underscore the progression of academic preparation. Grade 12 students consistently achieved higher scores across all evaluated metrics compared to their counterparts in Grades 10 and 11. The overall Total Score (ranging from 400 to 1600) increased

progressively with grade level, with Grade 12 students scoring an average of 1283 (SD ± 155), compared to 1208 (SD ± 178) in Grade 11 and 1138 (SD ± 192) in Grade 10.

A similar trend was observed in the Reading Test Score (ranging from 10 to 40). Grade 12 students achieved an average score of 30.3 (SD ± 3.2), reflecting a notable improvement over the scores in Grade 11 (27.8 ± 3.9) and Grade 10 (24.9 ± 3.4). For the Writing Test Score (also ranging from 10 to 40), Grade 12 students averaged 28.7 (SD ± 3.4), significantly outperforming Grade 11 (26.8 ± 4.0) and Grade 10 (24.4 ± 3.7).

Gender differences, though smaller in magnitude compared to grade-level differences, were evident in Writing Test Scores, where female students consistently outperformed male students across all grades. For example, in Grade 12, female students achieved a Writing Test Score average of 29.0 (SD ± 3.2) compared to 28.5 (SD ± 3.7) for male students. However, these differences were less pronounced in other metrics, such as the Total Score and Reading Test Score, where performance was relatively balanced between genders.

Table 2. Performance by Grade and Gender (Mean ± SD)

Metric	Grade 10	Grade 11	Grade 12
Total Score (400–1600)	1138 ± 192	1208 ± 178	1283 ± 155
Reading Test Score (10–40)	24.9 ± 3.4	27.8 ± 3.9	30.3 ± 3.2
Writing Test Score (10–40)	24.4 ± 3.7	26.8 ± 4.0	28.7 ± 3.4

The results indicate that students in higher grades consistently outperformed their peers in lower grades across all performance metrics. Furthermore, although the differences were modest, female students achieved slightly higher scores than male students in Writing Test Scores.

Multivariate Analysis

Multivariate Analysis of Variance (MANOVA) demonstrated significant effects for both grade and gender on test performance metrics. The effect of grade level was particularly strong ($F(2,298) = 55.4$, $p < 0.001$, $\eta 2 = 0.25 F(2, 298) = 55.4$,

$p < 0.001$, \eta^2 = 0.25F (2,298) =55.4, $p<0.001$, $\eta2=0.25$), accounting for 25% of the variance in performance. Gender had a smaller but statistically significant effect ($F(1,298)=6.3$, $p=0.012$,$\eta2=0.03F(1, 298) = 6.3$, $p = 0.012$, \eta^2 = 0.03$F(1,298)=6.3$, $p=0.012$, $\eta2=0.03$).

Table 3. MANOVA Results

Variable	Wilks' Lambda	F-value	p-value	Effect Size (η2\eta^2η2)
Grade	0.63	55.4	<0.001	0.25
Gender	0.97	6.3	0.012	0.03

Interaction effects between grade and gender were not statistically significant ($F(2,298)=1.2$,$p=0.31F(2, 298) = 1.2$, $p = 0.31F(2,298)=1.2$,$p=0.31$), indicating that performance trends were consistent for both genders across grades.

Post Hoc Comparisons

Post hoc analyses revealed that Grade 12 students significantly outperformed their peers in Grades 10 and 11 across all metrics ($p<0.001p < 0.001p<0.001$). The difference between Grades 10 and 11 was also statistically significant ($p<0.05p < 0.05p<0.05$), but less pronounced. Gender differences were significant only for Writing Test Scores ($p=0.03p = 0.03p=0.03$), where female students achieved slightly higher scores than male students.

Discussion

This study examined differences in standardized test performance across high school grade levels and genders. The findings revealed significant grade-level effects, with higher grades demonstrating better performance and modest gender differences, particularly in Writing Test Scores. These results provide insights into the academic progression and assessment outcomes of high school students.

The strong influence of grade level on test performance aligns with recent research indicating that academic progression enhances standardized test outcomes.

Older students benefit from increased exposure to curriculum content, test preparation resources, and cognitive maturation, all of which contribute to improved performance. Longitudinal analyses of standardized testing data reveal that cumulative educational experiences can significantly affect achievement levels (Matheny et al., 2022).

Gender differences, though statistically significant, were less pronounced compared to grade-level effects. Female students outperformed male students in Writing Test Scores, consistent with research suggesting that female students excel in language-focused tasks, often due to stronger verbal and communication skills developed during adolescence (Hirnstein et al., 2022). Additionally, the format of standardized tests can influence gender-based performance differences, with female students performing better on open-ended or writing-intensive sections and male students excelling in multiple-choice formats (Reardon et al., 2018). However, research highlights a narrowing gender gap in standardized test performance, indicating that broader sociocultural and educational interventions may be contributing to more equitable outcomes (Wang & Degol, 2017; Charlesworth & Banaji, 2019).

The absence of significant interaction effects between grade and gender suggests that performance trends across grades are consistent for both male students and female students. This finding supports the notion that while demographic factors such as gender can influence individual performance, the impact of academic progression is more substantial in shaping standardized test outcomes (Pope & Sydnor, 2010).

Implications

The findings have several practical implications for educational practices and assessment design.

Educational Practices: The substantial grade-level differences observed in this study underscore the importance of targeted test preparation strategies. Schools should prioritize integrating test preparation into the curriculum earlier, particularly in Grades 10 and 11, to bridge the performance gap with Grade 12. Evidence suggests that targeted interventions, such as personalized learning plans and practice tests, can significantly improve student outcomes in standardized assessments (Ingkavara et al.,

2022; Major et al., 2021). Furthermore, addressing modest gender differences in Writing Test Scores through gender-sensitive teaching approaches could help ensure equitable skill development.

Assessment Design: This study highlights the reliability of multiple-choice questions in evaluating foundational and language-related skills. However, their limited scope in capturing higher-order cognitive abilities, such as critical thinking and problem-solving, calls for complementary assessment formats. Research indicates that combining multiple-choice questions with open-ended tasks enhances the validity of assessments by measuring a broader range of competencies (Yusuf Oc & Hassen, 2024; Yang et al., 2019). Additionally, the influence of test format on gender achievement gaps underscores the importance of designing assessments that minimize bias and promote inclusivity (Reardon et al., 2018).

Limitations

Several limitations should be considered when interpreting the findings. First, the use of imputation for missing data, while necessary to preserve the dataset, may have introduced slight biases and reduced variability. Second, the smaller sample size for Grade 10 students limited the statistical power for comparisons involving this group. Finally, the reliance on multiple-choice assessments, while effective for evaluating specific skills, may not fully capture students' broader academic competencies or critical thinking abilities.

Future Research

Future studies should explore longitudinal trends in test performance to assess how academic progression influences outcomes over time. Incorporating factors such as socioeconomic status, access to educational resources, and extracurricular engagement would provide a more nuanced understanding of performance disparities. Additionally, examining the impact of alternative assessment formats, such as essays or performance-based tasks, could offer insights into skills not adequately measured by standardized tests. There is a need to evaluate a

combination of test scores and non-cognitive skills to predict long-term academic and career success (Allensworth & Clark, 2020).

Conclusion

This study examined test performance across high school grade levels and genders, revealing significant grade-level effects and modest gender differences. Higher grades demonstrated progressively better performance, underscoring the importance of academic progression in shaping standardized test outcomes. Female students outperformed male students in Writing Test Scores, reflecting broader trends in language-related tasks, while other metrics showed minimal gender differences.

These findings have important implications for educational practices and assessment design. Schools should integrate test preparation into the curriculum earlier, particularly in Grades 10 and 11, to bridge the performance gap with Grade 12. Additionally, gender-sensitive instructional approaches can help address specific disparities, particularly in language-focused domains.

While the current study provides valuable insights, limitations such as the small sample size for Grade 10 and reliance on multiple-choice assessments highlight the need for further research. Future studies should explore longitudinal trends, consider socioeconomic and educational resource factors, and incorporate alternative assessment formats to capture a broader range of competencies.

References:

- Allan, E. & Driscoll, D. (2014). The three-fold benefit of reflective writing: Improving program assessment, student learning, and faculty professional development. Assessing Writing. 21. 37–55. 10.1016/j.asw.2014.03.001.
- Allensworth, E., & Clark, K. A. (2020). Test scores don't stack up to GPAs in predicting college success. *Educational Researcher*
- Bacon, D. R. (2003). Assessing learning outcomes: A comparison of multiple-choice and short-answer questions in a marketing context. *Journal of Marketing Education, 25*, 31-36.

- Black, P., Harrison, C., & Lee, C. (2003). *Assessment for learning: Putting it into practice*. UK: McGraw-Hill Education.
- Brown, A. E., et al. (2022). The impact of personalized test preparation strategies on SAT outcomes. *Assessment in Education, 29*(3), 301–315.
- Cahill, D. R., & Leonard, R. J. (1999). Missteps and masquerade in American medical academy: Clinical anatomists call for action. *Clinical Anatomy*. 12: 220-222
- Charlesworth, T. E. S., & Banaji, M. R. (2019). Gender in Science, Technology, Engineering, and Mathematics: Issues, Causes, Solutions. *Journal of Neuroscience*, *39*(37), 7228–7243. https://doi.org/10.1523/JNEUROSCI.0475-18.2019.
- Cheryl, A. M., David, O. D., Bart, K., & Nagasawami, S. V. (2017). Analysis of testing with multiple choice versus open-ended questions: Outcome-based observations in an anatomy course. *American Association of Anatomists Sci. Education*, 11, 254-261.
- Cooney, T. J., Sanchez, W. B., Leatham, K., & Newborn, D. S. (2004). *Open-ended assessment in math: A searchable collection of 450+ questions*. "Open-ended Assessment in Math."
- Graesser, A. C., Ozuru, Y., & Sullins, J. (2010). What is a good question? In M. McKeown & L. Kucan (Eds.), *Bringing reading research to life* (pp. 112–141). New York, NY: Guilford Press
- Hirnstein, M., Stuebs, J., Moè, A., & Hausmann, M. (2022). Sex/Gender Differences in Verbal Fluency and Verbal-Episodic Memory: A Meta-Analysis. *Perspectives on Psychological Science*, *18*(1), 174569162210821. https://doi.org/10.1177/17456916221082116.
- Ingkavara, T., Panjaburee, P., Srisawasdi, N., & Sajjapanroj, S. (2022). The use of a personalized learning approach to implementing self-regulated online learning. *Computers and Education: Artificial Intelligence*, *3*, 100086. https://doi.org/10.1016/j.caeai.2022.100086
- Karadağ, N. (2014). *Açık ve uzaktan eğitimde ölçme ve değerlendirme: Mega üniversitelerdeki uygulamalar. [Assessment in open and distance education:*

Practices in mega universities.]. (Publication No: 363040) [Doctoral dissertation, Anadolu University].

- Klufa, J. (2015). Multiple choice question tests–advantages and disadvantages. *Mathematics and Computers in Sciences and Industry Journal, 3*, 91-97.
- Kwon, O. N., Park, J. H., & Park, J. S. (2006). Cultivating divergent thinking in mathematics through an open-ended approach. *Asia Pacific Education Review, 7*, 51-61.
- Lee, H-S., Liu, O., & Linn, M. C. (2011). Validating measurement of knowledge integration in science using multiple-choice and explanation items. *Applied Measurement in Education, 24*(2), 115–136.
- Magliano, J., Millis, K., Ozuru, Y., & McNamara, D. (2007). A multidimensional framework to evaluate reading assessment tools. In D. S. McNamara (Ed.), Reading comprehension strategies: Theories, interventions, and technologies. Lawrence Erlbaum Associates Publishers, London (p. 107–136).
- Major, L., Francis, G. A., & Tsapali, M. (2021). The effectiveness of technology-supported personalised learning in low- and middle-income countries: A meta-analysis. *British Journal of Educational Technology, 52*(5).
- Martinez, M. E. (1999). Cognition and the question of test item format. *Education Psychology*, 34, 207-218.
- Matheny, K. T., Reardon, S. F., & Fahle, E. M. (2022). Uneven Progress: Recent Trends in Academic Performance Among U.S. School Districts. *CEPA Working Paper No. 22-02*
- Melovitz Vasan, C. A., DeFouw, D. O., Holland, B. K., & Vasan, N. S. (2017). Analysis of testing with multiple choice versus open-ended questions: Outcome-based observations in an anatomy course. *Anatomical sciences education, 11* (3), 254-261.
- Oksuz, Y., & Guven Demir, E. (2019). Acik uclu ve coktan secmeli basari testlerinin psikometrik ozellikleri ve ogrenci performansi acisindan karsilastirilmasi. (Comparison of Open-Ended Questions and Multiple Choice Tests in terms of Psychometric Features and Student Performance). Hacettepe Universitesi Egitim Fakultesi Dergisi, 34(1), 259-282. doi: 10.16986/HUJE.2018040550

- Ozuru, Y., Best, R., Bell, C., Witherspoon, A., & McNamara, D. S. (2007). Influence of question format and text availability on the assessment of expository text comprehension. *Cognition and Instruction, 25*, 399 – 438.
- Palmer, E. J., & Devitt, P. G. (2007). Assessment of higher order cognitive skills in undergraduate education: modified essay or multiple-choice questions? *BMC Medical Education, 7*, 49.
- Palmer, E. J., & Devitt, P. G. (2007). Assessment of Higher Order Cognitive Skills in Undergraduate Education: Modified Essay or Multiple-Choice Questions? BMC Medical Education, 7, 49. http://dx.doi.org/10.1186/1472-6920-7-49
- Pope, D. G., & Sydnor, J. R. (2010). A new perspective on stereotypical gender differences in test scores. *Journal of Economic Perspectives, 24*(2), 95–108.
- Rauch, D. P., & Hartig, J. (2010). Multiple-choice versus open-ended response formats of reading test items: A two-dimensional IRT analysis. *Psychological Test and Assessment Modelling, 52*(4), 354–379.
- Reardon, S. F., Fahle, E. M., Kalogrides, D., Podolsky, A., & Zarate, R. C. (2018). The relationship between test item format and gender achievement gaps on state standardized tests. *Educational Researcher*
- Schuwirth, L. W. T., & van der Vleuten, C. P. M. (2003). Different written assessment methods: what can be said about their strengths and weaknesses? *Med Educ, 38* (9), 974-979.
- Shohamy, E. (1984). Does the testing method make a difference? The case of reading comprehension. *Language Testing,1 (2)*, 147-170.
- Turgut, M.F., ve Baykul, Y. (2012). Egitimde olcme ve degerlendirme. Ankara: Pegem Akademi Yayincilik.
- Walsh, C. M., & Seldomridge, L. A. (2006). Critical thinking: Back to square two. *Nursing Education, 45*, 212-219.
- Wang, M.-T., & Degol, J. L. (2017). Gender Gap in Science, Technology, Engineering, and Mathematics (STEM): Current Knowledge, Implications for Practice, Policy, and Future Directions. *Educational Psychology Review, 29*(1), 119–140. https://doi.org/10.1007/s10648-015-9355-x

- Yang, B. W., Razo, J., & Persky, A. M. (2019). Using Testing as a Learning Tool. *American Journal of Pharmaceutical Education*, *83*(9), 7324. https://doi.org/10.5688/ajpe7324
- Yusuf Oc, & Hassen, H. (2024). Comparing The Effectiveness of Multiple-Answer And Single-Answer Multiple-Choice Questions In Assessing Student Learning. *Marketing Education Review*, 1–14. https://doi.org/10.1080/10528008.2024.2417106