Research Report

Influence of Selected-Response Format Variants on Test Characteristics and Test-Taking Effort: An Empirical Study

ETS RR–22-01

Hongwen Guo
Joseph A. Rios
Guangming Ling
Zhen Wang
Lin Gu
Zhitong Yang
Lydia O. Liu

December 2022

ETS Research Report Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

Influence of Selected-Response Format Variants on Test Characteristics and Test-Taking Effort: An Empirical Study

Hongwen Guo[1], Joseph A. Rios[2], Guangming Ling[1], Zhen Wang[1], Lin Gu[1], Zhitong Yang[1], & Lydia O. Liu[1]

1 ETS, Princeton, NJ
2 University of Minnesota, Minneapolis, MN

Different variants of the selected-response (SR) item type have been developed for various reasons (i.e., simulating realistic situations, examining critical-thinking and/or problem-solving skills). Generally, the variants of SR item format are more complex than the traditional multiple-choice (MC) items, which may be more challenging to test takers and thus may discourage their test engagement on low-stakes assessments. Low test-taking effort has been shown to distort test scores and thereby diminish score validity. We used data collected from a large-scale assessment to investigate how variants of the SR item format may impact test properties and test engagement. Results show that the studied variants of SR item format were generally harder and more time consuming compared to the traditional MC item format, but they did not show negative impact on test-taking effort. However, item position had a dominating influence on nonresponse rates and rapid-guessing rates in a cumulative fashion, even though the effect sizes were relatively small in the studied data.

Keywords New item type; test-taking effort; low-stakes assessment

doi:10.1002/ets2.12345

The selected-response (SR) item format is the best choice for test developers interested in efficient and effective assessments, supported by nearly a century of research and development activities (Downing, 2006). The SR item format includes the traditional multiple-choice (MC) item and many variants. The traditional MC item usually has three to five options, and test takers are asked to select one from multiple options (labeled as MC.1 for convenience; Lord, 1980; Rodriguez, 2005). Because of the flexibility and versatility in a wide range of testing applications, the SR format can be used to test cognitive domains ranging from simple recall of knowledge to high-level problem solving, synthesis, or evaluation (Downing, 2006). However, random guessing on SR items, particularly on traditional MC.1 items administered on low-stakes assessments (e.g., those that evaluate institutional performance), is arguably a major weakness of this item type (Martinez, 1999; Meara & Buxton, 1987). On a low-stakes assessment, test takers may exhibit low test engagement and thus have noneffortful responses, such as rapid guessing or skipping questions, which lead to construct-irrelevant variance (Sireci & Zenisky, 2006) and distorted test scores and thereby diminish score reliability and validity (Goldhammer et al., 2017; Rios et al., 2017; Wise, 2017; Wise & DeMars, 2006). The standards written for testing professionals (American Educational Research Association et al., 2014) highlight the responsibility of test designers to provide "evidence of validity, reliability/precision, and fairness" for each intended use (p. 195), and they further state that "test results should be used in conjunction with information from other sources when the use of additional information contributes to the validity of the overall interpretation" (p. 213). Therefore collecting data on students' test-taking effort may help with score validity.

With advances in technologies, test developers have been developing technology-enhanced items to simulate more authentic scenarios, measure more complex constructs, and, at the same time, attempt to improve engagement with the assessment (Jiao & Lissitz, 2017; Veldkamp & Sluijter, 2019). With new opportunities, however, there are challenges: In comparison with traditional items, these technology-enhanced items are typically less accessible to all test takers, more expensive to develop, and likely harder to implement in large-scale contexts (such as the game-based assessment; DiCerbo, 2020); in addition, the resulting data may be harder to analyze and validate with classical approaches (Wools et al., 2019).

Corresponding author: H. Guo, E-mail: hguo@ets.org

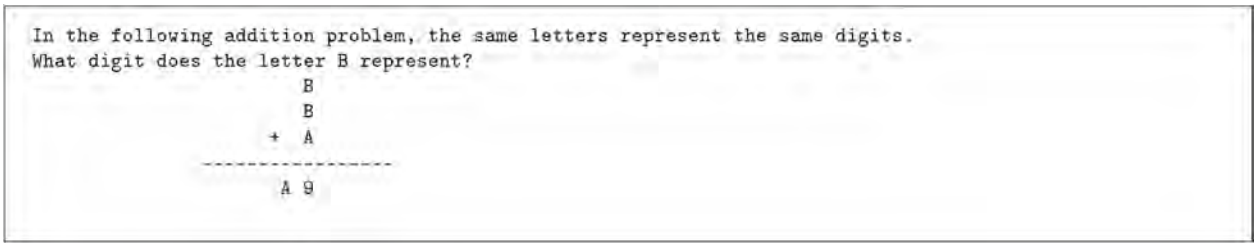


Figure 1 An example of the variant of SR format (multistep item; MC.s) on a mathematics test. To solve the problem, one needs to know that A must be an odd number and less than 3, that is, $A = 1$, and then solve for B. Key is 9.

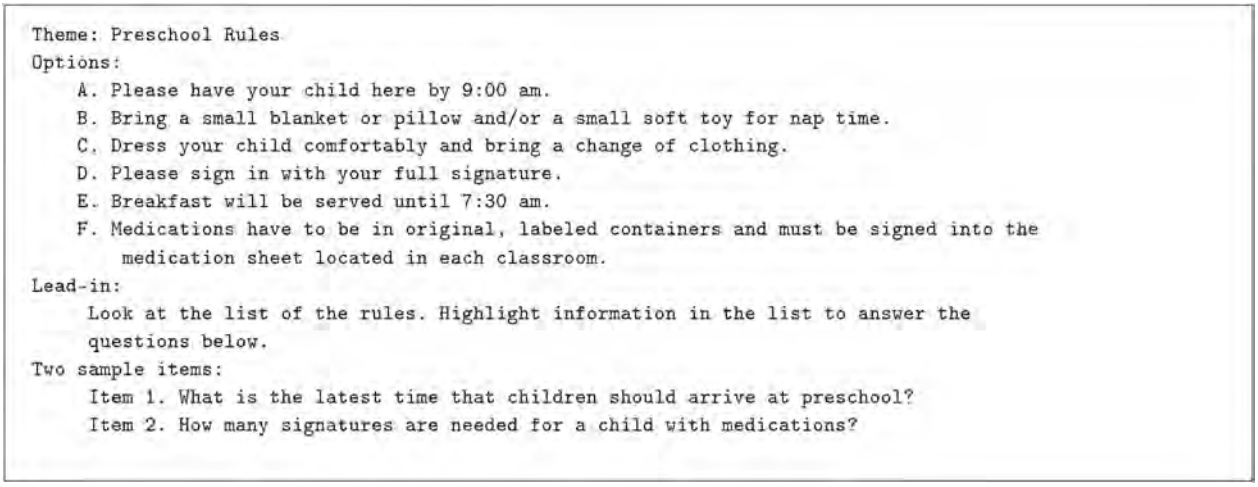


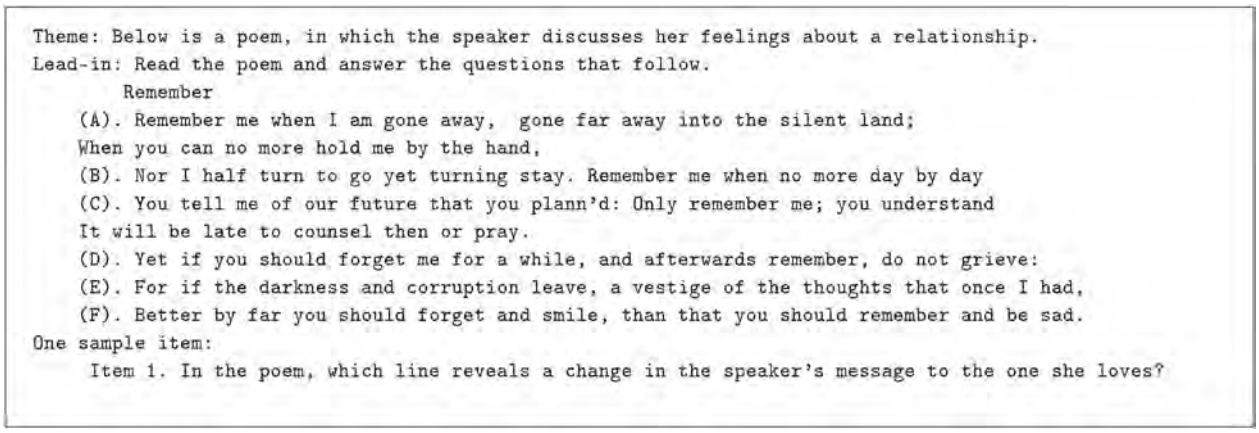
Figure 2 An example of an extended match format that contains the theme, options, lead-in, and two items. Adapted from a Programme for the International Assessment of Adult Competencies (PIACC) item (Organisation for Economic Co-operation and Development [OECD], 2020).

However, variants of the SR item format may be the middle ground between using the traditional MC items and the technology-enhanced items on educational assessments. While the most widely used assessments are built using the traditional select-one MC items (Lord, 1980; Rodriguez, 2005), some variants of the SR item format ask test takers to select two or all that apply from available options (labeled as MC.2 or MC.all). The third variant is to increase the number of options of MC.1 items, say, from four to eight, or even more. The fourth variant is to ask test takers to go through a series of steps involving making selections, resulting in multistep selections (MC.s). As an example, Figure 1 represents a MC.s question.

B could be any digit for 0–9. However, if one knows that A must be an odd number and less than 3, then $A = 1$; the next step is to solve $2B + 1 = 19$ for B.

A more unique variant of the SR item format, the extended matching (EM) format, is particularly promising in that it may be able to assess high-level knowledge and skills, such as problem solving, evaluation, or application (Case & Swanson, 1998; Downing, 2006; Haladyna & Rodrigues, 2013). The EM format uses a list of options linked to a list of items. It has four components: a theme, a set of options, a lead-in statement, and a set of item stems (refer to Figure 2 for an example). In particular, in this EM item set, Item 1 is a MC.1 item; Item 2 also has the select-all-that-apply feature (i.e., a MC.all embedded in the EM format).

The highlighting text format described by Sireci and Zenisky (2006) is another type of EM item (which they call the extended MC item), in which the test taker is asked to select a sentence from a passage that best matches what the item stem requires (such as the main idea of the paragraph or the meaning of a specialized term). When the sentences are labeled with (A), (B), …, the highlighting text format looks like the EM one (refer to Figure 3 for another example). Therefore, in this study, we use EM format to denote both EM and extended multiple choice formats. More variants of the SR item format and new items types can be found in Downing (2006) and Sireci and Zenisky (2006).


Theme: Below is a poem, in which the speaker discusses her feelings about a relationship.
Lead-in: Read the poem and answer the questions that follow.

Remember

(A). Remember me when I am gone away, gone far away into the silent land;
When you can no more hold me by the hand,
(B). Nor I half turn to go yet turning stay. Remember me when no more day by day
(C). You tell me of our future that you plann'd: Only remember me; you understand
It will be late to counsel then or pray.
(D). Yet if you should forget me for a while, and afterwards remember, do not grieve:
(E). For if the darkness and corruption leave, a vestige of the thoughts that once I had,
(F). Better by far you should forget and smile, than that you should remember and be sad.

One sample item:
Item 1. In the poem, which line reveals a change in the speaker's message to the one she loves?

Figure 3 An example of the highlighting text format. Adapted from the Smarter Balanced Assessment (Smarter Balanced Assessment Consortium, 2015). Answer is D.

Well-trained and experienced item writers typically find it relatively easy to produce effective EM items (Sireci & Zenisky, 2006). Besides being efficient, effective, and able to assess higher order knowledge, tests with EM and the aforementioned variants (MC.1 with a large number of options, MC.2, MC.all, MC.m) are typically not technologically demanding when delivered on a computer. Furthermore, they are accessible to all test takers, including those with disabilities and those who have to use their own nonstandard electronic devices in a nonstandard setting (such as at the student's own home), in a similar way to the technology-enhanced items (Wools et al., 2019). In addition, these SR variants can be presented on paper when necessary. One more advantage of these SR variants is that, compared to the traditional MC.1 items, the chance correct rates from random guessing are much smaller, and answers to them may not be easily online searchable, which in turn may improve score validity. However, Wood (2003) showed in an experiment that test takers previously familiar with traditional MC items did relatively poorly the first time they were administered an EM item. Previous studies also showed that some variants of the SR format were likely more difficult and time consuming (Budescu & Nevo, 1985; DeMars, 2000; Guo et al., 2018; Guo et al., 2020; Rodriguez, 2005; van der Linden, 2009; Wise, 2017; Wise et al., 2009). Therefore it is unclear whether variants of the SR format may discourage test-taking effort, particularly when noneffortful responses were caused by unmotivated students who perceived a low probability of success (Freund et al., 2011; Wise & DeMars, 2005).

Given the promises presented by variants of the SR item format and increasing interest in using such an item format in low-stakes educational assessments (National Center for Education Statistics [NCES], 2019; OECD, 2019; Smarter Balanced Assessment Consortium, 2015), it is important to conduct more studies to assess the SR item format and its potential statistical issues (Johnson & Morgan, 2016). More importantly, with additional data recorded from computer-based assessments (log data, such as item response times [RTs], which capture traces of students' test-taking behaviors), researchers are able to investigate what impact these variants of the SR item format may have on item statistics and students' test engagement and what the opportunities and threats are, as suggested by Wools et al. (2019).

In this study, we used data collected from a large-scale low-stakes test to evaluate the performance of some variants of the SR item format. The test included mostly EM item sets with a few discrete items, and items in the EM sets were either MC.1, MC.2, MC.all, or MC.m. The variants of the SR item format were implemented purposely to simulate realistic elements of critical-thinking scenarios that the test intended to measure (Downing, 2006; Halpern, 2014). Because of computer delivery, the data contained both item responses and item RTs.

With a limited number of items on each test form, we aggregated item characteristics and investigated the following four factors related to how items were presented on the test. The first factor was item position. Previous studies (Debeer & Janssen, 2013; Setzer et al., 2013; Weirich et al., 2017; Wise, 2017) showed that items in later positions on a test appeared to be more difficult than in earlier positions for various reasons, such as fatigue, frustration, and low test-taking motivation. However, it is unclear how item position may influence test takers' test-taking effort in terms of either rapid guessing or omitting these items. We anticipated that the test takers may show higher rates of rapid responding and omitting behaviors on items presented at a later position on the test. The second factor we considered was the number of options associated with an item. As discussed earlier, previous studies (Budescu & Nevo, 1985; Guo et al., 2020; Rodriguez, 2005; van der

Linden, 2009; Wise, 2017) have shown that, when the number of response options exceeds five (larger than the traditionally used numbers, such as three, four, and five), the item's difficulty increases, which may lead to more noneffortful responses (see Rios & Guo, 2020). The third factor investigated was whether the item was the EM set leader. The EM set leader refers to the first item in each EM item set; we were interested in the set leaders because their RT covered reading the components of theme, options, and lead-in, and we expected that they were the most time-consuming items; however, it was unknown whether these items would be associated with increased noneffortful responses. The fourth factor was whether the variant of the SR format was select-one MC or not (i.e., MC.1 vs. non-MC.1). Again, because the non-MC.1 would be more difficult and have a low probability of a correct response when guessing, we were interested in how this characteristic may influence test engagement. Note that, because most items on the test were in EM item sets, it was not possible to study EM versus non-EM items.

Our study focused on two research questions: How these four item presentation factors impact (a) test takers' engagement on the test and (b) item statistics, such as item difficulty and item discrimination. To measure test engagement, we examined both nonresponses and rapid-guessing responses, which are likely to be indicators of low test-taking effort. Rapid-guessing responses were identified by a hybrid flagging rule, which combined a parametric version of the visual inspection of RT distribution (VI; Schnipke, 1995; Schnipke & Scrams, 2002; Wise & DeMars, 2006) and the cumulative probability (CUMP) procedure (Guo et al., 2016; Rios & Guo, 2020), which used both item RT and item response accuracy for classification.

Using two relatively large data sets, the current study took an additional focus on the impact of the factors related to the SR format, as well as item position, on test-taking effort, which are mostly absent from previous studies. Answers to the research questions may provide useful information for future test development and applications of SR item format variants in educational assessments.

Method

Data

The studied data were obtained from an operational critical-thinking test in a higher education student learning outcome assessment suite. These computer-based assessments are used by colleges and universities for accreditation and accountability, to guide curriculum improvement and to measure students' growth and development. Therefore they are low stakes to test takers (i.e., test takers have no personal consequences for their test performance). The critical-thinking test comprised 26 items per form administered in a 45-minute testing session. In this study, we used two of the test forms (labeled as Form 1 and Form 2).

The two test taker samples who took Form 1 and Form 2 were college students across approximately 60 colleges and universities in the United States and Canada. The sample size, percentage of female students, and percentage of students who majored in science, technology, engineering, and mathematics (STEM) are presented in Table 1. Also shown are the percentage of students who studied in the United States and the means (standard deviations) of their standardized test scores for admission (in a scale range of 400 – 1,600). Table 1 shows that the two samples were quite comparable. For example, the *t*-test of no mean difference in admission scores returned a *p*-value of .68.

Variants of the SR item format were used in the test to simulate elements of authenticity and engage test takers to interact with testing materials. For example, to mimic realistic scenarios of critical thinking in the information age, in which students need to recognize evidence-and-conclusion and fact-and-opinion relationships, test takers were asked to mark up a text in an EM item set with a list of 12 statements (Halpern, 2014).

The overall item test information of the two forms is summarized in Table 2. The first column is item position, where Items 1, 9, 15, and 19 are the set leaders associated with four EM item sets. The second column (number of options) shows that the items have different numbers of options, either four or more than four options. The third column (item format)

Table 1 Sample Information of the Two Test Forms

Form	Sample size	Female (%)	Major (%)	United States (%)	Admission score
1	7,296	43	45	63	1169 (216)
2	7,297	41	44	63	1166 (213)

Table 2 Item and Test Information for Forms 1 and 2

	Form 1					Form 2				
	No.Op	Format	AIS	r-Bis	MRT	No.Op	Format	AIS	r-Bis	MRT
Item										
1	>4	MC.1	0.81	0.62	118	4	MC.1	0.69	0.39	171
2	>4	MC.1	0.9	0.57	51	4	MC.1	0.53	0.44	77
3	>4	MC.1	0.7	0.54	64	4	MC.1	0.65	0.6	82
4	>4	MC.all	0.51	0.44	68	4	MC.1	0.46	0.34	80
5	4	MC.1	0.56	0.33	67	4	MC.1	0.62	0.52	65
6	3	MC.all	0.33	0.29	80	4	MC.1	0.43	0.41	42
7	4	MC.1	0.57	0.54	85	4	MC.1	0.54	0.48	85
8	4	MC.1	0.53	0.58	90	4	MC.1	0.57	0.23	72
9	4	MC.1	0.54	0.51	132	4	MC.1	0.74	0.58	116
10	4	MC.1	0.63	0.54	51	4	MC.1	0.57	0.36	61
11	4	MC.1	0.57	0.51	56	>4	MC.1	0.47	0.61	74
12	3 + 3	MC.s	0.47	0.45	43	3	MC.all	0.42	0.48	51
13	4	MC.1	0.56	0.56	38	4	MC.1	0.32	0.32	38
14	>4	MC.1	0.41	0.63	61	>4	MC.all	0.29	0.51	73
15	4	MC.1	0.66	0.59	125	4	MC.1	0.71	0.61	108
16	4	MC.1	0.51	0.41	56	4	MC.1	0.66	0.59	39
17	4	MC.1	0.61	0.57	42	4	MC.1	0.49	0.34	48
18	4	MC.1	0.5	0.57	43	4	MC.1	0.54	0.59	31
19	>4	MC.2	0.18	0.66	125	>4	MC.all	0.13	0.46	124
20	4	MC.1	0.54	0.49	62	4	MC.1	0.45	0.37	40
21	4	MC.1	0.52	0.52	59	>4	MC.1	0.36	0.5	49
22	4	MC.1	0.48	0.52	41	>4	MC.1	0.31	0.38	58
23	>4	MC.1	0.42	0.62	52	>4	MC.all	0.15	0.43	46
24	>4	MC.2	0.19	0.5	54	4	MC.1	0.33	0.58	36
25	4	MC.1	0.4	0.16	57	4	MC.1	0.39	0.5	58
26	4	MC.1	0.28	0.24	58	4	MC.1	0.36	0.43	57
Mean			0.51	0.50	68.38			0.47	0.46	68.50
SD			0.16	0.12	27.77			0.16	0.10	32.21

Note. AIS = average item score. MRT = median response time. No.Op = number of options. r-Bis = item biserial correlation. MC.1 asks test takers to select one option. MC.all asks test takers to select all options that apply (Items 4 and 6). MC.2 asks test takers to select two options (Items 19 and 24). MC.s consists of two steps of "select one from three options" (Item 12). Items 1, 9, 15, and 19 are item set leaders on each form. Items 7, 8, 25, and 26 are discrete items that are not embedded in the extended matching (EM) format.

shows whether items are MC.1 format or non-MC.1 format. The fourth column shows the average item scores, which is equal to the proportion correct, since all items were dichotomously scored. The mean of the average item scores (standard deviation) on the test was 0.51 for Form 1, which indicates that the test difficulty was moderate. The fifth column shows the classical item discrimination power (Drasgow, 1986). The average biserial coefficient (*SD*) on the test was .50 (.12) for Form 1. Noticeably, the last two items on Form 1 have low biserial coefficients and need further examination. The sixth column shows the median RT in seconds. As expected, item set leaders were more time consuming.

The last five columns in Table 2 show the item and test information of Form 2, which is similar to Form 1 in terms of summary statistics, numbers of EM sets and their positions in the test form, and content specifications. We observed that items with more than four options, set leaders, and non-MC.1 were scattered in different positions on the forms. Based on available data, Form 1 had an internal consistency reliability (Cronbach's alpha) of .78 and a standard error of measurement (SEM) of 2.26, and Form 2 had a reliability of .73 and a SEM of 2.30.

Flagging Rule for Rapid Responses

As mentioned earlier, to evaluate students' test engagement, we investigated two indicators of low test-taking effort: non-responses and rapid-guessing responses. Nonresponses include both omitted items (item RT was nonzero but no response was provided) and not-reached items (neither item RT nor response was presented). To identify rapid-guessing responses, many procedures have been developed (Wise, 2017). Procedures that use both item RT and response accuracy (RTRA)

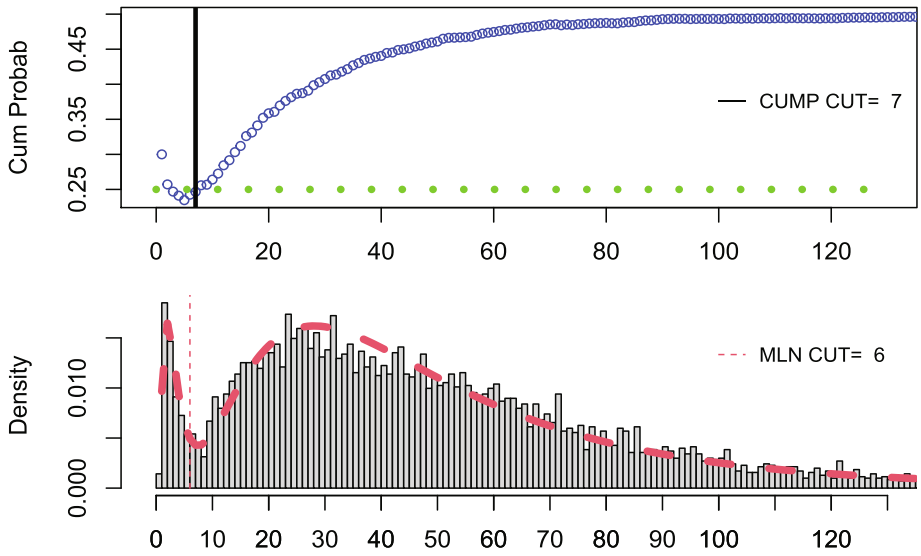


Figure 4 Item plot and the thresholds for Item A. The upper panel shows the cumulative probability (CUMP) method to identify the threshold (CUMP CUT = 7), and the lower panel shows the mixture of lognormal distribution (MLN) method to identify the threshold (MLN CUT = 6).

may result in a valid classification, so that the flagged responses have a correct rate close to the chance score of the item; among them, the CUMP procedure (Guo et al., 2016) makes use of the cumulative probability to mitigate the disadvantage of the RTRA procedures that require substantial response data across the entire RT distribution for a studied item. The CUMP method identifies the RT threshold for an item, at which the cumulative proportion correct rate begins to be consistently above the chance rate for the studied item. Item RTs smaller than the threshold are flagged as rapid-guessing responses (Guo et al., 2016; Rios & Guo, 2020). In the upper panel of Figure 4, the *x*-axis stands for item RT in seconds, and the *y*-axis shows the proportion correct. The thick blue curve made of dots stands for the cumulative probability (CUMP). As more test takers accumulated along testing time, we observed that the CUMP converged to the item difficulty ($P^+ = .50$). The horizontal dotted green line is the item chance score of .25 (because the item has four options). The CUMP curve and the chance line intercept when RT is 7 seconds (represented by the solid vertical black line). Therefore the CUMP procedure returns a threshold of 7 seconds for this item.

One limitation of the CUMP procedure, however, is that threshold cannot be identified for an item when the proportion correct rate is always above or below the chance rate (i.e., a very easy or very hard item).[1] The upper panel in Figure 5 shows such an example, in which the CUMP procedure cannot identify a threshold because the CUMP curve is above the chance score of .25 in the RT range. Students' guessing seemed to have a slightly higher probability than chance on this item (note that the key is Option 3; Attali & Bar-Hillel, 2003). When this is the case, an alternative method can be used, which is the mixture of lognormal distribution (MLN) procedure.

The MLN procedure can be viewed as a parametric version of the VI procedure (e.g., refer to Meyer, 2010; Wise & DeMars, 2006). It assumes that the item RT distribution is bimodal, which can be modeled by a mixture of two lognormal distributions. The lower mode of the RT distribution represents rapid responding, and the upper mode indicates effortful responding. Similar to the CUMP procedure, the MLN procedure can be implemented in an automated process that utilizes an empirical RT distribution, fits a lognormal mixture function, and then locates the lowest point between the two modes of the mixture distribution.

In the lower panels of both Figures 4 and 5, the *x*-axis stands for RT and the *y*-axis for relative frequency. The background histogram is the observed RT distribution, and the dashed red curve is the mixture of two lognormal distributions obtained from the MLN procedure, which returns a threshold of 6 seconds for Item A and 23 seconds for Item B, respectively.

A hybrid approach can be applied on each item by evaluating the two thresholds produced by the CUMP and MLN methods. When the two thresholds were different, as is the case for Item A in Figure 4, to be conservative, we used the smaller one as the threshold to flag rapid-guessing responses for the item. For interested readers, the MLN method

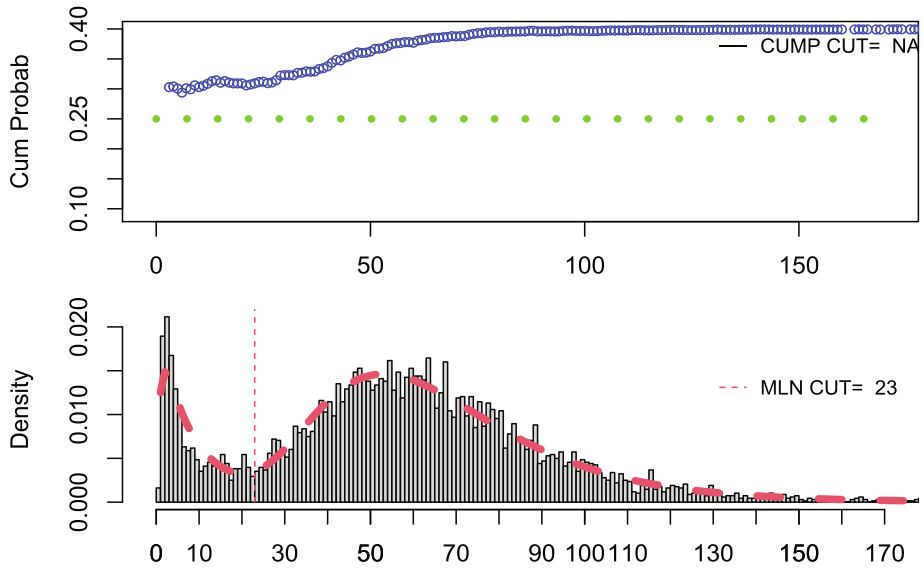


Figure 5 Item plot for the threshold for Item B. The threshold cannot be identified by the cumulative probability approach, but the mixture of lognormal distribution approach returns a threshold of 23 seconds.

is described in greater detail in Appendix A. A similar approach was used by Rios and Guo (2020) to analyze group differences in test-taking effort on an international assessment. The hybrid approach worked well in other large-scale national and international assessments to flag rapid responses that reflected random guessing (Ercikan et al., 2020; Guo & Ercikan, 2021a, 2021b).

Impact on Item Statistics

To evaluate the impact of item presentation on item statistics, we examined item difficulties and item biserial coefficients among different SR variants from the classic test theory (CTT) point of view. Results from the two-parameter logistic (2PL) item response theory (IRT) model generally agreed with findings from CTT (see Appendix B).

Item RTs were examined as well among different variants of the SR format. Results from the lognormal RT (LNRT) model (Fox & Marianti, 2016; van der Linden, 2006),[2] presented in Appendix A, generally agreed with those from median RTs.

Impact on Test Engagement

The impacts of the four item presentation factors on test engagement were presented in descriptive statistics and evaluated in multiple linear regression models. More specifically, the rapid-guessing rates and nonresponse rates of items were regressed, respectively, on item presentation (item position, number of item options, set leader, and item format).

Note that it was often hypothesized that once a test taker started rapid-responding to items, the test taker might have a tendency to do so with all later items on the test (Schnipke & Scrams, 2002). Therefore the observed rapid-response rates may have nonignorable dependency, sometimes a strong trend, which violates assumptions in a regression or analysis of variance (ANOVA) analysis.

To reduce the dependency among the observed rapid-guessing response rates, we used the differencing approach that is commonly used for removing stochastic dependency and a nonlinear trend to obtain a stationary sequence for estimation and prediction. Such a step is also necessary to obtain meaningful sample statistics, such as means and variances of a sequence and its correlations with other variables (Chatfield, 2004; Guo et al., 2017).

More specifically, let $\{r_1, r_2, \ldots, r_J\}$ be the sequence of the rapid-response rates for items in the position of $j = 1, 2, \ldots, J$, where J is the last item on the test. The sequence of the differenced rapid-response rate is defined as $\{d_1, d_2, \ldots, d_{J-1}\}$, where $d_j = r_{j+1} - r_j, j = 1, 2, \ldots, J-1$. Once the dependency is removed, regression can be

conducted appropriately on the differenced rates. The same approach was applied to analyze the nonresponse rates as well.

Results

In this section, we first present the relationship between item presentation and item statistics, and then we show how item presentation impacted nonresponse rates and rapid-guessing rates. Impacts of removing rapid-guessing responses on IRT item parameters are evaluated in Appendix B.

Again, item presentation includes four factors: item position (varied from 1 to 26), number of item options (coded as 0 = four options, 1 = more than four options), set leader (coded as 0 = nonleader, 1 = leader), and item format (coded as 0 = MC.1, 1 = non-MC.1). Both numerical results and visual presentations are reported in the following sections to assist in understanding of the statistics.

To obtain more reliable item analysis results,[3] we combined the items from the two test forms, and thus the following results are based on 52 items.

Item Presentation and Item Statistics

Figure 6 shows how descriptive item statistics differed with regard to each of the four studied factors; the first row shows the average item score in relation to the four factors, the second row shows the item biserial coefficient, and the third row shows the median item RT. The first column of Figure 6, made of scatterplots, shows the descriptive item statistics (on the *y*-axis) against item position (on the *x*-axis). In each scatterplot, the red solid line is the simple linear regression line, and the dotted blue line is the nonparametric regression line. The numeric value on the upper center is the Spearman correlation coefficient between the two studied variables, and the symbols for significance of correlation coefficients are, ∗, ∗∗, and ∗∗∗ for *p*-values less than .10, .05, .01, and .001, respectively. The second to fourth columns of Figure 6 show the box plots of the descriptive item statistics with respect to item leader, number of options, and item format, respectively. In each box plot, the red triangular dots represent the means, and the numeric value in the upper center is the *p*-value of the *t*-test of equal means of the two item groups. Note that the *t*-tests reported in this figure and following figures may not have much power because of the relatively small numbers of items in the comparisons.

As can be observed from the first row of Figure 6, items in the later positions had lower average item scores, with a Spearman correlation coefficient of −.62 between item difficulty (AIS) and item position. Item leaders did not show significant difference on average from nonleaders in item difficulty, and the effect size (the mean difference) was .08. Items with more than four options were somewhat harder than items with four options or fewer on average, and the effect size was −.12. Items with the nontraditional format (non-MC.1, such as MC.2, MC.all, MC.s) were significantly harder than the traditional format (MC.1) on average, with an effect size of −.23. The second row of this figure shows that, on average, items with more than four options (effect size = .05) and the nontraditional format (effect size = −.02) did not have a significant impact on item biserial coefficients, except that item leaders showed significantly higher average discrimination power than the nonleader items, with an effect size of .09. The last row shows that items in later positions had shorter median RTs on average than items in the earlier positions, and the Spearman correlation coefficient was −.45. Also, as expected, set leaders had significantly longer average median RTs, and the effect size was about 70 seconds. However, the number of response options (effect size = 4.38 seconds) and item format (effect size = 7.22 seconds) did not significantly influence item RT on average.

Item Presentation and Test Engagement

In this section, we first present results related to rapid-guessing responses, and then show those related to nonresponses. Again, items from both test forms were used.

We used the hybrid procedure to identify the item RT threshold to flag rapid-guessing responses for each item. To be conservative, the threshold was set to be the minimum of the CUMP threshold and the MLN threshold for each item.[4]

Figure 7 shows, from left to right, the scatterplots of the rapid-guessing response rates (multiplied by 100) against item position, box plots of rapid-guessing response rates by set leader, number of item options, and item format, respectively. The flagged rapid-guessing response rate increased with item position in a curvilinear fashion, and the magnitude of the

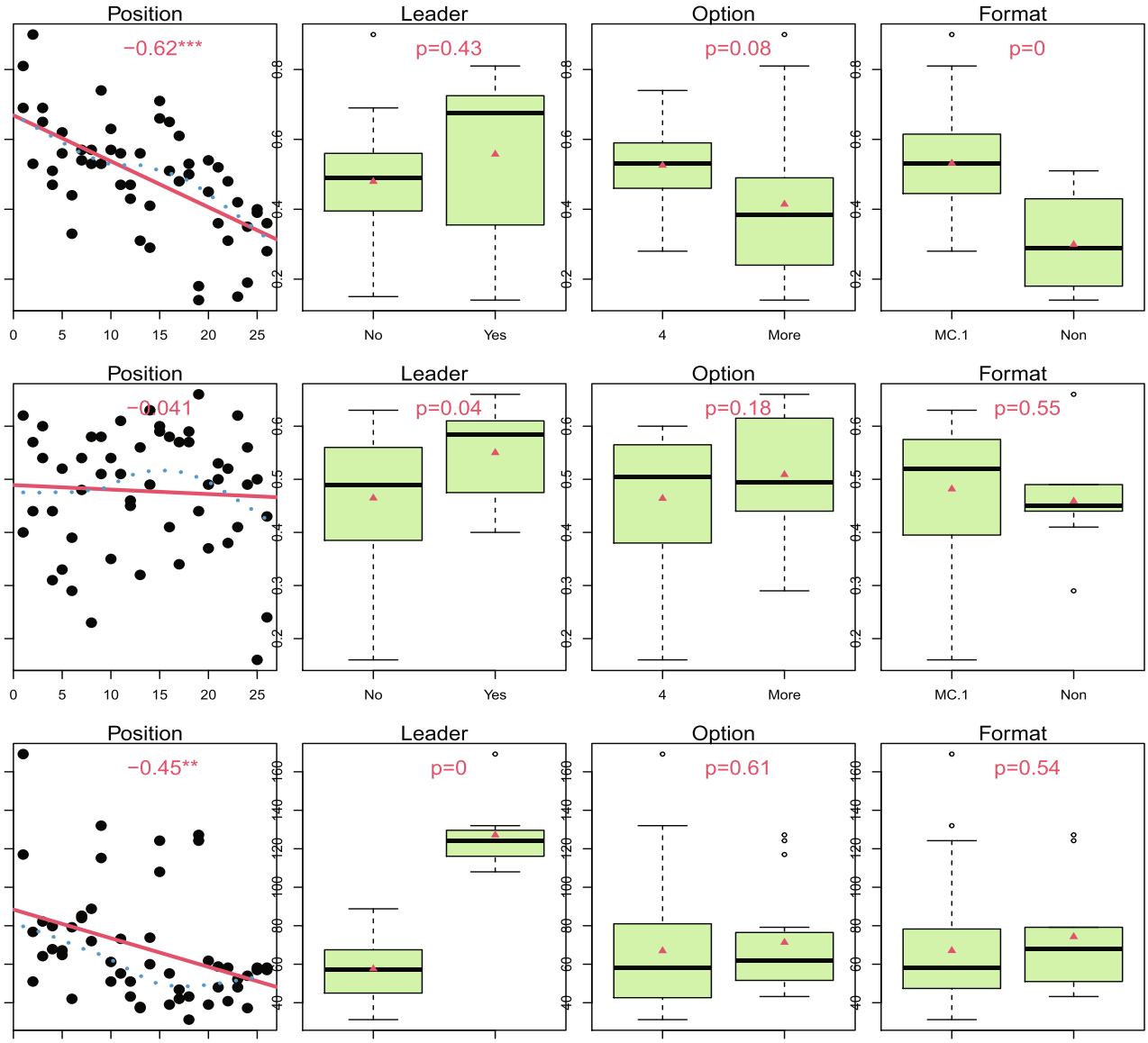


Figure 6 Relationships between descriptive item statistics (row-wise) and item presentation (column-wise). The first, second, and third rows are average item score, item biserial correlation, and median RT, respectively. The first, second, third, and fourth columns are item position, set leader, number of item options, and item format, respectively. In the first column, the numeric value above the scatterplot is the Spearman correlation coefficient; in the next three columns, the *p*-value (rounded to the second decimal place) is for testing equal means of the two item groups.

rapid-response rate was high (reaching 25%). From Figure 7, we also observed that, on average, set leader, number of options, and item format did not make significant differences in the rapid-response rates, as indicated by the *p*-value of the *t*-test in the top center of each box plot and the effect size (−0.82%, 1.32%, and 1.82% for item leader, item with more options, and non-MC.1 item, respectively).

As mentioned in the Method section, researchers have observed that, once low-engaged test takers start rapid-responding to an item, they might do so to subsequent items (Schnipke & Scrams, 2002). In fact, the Durbin–Watson test (Durbin & Watson, 1951) showed that the residuals in the regression model of the rapid-response rate against item presentation were autocorrelated, a sign of strong dependency of the rapid-response rate and item position (the *p*-value was smaller than .001).

Hence we used the differenced rapid-response rates in the regression analysis to remove the strong dependency. The Durbin–Watson test showed that the residuals were not significantly autocorrelated in the multiple linear model of the

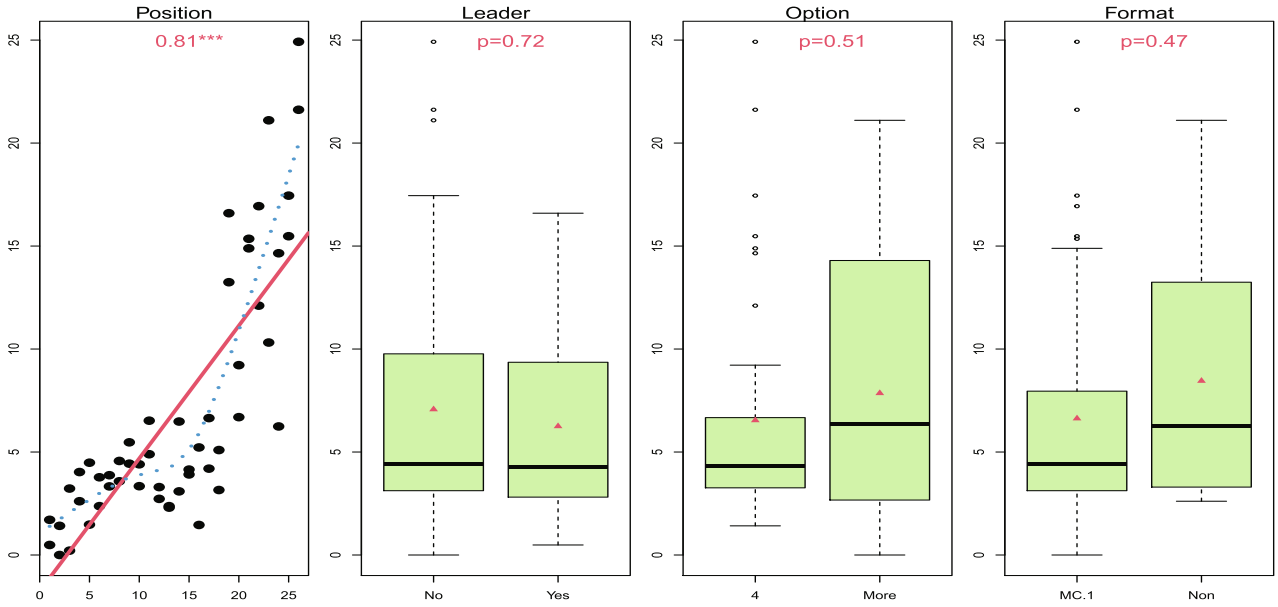


Figure 7 Flagged rapid-response rates with item position, item option, item format, and set leader.

Table 3 Linear Regression of the Differenced Flagged Rate Against Set Leader, Number of Options, and Item Format

	Estimate	*SE*	*t*-value	Pr(>|t|)
(Intercept)	7.85	2.50	3.14	0.00
Position	0.10 (0.19)	0.07	1.45	0.15
Leader	−3.24 (−0.30)	1.40	−2.31	0.03
Option	−1.46 (−0.17)	1.33	−1.10	0.28
Format	−2.23 (−0.22)	1.63	−1.37	0.18

Note. Values in parentheses are the standardized coefficients of the regression.

differenced rapid-response rate on the four item presentation factors (the *p*-value was larger than .12). These results indicate that, collectively, the rapid response on a previous item tends to be integrated into that on the next item in position in a cumulative manner.

The linear regression results of the differenced flagged rapid-response rate against item position, set leader, number of options, and item format are shown in Table 3. Being a set leader significantly decreased the differenced rapid-guessing response rate by −3.24%; that is, on average, when the item is a set leader, the rapid-response rate on this item is that on the previous item minus 3.24%, which may slow down the accumulation of the rapid-response rate. Other factors did not show significant influence.

We then further investigated the relationships between item nonresponse rates and item presentation. Figure 8 shows the nonresponse rate (multiplied by 100) association with item presentation, displayed in the same way as in Figure 7.

We observed in Figure 7 that the nonresponse rate increased with item position, with a curvilinear relationship, but the magnitude was smaller than that in the rapid-response rates. From Figure 7, we again observed that set leader, number of options, and item format did not make significant differences in the nonresponse rate, as indicated by the *p*-value of the *t*-test in the top center of each box plot and the ef ect size (−0.61%, 0.72%, and 0.22% for item leader, item with more options, and non-MC.1 item, respectively).

Similar to the rapid-response rates, the Durbin – Watson test showed that the residuals were autocorrelated in the multiple linear model of the nonresponse rate (*p*-value is smaller than .001); those of the differenced nonresponse rate were not significantly autocorrelated (*p*-value is .18).

Table 4 shows that, after differencing the nonresponse rate, item position still had significantly positive impact (0.04% for each increased position); that is, on average, the nonresponse rate on the next item is that on the previous item plus 0.04%. Items with more than four options had significantly negative impact (−0.32%); that is, when the item has more

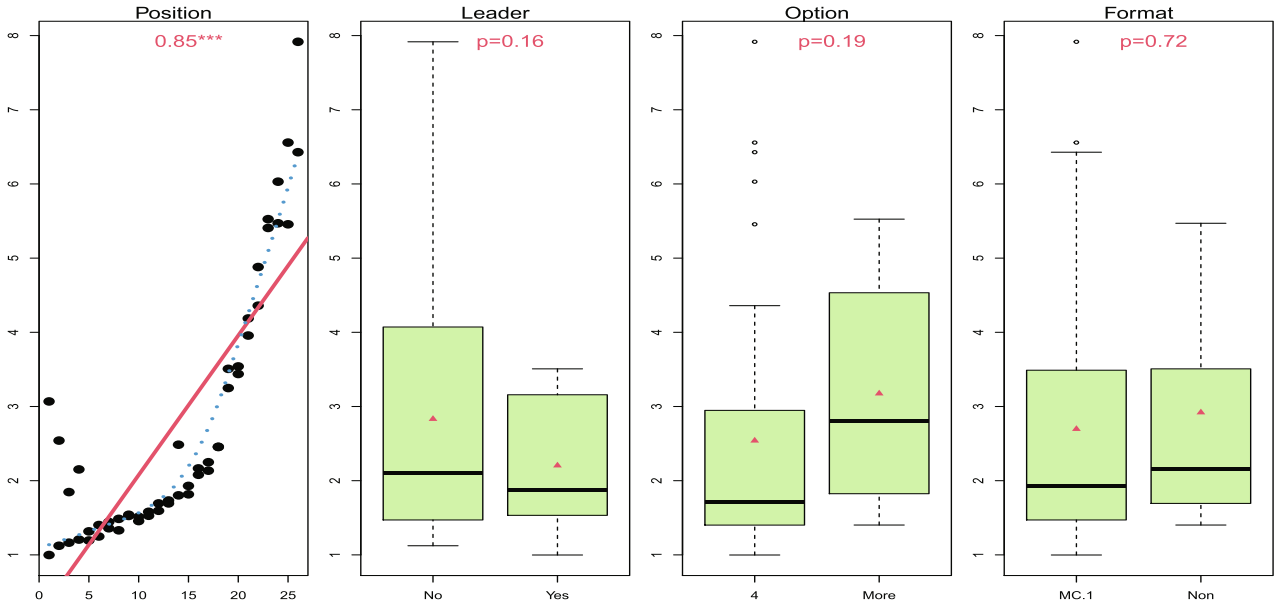


Figure 8 Nonresponses by item position, item option, and item leader.

Table 4 Linear Regression of the Differenced Nonresponse Rate Against Set Leader, Number of Options, and Item Format

	Estimate	*SE*	*t*-value	Pr(>|t|)
(Intercept)	0.35	0.21	1.64	0.11
Position	0.04 (0.64)	0.01	6.38	0.00
Leader	−0.09 (−0.08)	0.12	−0.77	0.45
Option	−0.32 (−0.35)	0.11	−2.88	0.01
Format	−0.09 (−0.08)	0.14	−0.68	0.50

Note. Values in parentheses are the standardized coefficients of the regression.

than four options, the nonresponse rate on this item is that on the previous item minus 0.32%. Set leader and item format did not have a significant impact. Note that the largest nonresponse rate was 8%, so the effect sizes (estimated coefficients) seemed relatively small.

Relationship Between Rapid Responses, Total Score, and Total Response Time

The summary statistics and correlations among test takers' total numbers of flagged rapid responses, total raw scores,[5] and total RTs are presented in Table 5 for the two test forms, respectively.

Overall, Table 5 shows that the average numbers of flagged items are 1.5 and 1.9 for Form 1 and Form 2, respectively. Note that the distribution of the number of flagged items is extremely skewed to the right, and the median number of flagged items is zero; that is, the majority of students had one or two items flagged. The average total scores are 13 and 12 out of 26 on Form 1 and Form 2, respectively, indicating a moderate test difficulty for the student samples. On the 45-minute-long test, the average total times students spent were 1,901 seconds and 1,933 seconds (about 32 minutes) for Form 1 and Form 2, respectively.

In addition, there was a significantly negative association between the number of flagged rapid-guessing responses and the student's test score or total RT (Spearman correlation coefficient = −.44 or −.38 on Form 1 and −.42 or −.44 on Form 2). That is, higher numbers of rapid responses were associated with lower scores and shorter total RTs. As expected on the low-stakes assessments, a significantly positive association was observed between the total score and the total RT as well (Spearman correlation coefficient = .33 on Form 1 and .34 on Form 2), which may indicate that those students who spent less time (possibly with less effort) on the test were likely to get lower scores. Note that, upon inspecting the scatterplots, the associations between scores and RTs were weaker for students who scored 15 points or higher than those

Table 5 Summary and Correlation of Flagged Responses, Total Scores, and Total Time for Test Takers

	Flagged no.	Total score	Total time
Form 1			
Mean	1.536	13.02	1901
SD	3.56	5.07	582.18
Form 2			
Mean	1.925	11.94	1933
SD	3.96	4.60	603.66
Spearman correlation			
Form 1			
Flagged no.		−0.44	−0.38
Total score			0.33
Form 2			
Flagged no.		−0.42	−0.44
Total score			0.34

who scored below 15, further supporting that there was a positive relationship between score and time effort for low performers.

Summary and Discussion

To simulate realistic scenarios, construct engaging items, and measure high-order knowledge for low-stakes assessments, researchers have been developing many variants of the SR item format that are different from the traditional MC (MC.1) items with three, four, and five options (NCES, 2019; OECD, 2019; Smarter Balanced Assessment Consortium, 2015). Besides improved construct representation, these items are easy to develop and implement and have good accessibility, compared to the technology-enhanced items. However, as it is desirable that these items can engage test takers on low-stakes assessments, such as NAEP, PIAAC, PISA, TIMSS, and state accountability assessments, it is important to evaluate the psychometric properties of these nontraditional item types.

In this study, using data collected from two test forms of a low-stakes assessment, we explicitly examined how these SR variants may impact test characteristics and test takers' engagement. To analyze test engagement, the hybrid approach (Guo & Ercikan, 2021a, 2021b; Rios & Guo, 2020) was used to obtain the RT thresholds for flagging rapid-guessing item responses.

Not surprisingly, non-MC.1 items were harder on average than the single-selection MC.1 items on the test, reflected in both classical and IRT item difficulty. Items with more options tended to be more difficult than those with fewer options, but with a large variation; these results support findings from previous literature (Budescu & Nevo, 1985; Rodriguez, 2005; Wise, 2017; Wood, 2003). The most statistically significant factor in item difficulty was associated with item position: Items in later positions were harder than those in earlier positions, sometimes by test design, and potentially by low test engagement, time pressure, fatigue, and other factors, as discussed by Debeer and Janssen (2013) and Weirich et al. (2017). However, item position did not have a significant association with item discrimination. We also found that variants of SR format (MC.1 or non-MC.1) and numbers of options did not have much impact on item discrimination, where set leaders may have slightly higher discrimination power.

As for item RTs, the most significant factor was whether the studied item was an EM set leader, as expected. Set leaders had significantly longer RTs, because test takers had to read the theme, options, and lead-in of the EM format, beyond the item stem. Item position had a somewhat negative impact on item RT; that is, later items had shorter RT overall, which also pointed to potential low test engagement (Schnipke & Scrams, 2002; Wise, 2017). Number of options and item format (MC.1 vs. non-MC.1) did not show significant impact on item RTs. Nevertheless, there may be room for adjusting the large number of options so that the distractors would be functioning better (Guo et al., 2020; Haladyna & Rodrigues, 2013).

For test engagement, the different variants of the SR format (such as large number of options, non-MC.1 format, and item leader) showed limited impact, while item position had the most significant impact. That is, test takers produced significantly more noneffortful (nonresponses or rapid guessing) responses on the later items than on the earlier items on the test. As discussed in the literature, rapid guessing and skipping items were individual student behaviors with large

variations: Some test takers sparsely produced noneffortful responses here and there, switching between effortful/solution and noneffortful response behaviors, and some started with solution behaviors, then switched to noneffortful behaviors, but never switched back. Regardless of the noneffortful responding patterns, prior research has suggested that test engagement decreases across item position (Goldhammer et al., 2017; Wise et al., 2009; Wise & DeMars, 2005) and can bias item parameter estimates and group performance (DeMars & Wise, 2010; Goldhammer et al., 2017; Rios et al., 2017; Wise, 2017). Nevertheless, empirical results from the current study highlight that these noneffortful responses from different individuals may have a distinctive pattern collectively; that is, a noneffortful response on the previous item tends to be integrated into the next item in position in a cumulative manner, and thus these may inflate item difficulty and diminish item discrimination power, particularly for items in a later position.

We also found that the numbers of flagged rapid-guessing responses were negatively and significantly associated with both test scores and test RTs, and test scores and test RTs were positively associated, pointing to a potential test disengagement issue on this low-stakes assessment (Ercikan et al., 2020; Wise, 2017; Wise & DeMars, 2005). Note that, in the studied data, only about 15% of students had more than two rapid-guessing responses on each test form. Given the relatively small portions of students who had many rapid responses, we observed very small differences between item statistics/parameters using full data and those using partial data after removing flagged rapid responses (i.e., treating them as missing in the IRT calibration). The correlation coefficients between the two sets of IRT item parameter estimates are about .99, and those of the LN RT parameters are also above .90, with only slight improvement on model fit (refer to Appendix B). Hence rapid guessing had limited impact on the test's psychometric properties in terms of item parameter estimation and score changes, which agrees with previous case studies on large-scale low-stakes assessments (Debeer et al., 2014; Goldhammer et al., 2016; Setzer et al., 2013).

Overall, findings from the current study generally agree with previous studies that nontraditional SR items (non-MC.1 and large numbers of options) are harder than traditional MC.1 items. However, we found that the studied SR variants may not decrease test-taking effort in terms of the nonresponse rate and rapid-guessing rate.

One limitation of the current study is that the data are observational, collected from an operational testing program, instead of experimental. Hence the four item presentation factors could not be manipulated in our investigation, and thus the impact of the four factors on item statistics and test engagement may have been confounded. Another limitation is that, even though findings from the current study are similar between the two forms of the studied test, the total number of items is relatively small compared to other large-scale assessments (such as NAEP and PISA, which typically have 100–200 items on main subjects in one administration), so more evidence is needed to generalize the findings to other low-stakes tests. In addition, as the studied test population comprised college students, the rapid-response behaviors may not be generalizable to younger students, as observed by Wise (2020). Nevertheless, the statistical approaches proposed in the current study are applicable to further studies.

Implications

Findings from the current study show that, even though the SR variants may be harder and more time consuming, they did not show negative impact on test engagement. Given the increasing interest in developing new item types for assessing educational outcomes (NCES, 2019; OECD, 2019; Smarter Balanced Assessment Consortium, 2015), these findings are encouraging for practitioners and stakeholders who are interested in using SR variants to assess higher level cognitive skills and reduce random-guessing noise.

However, the current study also demonstrated that items in the later positions on the test are significantly associated with increased item difficulty, shortened response item times, and higher rates of noneffortful responses on the low-stakes assessment. When these noneffortful response rates are high, particularly when they are different for different student groups, score validity and score comparability may be undermined for group comparison and for educational program evaluation.

These findings have several implications for educational practitioners who work on low-stakes assessments. From the perspective of test design, items in different content areas need to be spread out on the test, because higher rates of noneffortful responses in later item positions may impact content representative of the assessments. This may also imply that the use of the EM item set with a long list of items may be limited. When test forms are balanced and randomized with short item blocks, as is done with NAEP (NCES, 2020), the impact of noneffortful responses on test properties can also be mediated, which in turn may improve score comparability among student groups. In addition, if low engagement is motivated

by foreseeing unsuccess, adaptive test designs, such as the multistage test design, may improve students' engagement on assessments (Yamamoto et al., 2018). From the perspective of psychometrics, when noneffortful responses are observed in test data, practitioners can apply the methods proposed in the current study (i.e., identifying noneffortful responses, treating them as missing) to evaluate the impact of those responses on test properties and decide whether further data cleaning is necessary to improve item calibration and score reporting (Guo & Ercikan, 2021a, 2021b; Rios et al., 2017; Rios & Guo, 2020; Wise, 2017). In addition, statistical methods, similar to differential item functioning, can be used to test the significance of the differences in rapid-response rates for groups of similar performance (Ercikan et al., 2020; Guo & Ercikan, 2021a)). Practitioners can also investigate the usefulness of proctor notification as a means of test-taking effort monitoring on digitally based assessments (Wise et al., 2019). For general discussion on interventions to improve test-taking effort, interested readers can refer to a recent meta-analysis by Rios (2021).

In summary, with the increasing interest in developing new item types in large-scale state, national, and international assessments, it is clear that more studies are necessary to evaluate the performance of these item formats and their impact on test engagement with data collected from different digitally based assessments.

Notes

1 These may occur when the item has very unattractive distractors or when the test takers tend to choose an option in the middle position (Attali & Bar-Hillel, 2003).
2 Note that the joint model of LNIRT (Fox & Marianti, 2016; van der Linden, 2007) was not reported because of anticipated misfit (caused by the aberrant responses and the bimodal item RT distributions commonly observed in the low-stakes assessment data; refer also to Figures 4 and 5) and because a test taker's testing speed was not part of what the test was designed to measure.
3 If the two college student samples were considered to be randomly equivalent (refer to Table 1), item statistics were comparable. However, scores on the two forms were not comparable because of form differences. In practice, the two form scores were equated.
4 On both forms, the thresholds for about 10 items could not be identified by the CUMP method, while the MLN methods identified thresholds for all items. For items having both CUMP and MLN thresholds, the two methods had a nearly equal chance to be selected for flagging the rapid responses.
5 Nonresponses were treated as incorrect in the total raw score.

References

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

Attali, Y., & Bar-Hillel, M. (2003). Guess where: The position of correct answers in multiple-choice test items as a psychometric variable. *Journal of Educational Measurement*, *40*(2), 109–128. https://doi.org/10.1111/j.1745-3984.2003.tb01099.x

Budescu, D., & Nevo, B. (1985). Optimal number of options: An investigation of the assumptions of proportionality. *Journal of Educational Measurement*, *22*(3), 183–196. https://doi.org/10.1111/j.1745-3984.1985.tb01057.x

Case, S., & Swanson, D. (1998). *Constructing written test questions for the basic and clinical sciences*. National Board of Medical Examiners.

Chatfield, C. (2004). *The analysis of time series* (6th ed.). Chapman and Hall/CRC Press.

Debeer, D., Buchholz, J., Hartig, J., & Janssen, R. (2014). Student, school, and country differences in sustained test-taking effort in the 2009 PISA reading assessment. *Journal of Educational and Behavioral Statistics*, *39*(6), 502–523. https://doi.org/10.3102/1076998614558485

Debeer, D., & Janssen, R. (2013). Modeling item-position effects within an IRT framework. *Journal of Educational Measurement*, *50*(2), 164–185. https://doi.org/10.1111/jedm.12009

DeMars, C. (2000). Test stakes and item format interactions. *Applied Measurement in Education*, *13*(1), 55–77. https://doi.org/10.1207/s15324818ame1301_3

DeMars, C. E., & Wise, S. L. (2010). Can differential rapid-guessing behavior lead to differential item functioning? *International Journal of Testing*, *10*(3), 207–229. https://doi.org/10.1080/15305058.2010.496347

DiCerbo, K. (2020). *How can we use the pandemic as an opportunity to advance assessment innovation?* [Audio podcast]. https://www.listennotes.com/podcasts/beyond-multiple/bmc2020-conference-kickoff-6JCPK4VjXS8/

Downing, S. M. (2006). Selected-response item format in test development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 287–301). Erlbaum. https://doi.org/10.4324/9780203874776.ch12

Drasgow, F. (1986). Polychoric and polyserial correlations. In N. Johnson & S. Kotz (Eds.), *Encyclopedia of statistical sciences* (pp. 68 – 74). Wiley.

Durbin, J., & Watson, G. S. (1951). Testing for serial correlation in least squares regression, II. *Biometrika*, *38*(1/2), 159 – 179. https://doi.org/10.2307/2332325

Durrett, R. (2010). *Probability: Theory and examples*. Cambridge University Press.

Ercikan, K., Guo, H., & He, Q. (2020). Use of response process data to inform group comparison and fairness research. *Educational Assessment*, *25*(3), 179 – 197. https://doi.org/10.1080/10627197.2020.1804353

Fox, J.-P., & Marianti, S. (2016). Joint modeling of ability and differential speed using responses and response times. *Multivariate Behavioral Research*, *51*(4), 540 – 553. https://doi.org/10.1080/00273171.2016.1171128

Freund, P. A., Kuhn, J.-T., & Holling, H. (2011). Measuring current achievement motivation with the QCM: Short form development and investigation of measurement invariance. *Personality and Individual Differences*, *51*(5), 629 – 634. https://doi.org/10.1016/j.paid.2011.05.033

Goldhammer, F., Martens, T., Christoph, G., & Lüdtke, O. (2016). *Test-taking engagement in PIAAC* (Education Working Paper No. 133). OECD Publishing.

Goldhammer, F., Martens, T., & Lüdtke, O. (2017). Conditioning factors of test-taking engagement in PIAAC: An exploratory IRT modeling approach considering person and item characteristics. *Large-Scale Assessments in Education*, *5*, Article 18. https://doi.org/10.1186/s40536-017-0051-9

Guo, H., & Ercikan, K. (2021a). Differential rapid-responding across language and cultural groups. *Educational Research and Evaluation*, *26*(5 – 6), 302 – 327. https://doi.org/10.1080/13803611.2021.1963941

Guo, H., & Ercikan, K. (2021b). *Using response-time data to compare the testing behaviors of English language learners (ELLs) to other test-takers (non-ELLs) on a mathematics assessment* (Research Report No. RR-21-25). ETS. https://doi.org/10.1002/ets2.12340

Guo, H., Ling, G., & Frankel, L. (2020). *Using existing data to inform development of new item types* (Research Report no. RR-20-01). ETS. https://doi.org/10.1002/ets2.12284

Guo, H., Rios, J. A., Haberman, S., Liu, O. L., Wang, J., & Paek, I. (2016). A new procedure for detection of students' rapid guessing responses using response time. *Applied Measurement in Education*, *29*(3), 173 – 183. https://doi.org/10.1080/08957347.2016.1171766

Guo, H., Robin, F., & Dorans, N. (2017). Detecting item drift in large-scale testing. *Journal of Educational Measurement*, *54*(3), 265 – 284. https://doi.org/10.1111/jedm.12144

Guo, H., Zu, J., & Kyllonen, P. (2018). *A simulation-based method for finding the optimal number of options for multiple-choice items on a test* (Research Report No. RR-18-22). ETS. https://doi.org/10.1002/ets2.12209

Haladyna, T. M., & Rodrigues, M. C. (2013). *Developing and validating test items*. Routledge. https://doi.org/10.4324/9780203850381

Halpern, D. F. (2014). *Thought and knowledge: An introduction to critical thinking*. Psychology Press.

Jiao, H., & Lissitz, R. W. (Eds.). (2017). *Technology enhanced innovative assessment: Development, modeling, and scoring from an interdisciplinary perspective*. Information Age.

Johnson, R., & Morgan, G. (2016). Item types, response formats, and consequences for statistical investigations. In K. Schweizer & C. DiStefano (Eds.), *Principles and methods of test construction: Standards and recent advances* (pp. 82 – 103). Hogrefe.

Lord, F. (1980). *Applications of item response theory to practical testing problems*. Erlbaum.

Martinez, M. E. (1999). Cognition and the question of test item format. *Educational Psychologist*, *34*(4), 207 – 218. https://doi.org/10.1207/s15326985ep3404_2

Meara, P., & Buxton, B. (1987). An alternative to multiple choice vocabulary tests. *Language Testing*, *4*(2), 142 – 154. https://doi.org/10.1177/026553228700400202

Meyer, J. P. (2010). A mixture Rasch model with item response time components. *Applied Psychological Measurement*, *34*(7), 521 – 538. https://doi.org/10.1177/0146621609355451

National Center for Education Statistics. (2019). *Advancements in assessments*. https://nces.ed.gov/nationsreportcard/about/advancements.aspx

National Center for Education Statistics. (2020). *Student booklet block design*. https://nces.ed.gov/nationsreportcard/tdw/instruments/cog_blockdesign.aspx

Organisation for Economic Co-operation and Development. (2019). *PISA 2015 assessment and analytical framework*. http://www.oecd.org/education/pisa-2015-assessment-and-analytical-framework-9789264255425-en.htm

Organisation for Economic Co-operation and Development. (2020). *Organisation for economic co-operation and development skill surveys: Literacy — Sample items*. http://www.oecd.org/skills/piaac/

Rios, J. (2021). Improving test-taking effort in low-stakes group-based educational testing: A meta-analysis of interventions. *Applied Measurement in Education*, *34*(2), 85 – 106. https://doi.org/10.1080/08957347.2021.1890741

Rios, J., & Guo, H. (2020). Can culture be a salient predictor of test-taking engagement? An analysis of differential noneffortful responding on an international college-level assessment of critical thinking. *Applied Measurement in Education*, *33*(4), 263 – 279. https://doi.org/10.1080/08957347.2020.1789141

Rios, J. A., Guo, H., Mao, L., & Liu, O. L. (2017). Evaluating the impact of noneffortful responses on aggregated scores: To filter unmotivated examinees or not? *International Journal of Testing*, *1*(1), 74–104. https://doi.org/10.1080/15305058.2016.1231193

Rodriguez, M. (2005). Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. *Educational Measurement: Issues and Practices*, *24*(2), 3–13. https://doi.org/10.1111/j.1745-3992.2005.00006.x

Schnipke, D. L. (1995). *Assessing speededness in computer-based tests using item response times* [Unpublished doctoral dissertation]. Johns Hopkins University, Baltimore, MD.

Schnipke, D. L., & Scrams, D. J. (2002). Exploring issues of examinee behavior: Insights gained from response-time analyses. In C. N. Mills, M. Potenza, J. J. Fremer, & W. Ward (Eds.), *Computer-based testing: Building the foundation for future assessments* (pp. 237–266). Erlbaum.

Setzer, J. C., Wise, S. L., van den Heuvel, J. R., & Ling, G. (2013). An investigation of examinee test-taking effort on a large-scale assessment. *Applied Measurement in Education*, *26*(1), 34–49. https://doi.org/10.1080/08957347.2013.739453

Sireci, S. G., & Zenisky, A. L. (2006). Innovative item formats in computer-based testing: In pursuit of improved construct representation. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 329–347). Erlbaum. https://doi.org/10.4324/9780203874776.ch12

Smarter Balanced Assessment Consortium. (2015). *Smarter Balanced Assessment 8th grade ELA*.

van der Linden, W. J. (2006). A lognormal model for response times on test items. *Journal of Educational and Behavioral Statistics*, *31*, 181–204. https://doi.org/10.3102/10769986031002181

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, *72*, 287. https://doi.org/10.1007/s11336-006-1478-z

van der Linden, W. J. (2009). Conceptual issues in response-time modeling. *Journal of Educational Measurement*, *46*(3), 247–272. https://doi.org/10.1111/j.1745-3984.2009.00080.x

Veldkamp, B. P., & Sluijter, C. (Eds.). (2019). *Theoretical and practical advances in computer-based educational measurement, methodology of educational measurement and assessment*. Springer. https://doi.org/10.1007/978-3-030-18480-3_1

Weirich, S., Hecht, M., Penk, C., Roppelt, A., & Böhme, K. (2017). Item position effects are moderated by changes in test-taking effort. *Applied Psychological Measurement*, *41*(2), 115–129. https://doi.org/10.1177/0146621616676791

Wise, S., Kuhfeld, M., & Soland, J. (2019). The effects of effort monitoring with proctor notification on test-taking engagement, test performance, and validity. *Applied Measurement in Education*, *32*(2), 183–192. https://doi.org/10.1080/08957347.2019.1577248

Wise, S. L. (2017). Rapid-guessing behavior: Its identification, interpretation, and implications. *Educational Measurement: Issues and Practice*, *36*(4), 52–61. https://doi.org/10.1111/emip.12165

Wise, S. L. (2020). The impact of test-taking disengagement on item content representation. *Applied Measurement in Education*, *33*(2), 83–94. https://doi.org/10.1080/08957347.2020.1732386

Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment*, *10*(1), 1–17. https://doi.org/10.1207/s15326977ea1001_1

Wise, S. L., & DeMars, C. E. (2006). An application of item response time: The effort-moderated IRT model. *Journal of Educational Measurement*, *43*, 19–38. https://doi.org/10.1111/j.1745-3984.2006.00002.x

Wise, S. L., Pastor, D. A., & Kong, K. J. (2009). Correlated of rapid-guessing behavior in low- stakes testing: Implications for test development and measurement practice. *Applied Measurement in Education*, *22*(2), 185–205. https://doi.org/10.1080/08957340902754650

Wood, E. J. (2003). What are extended matching sets questions? *Bioscience Education*, *1*(1), 1–8. https://doi.org/10.3108/beej.2003.01010002

Wools, S., Molenaar, M., & Hopster-den Otter, D. (2019). The validity of technology enhanced assessments—Threats and opportunities. In B. P. Veldkamp & C. Sluijter (Eds.), *Theoretical and practical advances in computer-based educational measurement, methodology of educational measurement and assessment* (pp. 3–20). Springer.

Yamamoto, K., Shin, H. J., & Khorramdel, L. (2018). Multistage adaptive testing design in international large-scale assessments. *Educational Measurement*, *37*(4), 16–27. https://doi.org/10.1111/emip.12226

Appendix A

The Lognormal Response Time Model and the Mixture of Lognormal Distribution Method

The LNRT model (Fox & Marianti, 2016) posits a normal density for the distribution of the logarithm-transformed RT, $Y_i = \ln T_i$, for item i as

$$f\left(\ln t_i; \tau, \alpha_i, \beta_i\right) = \frac{a_i}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left[\alpha_i\left(\ln t_i - \left(\beta_i - \tau\right)\right)\right]^2\right\}, \tag{A1}$$

where the mean of the distribution is $\mu_i = \beta_i - \tau$, the standard deviation is the reciprocal of α_i, β_i is the item time intensity or time consumingness, and τ is the person speed (a larger τ indicates a shorter time on the item). Identifiability/constraints are imposed on $\sum_{j=1}^{N} \tau_j = 0$, where N is the number of total test takers.

The MLN method assumes that the item RT distribution is bimodal and can be modeled by a mixture of two lognormal distributions. The lower mode of the RT distribution represents rapid responding, and the upper mode indicates effortful responding. More specifically, let $Y = \ln(T)$ be the logarithm transformation of the original RT T, and let $Y_1 \sim N(\mu_1, \sigma_1) = f_1(y)$ and $Y_2 \sim N(\mu_2, \sigma_2) = f_2(y)$ denote the components of two normal distributions (assuming $\mu_1 < \mu_2$). The distribution of Y is represented by

$$f(y) = \pi_1 f_1(y) + \pi_2 f_2(y), \tag{A2}$$

where π_1 and π_2 are the proportions of the two normal density functions. Then the density function of the original RT T (Durrett, 2010) is

$$\begin{aligned} g(t) &= \frac{f(\ln t)}{t} \\ &= \pi_1 \frac{f_1(\ln t)}{t} + \pi_1 \frac{f_2(\ln t)}{t}. \end{aligned} \tag{A3}$$

The threshold for flagging rapid-guessing responses is defined as the time point $t \in (\mu_1, \mu_2)$, where $g(t)$ reaches the minimum value (which is also the time point where f_1 and f_2 intercept).

Appendix B

Model-Based Results

For comparison purposes between the descriptive item statistics and those from parametric IRT models, we evaluated the relationship between the item IRT parameter estimates and the item presentation (refer to Figures B1 and B2).

The 2PL model was fit to the data, and item fits were not ideal but seemed acceptable. For example, the G^2 tests for item fit were all statistically significant, but the magnitudes of item misfit (root mean square error of approximation) were around 0.05 or less. The LNRT model (Fox & Marianti, 2016) was also fit to timing data to examine the impact of item presentation on RT. In view of the bimodal item RT distributions for almost all items, we did not expect good fit. However, because the left (rapid guessing) mode usually had a relatively smaller proportion (refer to Figures 4 and 5), item fit seemed acceptable (the LNRT program reported no misfitting items at the 5% significance level).

The first two rows in Figure B1 show the item parameter and item presentation interactions. This figure and other similar figures can be read in the same way as Figure 6. Somewhat similar to Figure 6, item discrimination parameter a is not significantly impacted by different set leader, number of options, and format; it is negatively associated with item position without statistical significance. Similar to Figure 6, the item difficulty parameter b is significantly associated with item position: Items in the later positions are harder, and none of the other three item presentation factors had statistically significant impact on b.

The last two rows in Figure B1 show that the item position has a somewhat significantly negative impact on both item RT discrimination and time intensity, and set leader has a significant impact on both RT parameters as well. Noticeably, number of options and item format do not have much impact on RT parameters, which again agrees with those observations in Figure 6. Those on Form 2 in Figure B2 show similar patterns to Form 1.

In addition, we compared the item parameter estimates and item RT parameter estimates before and after screening out rapid-guessing responses (i.e., treated them as missing). The item parameter estimates (a and b) and item intensity parameters (β) are minimally impacted by rapid guessing in view of the strong correlation (the Pearson correlation coefficient $\geq$.99). Item RT discrimination parameters (α) were also highly correlated to the original α (the Pearson correlation coefficient $\geq$.91). Note that in IRT calibration, when using data without rapid-guessing responses (i.e., the rapid-guessing responses were coded as missing), the log likelihood of the 2PL IRT fit slightly increased, in agreement with previous findings (Guo et al., 2016; Rios et al., 2017; Wise & DeMars, 2006).

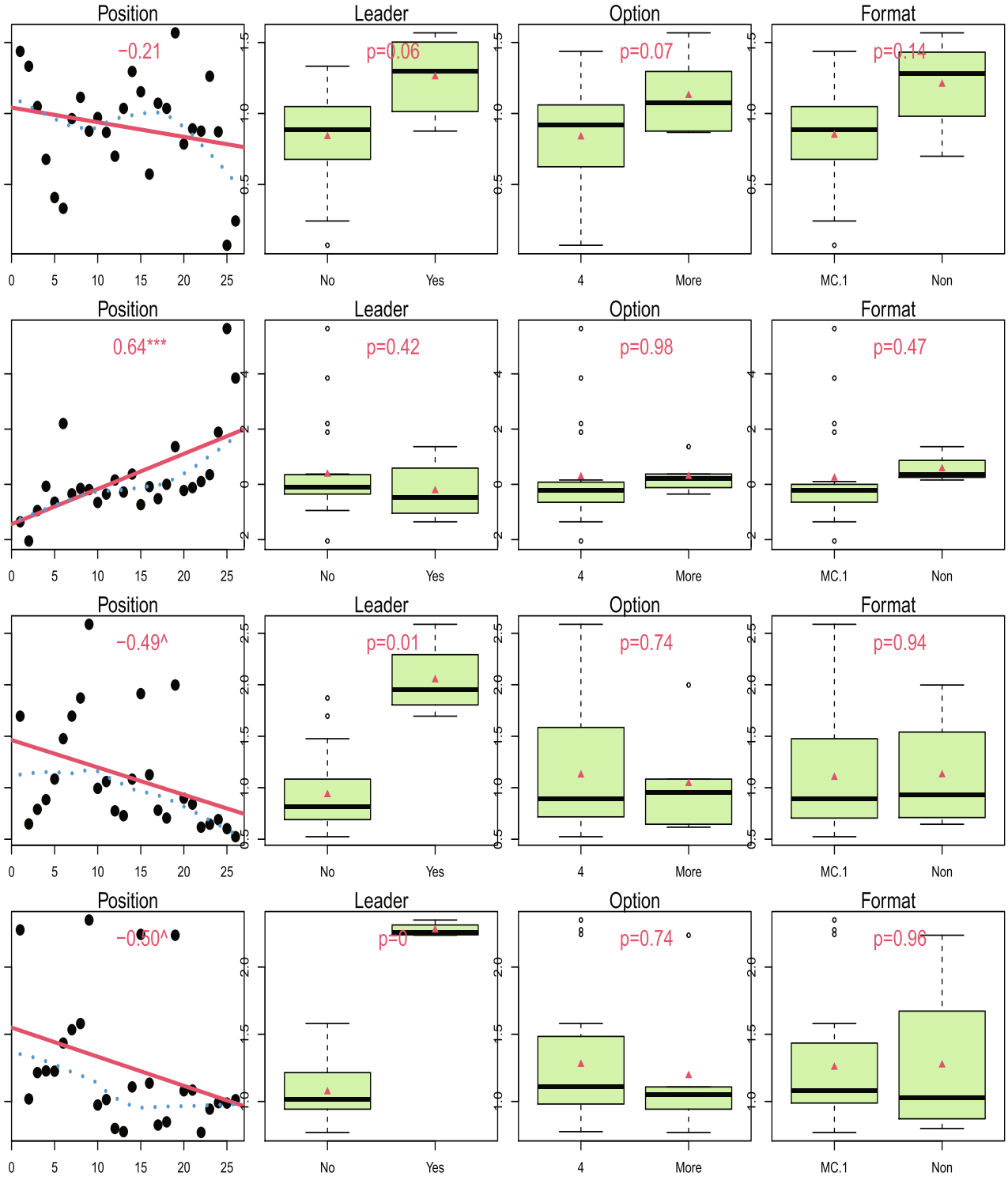


Figure B1 Item discrimination parameters *a* (first row), item difficulty parameter *b* (second row), response time discrimination parameter alpha (third row), and response time intensity parameter beta (fourth row) with item position, item option, and item leader on Form 1.

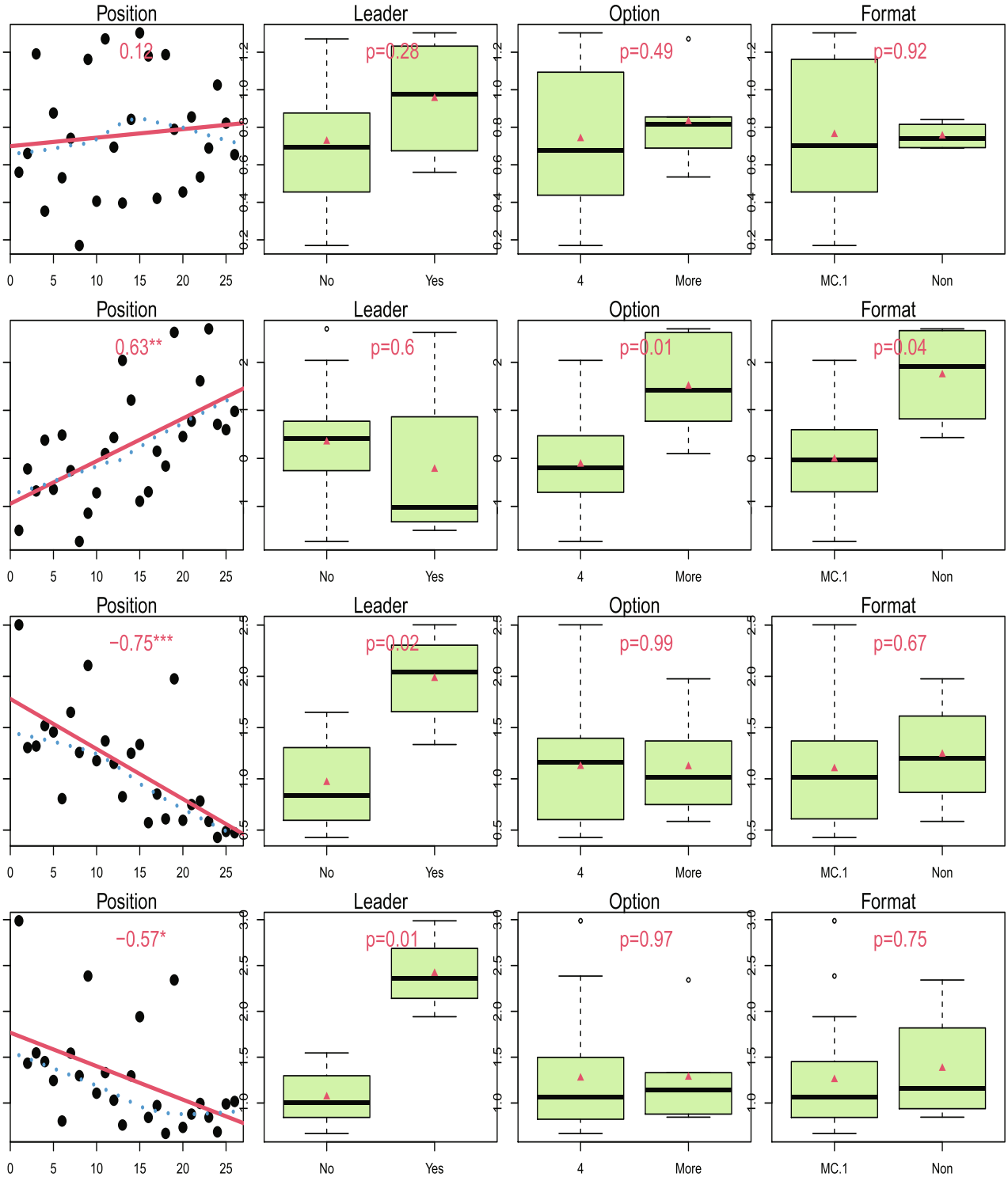


Figure B2 Item discrimination parameters *a* (first row), item difficulty parameter *b* (second row), response time discrimination parameter alpha (third row), and response time intensity parameter beta (fourth row) with item position, item option, and item leader on Form 2.

Suggested citation:

Guo, H., Rios, J. A., Ling, G., Wang, Z., Gu, L., Yang, Z., & Liu, L. O. (2022). *Influence of selected-response format variants on test characteristics and test-taking effort: An empirical study* (Research Report No. RR-22-01). ETS. https://doi.org/10.1002/ets2.12345

Action Editor: Shelby Haberman

Reviewers: Ru Lu and Gautam Puhan