education sciences

MDPI

Article

Exploring Relationships among Test Takers' Behaviors and Performance Using Response Process Data

Sergio Araneda [1], Dukjae Lee [1], Jennifer Lewis [1], Stephen G. Sireci [1,*], Jung Aa Moon [2], Blair Lehman [2], Burcu Arslan [2] and Madeleine Keehner [2]

[1] College of Education, University of Massachusetts Amherst, Amherst, MA 01003, USA; saraneda@umass.edu (S.A.); dlee@umass.edu (D.L.); jlewi0@umass.edu (J.L.)

[2] Educational Testing Service, Princeton, NJ 08544, USA; jungaamoon@gmail.com (J.A.M.); blehman@ets.org (B.L.); barslan@ets.org (B.A.); mkeehner@ets.org (M.K.)

* Correspondence: sireci@acad.umass.edu

Abstract: Students exhibit many behaviors when responding to items on a computer-based test, but only some of these behaviors are relevant to estimating their proficiencies. In this study, we analyzed data from computer-based math achievement tests administered to elementary school students in grades 3 (ages 8–9) and 4 (ages 9–10). We investigated students' response process data, including the total amount of time they spent on an item, the amount of time they took to first respond to an item, the number of times they "visited" an item and the number of times they changed their responses to items, in order to explore whether these behaviors were related to overall proficiency and whether they differed across item formats and grades. The results indicated a non-linear relationship between the mean number of actions and proficiency, as well as some notable interactions between correctly answering an item, item format, response time, and response time latency. Implications for test construction and future analyses in this area are discussed.

Keywords: assessment; log data; response processes; test development; validity

check for updates

Citation: Araneda, S.; Lee, D.; Lewis, J.; Sireci, S.G.; Moon, J.A.; Lehman, B.; Arslan, B.; Keehner, M. Exploring Relationships among Test Takers' Behaviors and Performance Using Response Process Data. *Educ. Sci.* **2022**, *12*, 104. https://doi.org/10.3390/educsci12020104

Academic Editor: Eila Jeronen

Received: 28 December 2021
Accepted: 27 January 2022
Published: 3 February 2022

1. Introduction

Educational assessments are increasingly being delivered via computer and collect a wide variety of information from students during the testing process. The end result for most students is a single test score, or perhaps several sub-scores, describing students' performance on specific domains within a subject area. These overall scores and sub-scores are used to make inferences about students' knowledge, skills, and abilities. However, these scores are only coarse indicators of the proficiencies we are trying to measure. Therefore, there is great interest in gaining a better understanding of students' proficiencies from educational assessments. Part of that interest is in improving assessments to better capture student behaviors that are reflective of their knowledge and skills on the constructs targeted by the assessments.

When educational tests are administered on a computer, students exhibit behaviors that can be digitally captured—some of which are related to their proficiency, and some are not. These potentially informative behaviors are ignored by traditional views of assessing students. Some researchers have proposed models for using the amount of time students take to respond to test items in scoring (e.g., van der Linden [1,2]), but clearly there are other behaviors students exhibit on computer-based tests (CBTs) that could be of interest. Such behaviors include accessing available reference or resource material (e.g., on-screen calculators), navigating between and back to items, changing responses, and skipping items. These data can be referred to as "response process data" because they describe the actions students exhibit as they navigate through a testing experience [3–6].

Response process data can take a variety of forms, only some of which are relevant to the constructs intended to be measured. In principle, there are many ways in which these processes could be categorized. Figure 1 illustrates one theoretical model of the universe of test-taker behaviors that can be captured in a CBT; only some are considered to reflect the targeted construct, and typically only a subset of these are captured in scoring. Traditionally, scored behaviors are limited to direct responses to items (e.g., selecting an answer option); however, behaviors that are not traditional item responses can also be defined and evaluated according to a rubric (e.g., how appropriately a student interacts with a simulated task or effectively plots points on a graph). These two types of scorable response processes include everything that falls within the inner two ellipses in Figure 1.

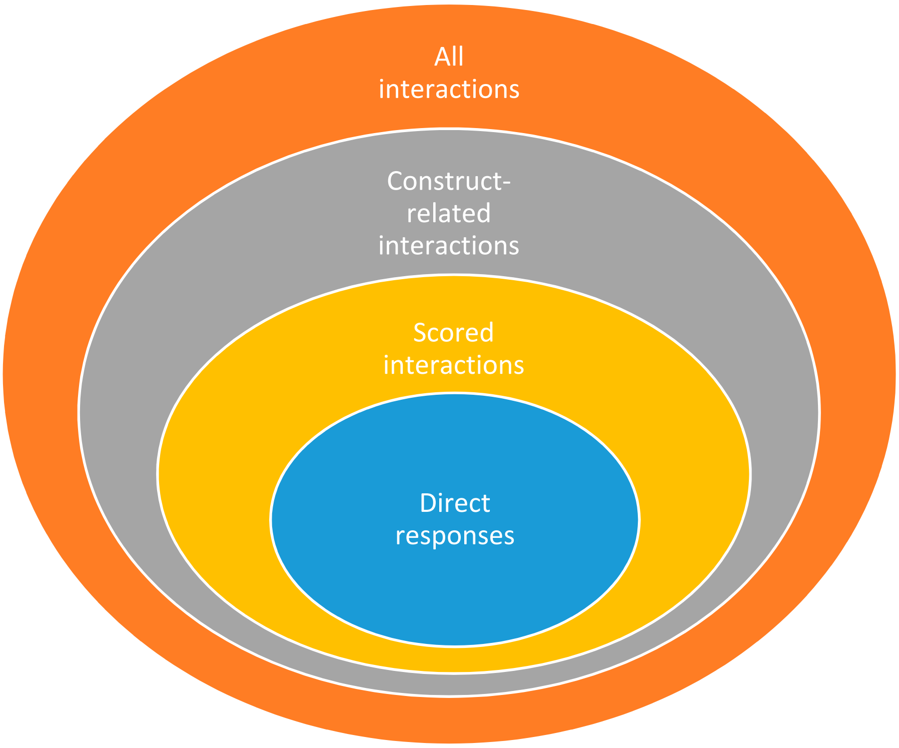


Figure 1. Theoretical partitioning of response behaviors on a computer-based test.

Other types of student actions and events are not scored but may still be related to the constructs measured, and hence they have the potential for increasing representation of the construct (the grey ellipse in Figure 1). For example, the *order* in which a student makes selections in a multiple-response item (rather than the selections per se) may reveal something about their cognitive processes in relation to the target construct. Finally, other test-taker actions and events not considered to be related to the target construct(s) may still reflect cognitive, affective, and behavioral phenomena that impact test scores (e.g., confidence, motivation, prior technological experience, strategies, self-regulation, etc.). Such cognitions can be revealed in observable processes, such as how students navigate through the testing system. Thus, the entire set of students' responses to items on CBTs can be partitioned into those that are construct-relevant and those that are not; not all of the construct-relevant responses students exhibit are leveraged in gauging students' proficiencies.

In addition to providing information about student proficiency, response process data may be helpful for providing "forensic" information for evaluating the design characteristics of the test itself. This information may be helpful for identifying threats to validity or design strengths, including low-level general issues such as usability of the interface or the use of tools and supports. Other process information that arguably falls into this category includes the interactivity patterns observed for the different CBT item formats. For example, if some item formats are responded to more (or less) quickly by students, or facilitate more student engagement, they may contribute to a more valid and efficient assessment. Thus, there is great potential in exploring students' response process behavior for both gauging students' proficiencies and evaluating the effectiveness of various item formats. In this article, we analyze students' log data from a computer-based test to learn more about their response processes and how they can be used to better understand student proficiencies and assessment characteristics. Before describing our exploration of these potentials in the present study, we first present a brief review of some relevant research in this area.

2. Previous Research on Analyzing Students' Response Process Data

Research into students' response behaviors while taking CBTs is relatively recent. Some research has focused on the *total* amount of time students take to respond to an item, while other research has focused on the amount of time it takes students to make a *first* response to an item. The consistency of these behaviors across item formats has also been studied. In addition, some researchers have looked at the total number of response actions as well as answer changing behavior.

2.1. Response Time

Li et al. [7] claimed response time (RT) "offers a promising window into test takers' cognitive processes and hence the construct(s) being measured . . . [and] offers an opportunity to build validity evidence for a test" (p. 159). RT has been used to evaluate validity by examining RT differences across various item formats. For example, multiple-choice (MC) items are frequently used because they are efficient for measuring a wide variety of content in a short amount of time. However, MC items are often limited to eliciting lower levels of cognitive processes [8]. For this reason, many CBT item formats have been proposed to address the limitations of MC items while maintaining their advantages [9,10].

Some studies have explored RT across different item formats, such as MC and technology-enhanced items (TEIs). Jodoin [11] explored the time test takers spent on solving traditional MC items and two TEI formats (drop-and-connect and create-a-tree items). He found examinees spent less time responding to MC items, and the TEIs provided less information per unit of testing time than the MC items, suggesting MC items had greater relative efficiency. Wan and Henly [12] explored the average time test takers spent on answering different item formats, using data from a statewide science achievement test. They focused on three TEI formats: figural response, short constructed response (SCR), and extended constructed response (ECR). The results indicated the figural response items were equally efficient to MC items, whereas CR items, and in particular ECRs, were less efficient overall due to the longer completion time.

Other research has examined how more fine-grained design features within an item format affect RT. For example, Moon et al. [13] examined how long test takers spent responding to different item formats under uncertainty. They focused on response behavior for variations of the "grid" item format (see Appendix A), which requires test takers to select appropriate grid cells, with options presented in a table of rows and columns. Grid items were presented in three formats: forced choice (FC), which allows only one selection per row (similar to true–false items); non-forced choice (NFC), which allows more than one selection per row; and all possible options (APO), which includes all possible options as grid cells. Test takers spent significantly more time on the FC relative to the NFC and APO formats, suggesting that, even within the same grid item format, fine-grained design features such as visual layouts of answer options affect test takers' response processes. Similarly, Arslan et al. [14] investigated the effect of drag-and-drop item design features on test-taker macro-level (e.g., RT) and micro-level (first response latency, transition pauses, dragging time) measures by constructing different design variants of content-equivalent mathematics items. They found test-taker macro- and micro-level performance measures were significantly affected by design variants. They concluded that some observed process differences reflected low-level effects of design changes, such as the need for longer mouse movements to complete the necessary steps, whereas others reflected the cognitive effects of design changes, altering the cognitive steps in which test takers engaged.

2.2. First-Response Latency (FRL)

Several studies have investigated whether students who take a longer amount of time to make their first response to an item have higher overall proficiency. In addition to reading the item stem and constructing a mental representation of the problem, first-response latency (FRL) may include the time test takers use to plan how to solve questions. In this view, longer planning time, reflected in longer FRL, may lead to better performance in the response portion of the task, with more efficient strategies and fewer trials and errors (e.g., Albert and Steinberg [15]; Mitchell and Poston [16]; see Eichmann et al. [17] for considering planning interval beyond FRL). Alternatively, longer FRLs may reflect greater difficulty due to construct-relevant factors such as prior knowledge or construct-irrelevant factors such as reading load. In this view, longer FRL should be associated with poorer item performance.

In the present study, we were interested in whether FRL differs systematically across item formats, which, to our knowledge, has not been previously studied. We hypothesized test takers might spend different amounts of time before making an initial response, depending on the item format, and that these differences may be related to particular cognitive or behavioral affordances of the item formats. For instance, test takers may take longer to plan their response in item formats requiring more complex actions than in those requiring simpler actions. Some item formats such as "inline choice" (see Appendix A) may elicit relatively shorter FRL compared with other item formats, because answer options in inline choice items are initially hidden within a "closed" interactive drop-down menu, forcing test takers to quickly initiate an "open menu" action to view the answer options. The magnitudes of the differences between item formats, if found, may also differ for students of different ages. For example, they may be greater for lower grade levels, assuming that students at this level are relatively less familiar with the technology.

2.3. Number of Actions

Some research in the domain of complex problem solving suggests the number of actions test takers perform on an item is related to item difficulty. For example, Goldhammer et al. [18] investigated the relationship between the number of actions performed during task completion and task success using process data from the Programme for the International Assessment of Adult Competencies (PIAAC). They hypothesized that number of actions is an indicator of test takers' engagement to the task. They further hypothesized that performance should be low for test takers who performed either a very small or a very large number of actions—those who were less engaged and therefore performed few actions—and those who worked hard on solving the tasks but became disoriented as time went by, resulting in a high number of actions. In contrast, those who were moderately engaged in the task, and who therefore performed a moderate number of actions, should perform well because they concentrated on the task more efficiently. These hypotheses specify a nonlinear relationship between number of actions and task performance (performance increases with more actions but decreases after a certain number of actions). Their results showed an inverted U-shape relationship between number of actions and task performance for low complexity tasks, which was consistent with their hypothesis. However, the relationship showed a monotonically increasing S-shape for high-complexity tasks, implying more actions were related to a higher probability of correctly answering tasks.

In the present study, we examined a similar relationship by comparing proficiency versus mean number of actions performed by each student across all items, and across items of a given format. The stimuli in our research were discrete TEIs, not complex PIAAC-like tasks. We also investigated whether the relationship between proficiency and number of actions was consistent across item formats. If a systematic relationship was found, it may indicate different levels and types of engagement with the items, and if it differed according to item format, it may be informative with respect to the factors influencing engagement in digital assessment.

2.4. Summary and Introduction to the Present Study

Previous research into test takers' response behavior has suggested differences in the amount of time it takes them to respond to different item formats, as well as nonlinear relationships between the number of actions on an item and performance. We note that the degree to which these response behaviors are consistent across different CBT item formats has not been widely studied, and it is possible there are differences in FRL across item formats. Therefore, in the present study, we analyze data from a computer-based mathematics achievement test at two grade levels. We explore various hypotheses to understand student cognition and behavior and to evaluate the quality of different item formats. The specific research questions addressed were as follows:

What is the relationship between students' proficiency, response time, response time latency, and other response behaviors?

Do students take different amounts of time to respond to different item formats?

Do students take different amounts of time to make an initial response to different item formats?

Are the relationships among students' response behaviors consistent across grade levels?

3. Method

3.1. Data

The data analyzed here come from a field test for a summative mathematics assessment administered to students in grades 3 (5854 students, mean age 9.2 years, SD = 0.67) and 4 (4568 students, mean age 10.2 years, SD = 0.65). A summary of the item formats investigated is presented in Table 1. Examples of item formats are provided in Appendix A. We focused on item formats that had at least five items that were field-tested in each grade. The grade 3 exam comprised 232 items, and the grade 4 exam comprised 237 items. Items were grouped into different test forms, with students responding to only 22–24 items on a form. There were 23 test forms for grade 3 and 25 forms for grade 4. The sample sizes for each item ranged from 212 to 1629. The median sample size for each item format within each grade is reported in the last column of the table (see Figures A1 and A2). Although the sample sizes for some items were in the low 200s, the average sample sizes for items were all above 400. Items on the test were presented individually (one item per screen). Students could skip an item entirely or give a partial answer, could return to previous items using arrow buttons or via a "review" panel, or change a previous response. Unfortunately, no data were available for interactions with the review panel; however, we could trace students' navigation across items in consecutive and nonconsecutive orderings.

Table 1. Number of items and item formats.

Item Format	Grade	# of Items	Sample Size Range	Median of Range
Grid	3	7	244–540	520
	4	9	235–856	447
Inline Choice MS	3	13	264–558	527
	4	11	223–881	432
Multiple Choice MS	3	34	244–560	519.5
	4	25	223–886	443
Multiple Choice SS	3	81	246–1626	523
	4	94	224–889	441
Match MS	3	11	512–557	525
	4	9	235–445	438

SS: Single Selection; MS: Multiple Selection.

3.2. Defining Response Process Variables

As described by Keehner and Smith [19], it is difficult to identify and create construct-relevant variables from the myriad of log data available from a CBT. In this study we created several variables for investigation. Operational definitions are as follows.

Item response time (RT): Response time (RT) was defined as the total amount of time a student spent on an item. If a student visited an item multiple times, the time spent on the item was aggregated across visits. Time spent on specific visits was also calculated and was used in calculating other variables.

Number of visits: Given that students could view an item, move to another item without responding to it, and circle back to the item at a later time, students could return to an item more than once. Each time a student viewed an item, we marked it as a "visit". We calculated a "number of visits" variable to record the number of times a student navigated to an item.

Number of actions: An action is a register in the process data file that reports a choice made by the student in response to an item. Examples of actions include selecting or unselecting an option in a MC multiple select item, dropping a drag element in a drag-and-drop item, selecting an option in an inline choice item, etc. All such actions were summed for a student interacting with an item. Actions across multiple visits to an item were included. It should be noted the number of actions required to answer an item differed across item formats and items within a format. This minimum number of actions required to answer an item was used as a covariate in the statistical analyses.

Number of changes: This variable is a subset of the number of actions variable, but it does not include the first click on each option of an item. Given that students were allowed to change their answers to an item, we calculated the total number of changes to an item across all items to which a student responded, as well as the mean number of changes per item.

First-response latency: We computed first-response latency (FRL) by calculating the amount of time between when a student viewed an item and conducted a first action on the item. Thus, FRL represented the latency between first viewing an item on a visit and making a first response to an item. If a student merely viewed an item and went on to the next item without making a response, that time was not counted in FRL.

Proficiency**:** Proficiency was computed using item response theory (IRT). Students' responses to the selected-response items were scored dichotomously and calibrated using the one-parameter IRT model (Hambleton, Swaminathan, and Rogers [20]), which is of the form

$$p\left(X_{ij} = 1\middle|\theta_j b_i\right) = \frac{e^{\theta_j - b_i}}{1 + e^{\theta_j - b_i}} \tag{1}$$

where Xij is the response of student j to item i, θ_j is the proficiency estimate of student j, and b_i is the difficulty of item i. The multiple-select selected-response items (e.g., select all correct options) were amenable to partial credit scoring. We used partial credit scoring in these situations according to "multiple true/false" (MTF) scoring, where each option is dichotomously scored, and the item score is the total number of correct responses to options divided by the number of options in an item. However, we added two adjustments to the MTF scoring: if a student did not select any response option, or if the student selected all response options, they received a score of zero. For multiple-select items that were amenable to partial credit scoring and for the constructed-response items, the partial credit IRT model was used [21]. This model is of the form

$$p\left(X_{ij} = k\middle|\theta_j a_i b_{ik}\right) = \frac{e^{\sum_0^k (\theta_j - b_{ik})}}{\sum_{j0}^{m_i} e^{\sum_{x=0}^{j} (\theta_j - b_{ik})}} \tag{2}$$

where k is a given score on item i, m is the maximum score on the item, and b_{ik} is the step difficulty of score k. All IRT scoring was conducted using FlexMIRT [22].

4. Data Analyses

Our analyses of the students' data involved computing descriptive statistics, correlations, and conducting statistical tests of mean differences. For the analyses regarding students' RT, we took the log of the students' RT before conducting statistical analyses, due to the extreme positive skewness of the data.

Linear Mixed-Effects Model Analyses

Linear mixed-effects models (LMEMs) were used to test hypotheses involving differences in a dependent variable (i.e., RT, FRL) by item formats and dichotomized item score (i.e., whether a student correctly or incorrectly answered an item). For polytomous items, students who earned less than half the maximum points on the item were scored zero, and students who earned at least half of the maximum points were scored as 1.0.

Two LMEMs were used—one for total item RT, the other for FRL. The models included item format and dichotomized item score as fixed effects and random intercepts for examinees and items. This model is essentially equivalent to a repeated measures two-way ANOVA when there are no missing data. The minimum number of responses required to answer an item varied across the multiple-select items, and so we included that variable as a covariate in the analysis (number of expected choices). For total item response time, the model was

$$\begin{aligned} log(RT_{ei}) = \beta_0 + &\sum_{it\ \epsilon\ IT} \beta_{1it} 1_{it}\ (i) + \beta_{correct} 1_{correct}\ (e,i) \\ &+ \sum_{it\ \epsilon\ IT} \beta_{2it} 1_{it}(i) 1_{correct}\ (e,i) + \beta_3 Nb_{Expected\ opt}(i) + \epsilon(e) \\ &+\epsilon(i) + \epsilon \end{aligned} \tag{3}$$

where

$1_{it}(i)$ = one, if item format of item i = it, zero if not.

$1_{correct}(e,i)$ = one, if examinee e responded correctly item i, zero if not.

$Nb_{Expected\ opt}\ (i)$ = number of expected choices marked in the correct response of item i.

$\epsilon(e)$ = random effect linked to examinee e.

(i) = random effect linked to item i.

IT = set of all item formats considered in the analysis

For FRL as the dependent variable, the model was

$$\begin{aligned} log(FRL_{ei}) = \beta_0 + \sum_{it\ \epsilon\ IT} \beta_{1it} 1_{it}\ (i) + \beta_{correct} 1_{correct}\ (e,i) \\ + \sum_{it\ \epsilon\ IT} \beta_{2it} 1_{it}(i) 1_{correct}\ (e,i) + \beta_3 Nb_{Expected\ opt}(i) + \epsilon(e) + \epsilon(i) + \epsilon \end{aligned} \tag{4}$$

where the same notation is used as for the previous equation.

The R package lme4 was used in all analyses [23]. For the ANOVA analyses, we used Satterthwaite's method for the approximation of degrees of freedom. Models with random slope on the item type by examinee were explored, but all those models failed either because they were singular or because they were not achieving convergence. The alpha level for statistical significance was set at $p < 0.01$.

A procedure for outlier removal was performed as well, to improve the normality of the residuals. We removed all outliers outside the area between the first quartile minus 1.5 times the interquartile range and the third quartile plus 1.5 times the interquartile range.

5. Results

In this section, we summarize the results conducted for all analyses. We begin with the relationships among response process behavior and proficiency and then report results regarding comparisons across RT, FRL, and item format.

5.1. Students' Proficiency and Response Actions

Table 2 presents the correlations among proficiency and response process variables for grade 3 (lower triangle) and grade 4 (upper triangle). The pattern of correlations was largely consistent across grade levels. The highest correlations were between number of actions and number of changes, which is not notable since the latter is a subset of the former. After that, the highest observed correlations were for FRL and RT (r = 0.81 for grade 3 and 0.78 for grade 4). Proficiency was moderately correlated with both RT and FRL (ranging from 0.11 to 0.13 across grades), which provides modest support for the hypothesis that students who take longer to make their first response action have higher overall proficiency. RT, but not FRL, was moderately correlated with mean number of actions, visits, and changes for each grade level (ranging from r = 0.11 to 0.19 across variables and grades). FRL was not correlated with mean number of actions, visits, or changes. These relationships were expected, given that actions, visits, and changes will add to the RT for an item, but not to FRL.

Table 2. Correlations among proficiency and response behaviors: grade 3 (lower triangle) and grade 4 (upper triangle).

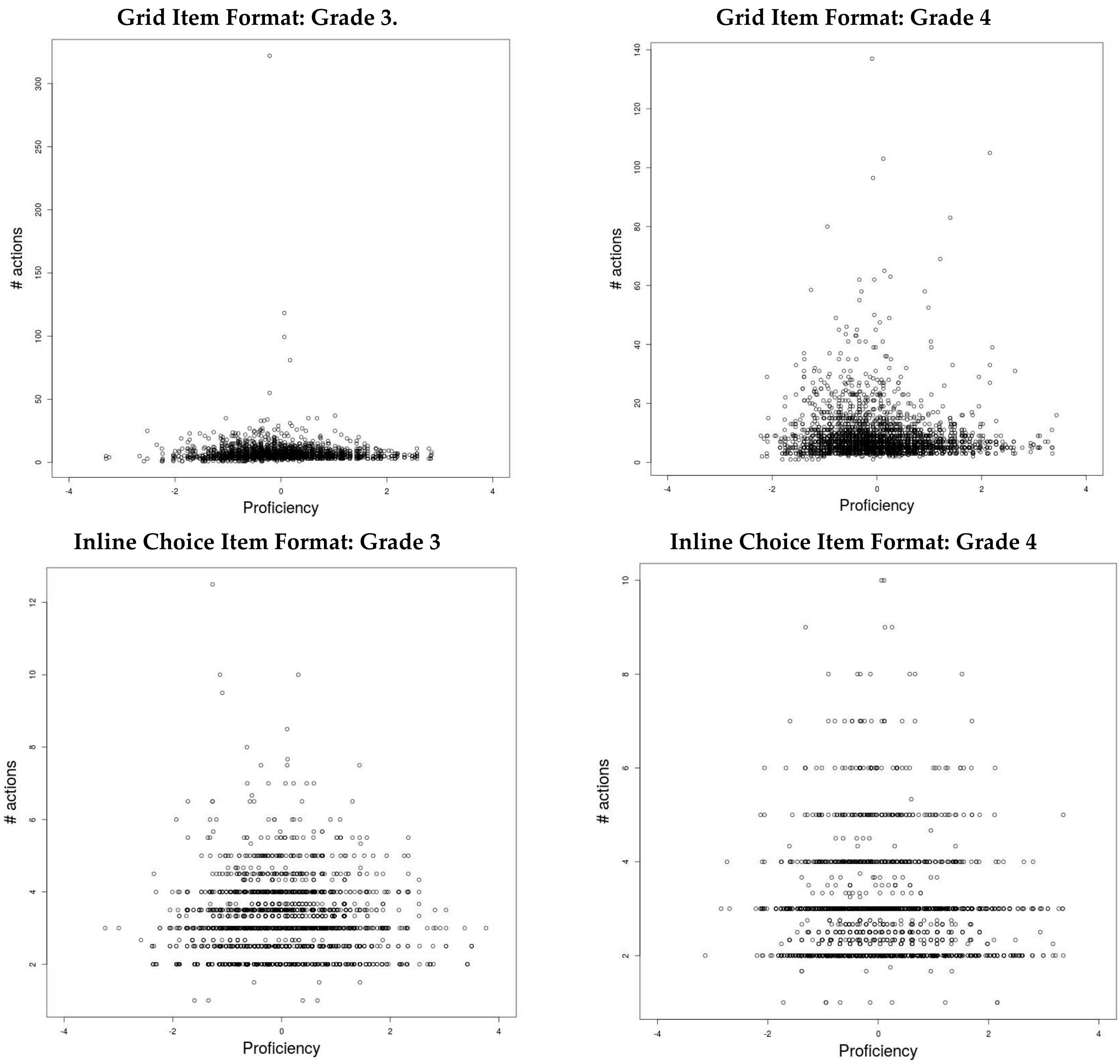
	Mean # of Actions	Mean # of Visits	Mean # of Changes	FRL	Total RT	Proficiency
Mean # of actions	–	0.16 **	0.98 **	−0.01	0.11 **	−0.04
Mean # of visits	0.17 **	–	0.14 **	−0.01	0.19 **	−0.02
Mean # of changes	0.96 **	0.17 **	–	0.00	0.11 **	−0.06 **
FRL	0.00	−0.01	0.00	–	0.78 **	0.12 **
Total RT	0.17 **	0.19 **	0.15 **	0.81 **	–	0.11 **
Proficiency	−0.001	−0.04 **	−0.03 *	0.13 **	0.13 **	–

Notes: RT = response time; * $p < 0.05$; ** $p < 0.001$.

Scatterplots of the variables were evaluated for nonlinear relationships. None were noted in the scatterplots of the relationships among proficiency, FRL, and RT and thus are not presented here. However, nonlinear relationships were observed between proficiency and number of actions. These scatterplots are presented in Figure 2, with a separate plot for each item format. A nonlinear pattern, where students of low and high proficiency had relatively fewer actions than students of middle proficiency, was consistent across grades and item formats. The Inline Choice, MC-SS, and Match MS item formats best illustrated this non-linear, quadratic relationship. The grid item format had the least conformity to this pattern, particularly for grade 3.

5.2. Linear Mixed-Effects Modeling Results

In this section, we report the results related to whether students took longer to respond to different item formats, took longer to make an initial response across different item formats, and whether there were differences across students who correctly answered the item versus those who did not. Before reporting the results of these LMEMs, we first report the descriptive statistics. The descriptive statistics for total item RT are reported in Tables 3 and 4 for grades 3 and 4, respectively. Both raw and log-transformed average item RT is reported. For both grades, multiple-choice single-select (MC-SS) items had the shortest mean RTs (results were nearly identical using the median), and the match multiple-select (Match-MS) items had the longest, with the difference in median response time across these two formats being about 35 s for both grades. The grid item format had the second-shortest average response time for 4th grade, but had the second-longest average response time for 3rd grade. Thus, some differences are noted in the rank-order of average RT for different item formats across grades.



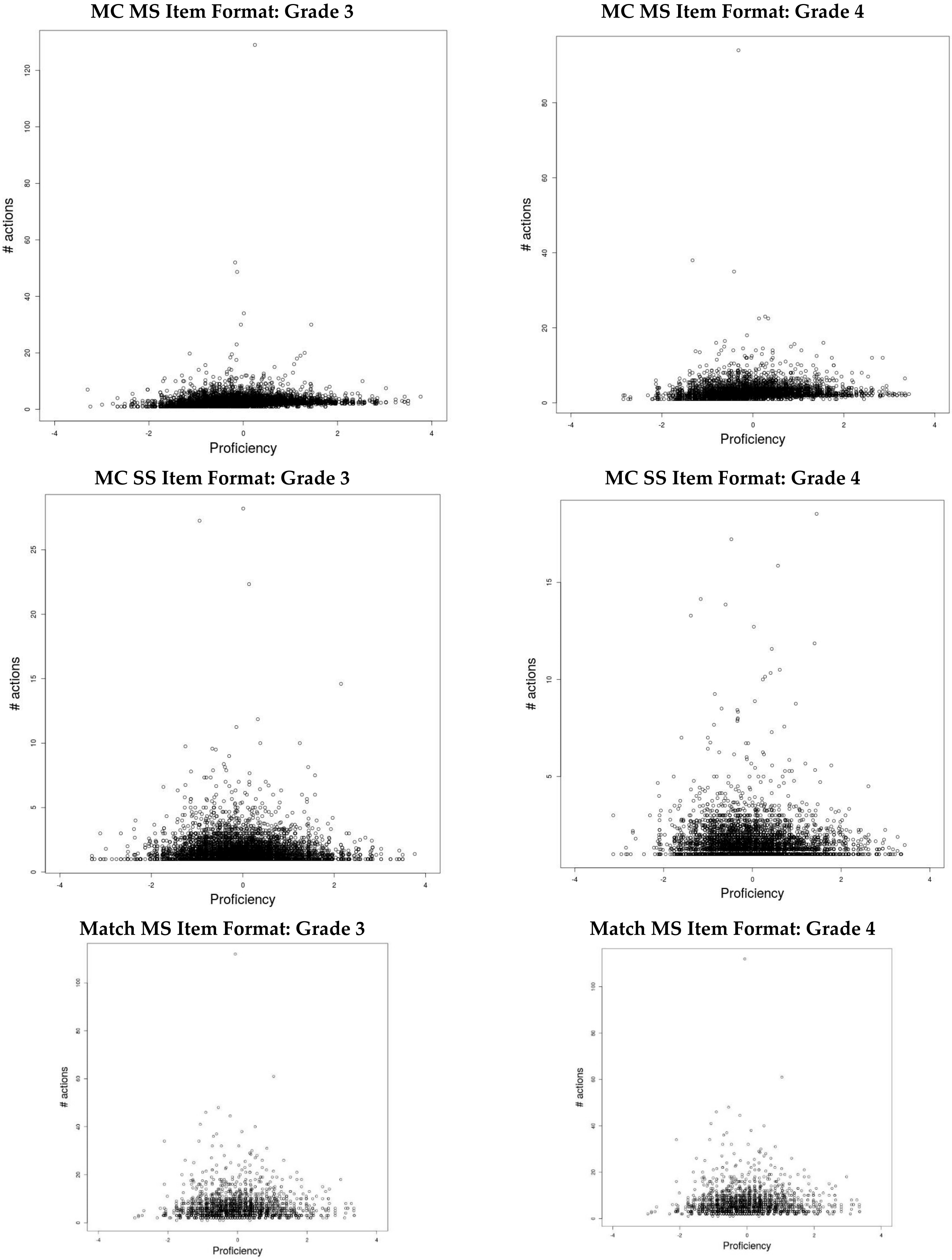
Grid Item Format: Grade 3.

Grid Item Format: Grade 4

Inline Choice Item Format: Grade 3

Inline Choice Item Format: Grade 4

Figure 2. *Cont*.



MC MS Item Format: Grade 3

MC MS Item Format: Grade 4

MC SS Item Format: Grade 3

MC SS Item Format: Grade 4

Match MS Item Format: Grade 3

Match MS Item Format: Grade 4

Figure 2. Scatterplots of proficiency (horizontal) versus number of actions (vertical) by item format.

Table 3. Descriptive statistics for average item response time: grade 3.

Item Format	Item Response Time (RT)				Log(RT)		
	N *	Mean	SD	Median	Mean	SD	Median
Match MS	5835	103.4	168.1	71.7	4.2	0.9	4.3
Grid	3118	86.4	104.2	61.8	4.1	0.9	4.1
Inline Choice MS	6675	77.7	108.1	55.1	4.0	0.9	4.0
Multiple Choice MS	16,552	74.3	117.5	46.5	3.8	1.0	3.8
Multiple Choice SS	42,925	68.8	101.4	45.6	3.7	1.0	3.8

* N indicates number of observations for examinee and item pairs.

Table 4. Descriptive statistics for average item response time: grade 4.

Item Format	Item Response Time (RT)				Log(RT)		
	N *	Mean	SD	Median	Mean	SD	Median
Match MS	3743	101.7	164.3	63.6	4.2	0.9	4.2
Multiple Choice MS	11,363	85.3	114.3	55.6	4.0	1.0	4.0
Inline Choice MS	5467	78.6	81.9	59.9	4.0	0.9	4.1
Grid	4415	76.9	93.6	53.7	4.0	0.9	4.0
Multiple Choice SS	40,622	66.6	93.1	43	3.7	1.1	3.8

* N indicates number of observations for examinee and item pairs.

Descriptive statistics for total RT, broken down by correct and incorrect response for grade 3, are presented in Table 5. The rank-ordering across item formats was identical for correct and incorrect responses. Correct answers had shorter average response times for the MC-SS and Match-MS items but not for the other item formats.

Table 5. Descriptive statistics of average item response time (seconds), grade 3 item format, by response correctness.

		Incorrect (0)				Correct (1)		
Item Format	N *	Mean	Median	SD	N *	Mean	Median	SD
Match-MS	3980	108.2	74.4	168.1	1809	94.9	68.2	169.5
Grid	2035	82.9	59.6	96.2	1056	94.6	66.8	118.1
Inline Choice MS	4069	77.3	55.6	113.7	2556	79.5	55.3	99.1
Multiple Choice MS	10,462	72.1	43.7	129.0	5953	79.5	51.6	95.0
Multiple Choice SS	20,717	71.3	47.7	107.8	21,857	67.1	44.4	94.6

* N indicates number of observations for examinee and item pairs.

The results of the LMEM analysis for Grade 3 are summarized in Table 6. The main effect for item format approached, but did not reach, statistical significance (p = 0.03); however, the correct answer main effect and interaction of correct answer and item format were statistically significant. As seen in Figure 3, the largest difference across students who got the item correct or incorrect were for the MC-MS item format, with students who correctly answered the item taking longer to answer.

Table 6. Analysis of variance for log completion time (Satterthwaite's method) item format by response correctness (grade 3).

Source	Num df	Den df	SS	MS	F	p
Item Format	4	147	5.11	1.28	2.72	0.031
Response Correctness	1	70,231	17.18	17.18	36.56	<0.001
Item format x Response Correctness	4	68,308	64.16	16.04	34.12	<0.001
Number of req. answers	1	145	1.12	1.12	2.39	0.12

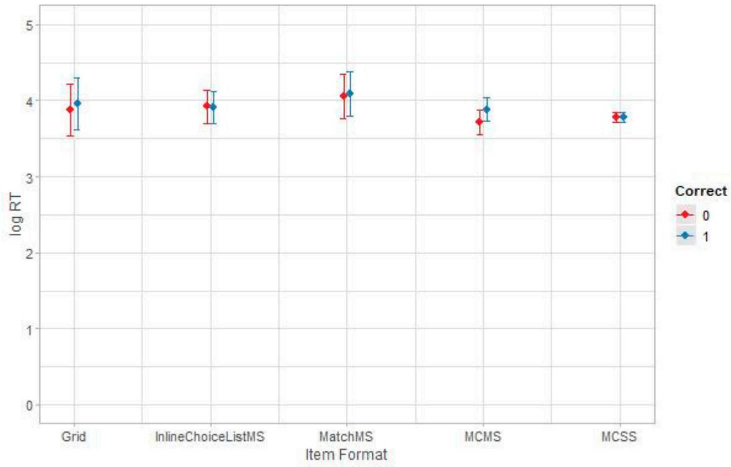


Figure 3. Predicted (marginal effects) log RT by item format and response correctness: grade 3. Error bars represent confidence intervals.

Descriptive statistics for RT, broken down by correct/incorrect response for grade 4, are presented in Table 7. Only the Match-MS item format had shorter average RT for students who correctly answered the item. The results of the LMEM analysis for grade 4 are summarized in Table 8. The main effect for item format was not statistically significant, but statistically significant differences were found for the main effect for correct answer and the interaction. Similar to grade 3, the MC-MS format had the largest difference across students who got the item correct/incorrect, with students who correctly answered taking longer (see Figure 4).

Table 7. Descriptive statistics of average item response time (seconds), grade 4 item format, by response correctness.

	Incorrect (0)				Correct (1)			
Item Format	**N ***	**Mean**	**Median**	**SD**	**N ***	**Mean**	**Median**	**SD**
Match-MS	2777	108.5	67.3	180.9	2777	82.6	54.5	97.6
Multiple Choice MS	7802	80.9	51.9	109.8	7802	96.9	66.0	123.2
Grid	3380	77.4	53.6	92.5	3380	76.4	55.4	92.2
Inline Choice MS	2949	74.2	55.0	85.1	2949	84.7	65.3	77.8
Multiple Choice SS	20,048	66.8	42.8	93.6	20,048	66.9	43.9	87.2

* N indicates number of observations for examinee and item pairs.

Table 8. Analysis of variance for log completion time (Satterthwaite's method) item format by response correctness (grade 4).

Source	Num df	Den df	SS	MS	F	p
Item Format	4	149	3.54	0.89	1.69	0.156
Response Correctness	1	61,761	25.53	25.53	48.71	<0.001
Item format x Response Correctness	4	60,128	79.11	19.78	37.74	<0.001
Number of req. answers	1	147	0.05	0.05	0.10	0.75

5.3. First Response Latency

Descriptive statistics for FRL for each item format are presented in Tables 9 and 10 for grades 3 and 4, respectively. Although MC-SS had the shortest average RT for grade 3 (Table 3), it had the longest FRL (Table 9). The same observation was noted for grade 4 (Tables 3 and 10). Thus, although students in both grades had shorter average RTs for this traditional item format, they took longer to make their first response. The grid item format had the shortest FRL for both grades. On average, FRL was about 18 seconds shorter for grid items than MC-SS for grade 3, and about 23 seconds shorter for grade 4. Similar to average response time, FRL was relatively shorter for the Match-MS item format for grade 4 than it was for grade 3.

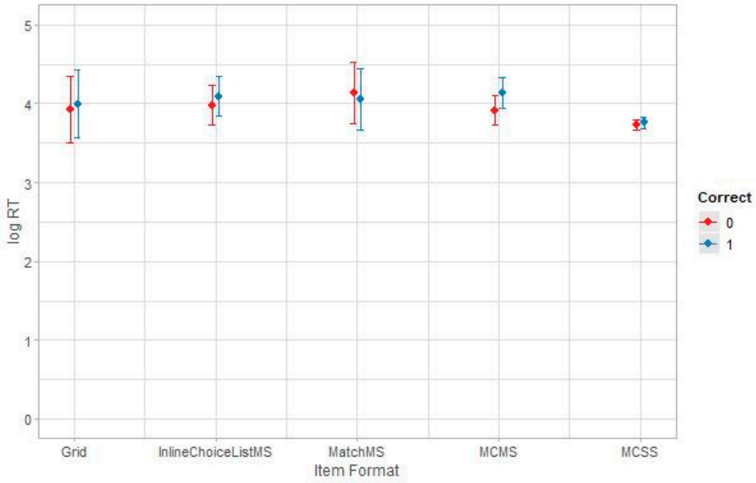


Figure 4. Predicted (marginal effects) log RT by item format and response correctness: grade 4. Error bars represent confidence intervals.

Table 9. Descriptive statistics for FRL by item format for grade 3.

Item Format	FRL				Log(FRL)		
	N *	Mean	SD	Median	Mean	SD	Median
Multiple Choice SS	42,515	51.73	71.20	33.7	3.36	1.19	3.52
Match-MS	5746	47.25	90.94	29.80	3.29	1.15	3.39
Multiple Choice MS	16,416	40.94	60.67	23.45	3.10	1.16	3.15
Inline Choice MS	6606	39.80	65.26	25.70	3.20	1.00	3.25
Grid	3090	32.58	45.90	20.50	2.90	1.14	3.02

* N is based on examinee/item pairs.

Table 10. Descriptive statistics for FRL by item format for grade 4.

Item Format	FRL				Log(FRL)		
	N *	Mean	SD	Median	Mean	SD	Median
Multiple Choice SS	40,240	50.19	63.33	31.60	3.31	1.23	3.45
Multiple Choice MS	11,274	48.31	63.52	30.00	3.26	1.21	3.40
Inline Choice	5427	46.09	50.11	32.70	3.35	1.07	3.49
Match-MS	3713	38.63	74.77	23.00	3.10	1.08	3.14
Grid	4390	26.99	37.73	15.70	2.68	1.16	2.75

* N is based on examinee/item pairs.

Descriptive statistics for FRL broken down by response correctness for grade 3 are presented in Table 11, and the results of the LMEM for this analysis are summarized in Table 12. For all item formats, students who correctly answered the item had longer FRL (see Figure 5). The only statistically significant finding was the main effect of correct response.

Table 11. Descriptive statistics of log FRL: item format by response correctness, grade 3.

		Incorrect(0)			Correct(1)		
Item Format *	**N**	**Pairs Examinee x Item**	**Mean**	**SD**	**Pairs Examinee x Item**	**Mean**	**SD**
Multiple Choice SS	34	20,488	3.36	1.24	21,823	3.40	1.11
Match-MS	11	3906	3.27	1.19	1805	3.36	1.04
Inline Choice MS	12	4013	3.15	1.04	2555	3.32	0.90
Multiple Choice MS	81	10,366	3.09	1.21	5951	3.17	1.01
Grid	7	2013	2.84	1.19	1056	3.06	0.96

Note. * Composite is polytomously scored, but dichotomized for the analysis.

Table 12. Analysis of variance for log FRL (Satterthwaite's method) item format by response correctness for grade 3.

Source	Num df	Den df	SS	MS	F	p
Item Format	4	148	6.10	1.52	2.12	0.081
Response Correctness	1	70,308	11.35	11.35	15.79	<0.001
Item format x Response Correctness	4	68,374	8.63	2.16	3.00	0.017
Number of req. answers	1	145	0.82	0.82	1.14	0.288

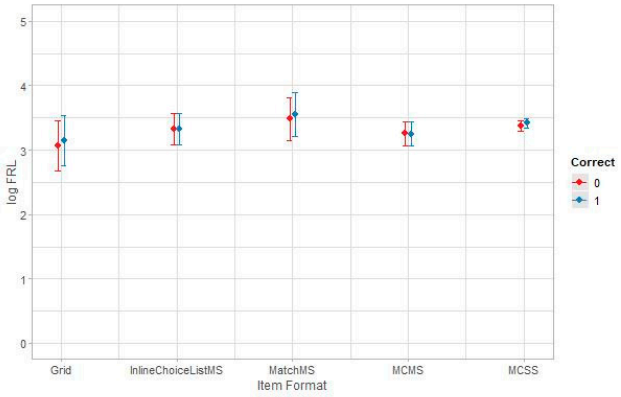


Figure 5. Predicted (marginal effects) log transformation of FRL by item format and response correctness: grade 3. Error bars represent confidence intervals.

Descriptive statistics for FRL broken down by correct/incorrect response for grade 4 are presented in Table 13; the results of the LMEM analysis are summarized in Table 14. Students who correctly answered the items had longer FRL for all item formats, except the Match-MS item format. In this case, the only statistically significant finding was the interaction effect of correct response and item format. The box plot summarizing these results is presented in Figure 6.

Table 13. Descriptive statistics of log FRL: item format by response correctness: grade 4.

		Incorrect(0)			Correct(1)		
Item Format *	**N**	**Pairs Examinee x Item**	**Mean**	**SD**	**Pairs Examinee x Item**	**Mean**	**SD**
Multiple Choice SS	25	19,850	3.26	1.27	20,212	3.37	1.15
Multiple Choice MS	94	7739	3.22	1.24	3474	3.40	1.06
Inline Choice MS	11	2917	3.20	1.11	2479	3.56	0.94
Match	9	2747	3.14	1.08	940	3.03	0.98
Grid	9	3358	2.68	1.17	997	2.74	1.05

Note. * Composite is polytomously scored, but dichotomized for the analysis.

Table 14. Analysis of variance for log FRL (Satterthwaite's method) item format by response correctness for grade 4.

Source	Num df	Den df	SS	MS	F	p
Item Format	4	149	6.25	1.56	1.91	0.112
Response Correctness	1	62,103	2.61	2.61	3.19	0.074
Item format x Response Correctness	4	60,454	36.63	9.16	11.18	<0.001
Number of req. answers	1	147	2.88	2.88	3.52	0.063

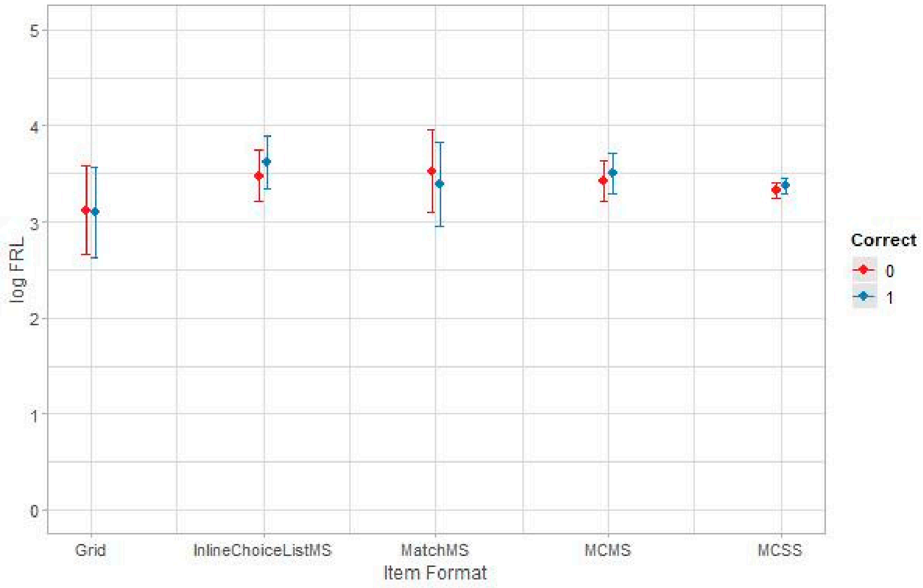


Figure 6. Predicted (marginal effects) log transformation of FRL by item format and response correctness: grade 4. Error bars represent confidence intervals.

6. Discussion

In this study, we analyzed data from a computer-based mathematics assessment across two elementary school grade levels to better understand the potential utility of a subset of students' response behaviors. Specifically, we were interested in (a) the amount of time it took students to respond to items, (b) the amount of time they took to make an initial response to an item, (c) the numbers of actions taken on items, (d) the degree to which these variables provided information about proficiency, and (e) the similarity of these relationships across the two grades.

With respect to our first research question (relationship between proficiency and response actions), we found moderate, positive correlations between total RT and proficiency (r = 0.13 for grade 3, r = 0.11 for grade 4) and between FRL and proficiency (r = 0.13 for grade 3, r = 0.12 for grade 4). These findings are similar to Albert and Steinberg [15] and Mitchell and Poston [16]. We also found for all item formats, except Match-MS in grade 4, students who correctly answered the item had longer FRL. This finding could indicate students who were more engaged with the items did better on them.

With respect to the relationship between number of actions on items and proficiency, the correlations were essentially zero, but there was clear evidence of nonlinear relationships. Goldhammer et al. [18] also found nonlinear relationships, where test takers with less success on low complexity tasks exhibited many fewer or many more actions on low-complexity tasks than students with greater success on the task. We found a more quadratic relationship, where students of relatively low and high proficiency exhibited a fewer number of actions relative to students of moderate proficiency. It should be noted we did not focus on single tasks as in Goldhammer et al.; instead, we collapsed across all items within an item format family, but it is still notable that nonlinearities were observed across all item formats.

With respect to our second and third research questions (differences across item formats), differences of up to 35 s on average were noted for RT, and up to 20 s for FRL. However, when focusing on the statistical results, the main effect for item format was diluted by the variations in RT or FRL within an item format due to our inclusion of the random effect for items in the models. There were some interesting interactions in that the MC-MS format had the longest average RT for both grade levels (see Figures 3 and 4) for students who correctly answered the item. This finding may indicate students who did not answer correctly did not spend sufficient time responding to the MC-MS items, possibly because the MC-MS items indicated the number of required selections (e.g., select the two correct responses), which may require more engagement to confirm the multiple selections. It is possible that students who had less engagement or less prior knowledge selected answer options that satisfied the number of options required by the item but did not invest the time needed to further evaluate those selections. The only other statistically significant finding was the Match-MS format had shorter FRL for grade 4 students who correctly answered the item. This item format may have been more familiar to grade 4 students, which allowed them to focus on solving the item rather than interacting with it to figure out how to record their responses.

With respect to differences across grade levels, the pattern of correlations among the process variables and proficiency was similar across grades, but as noted earlier, students who correctly answered Match-MS items had longer average FRL in grade 3, but shorter FRL in grade 4. In addition, for grade 3, students who correctly answered the grid items had relatively longer FRL than students who did not answer them correctly. In grade 4, students who correctly answered the inline choice items had relatively longer average FRL than students who incorrectly answered these items.

FRL may capture different cognitive processes, depending on the item format, and the cognitive processes employed by students may differ across grade levels. For some item formats, FRL may primarily reflect the time spent on solution processes, but in other item formats the time spent on planning and solution processes may extend well beyond the FRL period. For example, in the inline choice format, where answer options are initially hidden from view, we can speculate FRL will be relatively short because students are likely to make the first action (e.g., click a menu to view answer options) in the early phase when they are actively building a mental representation of the given problem. Thus, the action of clicking the menu is by definition part of the planning process.

If this assumption is true, one should be cautious when interpreting FRL results. For example, Albert and Steinberg [15] interpreted longer FRL as reflecting planning on how to solve items. Although the positive correlations between FRL and proficiency measures in the current study are consistent with prior research, the cognitive interpretation of the results may differ if FRL contains more than pure planning time. Conversely, for TEIs in which critical information must be actively discovered through interacting with the display (as in inline-choice items) or where the test taker can (re)organize or (re)construct the representation of information shown in the display (as in match items), the notion that planning is represented only or primarily by an initial period that does not contain external actions seems inadequate. If planning includes steps such as perceiving, processing, and building a mental representation of the information needed to respond, as well as doing the mental work involved in formulating and preparing a response, behavioral interactions with the external display are just as much part of planning and reasoning as non-visible internal mental processes such as reading the item stem or inspecting and interpreting a diagram. Future research could usefully investigate whether the current results hold when item content is controlled across item formats and explore the detailed cognitive processes captured by FRL in different item formats.

A corollary of this view is that different process metrics gathered from interactive items may belong in different parts of the theoretical model presented in Figure 1, depending on both the item format and the hypothesized phase of the solution process. For example, a click may represent simply accessing all of the information for an item (e.g., opening and inspecting a drop-down menu), it may represent an attempt to reason about the targeted construct (e.g., by comparing information across two semantically related menus), or it may represent the action of making the final selection. Thus, this same action can be categorized as not construct-related, or construct-related but not scorable, or potentially scorable target-construct evidence, depending on the specific context of the response process phase, the content of the item, and the affordances of the item format. Think-aloud protocols or other probes of response processes may be illuminating in this regard.

7. Limitations and Future Research

Our study analyzed process data gathered during the field-testing of a mathematics test, which included different item formats with varying degrees of interactivity and unique features. Therefore, an unavoidable limitation of the study was an inability to control the content and context of the items presented. An experimental study, with parallel items presented in the different formats, would control for the effects of item content. Such a controlled study would make it easier to see commonalities across items within each format category and to infer the effects of format differences (when content is held constant) on student cognition and behavior (e.g., see Arslan and Lehman [24], for a controlled experiment study in the ELA domain).

Our study also focused on only five item formats—those that were consistent and sufficiently numerous across grade levels. Other TEI formats used on the assessment, such as "composite" items, were not studied, since there were too few of them to draw conclusions. It is likely our choice of items focused only on item formats measuring lower complexity tasks, as opposed to the universe of possible TEIs that could be explored.

It is also possible students' response behaviors differ substantially by characteristics such as race, language, and culture. Thus, future research should explore the degree to which subgroups of students defined by demographic characteristics (e.g., race, sex, ethnicity, SES, disability, language proficiency) differ with respect to RT, FRL, and other test behaviors.

Another potential limitation of our study is the degree to which students were motivated to do well on the assessments. There may not have been strong motivation for students to try their best, or differential motivation across grades. The students who responded to these items were in elementary school, which may make this potential limitation less of a concern compared to higher grades, when students may become more easily bored or resistant; however, differential motivation could have affected the results. In our study, we did not exclude responses that could be considered disengaged.

With respect to future studies, evaluating the relative measurement precision of the different item formats would be informative. For example, the amount of item information provided per unit of testing time would be an important criterion for evaluating the relative precision of the different item formats that could be used to represent the construct tested in a given amount of testing time. Content validity studies evaluating the differential content representativeness of the item formats would also be illuminating, as would external evaluations of students' cognitive process while responding to items such as think-aloud protocols and cognitive interviews.

Another study of future interest would be to use the information learned in evaluating the different item formats to construct a new test using the most efficient formats from a measurement precision perspective, and those that maximize construct representation. The degree to which scores from this (presumably improved) test correlate more highly with other measures of the construct, and do not correlate with sources of construct-irrelevant variance, would be of interest.

Although our study had limitations, our goals, and the constraints under which we had to work, were more analogous to those encountered by assessment researchers seeking to draw useful conclusions from log data captured in operational settings. Thus, we hope our findings are helpful to practitioners and serve to complement controlled experimental studies on students' response process behaviors and item format effects.

Author Contributions: All authors contributed equally to this project. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by a grant from Educational Testing Service to the University of Massachusetts (contract # S18200000000164). The authors are grateful for this support. The opinions expressed in this paper are those of the authors and do not represent official positions of Educational Testing Service.

Institutional Review Board Statement: Educational Testing Service's Committee for Prior Review of Research (CPRR), through a Chair Review process, determined the de-identified secondary analyses conducted in this study to be Exempt under regulation 46.104(d)(4)(ii).

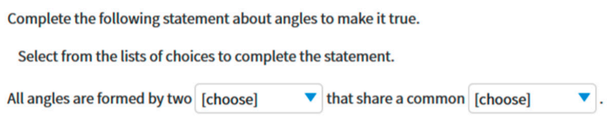
Informed Consent Statement: This study involved secondary analysis of anonymized data.

Data Availability Statement: The data analyzed in this study are not available to outside researchers due to privacy issues.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Examples of Item Formats

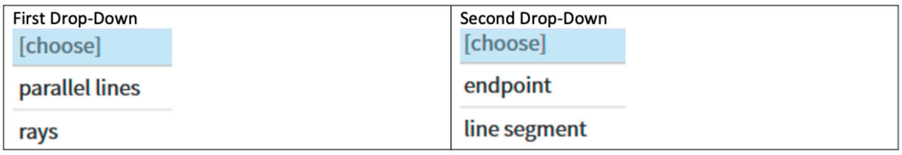
Sample Inline Choice Item



Complete the following statement about angles to make it true.

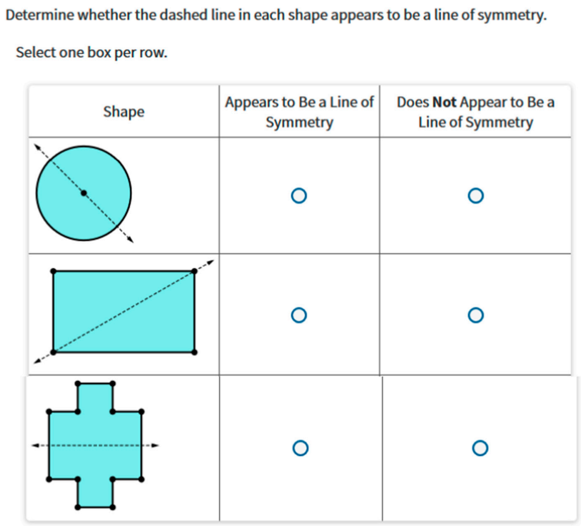
Select from the lists of choices to complete the statement.

All angles are formed by two [choose] ▼ that share a common [choose] ▼ .



First Drop-Down	Second Drop-Down
[choose]	[choose]
parallel lines	endpoint
rays	line segment

Figure A1. Sample Grid Item.

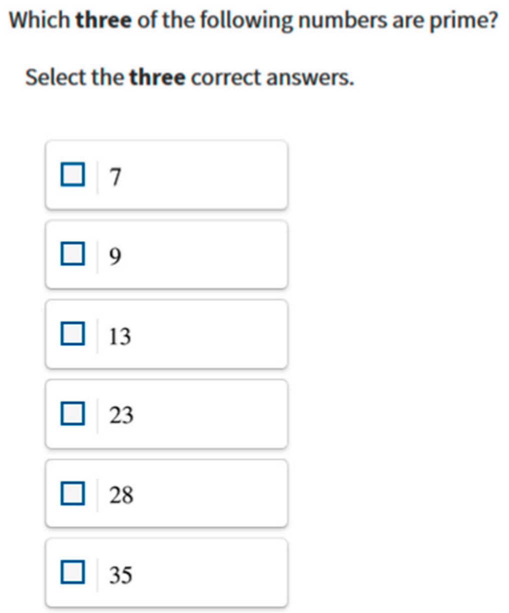


Determine whether the dashed line in each shape appears to be a line of symmetry.

Select one box per row.

Shape	Appears to Be a Line of Symmetry	Does **Not** Appear to Be a Line of Symmetry
	○	○
	○	○
	○	○

Figure A2. Sample Multiple-Choice Multiple-Select Item.



Which **three** of the following numbers are prime?

Select the **three** correct answers.

□ 7

□ 9

□ 13

□ 23

□ 28

□ 35

References

1. van der Linden, W.J. A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika* **2007**, *72*, 287–308. [CrossRef]
2. van der Linden, W.J. Conceptual issues in response-time modeling. *J. Educ. Meas.* **2009**, *46*, 247–272. [CrossRef]
3. Ercikan, K.; Pellegrino, J.W. Validation of score meaning using examinee response processes for the next generation of assessments. In *Validation of Score Meaning for the Next Generation of Assessments*; Routledge: New York, NY, USA, 2017; pp. 1–8.
4. Keehner, M.; Gorin, J.S.; Feng, G.; Katz, I.R. Developing and validating cognitive models in assessment. In *The Handbook of Cognition and Assessment: Frameworks, Methodologies, and Applications*; Wiley: Hoboken, NJ, USA, 2017; pp. 75–101. [CrossRef]
5. Tate, T.P.; Warschauer, M. Keypresses and Mouse Clicks: Analysis of the First National Computer-Based Writing Assessment. *Technology, Knowledge and Learning: Learning Mathematics, Science and the Arts in the Context of Digital Technologies*. 2019. Available online: https://doi-org.silk.library.umass.edu/10.1007/s10758-019-09412-x (accessed on 23 May 2021).
6. Waheed, H.; Hassan, S.-U.; Aljohani, N.R.; Hardman, J.; Alelyani, S.; Nawaz, R. Predicting academic performance of students from VLE big data using deep learning models. In *Computers in Human Behavior*; Elsevier: Amsterdam, The Netherlands, 2020; Volume 104.
7. Li ZBanerjee, J.; Zumbo, B.D. Response time data as validity evidence: Has it lived up to its promise and, if not, what would it take to do so? In *Understanding and Investigating Response Processes in Validation Research*; Zumbo, B.D., Hubley, A.M., Eds.; Springer: Cham, Switzerland, 2017; pp. 159–198.
8. Martinez, M.E. Cognition and the question of test item format. *Educ. Psychol.* **1999**, *34*, 207–218. [CrossRef]
9. Huff, K.L.; Sireci, S.G. Validity issues in computer-based testing. *Educ. Meas. Issues Pract.* **2001**, *20*, 16–25. [CrossRef]
10. Sireci, S.G.; Zenisky, A.L. Computerized innovative item formats: Achievement and credentialing. In *Handbook of Test Development*; Lane, S., Haladyna, T., Raymond, M., Eds.; National Council on Measurement in Education: Washington, DC, USA, 2016; pp. 313–334.
11. Jodoin, M.G. Measurement efficiency of innovative item formats in computer-based testing. *J. Educ. Meas.* **2003**, *40*, 1–15. [CrossRef]
12. Wan, L.; Henly, G.A. Measurement properties of two innovative item formats in a computer-based test. *Appl. Meas. Educ.* **2012**, *25*, 58–78. [CrossRef]
13. Moon, J.A.; Keehner, M.; Katz, I.R. Affordances of item formats and their effects on test-taker cognition under uncertainty. *Educ. Meas. Issues Pract.* **2019**, *38*, 54–62. [CrossRef]
14. Arslan, B.; Jiang, Y.; Keehner, M.; Gong, T.; Katz, I.R.; Yan, F. The effect of drag-and-drop item features on test-taker performance and response strategies. *Educ. Meas. Issues Pract.* **2020**, *39*, 96–106. [CrossRef]
15. Albert, D.; Steinberg, L. Age differences in strategic planning as indexed by the tower of London. *Child Dev.* **2011**, *82*, 1501–1517. [CrossRef] [PubMed]
16. Mitchell, C.L.; Poston, C.L. Effects of inhibiting of response on Tower of London performance. *Curr. Psychol.* **2001**, *20*, 164–168. [CrossRef]
17. Eichmann, B.; Goldhammer, F.; Greiff, S.; Pucite, L.; Naumann, J. The role of planning in complex problem solving. *Comput. Educ.* **2019**, *128*, 1–12. [CrossRef]

18. Goldhammer, F.; Naumann, J.; Rölke, H.; Stelter, A.; Tóth, K. Relating product data to process data from computer-based competency assessment. In *Competence Assessment in Education: Methodology of Educational Measurement and Assessment*; Leutner, D., Fleischer, J., Grünkorn, J., Klieme, E., Eds.; Springer: Cham, Switzerland, 2017.
19. Keehner, M.; Smith, L. Connecting actions, cognitions, and measurement: The role of cognitive science in NAEP TEL task development. In Proceedings of the Annual Meeting of the National Council on Measurement in Education, San Francisco, CA, USA, 30 April 2013.
20. Hambleton, R.K.; Shavelson, R.J.; Webb, N.M.; Swaminathan, H.; Rogers, H.J. *Fundamentals of Item Response Theory*; Sage: Newcastle, UK, 1991; Volume 2.
21. Masters, G.N. A Rasch model for partial credit scoring. *Psychometrika* **1982**, *47*, 149–174. [CrossRef]
22. Cai, L. *Flexible Multilevel Multidimensional Item Analysis and Test Scoring [Computer Software]*; flexMIRT R Version 3.51; Vector Psychometric Group: Chapel Hill, NC, USA, 2017.
23. Bates, D.; Maechler, M.; Bolker, B.; Walker, S. Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* **2015**, *67*, 1–48. [CrossRef]
24. Arslan, B.; Lehman, B. The effects of technology-enhanced item formats on student performance and cognition. In *The Annual Meeting of the National Council on Measurement in Education*; NCME: Gloucester County, NJ, USA, 2021.