ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
Research Article; 2022; 13(3); 194-211

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

Ergün Cihat ÇORBACI* Nilüfer KAHRAMAN**

Abstract

This study illustrates how eye-tracking data can be translated to "item process data" for multiple-choice test items to study the relationship between subjects' item responses and choice reading behavior. Several modes of analysis were used to test the hypothesized added value of using process data to identify choice reading patterns of subjects. In addition to the cross-sectional analises of agggregate measurements derived from the time series eye tracking data, Latent Growth Curve Model Analises were undertaken to test if the the shape of change observed in the sequential choice reading patterns differed for subjects depending on their responses to the item being correct or incorrect. Application data were from an experimental study and included seventy-one subjects' responses to two multiple-choice test items measuring reading comprehension ability in English as a second language. Analyses were carried out for one item at a time. For each item, first, each subject's recorded eye movements were coded into a set of Area of Interests (AOIs), segmenting the lines in the stem and the individual choices. Next, each subject's fixation times on the AOIs were time stamped into seconds, indicating when and in what order each subject's gaze had fixated on each AOI until a choice was marked as the correct answer, which ended the item encounter. A set of nested Latent Growth Curve models were considered for the choice-related AOIs to deliniate if distinct choice-process sequences were evident for correct and incorrect respoders. Model fit indices, random intercepts, slopes, and residuals were computed using the mean log fixation times over item encounter time. The results show that the LGM with the best model fit indicies, for both items, was the quadratic model using response variable as a covariate. Albeit limited due to the two-item – seventy-one subjects experimental setting of the study, the findings are promising and show that utilizing item-level process data can be very useful for defining distinct choice processing (task-oriented reading) patterns of examinees. Over all, the results warrant further study of choice derived AOIs using longitudinal statistical models. It is argued that, the screening methodology desribed in this study can be a useful tool to investigate speededness, distractor functioning, or even to flag subjects with irregular choice processing behavior, such as providing a direct mark on a choice, without any significant reading activity on any of the choices presented (i.e., whether cheating might have occurred.)

Keywords: Latent Growth Curve Modeling, eye-tracking, reading ability in English, multiple-choice items.

Introduction

The eye-tracking technology has been widely used for investigating how individuals read words or sentences and whether tracking their reading behavior while reading can be helpful to understand the cognitive processes functioning (Rayner, 1998). However, the use of time series eye-tracking data to improve educational assessment settings, where examinees are to answer questions given a text that is specifically constructed to measure reading comprehension ability (i.e., task-oriented reading), has been neglected to a great extent, which can potentially support and enrich reliability and validity studies

* Research Assistant, Gazi University, Faculty of Education, Ankara-Turkey, e.cihat.corbaci@gmail.com, ORCID ID: 0000-0002-7874-956X
** Professor Doctor, Gazi University, Faculty of Education, Ankara-Turkey, nkahraman@gazi.edu.tr, ORCID ID: 0000-0003-2523-0155

To cite this article:
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

Received: 22.04.2022
Accepted: 25.08.2022

focusing on various measurement processes (Solheim & Uppstad, 2011). As the Standards for Educational and Psychological Testing (American Educational Research Association, et al., 2014) states "The process of validation involves accumulating relevant evidence to provide a sound scientific basis for the proposed score interpretations". Eye movement data can be useful in the study of how examinees process the choices of a multiple-choice test item, that is, before providing a response (True/false). Identifying response processes patterns through measures derived from process data such as gaze-movements on a screen, although indirectly, could reveal item and domain specific features of test scores and uses. Moreover, different test-taker groups such as competent and incompetent test-takers may respond to a multiple-choice question using different patterns, which can be used to verify or falsify a proposed interpretation (Kane & Mislevy, 2017).

There are relatively few studies, in the literature, that underline the importance of investigating assessment-related aspects of such inquiries. Paulson and Henry (2002), for example, used eye-tracking movements to scrutinize claims (measure the reading comprehension process) asserted by the publisher on a reading assessment (Degrees of Reading Power, DRP) and investigated the reading processes of test-takers while taking DRP. They used a modified cloze setup of DRP that was intended to measure the process of reading by responding to the comprehension questions at the end of the passage. Tai et al. (2006) also used eye-tracking movements to investigate problem-solving behaviors within a group of subjects in three different disciplines while solving standardized multiple-choice questions. They analyzed the location of eye-gaze fixation, duration of fixation, scan paths, and duration between fixations as well as correct responses and latent response times, which consist of both quantitative and qualitative data analysis. Solheim and Uppstad (2011) also used eye-tracking to investigate problem-solving behaviors using a stimulus text comprised of both a verbal text and an illustration. They related correct-answer scores to gaze movement patterns arguing that subjects' gaze movement behavior revealed subjects' reading behavior. Tsai et al. (2012) examined students' visual attention when solving a multiple-choice science problem using an eye-tracker. They divided students into two groups: the high-score group and the low-score group, and unlike other studies, they investigated choices (distractors and the correct choice) in the multiple-choice questions. In addition, they stated that students paid more attention to the options they chose and to relevant areas and paid little attention to the irrelevant areas. Yaneva et al. (2022) demonstrated how to use multiple-choice questions to collect evidence for validity argument. They investigated how the presence of options in the multiple-choice question affect the response behavior of the students, what areas of item they viewed first and whether the options were processed in the same way, discussing validity inferences. Overall, considering the studies in the literature, it can be seen that measurement-based approaches used in education include traditional reading and task-oriented readings. However, it is critical that there is a methodological perspective that can assist in demonstrating the validity and reliability of such approaches.

To this end, this paper proposes and illustrates a two-stage methodology highlighting the necessity of an in-synch multi-stage data processing approach when integrating eye-tracking technology-derived (response-related) data into the conventional psychometric analysis that most often uses response data alone. Formulated to place a special emphasis on a data screening stage to be carried out prior to the actual data analysis stage, the proposed methodology is illustrated using real response data collected for a couple of multiple-choice items from a test measuring college students' reading comprehension ability in English as a Foreign Language. The application presented helps demonstrate that it pays off to investigate the inner-connectedness of research questions to the information available in the eye-tracking (device recordings, i.e., gaze durations and movements coded in milliseconds) and the conventional response data (i.e., given a question, markings for the correct choice given choices from A to E), as a priory to the final data analysis stage. It is argued that the stronger relationship is between the measurement variables created out of eye-tracking data and the desired interpretations, the easier it will be to make inferences about the findings for the integrated response process data (eye-tracking recordings + item responses). Underlining the importance of using a psychometric perspective when analyzing eye-tracking-aided item response data, the ultimate purpose of this study is to provide several modeling strategies that can help researchers capture construct-related information that might be available in eye-tracking data and to test the meaningfulness of its added value.

Method

Data

Application data were from a test experiment using items from a multiple-choice reading comprehension test in English. The data set included both the conventional item response data, i.e., the choice marked by the subjects as the correct answer (0/1), and the eye-tracked process data, i.e., fixation durations and sequences over item-encounter time given the area of interests, AOIs, for two separate multiple-choice test items. The test experiment, its subjects and how the AOIs were defined are described in the following text.

Test Experiment and Subjects

The eye-tracking apparatus used for data collection was the Tobii TX300 screen-based eye tracker, which performed binocular tracking with a sampling frequency 300Hz (Dell Desktop Computer, Intel Core i7-4790 @3.60 GHz, 16,0 GB of RAM). Seventy-one subjects took part in the test experiment. The subjects were non-native speakers of English. Each examinee took the exam individually and responded to test items appearing on a computer screen one at a time, in the same order. The items were from the released items of the Foreign (English) Language Exam that was administered by the Student Selection and Placement Center (OSYM) in 2018. This exam is held twice a year and consists of 80 multiple-choice items that includes different contents such as vocabulary, grammar, translation etc. to evaluate foreing language skills. In addition, this test evaluates only reading comprehension skills, which means there is no questions related to other skills such as listening, speaking and writing.

Processing the Eye-Tracking Data

In the test experiment, the layout on the test screen placed the stem part of the item on the left side of the screen, while the choices were placed on the right side. Figure 1 shows the ten areas of interest (AOI) used for Item 1: Direction, Line 1, Line 2, Line 3, Line 4, Choices A, B, C, D, and E. Figure 2 shows the nine AOIs used for Item 2: Direction, Line 1, Line 2, Line 3, Choices A, B, C, D, and E. As Figure 1 (Item 1-10 AOIs) and Figure 2 (Item 2 - 9 AOIs) illustrate, within the context of multiple-choice test items, unlike the AOIs for the choices A to E, the AOIs in the stem may differ drastically from one item to another due to item-specific features, such as the number of lines included or the number of words included in each line.

Figure 1

The Areas of Interest for the Multiple-Choice Item 1

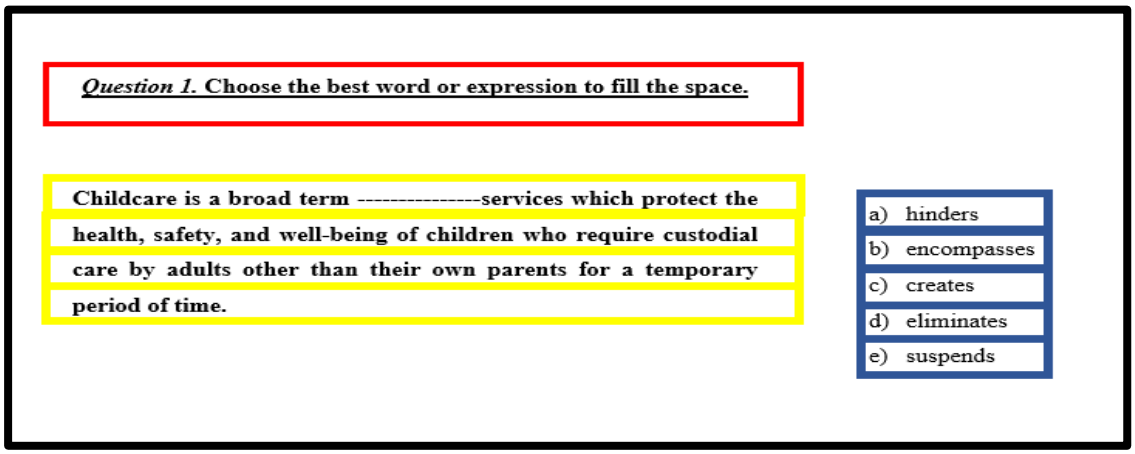
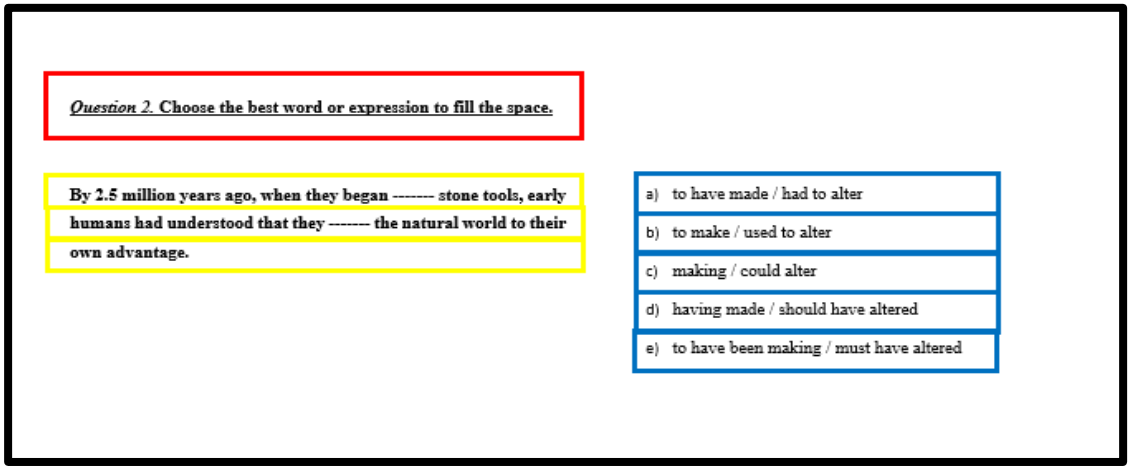


Figure 2

The Areas of Interest for a Multiple-Choice Item 2



In the data processing stage of this study, the AOIs presented in the corresponding figures above were used to binary code the eye-tracking data for each item. The data were exported using the defined item-AOIs: 300 rows of data per second (the data were collected with the sampling frequency of 300Hz, where a piece of new information was collected in each 3.34 ms). Considering that each student spent approximately one minute answering an item, a raw data file of about 18000 rows was obtained per item subject encounter.

Next, the data collected in the item experiments were screened to detect and remove corrupt, irrelevant, and inaccurate recordings from the dataset composed of (1) response (correct-incorrect coding, 1-0), and fixation durations over (2) lines within item stems (text part of the items) and (3) choices A to E. One of the fundamental problems in eye-tracking data is the desynchronization between gaze and stimuli resulting from poor calibration or visual impairment. To overcome this problem, as a first step, subjects who never looked at the relevant areas and therefore did not have any eye movements in these areas were determined through a careful analysis of the corresponding time graphs. In addition, the subjects with gaze-stimulus mismatches were identified by examining each examinee's scan paths one by one. For example, Figure 3 below shows the scan path, and Figure 4 shows the heatmap of a subject's encounter with an item illustrating how the desynchronization between the subject's gaze and stimuli might occur, rendering the validity of the information available in the resulting eye-tracking process data for this subject.

The dots, in Figure 3, show where the subject is looking (fixations), and the numbers in these dots show the order in which an individual looks/fixates, and the lines connecting these dots show the transitions (saccades). As one can see, the subject looked at the stem of the question, but the gaze movements were lower. Similarly, the subject actually looked at choice A, but did not seem to look at it because their eye movements were shifted downwards.

Figure 4 shows where the subject is looking and focusing. As in the scan path, the subject looked at the stem of the question, but the gaze movements were lower. Similarly, the subject actually looked at choice A, but did not seem to look at it because their eye movements were shifted downwards. In general, the subject's reading movements shifted below the lines and choices. Process data collected for each item-subject encounter were screened to spot desynchronizations.

Figure 3

The Scan Path for Item 1 for Subject 12011 Illustrating Desynchronization

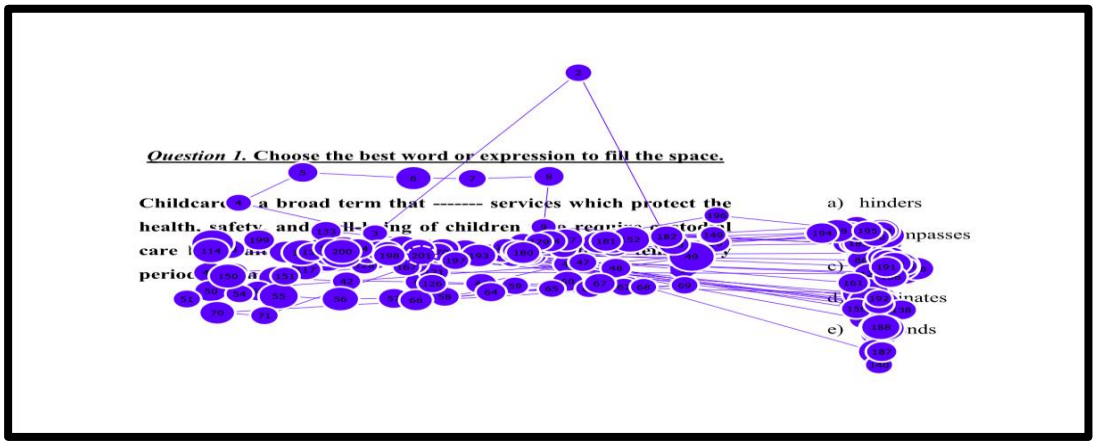


Figure 4

The Heatmap for Item 1 for Subject 12011 Illustrating Desynchronization

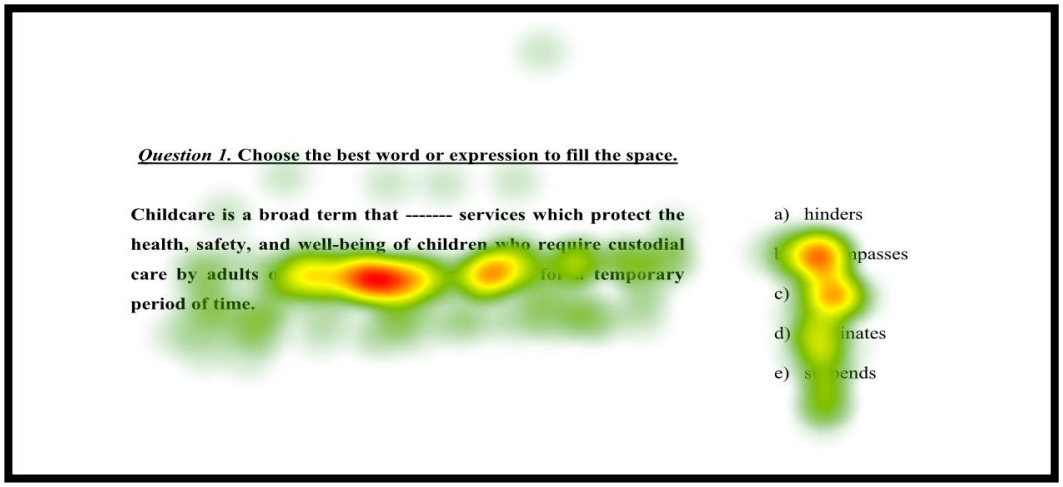


Figure 5 illustrates another subject's eye movements for Item 1. Unlike the Figure 3, Figure 5 shows that this subject's gaze movements and the lines are overlapping, which is desired for the data quality. Figure 6 illustrates the heatmap for the second subject for Item 1. As one can see, the subject's gaze movements and the lines are overlapping, and the subject focuses on the first line and choices; however, there is no downward or upward movement beyond the lines.

Figure 5

The Scan Path for Item 1 for Subject 39770 Illustrating Synchronization

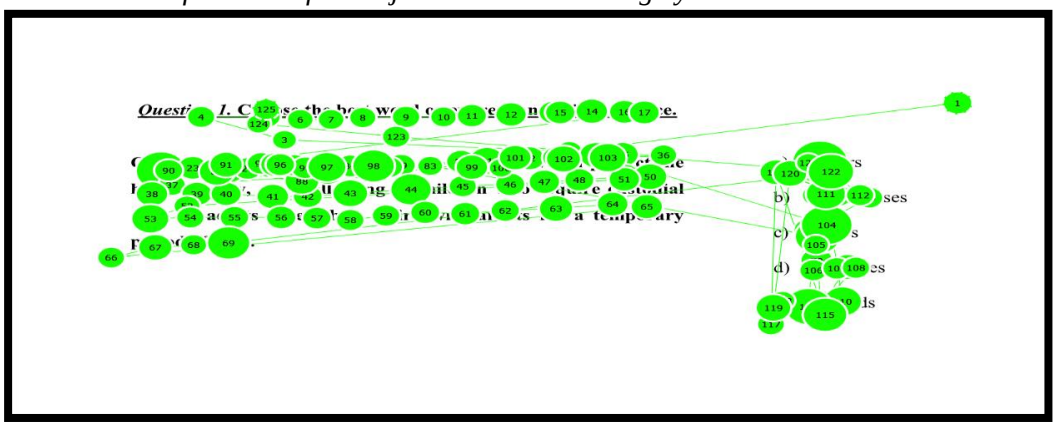
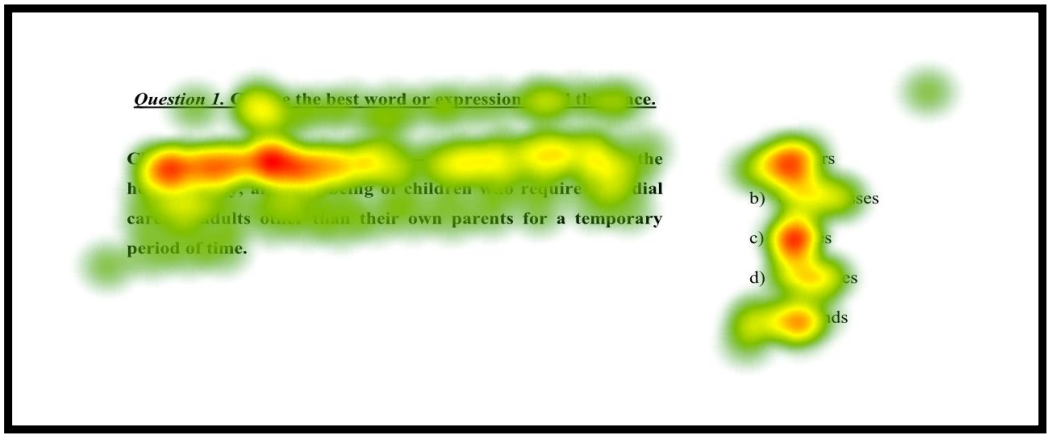


Figure 6

The Heat Map for Item 1 for Subject 39770 Illustrating Synchronization



In this study, utilized eye-tracking metrics were "fixation-related" recordings of each subject-item encounter. These metrics can be processed to obtain the number of fixations, fixation duration, total fixation duration, average fixation duration, fixation rate, regressive fixation, number of saccade, saccadic duration (see Holmqvist et al., 2011, for an elaborative list of measures). In this study, only fixation-related measures such as total gaze duration, total fixation duration, mean fixation duration and the mean fixation that was used to refer to the means that were computed from the choice sequence process data" were used.

Table 1 gives summary statistics calculated for the areas of interest (direction, line, and choices) determined for item 1. This table was created by aggregating time-stamped data (it can be seen in the Figures 3 and 5 (scan paths), which aggregated graphical version of these miliseconds into fixations and saccades, showing the hit sequence and fixation densities of eye-movements) provided by eye-tracking device, which collects data every 3.4 ms. The generated data set in Table 1 lists numerical values that can be thought of as their quantitative counterparts needed for statistical analysis and model fitting to describe and test the hypotheses about whether there was a within person change over the choices presented A to E, if so, which parameters would be needed to predict the shape of change implied by the data.

Table 1

Descriptive Statistics for Item 1 Based on Fixation Measures for Subject 76798

Areas of Interest	Hit Sequence	Time to First Fixation	Total Fixation Duration (ms)	Total Dwell Duration	Total Fixation Count
Direction	2	0,84	1,77	2,1	10
Line 1	1	0,4	10,91	12,25	59
Line 2	8	12,25	2,81	3,44	17
Line 3	9	18,67	1,93	2,28	9
Line 4	10	21,72	0,22	0,23	2
Choice A	3	5,56	3,96	4,66	17
Choice B	4	6,23	3,58	3,69	16
Choice C	6	7,64	2,58	2,78	12
Choice D	5	7,48	3,95	4,34	18
Choice E	7	10,19	2,41	2,8	11

Analysis of Eye-Tracking Process Data

As stated in the previous section, the eye-tracking device provides process data sampled at regular intervals. The process data in this research were collected each 3.4 ms using the eye-tracking technology and were restructured for the cross-sectional analysis and the latent modeling on each item. As a first step, summary statistics were computed for Item 1 and Item 2 by using total gaze duration and mean log fixation duration for the direction, stem and the choices. Then, it was tested whether these durations were differed for groups who responded correctly/incorrectly to the item. After that, for the latent modeling, mean log fixation times were calculated using time-stamped data by averaging each time the test-taker reads an area of interest (choice A, B, C, D and E), from starting to read an item until a response was provided. In this context, it was hypothesized that the default processing order was the given presentation order of item segments from A to E (within-person variance). It was hypothesized if this were the case for the sample, along with threshold and slope variances showing between-examinee variances.

Latent Growth Curve Modeling For Eye-Tracking Response Data

Although many techniques have been used to analyze eye-tracking data, the usefulness of Latent Growth Curve Modeling remains relatively unexplored. It is, therefore, worthwhile to explore what model formulations can offer for process data analytics and their connections to the latent structure intended to be measured. Permiting both within-person change over time and between-person variability, Latent Growth Curve Models are used to answer questions such as: "What is the shape of the mean trend over time?, Does the initial level predict the rate of change? Do two or more groups differ in their trajectories? etc." (Duncan et al., 2013; Muthen, 2001; Preacher at. al., 2008; Willett & Sayer, 1994).

In this study, LGCM applications were utilized to model change over response choices (AOIs for A to E) for one item at a time. For this purpose, a series of nested latent growth curve models were fit to item-level data to estimate (a) choice mean log fixation time trajectories and (b) variability around the initial status and overall rate of change. Relying on the conventional assumption that subjects read the choices in given presentation order, a linear model was considered first. The linear latent growth model curve model was estimated by Equation 1:

Level 1 within-subject model: $$y_{it} = \eta_{0i} + \eta_i a_{it} + \varepsilon_{it},$$

Level 2 between-subject model: $$\eta_{0i} = \beta_0 + \zeta_{0i}, \quad (1)$$

$$\eta_{1i} = \beta_1 + \zeta_{1i},$$

where *t* represents the choice set coded 0 to 4, *i* represents the subject (*I=1,2,…, N*), *y* represents the mean log fixation time influenced by the random effects η_{0i} and η_{1i}. The intercept η_{0i} describes a subject's initial mean log fixation time when reading the choices. The linear term η_{1i} describes the rate of change in mean log fixation time during the reading of the choices.

The quadratic latent growth curve model was considered next. The model was estimated by Equation 2:

Level 1 within-subject model: $$y_{it} = \eta_{0i} + \eta_{1i} a_{it} + \eta_{2i} a_{it}^2 + \varepsilon_{it},$$

Level 2 between-subject model: $$\eta_{0i} = \beta_0 + \zeta_{0i}, \quad (2)$$

$$\eta_{1i} = \beta_1 + \zeta_{1i},$$

$$\eta_{2i} = \beta_2 + \zeta_{2i},$$

where η_{2i} represents the quadratic term, and it describes the rate of acceleration in the rate of change over the course of reading choices.

The quadratic latent growth curve model with the item response (correct/incorrect) as a time-invariant covariate was considered next. The model was estimated by Equation 3:

Level 1 within-subject model: $$y_{it} = \eta_{0i} + \eta_{1i} a_{it} + \eta_{2i} a_{it}^2 + \varepsilon_{it},$$

Level 2 between-subject model: $$\eta_{0i} = \beta_0 + \gamma_0 X_i + \zeta_{0i}, \quad (3)$$

$$\eta_{1i} = \beta_1 + \gamma_1 X_i + \zeta_{1i},$$

$$\eta_{2i} = \beta_2 + \gamma_2 X_i + \zeta_{2i},$$

Where γ_0, γ_1 and γ_2 *y* represents the parameters for the association between the covariate, X_i, and each latent growth term.

Log transformations were used in model estimations to normalize mean fixation time distributions. The models were estimated with Mplus using the Maximum Likelihood estimator. Consistent with recommended practices (see Hu & Bentler, 1999; for a detailed discussion), more than one fit index was used in evaluating model fit. The models were compared using (a) the Comparative Fit Index (CFI; Bentler, 1990), and (b) the Tucker-Lewis Index (TLI; Bentler & Bonnet, 1980), where values range from zero to one and the values greater than 0.95 may be interpreted as an acceptable fit; (c) the root Mean Square Error of Approximation (RMSEA; Steiger, 1990), where values smaller than 0.05 indicate a good model fit (Browne & Cudeck, 1993); (d) Standardized Root Mean Square Residual (SRMR; Bentler, 1995), where values should be less than 0.05 for a good fit (Hu & Bentler, 1998), (e) Akaike Information Criterion (AIC; Akaike, 1974) the Bayesian Information Criterion (BIC; Schwarz, 1978), where smaller values indicate a better fit and are often used to test goodness-of-fit for a full model in comparison with a reduced one.

Results

For the data obtained from the eye tracking device to be valid and reliable, the correct positioning of the eye movements on the stimuli is one of the essential points to be considered. Even though the subjects in the study stated that they did not have any visual impairment and passed the calibration test, it was found that the desired quality of stimulus-eye movement harmony was somewhat flactuating for some subjects while steadily unusable for some others. A careful screening of the scanmaps and heat maps

revealed that, out of 71 subjects in the sample, eye-tracking recordings of 11 subjects were unusable for item 1, while only 6 of these subjects' eye-tracking recordings were also unusable for item 2.

Table 2 shows the summary data (gaze even duration and the mean fixation duration) calculated for the areas of interest (from question stem to lines) for all subjects in item 1 and item 2. The table show that the total-time subjects spent on Item 1 (gaze event duration) varied between 8,23 seconds and 74,76 seconds, and the mean duration was 36,07 seconds. There was a considerable variation in how much time the subjects spent answering the item. Similarly, there was an apparent variation in how much time the subjects spent on each choice and each line considering the standard deviations. The minimum time spent on some areas of interest, such as stem, choice A, and line 2, was zero, which means that at least one subject responded to the question without looking at these areas. Among the choice, the true choice (choice B) had the maximum time spent, and of the lines, the first line had the maximum time spent.

The total-time subjects spent on item 2 varied between 15,35 seconds and 97,42 seconds, and the mean fixation duration was 39,82 seconds. Similarly, there was an apparent variation in the duration the subjects spent responding to item 2, each choice, and each line, considering their standard deviations. The minimum time spent on some areas of interest, such as stem, choice B, choice E, and line 2, was zero, which means that at least one subject responded to the question without looking at these areas. Unlike item 1, the true choice (choice C) didn't have the maximum time spent, and of the lines, the first line had the maximum time spent.

Considering the right/wrong answer, which is one of the most basic features of a multiple-choice items, subjects were grouped into two groups (subjects who responded correctly vs. incorrectly) to explore whether the subjects who gave correct answers allocated their attention to different parts of the items differently from the subject who gave an incorrect answer.

Table 2

Descriptive Statistics for Item 1 and Item 2 Based on Gaze Event Duration and Mean Fixation Duration for All Subjects

			Fixation Durations								
Item 1											
	Total Gaze Dur.	**Stem**	**Line 1**	**Line 2**	**Line 3**	**Line 4**	**Cho. A**	**Cho. B***	**Cho. C**	**Cho. D**	**Cho. E**
Mean	36,07	1,93	8,13	5,58	2,65	0,68	1,79	2,8	2,64	2,18	1,69
Std. Deviation	16,01	1,80	5,08	3,95	2,01	0,69	1,44	1,4	1,82	1,73	1,64
Median	35,08	1,65	7,5	4,47	2,19	0,51	1,3	2,57	2,24	1,48	1,08
Minimum	8,23	0,00	0,1	0,18	0	0	0	0,62	0,43	0,16	0,19
Maximum	74,76	9,87	22,77	17,43	8,89	3,79	6,6	6,37	8,86	6,98	7,71
Item 2											
	Total Gaze Dur.	**Stem**	**Line 1**	**Line 2**	**Line 3**	**Line 4**	**Cho. A**	**Cho. B**	**Cho. C***	**Cho. D**	**Cho. E**
Mean	39,82	1,51	9,46	7,73	0,86	N/A	3,39	3,58	3,29	2,28	1,53
Std. Deviation	16,50	1,38	4,82	4,13	0,73	N/A	2,58	2,94	1,89	1,86	1,54
Median	39,99	1,35	8,87	6,88	0,73	N/A	2,77	3,04	2,89	1,63	1,08
Minimum	15,35	0,00	1,38	1,91	0	N/A	0,28	0	0,55	0,26	0
Maximum	97,42	6,12	22,48	19,56	4,36	N/A	10,33	17,04	10,4	11,44	7,58

* the correct answer

Table 3 illustrates the descriptive statistics for the two groups (Item 1). The *First Group* consisted of those who did not score any point on Item 1 (n = 36), and the *Second Group* consisted of those who scored one point on Item 1 (n = 24). The first group had higher *mean total gaze duration* and *mean fixation durations* for each variable except for *choice B*, the correct choice, than the second group. According to the Mann-Whitney U test results, there was a statistically significant difference between the first and second groups regarding total gaze duration, choice A, C, D, E, and Lines 1,2,3,4 ($p < .05$).

Table 4 also illustrates the decriptive statistics for two groups (Item 2). As in item 1, the *First Group* consisted of those who did not score any point on Item 2 (n = 40), and the *Second Group* consisted of those who scored one point on Item 2 (n = 25). The first group had higher *mean total gaze duration* and *mean fixation durations* for each variable except for *choice C*, the correct choice, than the second group. According to the Mann-Whitney U test results, there was a statistically significant difference between the first and second groups regarding total gaze duration, choices A, B, D, E, and Lines 1 and 2 ($p < .05$).

Table 3

Descriptive Statistics for Item 1 Based on Mean Fixation Duration in Terms of Correct-Incorrect Response Groups

Item 1	Fixation Durations										
	Total Gaze Dur.**	**Inst.**	**Line 1****	**Line 2****	**Line 3****	**Line 4****	**Opt. A****	**Opt. B**	**Opt. C****	**Opt. D****	**Opt. E****
Subjects who answered the item incorrectly											
Mean	42,39	2,17	9,4	6,89	3,09	0,83	2,14	2,78	3,44	2,85	2,17
Std. Deviation	13,89	1,92	5,05	4,31	2,06	0,75	1,57	1,46	1,83	1,88	1,75
Median	39,09	1,94	9,47	5,49	2,6	0,59	1,55	2,63	3,58	2,17	1,61
Minimum	19,05	0,00	0,31	1,46	0	0	0,16	0,62	0,88	0,41	0,21
Maximum	74,76	9,87	22,77	17,43	8,89	3,79	6,6	6,3	8,86	6,98	7,71
Subjects who answered the item correctly											
	Total Gaze Dur.**	**Inst.**	**Line 1**	**Line 2**	**Line 3**	**Line 4**	**Opt. A**	**Opt. B**	**Opt. C**	**Opt. D**	**Opt. E**
Mean	26,58	1,56	6,24	3,6	1,98	0,46	1,26	2,83	1,45	1,17	0,96
Median	21,54	1,39	5,05	3,14	1,76	0,28	0,91	2,49	1,18	1,1	0,41
Std. Deviation	14,39	1,56	4,6	2,21	1,76	0,53	1,05	1,34	0,98	0,71	1,16
Minimum	8,23	0,00	0,1	0,18	0	0	0	1,15	0,43	0,16	0,19
Maximum	59,15	7,02	17,84	9,45	7,17	1,75	4,11	6,37	4,68	3,03	4,96

* the correct answer = Choice B, ** $p < 0.05$

Table 4

Descriptive Statistics for Item 2 Based on Mean Fixation Duration for Terms of Correct-Incorrect Response Groups

Item 2	Fixation Durations									
	Total Gaze Dur.**	**Inst.**	**Line 1****	**Line 2****	**Line 3**	**Opt. A****	**Opt. B****	**Opt. C****	**Opt. D****	**Opt. E****
Subjects who answered the item incorrectly										
Mean	45,21	1,79	10,66	8,58	0,96	4,39	4,44	3	2,6	2,03
Std. Deviation	16,61	1,59	4,65	4,09	0,85	2,71	3,34	2,03	2	1,72
Median	42,30	1,37	9,19	7,46	0,74	3,87	3,66	2,76	1,89	1,58
Minimum	15,35	0,00	1,38	2,33	0	0,87	0	0,55	0,48	0,19
Maximum	97,42	6,12	22,48	19,56	4,36	10,33	17,04	10,4	11,44	7,58
Subjects who answered the item correctly										
	Total Gaze Dur.	**Inst.**	**Line 1**	**Line 2**	**Line 3**	**Opt. A**	**Opt. B**	**Opt. C**	**Opt. D**	**Opt. E**
Mean	31,35	1,06	7,56	6,39	0,69	1,82	2,23	3,74	1,78	0,76
Std. Deviation	12,52	0,83	4,56	3,92	0,45	1,25	1,36	1,58	1,54	0,75
Median	29,41	1,31	6,15	5,53	0,72	1,68	2,26	3,68	1,18	0,51
Minimum	16,32	0,00	1,57	1,91	0	0,28	0,16	1,71	0,26	0
Maximum	56,44	2,26	21,38	19,01	1,57	5,41	4,9	7,83	6,25	3,29

* the correct answer = Choice C, ** $p < 0.05$

Table 5 lists model fit statistics of item 1 for latent growth models identified in the research. Overall, the results suggest that the goodness-of-fit observed for the three models ranges from unacceptable to very good and that Model 2 with the quadratic term and Model 3 with the quadratic term and the response (0/1) as a covariate fit the data better than Model 1 with a linear term only. For Model 1, the fit statistics show that the model does not provide an adequate fit to the data as the RMSEA and SRMR far exceeds the acceptable fit range. For Model 2 and Model 3, the fit statistics suggest that these models have a very good fit to the data.

For comparing which one indicates a better model (Model 2 or Model 3), three information-based fit indices (Akaike's Information Criterion (AIC), Bayesian Information Criterion (BIC), and sample-size-adjusted BIC (SABIC)) were examined, where smaller values indicate a better model. These indices suggest that Model 3 fit the data better than its reduced counterparts. Besides, a likelihood ratio test was done to test the sufficiency of a complex model versus a smaller model. The significance value for this test ($9{,}737_{(2)}$) is 0.02, where one can accept a quadratic model with a covariate improved the model.

Table 6 shows random intercepts and slopes produced by different models. The results of Model 1 suggest that the subjects differed in two ways: in the estimates of their initial status (intercept) and their rates of change during the reading choices (slopes). Figure 7 plots intercepts and slope estimates for the five timing conditions (choices). This figure shows observed scores, Linear Model (Model 1) and Quadratic model (Model 2). It shows that the mean log fixation times of the subjects who took item 1 rose rapidly from Choice A to Choice B, while there were few increases from Choice B to Choice E.

Figure 8 illustrates the group-specific random coefficients (intercepts and slopes) produced by the Quadratic Model with Time-Invariant Covariate (Model 3) for the two response groups (correct/incorrect). Figure 8 also demonstrates that the correct-response group and incorrect response group differed in two respects: in the observed scores of their initial status (intercepts) and their rates of change during the fixation orders (slopes). While there is a downward trend in mean log fixation times after the correct choice B for the correct-choice group, there is an upward trend in the mean log fixation times from choice A to choice E for the incorrect-choice group.

Table 5

Model Fit Comparisons for Three Models for Item 1

	CFI	TLI	AIC	BIC	SABIC	RMSEA	SRMR
Model 1[a]	0, 96	0,96	272,89	293,88	262,38	0,14	0,13
Model 2[b]	0,99	0,99	266,77	296,09	252,06	0,07	0,05
Model 3[c]	0,99	0,99	253,26	288,87	235,40	0,06	0,05

[a]Linear Model,
[b]Quadratic Model,
[c]Quadratic Model with Time-Invariant Covariate

Table 6

Parameter Estimates for Three Models

		Intercept	Slope	Quadratic
Model 1	Mean	2.80 (0,00)	0.04 (0,02)	
	Variance	0.20 (0,00)	0.01 (0,04)	
Model 2	Mean	2.68 (0,00)	0.18 (0,00)	-0,03 (0,01)
	Variance	0.27 (0,00)	0.03 (0,40)	0,00 (0,23)
Model 3	Mean	2.82 (0.00)	0.20 (0.001	-0.03 (0.04)
	Variance	0.24 (0.00)	0.02 (0.46)	0.00 (0.25)

Figure 7

Observed Means and Expected Mean Log Fixation Times for Model 1 and Model 2

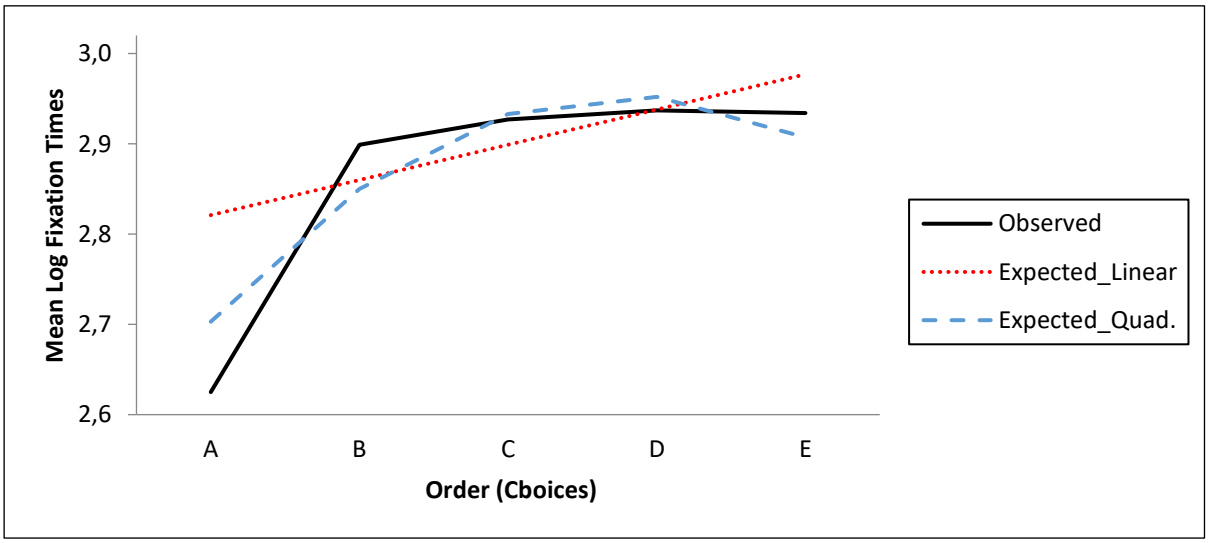


Figure 8

Estimated Mean Log Fixation Time for Correct-Response and Incorrect-Response Groups for Model 3

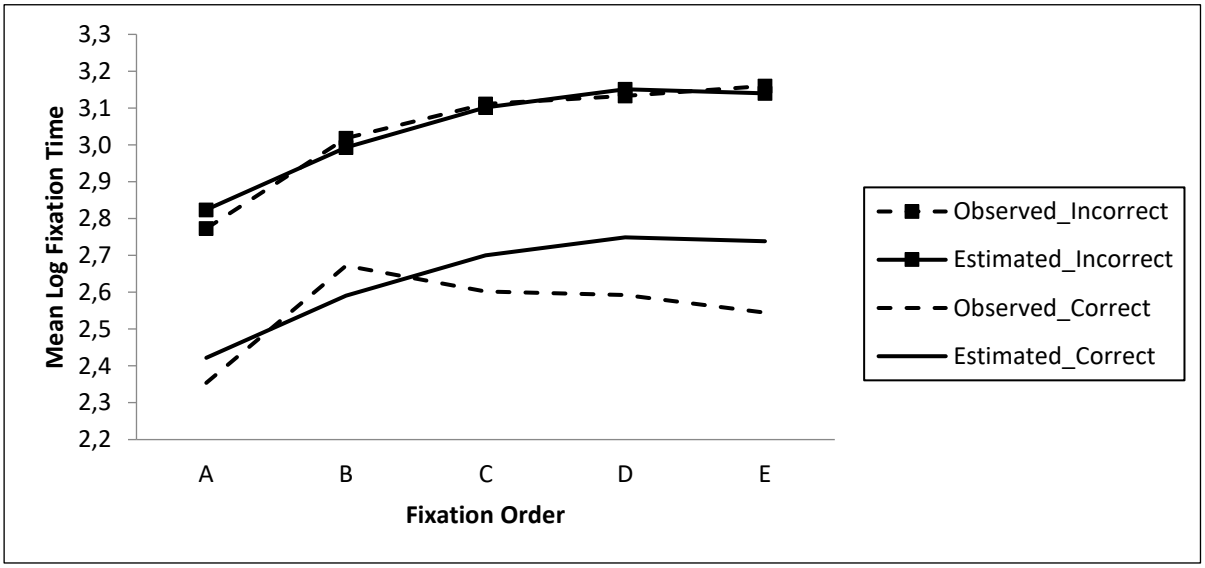


Table 8 lists model fit statistics of item 2 for latent growth models identified in the research. Overall, the results suggest that the goodness-of-fit observed for the three models ranges from unacceptable to very good and that Model 2 and Model 3 fit the data better than Model 1. For Model 1, the fit statistics show that the model does not provide an adequate fit to the data as the RMSEA and SRMR far exceeds the acceptable fit range. Although for Model 2 and 3, the statistics look similar, the fit statistics suggest that Model 3 has a very good fit to the data. To compare which one indicates a better model (Model 2 or Model 3), three information-based fit indices (Akaike's Information Criterion (AIC), Bayesian Information Criterion (BIC), and sample-size-adjusted BIC (SABIC)) were examined, where smaller values indicate a better model. These indices suggest that Model 3 fit the data better than its reduced counterparts. Besides, a likelihood ratio test was done to test the sufficiency of a complex model versus a smaller model. The significance value for this test ($6{,}214_{(2)}$) is 0.04, where one can accept a quadratic model with a covariate improved the model.

Table 9 shows random intercepts and slopes produced for Item 2 by different models. The results of Model 1 suggest that the subjects differed in two ways: in the estimates of their initial status (intercept) and their rates of change during the reading choices (slopes). Figure 9 illustrates that the mean log fixation times of the subjects who took item 2 rose rapidly from Choice A to Choice C, which is the correct choice, while there are few increases from Choice C to Choice E.

Table 8

Model Fit Comparisons for Three Models for Item 2

	CFI	TLI	AIC	BIC	SABIC	RMSEA	SRMR
Model 1[a]	0, 95	0,95	206,89	228,63	197,15	0,13	0,12
Model 2[b]	0,99	0,99	201,77	231,63	187,56	0,07	0,07
Model 3[c]	0,99	0,99	194,76	231,72	178,21	0,07	0,05

[a]Linear Model,
[b]Quadratic Model,
[c]Quadratic Model with Time-Invariant Covariate

Table 9

Parameter Estimates for Three Models for Item 2

		Intercept	Slope	Quadratic
Model 1	Mean	2.81 (0,00)	0.07 (0,00)	
	Variance	0.19 (0,00)	0.00 (0,02)	
Model 2	Mean	2.74 (0,00)	0.18 (0,00)	-0,02 (0,00)
	Variance	0.25 (0,00)	0.07 (0,02)	0,00 (0,02)
Model 3	Mean	2.91 (0.00)	0.09 (0.07)	-0.08 (0.45)
	Variance	0.19 (0.00)	0.06 (0.04)	-0.04 (0.03)

Figure 9

Observed Means and Expected Mean Log Fixation Times for Model 1 and 2

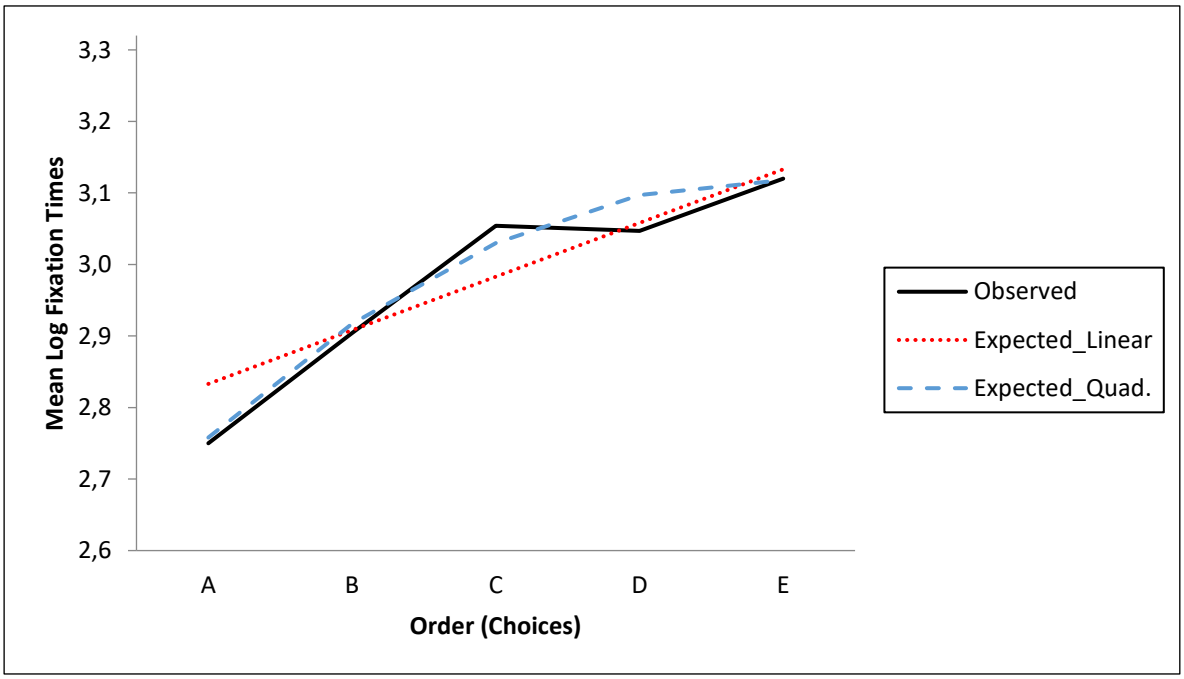


Figure 10

Model 3, Estimated Mean Log Fixation Time for Correct-Response and Incorrect-Response Groups

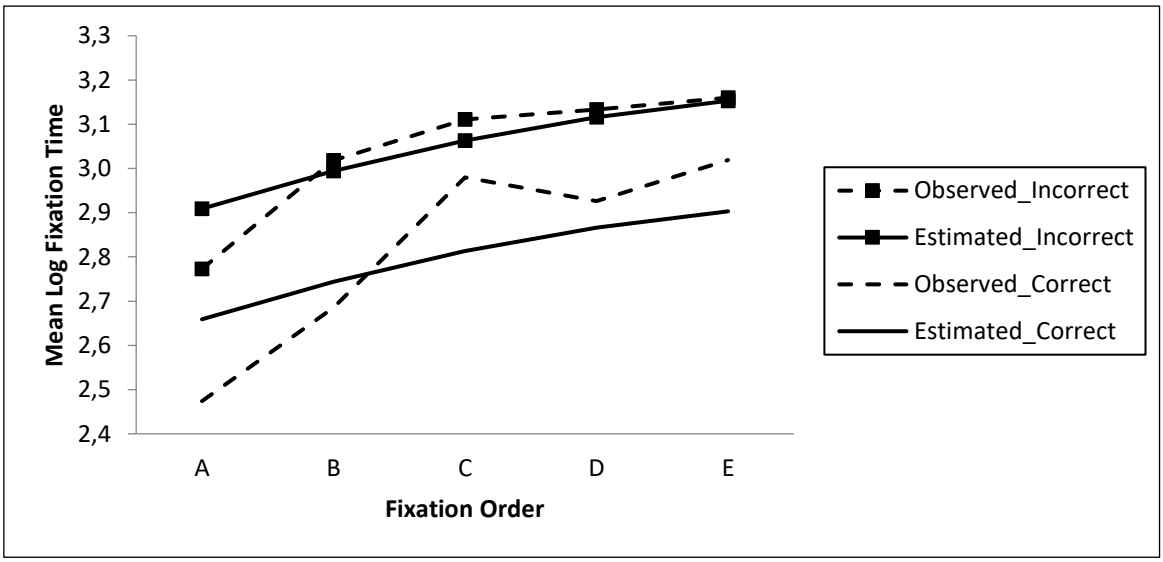


Figure 10 illustrates the group-specific random coefficients (intercepts and slopes) produced by the Quadratic Model with Correct/Incorrect Response Covariate (Model 3). Figure 10 demonstrates that the correct-response group and incorrect response group differed in two respects: in the observed scores of their initial status (intercepts) and their rates of change during the fixation orders (slopes). While there is a downward trend in mean log fixation times after the correct choice C for the correct-choice group, there is an upward trend in the mean log fixation times from choice A to choice E for the incorrect-choice group.

To sum up, subjects' reading behavior for choices was tested on two multiple-choice reading comprehension questions with a series of latent longitudinal models (Linear, quadratic and quadratic with a time-invariant covariate). For both item 1 and item 2, it was found that mean log fixation times increased quadraticly for subjects while reading over the choices A to E, meaning that subjects' reading speed was faster at first and slowed toward the final choice, choice E. The fact that the Quadratic model was reasonably acceptable for both items supports the finding in the previous sentence. Nonetheless, a better fit was observed for the quadratic model when the correct/incorrect response variable was added to the model as a time-invariant covariate. This finding reveals that the quadratic pattern estimated for the correct responders' was different than that of the incorrect responders, suggesting meaningful differences in both initial status and the change parameters for the two groups. These results are consistent with the expectation and imply that there were less of a searching behavior for subjects responding the item correctly after they reached to the correct coded choice.

Discussion and Conclusion

Eye-tracking is a technology that demonstrates great potential in educational assessment research because it provides time-series data that can be translated to item response process data for investigating the nature of the relationships between various cognitive processes and the performance of test-takers when expected to use these cognitive processes and has the potential to uncover the moment-to-moment processes of problem-solving behavior. Eye movements obtained from the eye-tracking devices are widely accepted to reflect cognitive processes for reading comprehension (Raney et al., 2014; Meziere et al., 2021); however, these cognitive processes cannot be directly inferred from eye-tracking data alone. In order to interpret eye-tracking data properly, theoretical and psychometrical models must always be the basis for designing experiments as well as for analyzing and interpreting eye-tracking based measurements.

Written from a psychometric perspective, this study illustrates that the first step of reaching accurate interpretations using eye-tracking enhanced data is to screen and translate time-series eye-movement data into item process data quantifying item surfaces or sections using binary-coded Areas of Interest variables. Within the context of this study, the Areas of Interest were multiple-choice item parts, each line in the question stem, and each choice. The results showed that the measures obtained from the subjects varied significantly in these areas when built as variables. For example, there was a considerable variation in how much time the subjects spent answering the items. Similarly, there was an apparent variation in how much time the subjects spent on each choice and each line, especially when the standard deviations were taken into account. This was taken as an indicator of the feasibility of the Areas of Interest utilized in this study.

In this study, special emphasis was given to the screening of eye-tracking recordings of subject-item encounters to spot recordings that were unreliable (eye movement recordings were not reliable for some subjects). The accuracy rate was approximately %85 and %92 for the two items studied in this study. This exemplifies that there could still be mismatches remaining, even after initial calibration tests (most, if not all, eye tracking devices have a calibration stage) were passed successfully. In order to overcome the problem of eye-gaze agreements being less than 100% accurate, our results suggest that, using a set of initial screening processes may help (involving analyses of graphs, scan paths, and heatmaps). Although ensuring reliability of the utilized eye movement data was of an interest in this study, our main goal was to use a novel approach, namely Latent Growth Modeling, to interpret the information that

might be available in item level eye movement data as it pertained to the particular choices presented. This was accomplished by binary coding time stamped fixations observed for each choice throughout examinee-item interaction time. Our findings reveal that through binary coding choice relevant time stamped eye movement data for choices A to E, it would be possible to map how fixation times (when ordered by choices A to E) changed over item-encounter time for each and all examinees. Albeit beyond the scope of this study, establishing a baseline trajectory for test items through the use of LGMs would be useful for flagging subjects responding an item correctly, yet, not showing any observed reading activity on any of the choices (possible cheating behavior). Another use could be for investigating item speededness by investigating if majority or a subgroup of examinees are running out of time before getting to the latter choices or if pseudo guessing has occurred.

To collect further evidence supporting validity arguments to be made about the accuracy of response item and test scores, an eye-tracking measure, total fixation duration, was calculated for the all the determined areas of interest (direction, lines and choices). It was clear that the test-takers paid their attention to these areas in a varying amount of time. In terms of the focus of this study, the fixation duration of the participants on each option were scrutinized, which shows that participicants who responded the item correctly spent more time on the correct answer/choice than the other choices (distractors). Also they fixated on the correct choice more than the participants who answered the item incorrectly, which is an indication of viewing inportant and relevant informaiton. Besides, participants gave more attention to some options (distractors) and less to others, which shows that some of the distractors were more related to construct to be measured while the others not. Of course, the fixation durations of partipicants on the options do not tell us the cognitive processes a participant uses while responding a multiple-choice quesiton; however, the distinction between these groups provides useful information on test validation.

The results from the latent growth models show that eye-tracking measurements obtained not only for the correct coded but also for the distracting choices may play an essential role in revealing the nature of within- and between-subject differences in reading behavior. Overall, the findings of this study show that the data obtained from the eye-tracking technology can be potentially useful in determining the patterns in reading behavior of subjects when responding to test items and testing if certain variables of interest, such as the response being correct or incorrect, explain some of the variances. Although not used in this study, model extensions can be easily formulated to include piecewise or cubic terms as well as additional time-invariant covariates such as gender, total score, etc., and time-varying covariates such as word count.

The scope of this study is limited in that it uses a small subject sample and only two multiple-choice item experiments. Another limitation is that this study focused on the fixation metrics while several other metrics are also provided by the eye-tracking technology, such as saccades. In addition, the caveat of this paper is that it uses a more of a psychometric perspective than a substantive one and that the future studies should focus more on the substantive issues. For future studies, researchers are recommended to use larger subject samples, and increased item experiments so that the placement order of the correct coded choices would not be limited. Albeit limited, the results of this study nicely illustrate that an effective integration of eye-tracking data into correct/incorrect response data may greatly enhance what we know about the item processing behaviors of subjects. The multi-stage data analysis approach was greatly useful for the findings from the initial screening and cross-sectional analyses were great pointers without which the longitudinal models could not have been easily estimated and interpreted. With larger sample sizes, multi-group latent growth curve models can be estimated to look into item and response reading patterns of subjects marking any of the five choices as the correct answer for a more detailed description of the subgroups being drawn to particular choices.

Declaration

Conflict of Interest: No potential conflict of interest was reported by the authors.

Ethical Approval: The study was approved by the Gazi University Ethics Committee (Research code: 2019-003, dated 08.01.2019/ 01)

This paper presents some of the results obtained during the Doctoral Thesis process that were partially supported via the BAP Project No 04/2019-01 under the supervision of Prof. Dr. Nilüfer Kahraman.

Author Contribution: Ergün Cihat Çorbacı-Conceptualization, implementation, methodology, analysis, writing & editing. Nilüfer Kahraman-Conceptualization, methodology, analysis, writing & editing, visualization.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716-723. https://doi.org/10.1109/TAC.1974.1100705

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107*(2), 238-246. https://doi.org/10.1037/0033-2909.107.2.238

Bentler, P. M. (1995). *EQS structural equations program manual* (vol. 6). Multivariate software. https://doi.org/10.4236/am.2014.510132

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological bulletin*, *88*(3), 588-606. https://doi.org/10.1037/0033-2909.88.3.588

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136−162). Sage.

Duncan, T. E., Duncan, S. C., & Strycker, L. A. (2013). *An introduction to latent variable growth curve modeling: Concepts, issues, and applications*. Routledge.

Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., & Van de Weijer, J. (2011). *Eye tracking: A comprehensive guide to methods and measures*. OUP Oxford.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1-55. https://doi.org/10.1080/10705519909540118

Kane, M., & Mislevy, R. (2017). Validating score interpretations based on response processes. In *Validation of score meaning for the next generation of assessments* (pp. 11-24). Routledge.

Mézière, D. C., Yu, L., Reichle, E., von der Malsburg, T., & McArthur, G. (2021). *Using eye-tracking measures to predict reading comprehension*. https://doi.org/10.31234/osf.io/v2rdp

Muthén, B. (2001). Second-generation structural equation modeling with a combination of categorical and continuous latent variables: New opportunities for latent class–latent growth modeling. In L. M. Collins & A. G. Sayer (Eds.), *New methods for the analysis of change* (pp. 291–322). American Psychological Association. https://doi.org/10.1037/10409-010

Öğrenci Seçme ve Yerleştirme Merkezi (2018). Retrieved on February 25, from https://www.osym.gov.tr/TR,15313/2018-yds-sonbahar-donemi-temel-soru-kitapciklarinin-yayimlanmasi--10.html

Paulson, E. J., & Henry, J. (2002). Does the Degrees of Reading Power assessment reflect the reading process? An eye-movement examination. *Journal of Adolescent & Adult Literacy*, *46*(3), 234-244. https://link.gale.com/apps/doc/A94123361/LitRC?u=anon~924799ab&sid=bookmark-LitRC&xid=2cf242f9

Preacher, K. J., Wichman, A. L., MacCallum, R. C., & Briggs, N. E. (2008). *Latent growth curve modeling* (No. 157). Sage.

Raney, G. E., Campbell, S. J., & Bovee, J. C. (2014). Using eye movements to evaluate the cognitive processes involved in text comprehension. *JoVE (Journal of Visualized Experiments)*, *83*, e50780. https://doi.org/10.3791/50780

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, *124*(3), 372-422. https://doi.org/10.1037/0033-2909.124.3.372

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461-464. https://doi.org/10.1214/aos/1176344136

Solheim, O. J., & Uppstad, P. H. (2011). Eye-tracking as a tool in process-oriented reading test validation. *International Electronic Journal of Elemantary Education*, *4*(1), 153-168. https://www.iejee.com/index.php/IEJEE/article/view/218

Steiger, J. H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research*, *25*(2), 173-180. https://doi.org/10.1207/s15327906mbr2502_4

Tai, R. H., Loehr, J. F., & Brigham, F. J. (2006). An exploration of the use of eye-gaze tracking to study problem-solving on standardized science assessments. *International Journal of Research & Method in Education*, *29*(2), 185-208. https://doi.org/10.1080/17437270600891614

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

Willett, J. B., & Sayer, A. G. (1994). Using covariance structure analysis to detect correlates and predictors of individual change. *Psychological Bulletin, 110*, 268-290. https://doi.org/10.1037/0033-2909.116.2.363

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, Online First*. https://doi.org/10.1007/s10459-022-10107-9