

Item Processing Patterns of Test-Takers in a Multiple-Choice Reading Comprehension Test: Insights from an Eye-Tracking Study

Derya Akbaş¹ , Ergun Cihat Çorbacı² , and Nilufer Kahraman³

Abstract

This study explores how test-takers process multiple-choice questions in a reading comprehension test using eye-tracking data. A total of 159 participants completed a 10-item English test, while their eye movements were recorded with the Smart Eye Aurora eye tracker. The study first examined item processing patterns using latent profile analysis and then compared test performance across these groups. To identify processing patterns, latent profile models (ranging from one to four classes) were tested for each item based on average log process times across defined Areas of Interest (AOIs) for text lines and answer choices. Results showed that a two-class model (fast- and slow-pacing) provided the best fit for five items, while a three-class model (fast-, moderate-, and slow-pacing) best fit the remaining items. Items with two subgroups were typically moderately difficult (short) or easy (long), while items with three subgroups varied in difficulty and length. Additionally, test-takers in fast-pacing group were more likely to answer items correctly and achieved higher total scores than those in other groups, particularly for highly discriminating and moderately difficult or easy items. Overall, these findings highlight the importance of examining item processing patterns to better understand how individuals interact with multiple-choice test items. Item characteristics—such as difficulty, discrimination, and length—play a crucial role in shaping processing behaviors, providing deeper insights into the cognitive aspects of eye movement patterns during test-taking.

Plain Language Summary

This study examined how test-takers process multiple-choice items in a reading comprehension test by using eye-tracking technology. A total of 159 students completed a 10-item English test while their eye movements were recorded. The goal was to identify distinct processing patterns and investigate how these patterns relate to test performance. Results showed that test-takers could be grouped into fast-, moderate-, and slow-pacing profiles depending on how they allocated their time across each item. Fast-pacing test-takers were generally more successful, particularly on items that were easier or moderately difficult and highly discriminating. In contrast, slow-pacing test-takers tended to perform less well. These findings highlight the importance of examining response processes to better understand how individuals interact with test items and how item characteristics shape their processing. They offer valuable insights for improving test design and instructional strategies.

¹Aydın Adnan Menderes University, Aydın, Türkiye

²Sinop University, Sinop, Türkiye

³Gazi University, Ankara, Türkiye

Corresponding Author:

Derya Akbaş, Faculty of Education, Department of Educational Sciences, Aydın Adnan Menderes University, Aydın, Efeler 09010, Türkiye.

Email: derya.akbas@adu.edu.tr



Keywords

response process, eye-tracking, multiple-choice, assessment, latent profile analysis

Introduction

Eye-tracking technology, which records where and in what order an individual gazes while interacting with stimuli (Carter & Luke, 2020; Holmqvist et al., 2011), has become an increasingly valuable tool for understanding cognitive aspects of response processing that are not directly observable (Just & Carpenter, 2018; Rayner, 1995). This technology is particularly useful in educational research, offering insights into how students process information within a given timeframe—for example, while engaging with educational material or performing a task (e.g., Hyönä, 2010; McCray & Brunfaut, 2018). Researchers use it not only to refine instructional and assessment strategies but also to gain deeper insights into cognitive processes, ultimately enhancing both learning and evaluation methods (Jarodzka et al., 2017; Was et al., 2016). Eye-tracking, by providing response process data, helps researchers explore how test-takers navigate through the test and what kind of strategies they employ while completing the test items. It also offers insights into the knowledge and competencies test-takers draw upon, whether they truly engage with items instead of guessing, and how they utilize the resources and information provided in the test (Ercikan & Pellegrino, 2017). This detailed understanding of test-taker behavior enhances the interpretation and use of test scores, ultimately improving assessment validity and fairness.

Eye-Tracking in Educational Assessment

In recent years, there has been growing interest in using eye-tracking technology to obtain response process data in educational assessment settings. One of the key advantages of eye-tracking is its ability to capture response processes that are not accessible through traditional data sources. For instance, Maddox et al. (2018) used eye-tracking in the OECD Programme for the International Assessment of Adult Competencies (PIAAC) and demonstrated how it can provide detailed insights into response processes, offering an explanatory depth beyond think-aloud protocols or log files.

Recent studies have used eye-tracking to explore test-takers' cognitive strategies by analyzing how they process assessment items. Y. Hu et al. (2017) analyzed students' eye movements while solving complex problems in the Programme for International Student Assessment (PISA) to understand problem-solving strategies used by low- and high-performing students. Similarly, Krstić et al. (2018) examined eye movement patterns in PISA reading tasks and identified notable differences between

students with high and low scores. Bax (2013) investigated the cognitive processing of test-takers completing the International English Language Testing System (IELTS) reading tasks. His study examined how successful and unsuccessful candidates approached reading items, revealing that successful test-takers demonstrated more expeditious reading behaviors, focusing efficiently on relevant portions of the text, while less successful ones spent a longer time on larger chunks of text and exhibited inefficient search patterns. Also, Langenfeld et al. (2020) explored test-takers' problem-solving strategies and cognitive load in a Graphic Literacy Assessment with multiple-choice items, showing that high performers used different strategies and engaged with graphics in distinct ways compared to low performers. These studies have provided results highlighting that efficient test-taking strategies might be a key differentiator between high- and low-performing test-takers across different test formats.

Another line of research has focused on how item characteristics, such as format and difficulty, influence response patterns. Bax and Chan (2019) examined the potential impact of item types (e.g., heading matching and multiple-choice) on students' reading behaviors, revealing that it significantly affects test-taking strategies. Similarly, Yaneva et al. (2021) examined how multiple-choice answer options influence medical students' response processing in questions assessing clinical reasoning. Their findings showed that students first examined the item stem before reviewing the answer choices, spending more time and fixating more frequently on the stem than on the options. Together, these studies highlight the value of eye-tracking in understanding response processes in assessments. They emphasize the role of cognitive strategies and item characteristics in shaping test performance, contributing to a deeper understanding of how individuals engage with test items and how assessments can be improved.

Research to date suggests that eye-tracking technology is particularly useful for identifying and describing eye movement patterns unique to test-takers responding to multiple-choice items, helping with the associated quality control and validity studies (Langenfeld et al., 2020; Zumbo & Hubley, 2017). By examining response processes through eye movements, researchers can assess whether test items engage test-takers as intended and whether the inferences drawn from test scores are justified (Ercikan & Pellegrino, 2017). Research specifically focused on distinct multiple-choice item processing patterns and their associations with test-takers' scores would provide nuanced insights into score meaning and

interpretation. Additionally, such research could enhance our understanding of the capabilities of the items that might be irrelevant to the construct, hence might be differentially affecting test-takers' performances (e.g., American Educational Research Association et al., 2014; Kane & Mislevy, 2017).

The Present Study

Previous research using eye-tracking to investigate response processes has often focused on different tasks or tests, comparing high- and low-performing test-takers in terms of their processing patterns. These studies have provided valuable insights into the association between response processes and performance. However, relatively few studies have examined how item characteristics influence response process patterns, leaving open questions about how specific item features shape test-takers' approaches.

The present study investigates test-takers' item processing patterns in a ten-item multiple-choice reading comprehension test using eye-tracking data. Specifically, it examines eye fixation patterns and their relationship with item and total test scores, while also evaluating item characteristics that may be associated with distinct response processes.

First, we applied Latent Profile Analysis (LPA) for each test item using segment-specific process times derived from time-stamped eye fixation metrics for defined Areas of Interest (AOIs). Each item included up to ten AOIs—five corresponding to the text lines preceding the choices and five representing the answer choices (A to E). Average process times were computed for each AOI, and potential subgroups of test-takers were identified for each item based on their processing patterns. Next, we tested models to examine the predictive effect of subgroups on item and total test scores, assessing whether processing patterns were associated with performance differences. Finally, we evaluated the items collectively to identify common characteristics among those with similar subgroup structures and significant subgroup performance differences.

The study addresses the following research questions:

1. Which distinct subgroups of test-takers can be identified based on eye fixation patterns for each item?
2. How do these subgroups differ in their item scores?
3. How do these subgroups differ in their total test scores?
4. What are the common characteristics of items with similar subgroup structures?
5. What are the common characteristics of items where subgroup differences in item and total test scores are observed?

This study extends prior research by modeling response processing heterogeneity and linking subgroup membership not only to item- and test-level performance outcomes but also to item features. Using LPA on fine-grained eye-tracking data, it reveals processing patterns that remain hidden in aggregate analyses, thereby offering a more nuanced account of test-taking behavior and contributing to the literature on response processes in educational assessment.

Methodology

Participants

A total of 159 participants were recruited from Gazi University through voluntary sign-up. The participants were enrolled in a variety of teacher-education programs, such as mathematics, Turkish, art, English, and biology, and were distributed across all four academic years (first through fourth). All participants were English as a Foreign Language (EFL) learners residing in Türkiye. They had normal or corrected-to-normal vision. The study was approved by the Gazi University Ethics Committee (No: 80287700-302.08.01-156936). Informed consent was obtained from all participants before the study.

Experimental Procedures

Eye-tracking data were collected using the Smart Eye Aurora screen-based eye tracker, which performed binocular tracking at a sampling frequency of 250 Hz. The study was conducted on an MSI computer (Intel Core i9-11900K @ 3.50 GHz, 64.0 GB RAM), with test items displayed on a 21.5-inch flat-panel monitor with a $1,920 \times 1,080$ resolution.

The experimental sessions took place in a research laboratory. At the start of each session, participants were informed about the procedure, followed by a five-point calibration to ensure the quality of the eye movement recordings. Each test-taker completed the test individually, responding to one item at a time in the same fixed order. The test screen layout positioned the item stem on the left side and the answer choices on the right. There was no time limit for the test; each session concluded when the participant submitted their responses.

Materials

Figure 1 shows the multiple-choice test items used in the study. These items were drawn from the publicly released item pool of a nationwide English as a foreign language exam administered by a governmental testing agency (The Measurement, Selection and Placement Center, 2018). Items were selected to represent a range

Item 1		Item 2	
AOI 1 <u>Choose the best option to complete the dialogue.</u> AOI 2 Professor: And have you seen a picture of a little devil frog. The male produces noisy sounds in pursuit of a partner. AOI 3 Student: _____ AOI 4 Professor: Well, he's a poisonous frog and has bright colours which warn predators that he's unsafe to eat. AOI 5 Student: How significant characteristics it has?	a) Do the sounds he makes differ or does he use the same kind of sounds? AOI 6 b) How do the partners react to these sounds when they hear them? AOI 7 c) Is it really difficult for the little devil frog to find a partner? AOI 8 d) But doesn't it make him noticeable and vulnerable to being attacked? AOI 9 e) Is it only males, or do females also produce such sounds before mating? AOI 10	AOI 1 <u>Choose the best word or expression to fill the space.</u> AOI 2 By 2.5 million years ago, when they began _____ stone tools, early AOI 3 humans had understood that they _____ the natural world to their AOI 4 own advantage.	a) to have made / had to alter AOI 6 b) to make / used to alter AOI 7 c) making / could alter AOI 8 d) having made / should have altered AOI 9 e) to have been making / must have altered AOI 10
Item 3		Item 4	
AOI 1 <u>Choose the best option to complete the given</u> AOI 2 Some birds build nests simply to protect their young _____	a) since many small animals that live above the ground make their homes in trees AOI 6 b) until they have learned how to look after themselves and survive in the wild on their own AOI 7 c) given that they make mud huts that are placed in all kinds of places, from underground to fence posts AOI 8 d) although sometimes a squirrel takes over an empty bird's nest and makes its home there AOI 9 e) while colonies of wasps build their nests from the top downwards AOI 10	AOI 1 <u>Choose the most accurate English translation of the sentences in Turkish.</u> AOI 2 Çarlık sistemleri Avrupa ve Asya ülkelerinde yaygın bir AOI 3 biçimde kullanılmaktadır. Bu da Birleşik Devletler'den farklıdır. AOI 4 farklı bir biçimde işlenmektedir.	
a) Although apprenticeship systems have been used widely in European and Asian countries, they operate in a somewhat different manner than those in the United States. AOI 6 b) European and Asian countries have been widely using apprenticeship systems that operate in a slightly different manner than the United States. AOI 7 c) Apprenticeship systems are widely used not only in European and Asian countries but also in the United States, but in a somewhat different manner in the former. AOI 8 d) Both European and Asian countries and the United States have been widely using apprenticeship systems, but the latter has been practicing it in a slightly different manner. AOI 9 e) There is a slight difference between the apprenticeship systems in European and Asian countries and those in the United States. AOI 10			
Item 5		Item 6	
AOI 1 <u>Choose the best option to complete the given sentence.</u> AOI 2 Trans fat may be as harmful to your health as saturated fat _____.	a) because it increases blood cholesterol levels, among other adverse effects AOI 6 b) although major sources are hardened vegetable fat such as margarine AOI 7 c) as long as foods high in saturated fat include red meats and dairy products. AOI 8 d) since a diet that has generous amounts of vegetables is naturally lower in fat AOI 9 e) even if a diet high in fat can also be unnecessarily high in calories AOI 10	AOI 1 <u>Choose the best word or expression to fill the space.</u> AOI 2 Childcare is a broad term that _____ services which protect the AOI 3 health, safety, and well-being of children who require custodial AOI 4 care by adults other than their own parents for a temporary AOI 5 period of time.	
a) hinders AOI 6 b) encompasses AOI 7 c) creates AOI 8 d) eliminates AOI 9 e) suspends AOI 10			
Item 7		Item 8	
AOI 1 <u>Choose the best word or expression to fill the space.</u> AOI 2 _____ the problems of obesity and disordered eating growing AOI 3 around the world, public health professionals have focused AOI 4 research efforts on identifying potential causes and treatments -- AOI 5 _____ these related problems.	a) About / in AOI 6 b) On / towards AOI 7 c) With / for AOI 8 d) Through / from AOI 9 e) Among / at AOI 10	AOI 1 <u>Choose the most accurate Turkish translation of the sentences in English.</u> AOI 2 More than a hundred different studies have shown that there AOI 3 really is a significant relation between happiness and physical AOI 4 health.	
a) Yüzyıl boyunca fazla kilo çalma sorunlarına ki mutluluk ile beden sağlığı arasında gerçekleşen önemli bir ilişki bulunabilir. AOI 6 b) Mutluluk ile beden sağlığı arasında önemli bir ilişki bulunabilir. Yüzyıl boyunca fazla kilo çalma sorunu ortaya çıkmıştır. AOI 7 c) Mutluluk ile beden sağlığı arasında gerçekleşen önemli bir ilişki olup olmadığının göstermek için yüzyıl boyunca fazla kilo çalma yapılmıştır. AOI 8 d) Yüzyıl boyunca fazla kilo çalma, mutluluk ile beden sağlığı arasında gerçekleşen önemli bir ilişki olduğunu göstermiştir. AOI 9 e) Yüzyıl boyunca fazla kilo çalmaya göre, mutluluk ile beden sağlığı arasında gerçekleşen önemli bir ilişki olabilir. AOI 10			
Item 9		Item 10	
AOI 1 <u>Choose the best word or expression to fill the space.</u> AOI 2 The idea of the Internet was initially to provide redundancy of AOI 3 communications _____ a catastrophic event like a nuclear blast, AOI 4 which might destroy a single connection or computer but not the AOI 5 entire network.	a) such as AOI 6 b) on behalf of AOI 7 c) in terms of AOI 8 d) in addition to AOI 9 e) in case of AOI 10	AOI 1 <u>Choose the best rephrased form of the given</u> AOI 2 The years following the 1789 revolution in France brought AOI 3 massive upheaval in French society, which, in turn, produced AOI 4 profound changes in dress and fashion.	
a) Although the revolution took place in France in 1789, it was only after many years that French society went through considerable changes in dress and fashion following this great disturbance in their society. AOI 6 b) The revolution in France in 1789 led to a serious disturbance among people in French society, and this was the reason why the French had to allow for fundamental and important changes in dress and fashion. AOI 7 c) If it were not for the revolution in France in 1789, dress and fashion in French society might not have changed considerably despite the enormous disturbance it caused in their society. AOI 8 d) What contributed to the revolution in France in 1789 were the major changes in dress and fashion, which resulted from the huge confusion experienced by French society at that time. AOI 9 e) It was not until after the revolution in France in 1789 that French society experienced dramatic changes in dress and fashion, due to the enormous turmoil generated by this period. AOI 10			

Figure 1. Items in the test and defined AOIs.

Note. The items were sampled from a large-scale test called 'Foreign Language Exam (YDS)' (The Measurement, Selection and Placement Center, 2018).

Table 1. Characteristics of the Items in the Test.

Items	Format	a	b	Word count within AOIs										Total
				Ins	L-1	L-2	L-3	L-4	A	B	C	D	E	
1	Complete dialogue	0.74	0.66	8	22	1	18	6	15	12	13	11	13	119
2	Fill in the blanks	0.69	0.52	10	11	10	2	-	6	5	3	5	7	59
3	Complete sentence	0.62	0.73	8	9	-	-	-	14	17	19	15	11	93
4	Turkish-English translation	0.57	0.74	11	8	7	4	-	25	21	27	27	20	150
5	Complete sentence	0.52	0.75	9	11	-	-	-	10	10	14	14	14	82
6	Fill in the blanks	0.47	0.43	10	10	9	11	3	1	1	1	1	1	48
7	Fill in the blanks	0.46	0.53	10	8	8	8	3	2	2	2	2	2	47
8	English-Turkish translation	0.41	0.77	11	10	9	1	-	17	16	18	15	15	112
9	Fill in the blanks	0.41	0.52	10	11	8	11	2	2	3	3	3	3	56
10	Rephrase sentence	0.35	0.37	8	9	9	6	-	33	36	31	30	30	192

Note. a = Item discrimination, b = Item difficulty, Ins = Instruction, L-1, ..., L-4 = first line, ..., fourth line, A, B, C, D, E = Choice A, B, C, D, E. Items were presented in descending discrimination order for convenience, but were given randomly during test experiment. Item discrimination = corrected item-total correlation. Item difficulty = proportion of correct responses. Cronbach α internal consistency coefficient = .85 for the test.

of skills such as grammar, vocabulary, translation, and reading comprehension. For each item, up to ten Areas of Interest (AOIs) were defined: Instruction, Line-1, Line-2, Line-3, Line-4, and Choices A, B, C, D, and E. As illustrated in Figure 1, the AOIs for the answer choices (A–E) remained consistent across items, whereas the AOIs in the stem varied due to item-specific features, such as differences in the number of lines or words per line.

We first calculated the psychometric properties of the test items using Classical Test Theory (Crocker & Algina, 2006). Table 1 provides item discrimination and difficulty indices along with item format and word count within AOIs. Item discrimination (a) is represented by the corrected item-total correlation coefficient, where higher values indicate that an item differentiates more effectively between high- and low-performing test-takers. Item difficulty (b) is represented by the proportion of correct responses, where higher values indicate that an item is easier, meaning that more test-takers responded correctly.

Data Extraction and Analysis

The dataset included item response data (i.e., incorrect or correct [0/1]), total test scores (ranging from 0 to 10), and a series of up to ten AOI process times derived from time-stamped fixations, binary-coded per AOI over item encounter times. After exporting the data based on the AOIs, we screened it for corrupt, irrelevant, or inaccurate recordings and excluded any problematic data from the dataset. For the analyses, mean log AOI process times per item were calculated using time-stamped fixation data. This was done by averaging the time a test-taker spent reading an AOI from the moment they started reading an item until they provided a response (Çorbacı & Kahraman, 2022). The dataset also included total

fixation durations, representing the sum of all time spent fixating on an item, that is, item response times.

Data analysis was conducted in two stages:

1. Preliminary Analysis: Descriptive statistics of total fixation duration were calculated. We then examined whether fixation durations differed between response groups using t-tests. Additionally, mean AOI process times within items were computed and compared across response groups to assess differences in AOI process times.
2. Latent Profile Analysis (LPA): We tested three latent profile analysis models (Figure 2), where AOI process times served as the manifest/observed variables of the latent variable. LPA, a variation of Latent Class Analysis for continuous variables, identifies hidden subgroups in the data by estimating the probability that individuals belong to different latent profiles (Ferguson et al., 2020). This method uses individuals' responses to a set of indicators to classify them into subgroups with similar response patterns (for more information, see Nylund-Gibson & Choi, 2018).

Profiles are represented by two sets of parameter estimations: latent class parameters and class-conditional parameters (for more information, see Hickendorff et al., 2018). Latent class parameters estimate the proportion of individuals in each class (class prevalence), while class-conditional parameters define the measurement model by describing the relationship between the latent variable and its indicators. Since LPA uses continuous indicators, it estimates profile-specific parameters (e.g., means and variances), that characterize each subgroup.

In Model-a (the unconditional model), we tested a series of Latent Profile Analysis (LPA) models, ranging

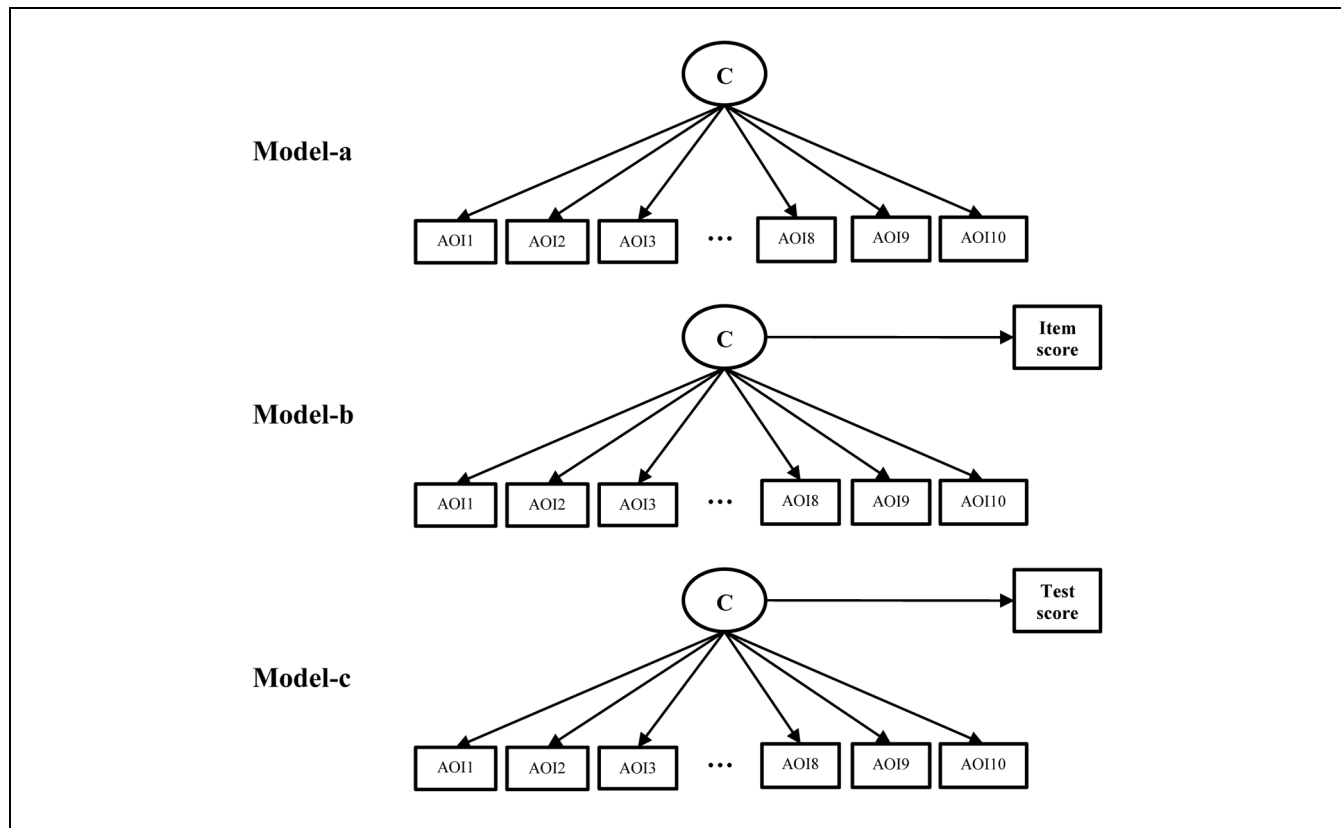


Figure 2. Graphical representation of the LPA models tested.

Note. *C* = latent class variable, *AOI1*,...*AOI10* = indicators of *C*, *Item score* and *Test score* are distal outcomes.

from one-profile to four-profile solutions per item, to identify distinct item processing patterns present in the data (Research Question 1). The unconditional model was used to determine the optimal number of profiles that best fit the data. To evaluate model fit, we used the Akaike Information Criterion (AIC) (Akaike, 1987), Bayesian Information Criterion (BIC) (Schwarz, 1978), Vuong-Lo-Mendell-Rubin Likelihood Ratio Test (VLMR-LRT) (Lo, Mendell & Rubin, 2001; Vuong, 1989), and Bootstrap Likelihood Ratio Test (BLRT) (McLachlan & Peel, 2000). A model with lower AIC and BIC values indicates a better fit. VLMR-LRT and BLRT compare the *K*-class model to the *K*-1 class model and provide a *p*-value to assess whether adding a class results in a statistically significant improvement in model fit (Nylund-Gibson & Choi, 2018). To assess classification precision, we examined the entropy value, which ranges from 0 to 1, where higher values (>0.80) indicate better classification accuracy (Clark, 2010).

In Model-b, we incorporated item scores as a distal outcome (i.e., a consequence of latent class membership) into the LPA model to determine whether item scores differed across test-takers in the identified profiles of item processing patterns (Research Question 2). To include

the distal outcome, we used the three-step approach (Vermunt, 2010), which consists of the following steps: (1) estimating the latent class model using only the latent class indicator variables (i.e., the unconditional model), (2) creating the most likely class variable based on the latent class posterior distribution obtained in the first step, and (3) regressing the most likely class on the predictor variables while accounting for misclassification from the second step (Asparouhov & Muthén, 2014). For distal outcomes, class-specific means and variances are estimated for each distal outcome to examine differences across profiles, followed by pairwise comparisons to determine whether the differences were statistically significant (Nylund-Gibson & Choi, 2018). This approach ensures that the measurement of the latent class variable remains unaffected by the inclusion of covariates or distal outcomes, as the measurement parameters are fixed at the values estimated in the unconditional model (Nylund-Gibson et al., 2014). For effect size, we calculated Cohen's *d* for each pairwise comparison and interpreted following Cohen's (1988) guidelines: small ($d = 0.20$), medium ($d = 0.50$), and large ($d = 0.80$).

In Model-c, we incorporated total test scores as a distal outcome to examine whether test scores varied across

Table 2. Descriptive Statistics of Total Fixation Duration Across Item Response Groups.

Items	Response groups	<i>n</i>	Mean	SD	Min	Max	t-test (df)
1	Incorrect	52	70.67	32.38	24.08	151.85	4.870* (150)
	Correct	100	48.79	22.49	16.61	141.66	
	Total	152	56.27	28.18	16.61	151.85	
2	Incorrect	64	42.98	19.20	9.12	116.79	3.688* (130)
	Correct	68	31.18	17.58	8.61	86.41	
	Total	132	36.90	19.25	8.60	116.79	
3	Incorrect	40	53.23	20.86	19.02	98.23	1.980* (144)
	Correct	106	43.98	26.60	9.32	174.86	
	Total	146	46.51	25.43	9.32	174.86	
4	Incorrect	36	67.49	33.76	12.06	155.75	4.299* (139)
	Correct	105	43.76	26.62	9.24	175.11	
	Total	141	49.82	30.32	9.24	175.11	
5	Incorrect	38	52.10	19.42	20.01	101.50	1.478 (151)
	Correct	115	45.76	23.95	9.84	142.91	
	Total	153	47.33	23.01	9.84	142.91	
6	Incorrect	79	35.62	13.58	6.24	90.39	4.503* (137)
	Correct	60	24.84	14.49	6.13	69.16	
	Total	139	30.97	14.92	6.13	90.39	
7	Incorrect	59	43.96	20.56	9.53	104.97	1.974 (124)
	Correct	67	37.08	18.51	7.88	97.77	
	Total	126	40.30	19.72	7.88	104.97	
8	Incorrect	34	52.01	19.14	14.32	86.59	1.964 (146)
	Correct	114	44.08	21.09	11.30	139.07	
	Total	148	45.90	20.87	11.30	139.07	
9	Incorrect	65	36.04	19.10	6.78	106.52	-0.636 (133)
	Correct	70	38.11	18.68	10.22	101.54	
	Total	135	37.11	18.84	6.78	106.52	
10	Incorrect	93	78.40	37.13	22.12	280.58	0.109 (145)
	Correct	54	77.72	34.96	25.18	187.13	
	Total	147	78.15	36.23	22.12	280.58	

Note. df = degrees of freedom.

* < .05.

the identified profiles (Research Question 3). Using the same three-step approach as in Model-b, we specified distal variables with unequal means and equal variances to account for potential differences across profiles. All analyses were conducted using Mplus (Muthén & Muthén, 1998–2017; for details on the three-step approach, see Asparouhov & Muthén, 2014; for illustrative applications of this approach, see Nylund-Gibson & Choi, 2018 and Nylund-Gibson et al., 2014).

Finally, we synthesized our findings and presented a comprehensive summary to determine whether there were any common characteristics among items with similar subgroup structures (i.e., profiles) and significant subgroup differences in item and total test scores (Research Questions 4 and 5, respectively).

Results

Preliminary Analysis: Descriptive statistics

Table 2 presents the descriptive statistics for total fixation duration per item, along with t-test results. The

results indicated that test-takers who answered incorrectly had significantly longer fixation durations than those who answered correctly for items 1, 2, 3, 4, and 6. These results indicate that incorrect responses were often associated with longer processing times, potentially reflecting inefficient strategy use. A similar pattern was observed for the remaining items, except for item 9, but the mean differences were not statistically significant.

Figure 3 presents the mean process time per AOI for each response group. The results indicated that, for most AOIs and items, test-takers who answered incorrectly had longer mean process times than those who answered correctly. The t-test results comparing AOI process times across response groups are not included here for simplicity, but followed a similar pattern to the total fixation duration comparisons reported in Table 2. For items 1, 2, 3, 4, and 6, significant differences were observed in process times across response groups for most AOIs. For the remaining items, differences in process times were either limited to a few AOIs or not statistically significant.

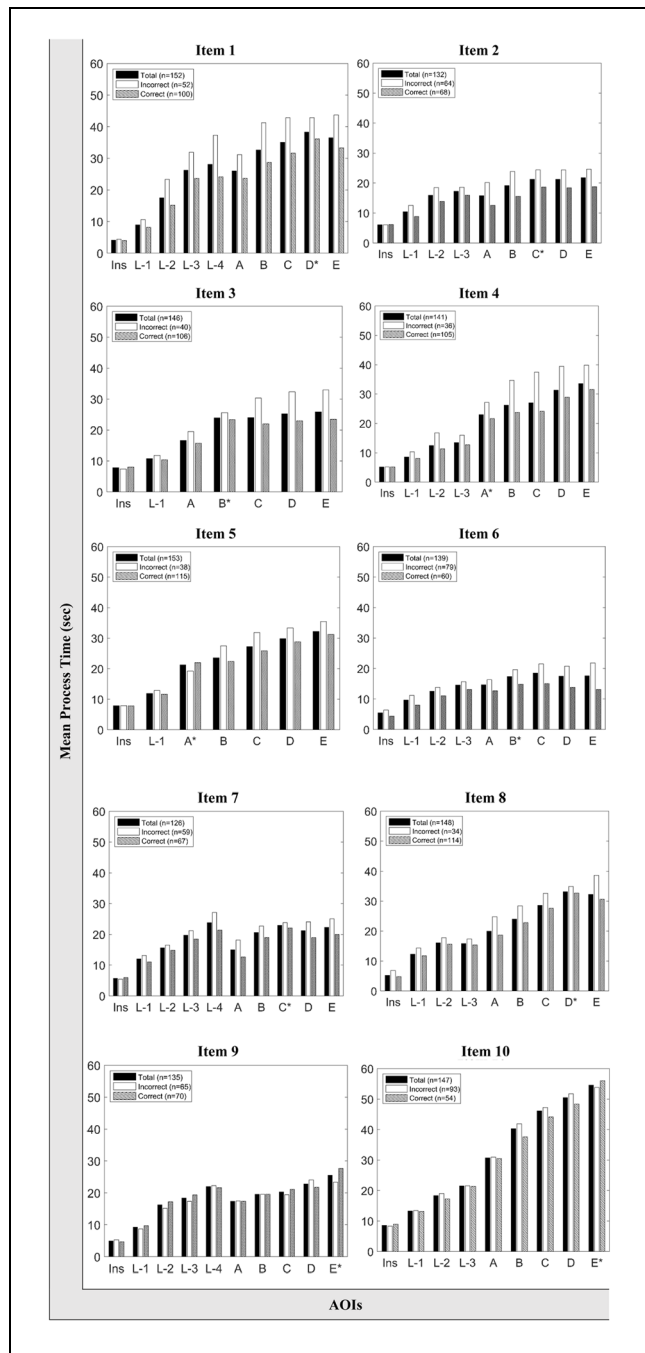


Figure 3. Mean AOI process time in terms of response groups.
 Note. AOI = Area of interest, Ins = Instruction, L-1, ..., L-4 = first line, ..., fourth line, A, B, C, D, E = Choice A, B, C, D, E.
 The choice with * shows the correct response.
 The graphs were constructed using MATLAB R2023b (The MathWorks Inc., 2023).

Latent Profile Analysis: Results of The Unconditional Model (Model-a)

In Model-a, we tested a series of LPA models, ranging from a one-profile to a four-profile model, to address the

first research question: identifying the presence of subgroups of item processing patterns. Table 3 presents the model fit indices.

The results showed that a two-profile model provided the best fit for items 1, 5, 6, 7, and 8, based on lower AIC and BIC values, significant VLMR-LRT and BLRT tests, high entropy, and interpretability of the identified profiles. While selecting the model, we also considered the smallest profile size (to avoid groups smaller than 5%) and ensured adequate separation between profiles to reflect meaningful differences in processing patterns. Based on these criteria, two distinct item processing patterns emerged among test-takers for these items. In contrast, a three-profile model provided the best fit for items 2, 3, 4, 9, and 10, indicating three item processing patterns for these items.

After determining the best-fitting model for each item, we interpreted the proportions of test-takers in each subgroup (latent class parameters) and their mean AOI process times (class-conditional parameters). These estimations are presented in Figure 4.

The results showed that in the first item, Profile-1, which included test-takers with lower mean AOI process times, was labeled as fast-pacing, while Profile-2, consisting of test-takers with higher mean process times, was labeled as slow-pacing. For the second item, three distinct profiles emerged. Profile-1 (fast-pacing) included test-takers with lower mean process times, Profile-2 (moderate-pacing) consisted of test-takers with moderate mean process times, and Profile-3 (slow-pacing) comprised test-takers with higher mean process times. Additionally, test-takers in the slow-pacing group spent more time on the answer choices than those in the fast- or moderate-pacing groups. Using this labeling approach, we categorized profiles based on mean process time, assigning fast-pacing to groups with lower process times, moderate-pacing to those with moderate times, and slow-pacing to those with higher times.

Figure 4 also presents the proportion of test-takers in each profile. For example, in the first item, 58% of test-takers belonged to the fast-pacing group, while 42% belonged to the slow-pacing group. Across all items, profiles contained a substantial number of test-takers, with the smallest group representing 10% of the sample.

Results of the Model with Item Score as a Distal Outcome (Model-b)

In Model-b, item scores were incorporated as a distal outcome to address the second research question, which examines differences in item scores across item processing groups. Table 4 presents the paired comparisons of item scores.

The results indicated that test-takers in the fast-pacing group had significantly higher mean item scores (i.e.,

Table 3. Fit Indices of LPA Models (Model-a).

Items	Indices	Model Alternatives			
		One-profile	Two-profile	Three-profile	Four-profile
1	AIC/BIC	2,057.98/2,118.45	1,438.40/1,532.14	1,137.70/1,264.70	1,034.63/1,194.90
	2 LL Difference	-	641.58	322.70	125.06
	VLMR/BLRT p	-	0.06/0.00	0.11/0.00	0.05/0.00
	Entropy	-	0.91	0.92	0.93
2	AIC/BIC	1,865.36/1,917.25	1,382.72/1,463.43	1,106.41/1,215.95	991.49/1,129.87
	2 LL Difference	-	502.65	296.31	134.91
	VLMR/BLRT p	-	0.40/0.00	0.04/0.00	0.07 /0 .00
	Entropy	-	.90	0.94	0.94
3	AIC/BIC	1,669.08/1,710.85	1,339.59/1,405.23	1,204.40/1,293.91	1,170.75/1,284.13
	2 LL Difference	-	345.49	151.18	49.66
	VLMR/BLRT p	-	0.01/0.00	0.02/0.00	0.56/0.00
	Entropy	-	0.87	0.89	0.87
4	AIC/BIC	2,175.40 /2,228.47	1,683.00/1,765.57	1,479.02/1,591.07	1,345.67/1,487.21
	2 LL Difference	-	512.40	223.98	153.35
	VLMR/BLRT p	-	0.00/0.00	0.04/0.00	0.02/0.00
	Entropy	-	0.93	0.93	0.95
5	AIC/BIC	1,498.24/1,540.67	1,043.49/1,110.16	875.78/966.70	780.60/895.75
	2 LL Difference	-	470.75	183.71	111.19
	VLMR/BLRT p	-	0.00/0.00	0.32/0.00	0.35/0.00
	Entropy	-	0.92	0.88	0.92
6	AIC/BIC	2,043.10/2,095.92	1,622.20/1,704.36	1,401.94 / 1,513.45	1,302.41/1,443.27
	2 LL Difference	-	440.90	240.26	119.53
	VLMR/BLRT p	-	0.01/0.00	0.08/0.00	0.07/0.00
	Entropy	-	0.92	0.92	0.92
7	AIC/BIC	1,987.31/2,044.04	1,464.26/1,552.19	1,274.91/1,394.03	1,136.79/1,287.12
	2 LL Difference	-	545.05	211.36	160.11
	VLMR/BLRT p	-	0.02/0.00	0.33/0.00	0.06/0.00
	Entropy	-	0.92	0.92	0.94
8	AIC/BIC	2,017.10/2,071.05	1,524.86/1,608.78	1,308.77/1,422.67	1,182.95/1,326.82
	2 LL Difference	-	512.24	236.09	145.82
	VLMR/BLRT p	-	0.02/0.00	0.08/0.00	0.09/0.00
	Entropy	-	0.92	0.93	0.94
9	AIC/BIC	2,136.24/2,194.35	1,641.87/1,731.94	1,418.50/1,540.52	1,300.24/1,454.22
	2 LL Difference	-	516.37	245.37	140.26
	VLMR/BLRT p	-	0.05/0.00	0.05/0.00	0.13/0.00
	Entropy	-	0.91	0.94	0.93
10	AIC/BIC	1,991.25/2,045.08	1,492.77/1,576.50	1,312.61/1,426.25	1,223.96/1,367.50
	2 LL Difference	-	518.49	20.15	108.66
	VLMR/BLRT p	-	0.17/0.00	0.05/0.00	0.14/0.00
	Entropy	-	0.89	0.92	0.91

Note. AIC = Akaike's Information Criterion; BIC = Bayesian Information Criterion; 2 LL Difference = 2 Times the Loglikelihood Difference; VLMR = Vung-Lo-Mendell-Rubin Likelihood Ratio Test; BLRT = Bootstrap Likelihood Ratio Test. The chosen models were written in bold.

correct response rates) than those in the slow-pacing group for items 1, 2, 3, 4, 6, and 8 ($p < .05$). These findings suggest that test-takers in the fast-pacing group were more likely to respond correctly compared to those in the slow-pacing group. Additionally, test-takers in the fast-pacing group had significantly higher mean item scores than those in the moderate-pacing group for items 2, 3, and 4, indicating that fast-pacing test-takers were more likely to respond correctly than moderate-pacing ones. However, no significant differences were observed in mean item scores between the moderate- and slow-pacing groups

($p > .05$). Effect sizes for the item score differences ranged from small to large, indicating varying magnitudes of practical significance across comparisons.

Results of the Model with Total Test Score as a Distal Outcome (Model-c)

In Model-c, total test scores were incorporated as a distal outcome to address the third research question, which examines differences in total test scores across item

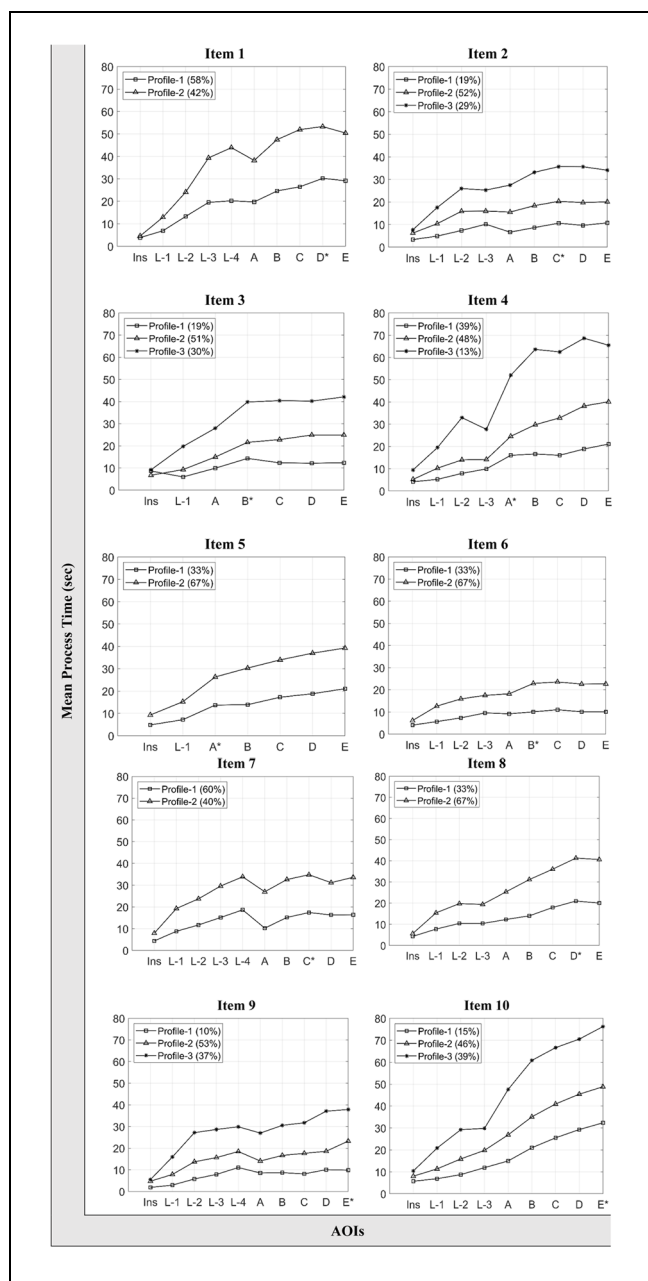


Figure 4. Mean AOI process time estimations conditioned on profiles.

Note. AOI = Area of interest, Ins = Instruction, L-1. ... L-4 = first line, ..., fourth line, A, B, C, D, E = Choice A, B, C, D, E.

The choice with * shows the correct response.

The graphs were constructed using MATLAB R2023b (The MathWorks Inc., 2023).

processing groups. Table 5 presents the paired comparisons of test scores.

The results indicated that test-takers in the fast-pacing group had significantly higher total scores than those in the slow-pacing group for items 1, 2, 3, 4, 6, and 8 ($p < .05$). Additionally, test-takers in the fast-pacing group

scored significantly higher than those in the moderate-pacing group for items 2, 3, and 4 ($p < .05$). For item 4, test-takers in the moderate-pacing group also had significantly higher total scores than those in the slow-pacing group ($p < .05$). The magnitude of effect sizes varied from small to large, reflecting varying levels of practical significance across these comparisons.

Common Characteristics of Items with Similar Group Structures and Observed Performance Differences

To address the fourth and fifth research questions, we examined the common characteristics of items with similar processing subgroups (i.e., profiles) and those with significant score differences among these groups.

Items tended to produce either two (fast- and slow-pacing) or three (fast-, moderate-, and slow-pacing) subgroups. Items with two groups were either moderately difficult ($b \sim 0.50$) or easy ($b > 0.60$), with moderately difficult ones being short (47 and 48 words) and easy items relatively longer (82, 112, and 119 words). Items with three subgroups, however, spanned a broader range of difficulty and length, including a difficult and very long item (192 words), moderately difficult and shorter items (56 and 59 words), and easy but longer items (93 and 150 words). These findings reveal that multiple processing subgroups can emerge across a range of difficulty levels and lengths, suggesting that the interaction of these factors may drive variation in processing patterns.

Significant item and test score differences among the subgroups were most often found on items that were moderately difficult or easy and also highly discriminating. In these cases, fast-pacing group was more likely to respond correctly to these items and had higher total test scores than the other groups. However, items that showed no subgroup differences varied widely in both length (e.g., 47–192 words) and difficulty. These results suggest that when items are both discriminating and moderate or low difficulty, response processing differences may align with performance differences.

Discussion

Using latent profile modeling of fine-grained eye movement data from a reading comprehension test consisting of ten multiple-choice items, this study presents empirical evidence that there may be distinct response processing patterns, and that these patterns may vary depending on items' psychometric properties (e.g., difficulty and discrimination parameters) or other characteristics such as word count.

Pertaining to the investigation of distinct response processing patterns (i.e., the first research question of the

Table 4. Equality Tests of Mean Item Scores Across Profiles (Model-b).

Items	Profiles	n	Mean item score	Std. error	χ^2 for paired comparisons		
					Fast vs. slow	Fast vs. moderate	Moderate vs. slow
1	Fast-pacing	88	0.780	0.046	12.141*	-	-
	Slow-pacing	64	0.490	0.066	(<i>d</i> = 0.611)		
2	Fast-pacing	25	0.886	0.064	23.741*	22.481*	0.587
	Moderate-pacing	70	0.458	0.063	(<i>d</i> = 1.169)	(<i>d</i> = 0.888)	(<i>d</i> = 0.158)
3	Fast-pacing	26	0.921	0.050	11.579*	7.322*	1.313
	Moderate-pacing	76	0.721	0.054	(<i>d</i> = 0.712)	(<i>d</i> = 0.468)	(<i>d</i> = 0.231)
4	Fast-pacing	44	0.608	0.078			
	Slow-pacing	55	0.877	0.045	6.434*	6.247*	1.006
5	Fast-pacing	68	0.688	0.058	(<i>d</i> = 0.846)	(<i>d</i> = 0.450)	(<i>d</i> = 0.277)
	Slow-pacing	18	0.554	0.119			
6	Fast-pacing	50	0.844	0.053	3.738	-	-
	Slow-pacing	103	0.707	0.045	(<i>d</i> = 0.317)		
7	Fast-pacing	46	0.691	0.074	17.289*	-	-
	Slow-pacing	93	0.302	0.050	(<i>d</i> = 0.796)		
8	Fast-pacing	76	0.595	0.057	2.837	-	-
	Slow-pacing	50	0.436	0.072	(<i>d</i> = 0.317)		
9	Fast-pacing	49	0.861	0.051	3.832*	-	-
	Slow-pacing	99	0.725	0.045	(<i>d</i> = 0.324)		
10	Fast-pacing	14	0.424	0.138	0.378	0.542	0.028
	Moderate-pacing	71	0.536	0.060	(<i>d</i> = 0.186)	(<i>d</i> = 0.221)	(<i>d</i> = 0.031)
10	Fast-pacing	50	0.520	0.073			
	Slow-pacing	22	0.458	0.111	1.062	0.435	0.263
10	Fast-pacing	67	0.372	0.061	(<i>d</i> = 0.266)	(<i>d</i> = 0.171)	(<i>d</i> = 0.093)
	Slow-pacing	58	0.326	0.064			

Note. *n* = number of test-takers assigned to the profiles; *d* = Cohen's *d* effect size.

**p* < .05.

study), our results showed that there may be two (fast- and slow-pacing) or three (fast-, moderate-, and slow-pacing) response processing subgroups, which differed in process times across all AOIs. Moreover, the results revealed that the slow-pacing groups spent more time on the choices compared to the stem AOIs, whereas the fast-pacing groups allocated similar amounts of time to both the choices and the stem AOIs. Overall, these findings indicated that the subgroups differed not only in the total time spent on items but also in how they distributed their time across different parts of the item, reflecting distinct cognitive approaches to response processing. Although the influences of item-related factors on process times and processing groups remain relatively unexplored, two processing groups are often distinguished in the literature. For example, Hicks (2019), using process data such as response time and the number of highlight events to examine reading strategies in the PIAAC testlets, reported two subgroups: one group that progressed through items quickly, resembling a skimming strategy, and another group that took longer, suggesting a full-reading strategy. Also, Evans and Stanovich (2013) describe a dual-process, which distinguishes between

fast, intuitive, and automatic processing (Type 1) and slow, reflective, and controlled processing (Type 2).

The item processing subgroups described in the present study nicely align with the two-group findings, and further present that, depending on the psychometric properties of the test items used, a 'moderate-processing' processing subgroup may emerge as a third group. We argue that the three-group structure, at least in part, might be related to how test-takers with varying performance levels might be engaging different modes of response processing when responding to items with varied psychometric properties, such as those with high versus low difficulty levels or other specific properties such as word count. The presence of a moderate-pacing group may reveal that some test-takers might be adopting a flexible processing strategy depending on an item's perceived difficulty and complexity (e.g., Nazareth et al., 2019; Mézière et al., 2023).

The findings of this study warrant future research investigating underlying cognitive strategies when processing items with varying psychometric properties and length. Longitudinal modeling (Çorbacı and Kahraman, 2022) as well as qualitative methods such as think-aloud

Table 5. Equality Tests of Mean Total Scores Across Profiles (Model-c).

Items	Profiles	n	Mean total score	Std. error	χ^2 for paired comparisons		
					Fast vs. slow	Fast vs. moderate	Moderate vs. slow
1	Fast-pacing	88	6.771	0.313	17.736*	-	-
	Slow-pacing	64	4.657	0.364	(<i>d</i> = 0.723)		
2	Fast-pacing	25	8.015	0.349	16.907*	21.930*	0.016
	Moderate-pacing	70	5.567	0.387	(<i>d</i> = 0.962)	(<i>d</i> = 0.837)	(<i>d</i> = 0.026)
3	Slow-pacing	37	5.483	0.507			
	Fast-pacing	26	7.651	0.375	17.430*	10.567*	2.088
4	Moderate-pacing	76	5.955	0.353	(<i>d</i> = 0.910)	(<i>d</i> = 0.599)	(<i>d</i> = 0.288)
	Slow-pacing	44	5.044	0.497			
5	Fast-pacing	55	7.335	0.319	33.346*	9.675*	9.532*
	Moderate-pacing	68	5.757	0.377	(<i>d</i> = 1.547)	(<i>d</i> = 0.563)	(<i>d</i> = 0.693)
6	Slow-pacing	18	3.705	0.539			
	Fast-pacing	50	6.317	0.412	0.995	-	-
7	Slow-pacing	103	5.798	0.300	(<i>d</i> = 0.173)		
	Fast-pacing	46	6.766	0.452	5.221*	-	-
8	Slow-pacing	93	5.479	0.310	(<i>d</i> = 0.427)		
	Fast-pacing	76	5.967	0.338	0.585	-	-
9	Slow-pacing	50	5.528	0.450	(<i>d</i> = 0.144)		
	Fast-pacing	49	7.779	0.237	48.082*	-	-
10	Slow-pacing	99	5.006	0.315	(<i>d</i> = 1.013)		
	Fast-pacing	14	5.992	0.702	0.092	0.096	0.000
11	Moderate-pacing	71	6.240	0.369	(<i>d</i> = 0.085)	(<i>d</i> = 0.082)	(<i>d</i> = 0.000)
	Slow-pacing	50	6.240	0.421			
12	Fast-pacing	22	6.607	0.661	1.248	0.468	0.413
	Moderate-pacing	67	6.089	0.343	(<i>d</i> = 0.275)	(<i>d</i> = 0.180)	(<i>d</i> = 0.119)
13	Slow-pacing	58	5.730	0.424			

Note. *n* = number of test-takers assigned to the profiles; *d* = Cohen's *d* effect size.

**p* < .05.

protocols (Noushad et al., 2023; Stark et al., 2024) could provide deeper insights into response processing behavior. Additionally, test-takers with higher metacognitive awareness may adapt their strategies to enhance performance while those with lower metacognitive awareness may struggle to regulate their approach and achieve lower scores (Bax, 2015; Fritzsche et al., 2018). Assessing metacognitive awareness through questionnaires or interviews could help determine whether specific response processes are associated with better self-regulation and adaptive test-taking behaviors.

Pertaining to the comparison of item and test scores across response processing groups (i.e., the second and third research questions of the study), our results indicated that test-takers in the fast-pacing group had higher scores compared to the other groups. Previous studies comparing eye movements of high- and low-performing test-takers support that high-performing test-takers tend to have shorter response times and fewer AOI fixations (e.g., Bartolomé et al., 2020) and exhibit longer fixations with fewer visits to problem information areas, resulting in shorter overall test durations (e.g., Y. Hu et al., 2017). Regarding the potential causes of differences, previous research has provided important insights. For example, Bax's study found that successful test-takers tended to

use conscious strategies to locate relevant information in the text, whereas unsuccessful test-takers struggled to do so efficiently (Bax, 2013). Similarly, Tsai et al. (2012) and Krstić et al. (2018) concluded that high-performing test-takers were better able to recognize and focus on relevant information necessary for correct responses. In contrast, low-performing test-takers had difficulties distinguishing and prioritizing relevant details, leading them to search through the item for clues rather than processing it strategically. Further supporting this, Yaneva et al. (2022) examined response processes in multiple-choice items and reported that incorrect responses were associated with a reverse processing order, where test-takers worked from the answer options to the stem rather than the other way around. Conversely, correct responses were linked to processing the stem first, spending more time carefully reading the problem, and making a more decisive selection of a response option.

Analyses incorporating test-takers' performance levels may help explain why test-takers in the slow-pacing group tended to answer incorrectly and had lower total test scores. Slow-pacing test-takers might struggle to locate relevant clues in the items or fully comprehend the stem due to unknown vocabulary or complex sentence structures. As a result, they may engage in a trial-and-

error approach, searching for the correct answer among the alternatives. The observed relationship between fast-pacing and higher test scores may reflect differences in processing efficiency, defined as the relationship between performance effectiveness and the amount of resources invested in processing (Eysenck & Calvo, 1992). High-performing test-takers may process information more effectively, enabling them to answer correctly in a shorter time. Additionally, test-takers in the fast-pacing group may exhibit better metacognitive awareness (Bax, 2015), allowing them to allocate cognitive resources more effectively. In contrast, slow-pacing test-takers may experience higher levels of anxiety (Derakshan & Eysenck, 2009), which can negatively impact performance by increasing cognitive load and response times. Efficiency advantage among fast-pacing test-takers could stem from better retrieval of relevant knowledge or more efficient problem-solving strategies. However, it is important to note that fast-pacing might not always be associated with efficient cognitive processing. As we defined response processing using only fixation-based metrics and did not include effort- or engagement-related variables, some instances of fast processing could reflect skipping or guessing without fully focusing on or comprehending the item. Indeed, fast-pacing may be an indicator of low effort and disengagement (Wise & Kong, 2005), warranting further study using engagement metrics.

Pertaining to the evaluation of the common characteristics of items with similar processing subgroups (i.e., the fourth research question of the study), our results showed that participants' response processing patterns might change depending on the items' psychometric properties and that the latent class structure underlying item processing behavior may reveal two or three subgroups. Items with two subgroups were typically moderate or low difficulty, with moderately difficult items tending to be very short and easier items relatively longer. In contrast, items with three subgroups covered a wider range of difficulty levels, from difficult to easy, and included both short items and much longer ones. Overall, these findings indicated that both difficulty and length contribute to subgroup variation, suggesting that the interaction of item-related factors may drive differences in response processing patterns. Longer items may encourage a more thorough reading strategy, while shorter items may be skimmed more quickly (e.g., Gerth & Festman, 2021). Beyond length, the complexity of an item—for instance, the presence of unknown words or complex sentences—might play a role in determining processing behaviors (Bartolomé et al., 2025). Such characteristics may increase demand in cognitive resources, leading to slower processing and potentially lower accuracy.

Pertaining to the evaluation of the common characteristics of items with observed score differences across groups (i.e., the fifth research question of the study), our results showed that items with high discrimination and moderate or low difficulty were most likely to reveal score differences across processing subgroups, with fast-pacing test-takers generally outperforming others. In contrast, some items did not differentiate among groups. For instance, one of the easiest items included weak distractors and an explicit clue to the correct answer (item 5), while others were among the shortest in the set, causing limited textual content (items 7 and 9). Another was the longest and most difficult item, but showed relatively low discrimination, meaning that processing differences did not translate into performance differences (item 10). These results suggest that item-related factors play a non-trivial role in the differences observed among item processing subgroups, with weak distractors, short text, or low item discrimination. Particularly, highly discriminating items are effective not only in differentiating test-takers based on their knowledge but also in distinguishing their response processing patterns. Overall, these findings reveal that response processing differences may nicely align with performance differences depending on the item's characteristics, such as difficulty, discrimination, and length, suggesting that item-related factors might be driving the variation observed in item processing patterns (e.g., Akgün & Ünalı, 2022; Arslan et al., 2020).

Inferences derived from a joint analysis of item and performance-related sources of differing item processing behavior carry important implications for both psychometric evaluation and educational practice. From a psychometric perspective, test developers can use item-level process indicators (e.g., time spent on stems vs. options or on distractors) to assess item quality, revise the items if necessary, and create more discriminative items that better reveal individual differences. From an educational standpoint, the findings point to response processing as a potential target for instructional support. Slow-pacing groups, for example, may benefit from strategy training in time management and efficient reading, such as techniques for locating key information and allocating time across item components. Process data can also be used to deliver individualized feedback that fosters metacognitive awareness and self-regulation. These data-driven insights, in turn, can guide the development of personalized training programs aligned with learners' individual processing profiles. Ultimately, leveraging eye-tracking data in this way can enhance both the validity of assessments and the effectiveness of instructional practices, contributing to more equitable and supportive learning and testing environments.

Limitations and Future Directions

While this study provides valuable insights into test-takers' response processing patterns, there are some caveats to consider. The results presented may not be fully generalizable due to the limited number of test items with specific characteristics, such as item difficulty and discrimination. Additionally, this study relied solely on fixation-based eye-tracking data, which may not fully capture the complexity of response processes. Future research could integrate both fixation- and saccade-based data to provide a more detailed understanding of how test-takers interact with multiple-choice items. Expanding the study to include a wider range of item characteristics, such as very difficult, low-discriminating, and extremely short or long items, would further enhance the findings.

Also, this study primarily focused on eye-tracking data, but incorporating multiple measures—such as neuroimaging techniques and self-report questionnaires—could offer a more comprehensive perspective on cognitive and emotional factors influencing response processes. Future research could also explore the role of affective variables, such as test anxiety and motivation, by incorporating physiological measures (e.g., heart rate monitoring) and self-report scales to assess their relationship with response processes and test performance. While the sample size was sufficient for the analyses conducted, future studies could benefit from a larger and more diverse participant pool, including test-takers with varying levels of reading comprehension proficiency. A broader sample would provide a more generalizable understanding of response processing patterns across different proficiency levels and demographic groups.

Moreover, all participants in this study were native Turkish speakers and learners of English as a foreign language. Although they were enrolled in different departments, all were undergraduate students from the Faculty of Education. While their sociocultural and linguistic backgrounds may appear relatively similar, some variability is inevitable. A further limitation of the study is that sociocultural factors and linguistic variables—such as native and foreign language proficiency—were not assessed. Previous research has emphasized that reading is a multifaceted construct, encompassing a wide range of competencies that collectively shape this skill (Koda, 2005). Therefore, unmeasured variables such as cultural background, language proficiency, and other linguistic characteristics may have influenced the response processes. Future studies would benefit from incorporating these variables to gain a deeper understanding of how test-taking strategies vary across linguistic and cultural contexts. In addition, future research could move beyond

exploratory questions by incorporating theory-driven hypotheses (e.g., Processing Efficiency Theory; Eysenck & Calvo, 1992), which would help to clarify how individual differences in response processes are shaped by underlying cognitive mechanisms.

Conclusion

This study uses latent profile analysis to address an important issue concerning the use of eye movement data to identify item response processing patterns in test-takers. The results indicated that two or three response processing patterns (fast-, moderate-, and slow-pacing) emerged depending on the difficulty and length of the items. Furthermore, the fast-pacing group exhibited higher test performance than others, particularly for highly discriminating and moderately difficult or easy items. The presented results suggest that although often overlooked, the psychometric properties of test items might indeed influence test-takers' item processing behavior. In particular, key characteristics such as difficulty, discrimination, and length play a crucial role in shaping the underlying response processing patterns. Our findings support that response processing patterns and their relationship with observed scores provide valuable insights into qualitative differences in multiple-choice item processing behaviors. Overall, our findings offer invaluable insights into response processing behavior, serving as an important source of evidence for evaluating the validity of inferences drawn from observed test scores (X. Hu & Aryadoust, 2024; Messick, 1995; Sousa et al., 2025). Beyond research implications, such analyses can support practical applications, including the development of items that more effectively reveal individual differences, the identification of distinct test-taker processing profiles, and the design of targeted feedback or instructional interventions that address specific processing behaviors. Studies aiming to relate specific eye movement patterns to proficiency measures of interest could greatly benefit from a psychometric analysis of choice to ensure the characteristics of the items used in test applications are well-suited for the intended inferences.

ORCID iDs

Derya Akbaş  <https://orcid.org/0000-0001-9852-4782>

Ergun Cihat Çorbacı  <https://orcid.org/0000-0002-7874-956X>

Nilüfer Kahraman  <https://orcid.org/0000-0003-2523-0155>

Ethical Considerations

The experimental methods were approved by the Gazi University Ethics Committee (No: 80287700-302.08.01-156936).

Consent to Participate

Informed consent was obtained from all participants before the study.

Author Contributions

Derya Akbaş, conceptualization; methodology; formal analysis; visualization; writing – original draft: Ergun Cihat Çorbacı, conceptualization; investigation; data curation; writing – review and editing: Nilufer Kahraman, conceptualization; data curation; funding acquisition; writing – review and editing.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This study was partially supported by the Scientific and Technological Research Council of Türkiye (TUBITAK) (Grant number 120K142).

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Data Availability Statement

The data that support the findings of this study are not publicly available.

GenAI Use Disclosure Statement

GenAI Use Disclosure Statement During the preparation of this manuscript, the authors used ChatGPT to assist with language editing and improving the clarity of the text. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the final version of the manuscript.

References

- Akaike, H. (1987). Factor analysis and AIC. *Psychometrika*, 52, 317–332. <https://doi.org/10.1007/BF02294359>
- Akgün, H., & Ünalı, A. (2022). Investigating cognitive processes in different item formats in reading tests through eye-tracking and verbal protocols. *Studies in Language Assessment*, 11(2), 98–129. <https://doi.org/10.58379/UIXK2220>
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (Eds.) (2014). *Standards for educational and psychological testing*. American Educational Research Association.
- Arslan, B., Jiang, Y., Keehner, M., Gong, T., Katz, I. R., & Yan, F. (2020). The effect of drag-and-drop item features on test-taker performance and response strategies. *Educational Measurement: Issues and Practice*, 39(2), 96–106. <https://doi.org/10.1111/emip.12326>
- Asparouhov, T., & Muthén, B. (2014). Auxiliary variables in mixture modeling: Three-step approaches using Mplus. *Structural Equation Modeling: A Multidisciplinary Journal*, 21(3), 329–341. <https://doi.org/10.1080/10705511.2014.915181>
- Bartolomé, J., Garaizar, P., & Bastida, L. (2020). *Validating item response processes in digital competence assessment through eye-tracking techniques*. In *Eighth International Conference on Technological Ecosystems for Enhancing Multiculturality*. <https://doi.org/10.1145/3434780.3436641>
- Bartolomé, J., Garaizar, P., Loizaga, E., & Bastida, L. (2025). Using eye-tracking data to examine response processes in digital competence assessment for validation purposes. *Applied Sciences*, 15(3), 1215. <https://doi.org/10.3390/app15031215>
- Bax, S. (2013). The cognitive processing of candidates during reading tests: Evidence from eye-tracking. *Language Testing*, 30(4), 441–465. <https://doi.org/10.1177/0265532212473244>
- Bax, S. (2015). *Using eye-tracking to research the cognitive processes of multinational readers during an IELTS reading test*. *IELTS Research Reports Online Series*, 2, 1–21.
- Bax, S., & Chan, S. (2019). Using eye-tracking research to investigate language test validity and design. *System*, 83, 64–78. <https://doi.org/10.1016/j.system.2019.01.007>
- Carter, B. T., & Luke, S. G. (2020). Best practices in eye tracking research. *International Journal of Psychophysiology*, 155, 49–62. <https://doi.org/10.1016/j.ijpsycho.2020.05.010>
- Clark, S. L. (2010). *Mixture modeling with behavioral data*. [Doctoral dissertation, University of California]. ProQuest Dissertations & Theses Global.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Routledge Academic.
- Çorbacı, E. C., & Kahraman, N. (2022). Latent growth modeling of item process data derived from eye-tracking technology: An experimental study investigating reading behavior of examinees when answering a multiple-choice test item. *Journal of Measurement and Evaluation in Education and Psychology*, 13(3), 194–211. <https://doi.org/10.21031/epod.1107597>
- Crocker, L., & Algina, J. (2006). *Introduction to classical and modern test theory*. Cengage Learning.
- Derakshan, N., & Eysenck, M. W. (2009). Anxiety, processing efficiency, and cognitive performance: New developments from attentional control theory. *European Psychologist*, 14(2), 168–176. <https://doi.org/10.1027/1016-9040.14.2.168>
- Ercikan, K., & Pellegrino, J. W. (2017). Validation of score meaning using test-taker response processes for the next generation of assessments. In K. Ercikan & J.W. Pellegrino (Eds.), *Validation of score meaning for the next generation of assessments* (pp. 1–8). Routledge. <https://doi.org/10.4324/9781315708591>
- Evans, J. S. B., & Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science*, 8(3), 223–241. <https://doi.org/10.1177/1745691612460685>
- Eysenck, M. W., & Calvo, M. G. (1992). Anxiety and performance: The processing efficiency theory. *Cognition and Emotion*, 6(6), 409–434. <https://doi.org/10.1080/02699939208409696>

- Ferguson, S. L. G., Moore, E. W., & Hull, D. M. (2020). Finding latent groups in observed data: A primer on latent profile analysis in Mplus for applied researchers. *International Journal of Behavioral Development*, 44(5), 458–468. <https://doi.org/10.1177/0165025419881721>
- Fritzsche, E. S., Händel, M., & Kröner, S. (2018). What do second-order judgments tell us about low-performing students' metacognitive awareness?. *Metacognition Learning*, 13, 159–177. <https://doi.org/10.1007/s11409-018-9182-9>
- Gerth, S., & Festman, J. (2021). Reading development, word length and frequency effects: An eye-tracking study with slow and fast readers. *Frontiers in Communication*, 6, 743113. <https://doi.org/10.3389/fcomm.2021.743113>
- Hickendorff, M., Edelsbrunner, P. A., McMullen, J., Schneider, M., & Trezise, K. (2018). Informative tools for characterizing individual differences in learning: Latent class, latent profile, and latent transition analysis. *Learning and Individual Differences*, 66, 4–15. <https://doi.org/10.1016/j.lindif.2017.11.001>
- Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., & Van de Weijer, J. (2011). *Eye tracking: A comprehensive guide to methods and measures*. Oxford University Press. <https://lup.lub.lu.se/record/1852359>
- Hu, X., & Aryadoust, V. (2024). A systematic review of eye-tracking technology in second language research. *Languages*, 9(4), 141. <https://doi.org/10.3390/languages9040141>
- Hu, Y., Wu, B., & Gu, X. (2017). An eye tracking study of high- and low-performing students in solving interactive and analytical problems. *Journal of Educational Technology & Society*, 20(4), 300–311. <http://www.jstor.org/stable/26229225>
- Hyönä, J. (2010). The use of eye movements in the study of multimedia learning. *Learning and Instruction*, 20(2), 172–176. <https://doi.org/10.1016/j.learninstruc.2009.02.013>
- Hicks, J. C. (2019). *The use of process data to examine reading strategies*. [Doctoral dissertation, The University of North Carolina]. ProQuest Dissertations & Theses Global.
- Jarodzka, H., Holmqvist, K., & Gruber, H. (2017). Eye tracking in educational science: Theoretical frameworks and research agendas. *Journal of Eye Movement Research*, 10(1), 1–18. <https://doi.org/10.16910/jemr.10.1.3>
- Just, M. A., & Carpenter, P. A. (2018). Using eye fixations to study reading comprehension. In D. E. Kieras & M. A. Just (Eds.), *New methods in reading comprehension research* (pp. 151–182). Routledge. <https://doi.org/10.4324/9780429505379>
- Kane, M., & Mislevy, R. (2017). Validating score interpretations based on response processes. In K. Ercikan & J. W. Pellegrino (Eds.), *Validation of score meaning for the next generation of assessments The use of response processes* (pp. 11–24). Taylor & Francis. <https://doi.org/10.4324/9781315708591>
- Koda, K. (2005). *Insights into second language reading: A cross-linguistic approach*. Cambridge University Press.
- Krstić, K., Šoškić, A., Ković, V., & Holmqvist, K. (2018). All good readers are the same, but every low-skilled reader is different: an eye-tracking study using PISA data. *European Journal of Psychology of Education*, 33, 521–541. <https://doi.org/10.1007/s10212-018-0382-0>
- Langenfeld, T., Thomas, J., Zhu, R., & Morris, C. A. (2020). Integrating multiple sources of validity evidence for an assessment-based cognitive model. *Journal of Educational Measurement*, 57(2), 159–184. <https://doi.org/10.1111/jedm.12245>
- Lo, Y., Mendell, N. R., & Rubin, D. B. (2001). Testing the number of components in a normal mixture. *Biometrika*, 88(3), 767–778. <https://doi.org/10.1093/biomet/88.3.767>
- Maddox, B., Bayliss, A. P., Fleming, P., Engelhardt, P. E., Edwards, S. G., & Borgonovi, F. (2018). Observing response processes with eye tracking in international large-scale assessments: Evidence from the OECD PIAAC assessment. *European Journal of Psychology of Education*, 33, 543–558. <https://doi.org/10.1007/s10212-018-0380-2>
- McCray, G., & Brunfaut, T. (2018). Investigating the construct measured by banked gap-fill items: Evidence from eye-tracking. *Language Testing*, 35(1), 51–73. <https://doi.org/10.1177/0265532216677105>
- McLachlan, G., & Peel, D. (2000). *Finite mixture models*. John Wiley & Sons.
- Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50(9), 741.
- Mézière, D. C., Yu, L., Reichle, E. D., Von Der Malsburg, T., & McArthur, G. (2023). Using eye-tracking measures to predict reading comprehension. *Reading Research Quarterly*, 58(3), 425–449. <https://doi.org/10.1002/rrq.498>
- Muthén, L., & Muthén, B. (1998–2017). *Mplus user's guide* (8th ed.). Muthén & Muthén.
- Nazareth, A., Killick, R., Dick, A. S., & Pruden, S. M. (2019). Strategy selection versus flexibility: Using eye-trackers to investigate strategy use during mental rotation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 45(2), 232–245. <https://doi.org/10.1037/xlm0000574>
- Noushad, B., Van Gerven, P. W. M., & de Bruin, A. B. H. (2023). Twelve tips for applying the think-aloud method to capture cognitive processes. *Medical Teacher*, 46(7), 892–897. <https://doi.org/10.1080/0142159X.2023.2289847>
- Nylund-Gibson, K., & Choi, A. Y. (2018). Ten frequently asked questions about latent class analysis. *Translational Issues in Psychological Science*, 4(4), 440–461. <https://doi.org/10.1037/tps0000176>
- Nylund-Gibson, K., Grimm, R., Quirk, M., & Furlong, M. (2014). A latent transition mixture model using the three-step specification. *Structural Equation Modeling: A Multidisciplinary Journal*, 21(3), 439–454. <https://doi.org/10.1080/10705511.2014.915375>
- Rayner, K. (1995). Eye movements and cognitive processes in reading, visual search, and scene perception. In J. M. Findlay, R. Walker, & R. W. Kentridge (Eds.), *Studies in visual information processing* (pp. 3–22). North-Holland.
- Sousa, F., Soares, A. C. B., Fontes, F. L. D. L., Oliveira, G. C., Arruda, L. S. N. D. S., & Borges, J. W. P. (2025). Use of eye tracking in instrument construction and/or validity studies: Scoping Review. *Journal of Measurement and Evaluation in Education and Psychology*, 16(2), 88–114. <https://doi.org/10.21031/epod.1569440>

- Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2), 461-464. <http://www.jstor.org/stable/2958889>
- Stark, L., Korbach, A., Brünken, R., & Park, B. (2024). Measuring (meta)cognitive processes in multimedia learning: Matching eye tracking metrics and think-aloud protocols in case of seductive details. *Journal of Computer Assisted Learning*, 40(6), 2985-3004. <https://doi.org/10.1111/jcal.13051>
- The MathWorks Inc. (2023). *MATLAB version: 23.2 (R2023b)*, Natick, Massachusetts: The MathWorks Inc. <https://www.mathworks.com>
- The Measurement, Selection and Placement Center (2018). Foreign Language Examination (Yabancı Dil Bilgisi Seviye Tespit Sınavı (YDS)). <https://dokuman.osym.gov.tr/pdfdokuman/2018/YDSILKBAHAR/CS/ING24072018.pdf>
- Tsai, M. J., Hou, H. T., Lai, M. L., Liu, W. Y., & Yang, F. Y. (2012). Visual attention for solving multiple-choice science problem: An eye-tracking analysis. *Computers & Education*, 58(1), 375-385. <https://doi.org/10.1016/j.compedu.2011.07.012>
- Vermunt, J. K. (2010). Latent class modeling with covariates: Two improved three-step approaches. *Political Analysis*, 18(4), 450-469. <https://doi.org/10.1093/pan/mpq025>
- Vuong, Q. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica*, 57(2), 307-333. <https://www.jstor.org/stable/1912557>
- Was, C., Sansosti, F., & Morris, B. (Eds.). (2016). *Eye-tracking technology applications in educational research*. IGI Global.
- Yaneva, V., Clauser, B. E., Morales, A., & Paniagua, M. (2021). Using eye-tracking data as part of the validity argument for multiple-choice questions: A demonstration. *Journal of Educational Measurement*, 58(4), 515-537. <https://doi.org/10.1111/jedm.12304>
- Yaneva, V., Clauser, B. E., Morales, A., & Paniagua, M. (2022). Assessing the validity of test scores using response process data from an eye-tracking study: a new approach. *Advances in Health Sciences Education*, 27, 1401-1422. <https://doi.org/10.1007/s10459-022-10107-9>
- Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education*, 18(2), 163-183. https://doi.org/10.1207/s15324818ame1802_2
- Zumbo, B. D., & Hubley, A. M. (Eds.). (2017). *Understanding and investigating response processes in validation research*. Springer.