Research Report

Using Existing Data to Inform Development of New Item Types

ETS RR–20-01

Hongwen Guo
Guangming Ling
Lois Frankel

December 2020

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

Using Existing Data to Inform Development of New Item Types

Hongwen Guo, Guangming Ling, & Lois Frankel

Educational Testing Service, Princeton, NJ

With advances in technology, researchers and test developers are developing new item types to measure complex skills like problem solving and critical thinking. Analyzing such items is often challenging because of their complicated response patterns, and thus it is important to develop psychometric methods for practitioners and researchers to analyze these new item types. In this study, we describe a generic approach that involves data-driven analyses and expert feedback from different research areas so that the analysis results can provide valuable information to test developers and researchers on how complex item types contribute to score reliability and validity and on how to make the test more efficient and reliable in measuring complex skills. A real data example was used to illustrate how to identify nonfunctioning options that might be removed from the test and whether partial credit for certain response selections can be considered.

Keywords Test reliability; partial credit; nonfunctioning option; data-driven simulation

doi:10.1002/ets2.12284

With advances in technology, researchers and test developers have been developing new item types to measure complex skills like problem solving and critical thinking. New item types include complex multiple-choice (MC) items, which have appeared on many assessments, including the National Assessment of Educational Progress (NAEP) and the Programme for International Student Assessment (PISA; National Center for Education Statistics, 2019; Organisation for Economic Co-operation and Development, 2019). It is often challenging to analyze these new items because of their complicated response patterns, and thus it is urgent to develop statistical and psychometric methods for practitioners and researchers to analyze these new items.

The traditional MC items require test takers to select one option from the four or five choices typically provided (denoted as MC.1 in this study). This practice has been shown to be effective and efficient in measuring test takers' ability, theoretically and empirically (Budescu & Nevo, 1985; Lord, 1980; Rodriguez, 2005; Rodriguez et al., 2014; Schneid et al., 2014). However, among others, the two drawbacks of MC.1 items are that they may have limited efficacy in assessing problem-solving or critical-thinking skills (Halpern, 2010; Ku, 2009; Liu et al., 2014) and they are prone to guessing, particularly for low-stakes assessments like NAEP and PISA (Goldhammer et al., 2017; Guo et al., 2016; Rios et al., 2017; Wise, 2017). To address the limitations of traditional MC.1 items, researchers call for complex item formats in assessments, including MC with multiple selections (MC.m) and multiple steps in selection (MC.s), which would offer the potential for a psychometrically sound assessment. In addition, these items (including MC.1 items) may be further modified to include more than five options, possibly to reduce guessing but also to present a more realistic problem-solving situation. For example, to simulate a realistic scenario in the digital age where abundant sources, true, false, or irrelevant, are all available, students are asked to evaluate these different sources and find the true ones; this would lead to an MC item with many options and possible multiple selections in multiple steps. While test developers are encouraged to make all, or as many as possible, of the distractors functional (i.e., providing adequate discrimination power and frequency of response of at least 5%, as defined by Haladyna & Downing, 1989a, 1989b, and Rodriguez, 2005), in practice, some of the options are rarely selected (Schneid et al., 2014). Haladyna and Downing (1988) defined the nonfunctioning option as having either no discrimination power or a frequency of response less than 5%. Nonfunctioning options may be implausible even to less proficient test takers and may increase the time spent on each MC question without making any contribution to item discrimination.

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

These modified MC items (i.e., MC.m, MC.s, and MC.1 with many options) may help to measure complex skills and reduce the chance of a random guess. However, excessive numbers of options or response selections may add little value to the test but noise (Lord, 1980). For example, we observed in an operational program that for an item with 12 options, test takers' responses contained more than 200 different response selections when they were asked to select all options that apply. The chance score of this item was undoubtedly near zero. However, most response selections were not functioning well and were chosen by only one or two test takers in a sample of nearly 7,000. Hence researchers may want to reduce the large numbers of options for these items.

Meanwhile, complex MC items are typically harder to develop and are more challenging to test takers. Certain response selections of test takers may contribute to their ability estimation even though the response selections are not completely correct (Bock, 1972; Frary, 1989; Thissen, 1976; Thissen et al., 1989), especially when these response selections show non-negligible positive association with ability based on data. Thus there is value to assigning partial credit to these response selections. Partial credit helps to extract more useful information from the complex response structure, and therefore it helps to increase test internal consistency without lengthening a conventional MC.1 test.

As is evident in most studies, reducing the number of options may also reduce item difficulty, item discrimination, and internal consistency reliability at various degrees (Budescu & Nevo, 1985; Guo et al., 2018; Rodriguez, 2005; Rodriguez et al., 2014). On the other hand, assigning partial credit may potentially increase the test reliability (Frary, 1989). For a specific testing program, we are interested in knowing to what extent the combination of the preceding two actions may impact the psychometric properties of the test based on available data. That is, with existing data, how can we find out which options were nonfunctioning and which response selections may be considered for partial credit? Before practitioners and researchers plan further field trial studies of the new item types with different numbers of options and/or partial credit to improve the test and develop new forms, can existing data shed light on what we can expect for the psychometric properties of a revised test design? In this report, we use empirical data to illustrate necessary psychometric and statistical procedures to analyze these complex MC items and to assist content experts in test development.

The report is organized as follows. Because the procedures are data driven, in the Data section, we introduce a data set collected from a higher education testing program that contains MC.1, MC.m, and MC.s items for illustration purposes. In this section, item response and response time are analyzed to evaluate how items functioned in the studied sample, with summary statistics and visual inspection. In the Methods section, we describe procedures and reasoning on which options are nonfunctioning and which response selections may be considered to have partial credit. Once decisions are made on item revision, we conduct simulations to assign responses from the remaining options to those test takers whose chosen options/response selections are removed; thus the potential impact of revised form on test properties can be evaluated. Most importantly, our approach in the current study takes into account content experts' opinions, in addition to data-driven analyses and psychometric principles. In the Results section, we present results of simulations. The impact of reduced number of options and partial credit assignment on the test is evaluated by changes in item difficulty, item discrimination power, test reliability, and test scores. In the last section, we summarize the results and discuss implications for the proposed item analysis approach.

All statistical analyses in the study were produced in R (R Core Team, 2019), and partial R codes are provided in Appendix A to assist practitioners and researchers in analyzing new item types.

Data

For illustrative purposes, we used a data set obtained from a large-scale higher education assessment operational program that measures critical-thinking skills. The test comprised 26 MC items administered in a 45-minutes testing session via computer. Therefore item raw responses, item scores, and item response times were available for our analysis. Our data contain responses from 7,296 U.S. and Canadian college students from approximately 60 colleges that used this assessment. Of the students, 43% were identified as female, and 48% identified as male; 44% of the students majored in science, technology, engineering, and mathematics subjects. In addition, 35% of the students studied in Canada, and 63% studied in the United States. The test is number-right scored; each item has a binary score of 1 if correct, and 0 otherwise. The test's internal consistency reliability (Cronbach's alpha) and standard error of measurement were .78 and 2.26, respectively.

A variety of nontraditional structural features and task/item types was included in the test to simulate elements of authenticity and to engage test takers to interact with the test and use their critical-thinking skills to make selections

Table 1 Item Information and Summary, Including Number of Option, Format, Difficulty, Polyserial, and Response Time Recorded in Seconds

Item	No. options	Item format	Average item score	Item polyserial	Median RT
1	>8	MC.1	.81	.62	118
2	>8	MC.1	.90	.57	51
3	8	MC.1	.70	.54	64
4	>8	MC.all	.51	.44	68
5	4	MC.1	.56	.33	67
6	3	MC.all	.33	.29	80
7	4	MC.1	.57	.54	85
8	4	MC.1	.53	.58	90
9	4	MC.1	.54	.51	132
10	4	MC.1	.63	.54	51
11	4	MC.1	.57	.51	56
12	3 + 3	MC.m	.47	.45	43
13	4	MC.1	.56	.56	38
14	>8	MC.1	.41	.63	61
15	4	MC.1	.66	.59	125
16	4	MC.1	.51	.41	56
17	4	MC.1	.61	.57	42
18	4	MC.1	.50	.57	43
19	8	MC.2	.18	.66	125
20	4	MC.1	.54	.49	62
21	4	MC.1	.52	.52	59
22	4	MC.1	.48	.52	41
23	8	MC.1	.42	.62	52
24	8	MC.2	.19	.50	54
25	4	MC.1	.40	.16	57
26	4	MC.1	.28	.24	58

Note. Items 1, 9, 15, and 19 are the first items (set lead) in their sets. >8 = items with larger than eight options are denoted; MC = multiple choice; MC.1 = traditional MC items that ask test takers to select one option; MC.2 = MC items that ask test takers to select two options; MC.all = MC items that ask test takers to select all options that apply; MC.m = MC items with multiple selections; MC.s = MC items that consist of two steps of selecting one option from three; RT = response time.

(Halpern, 2010; Ku, 2009; Liu et al., 2014). For example, in some items, test takers were provided a list of 12 statements and asked to select those that satisfied a requirement given in the instructions, so such an MC item would have 12 options. In other items, test takers were asked to go through a series of steps to make selections. Table 1 provides the item information and item summary statistics.

From Table 1, we observe that items on the test have different numbers of options (Column 1) and different formats (Column 2). Column 3 is the average item score between 0 and 1, which shows that items were presented in order of difficulty, from easy items to difficult items (note that a large value for the average item score indicates an easy item). Besides discrete items, there were four item sets; each item set was based on a common stimulus. Items 1, 9, 15, and 19 were the first items in their respective sets (set lead of an item set). As expected, Column 4 (item median response time) shows that test takers spent a much longer time on the four lead items than on other items, because the time included both reading the stimulus material and responding to the lead item.

Methods

Our analysis has four steps. The first step is a data-driven approach to flag options that are nonfunctioning or that are candidates for partial credit; the second step invites content experts to investigate the flagged options from content perspectives and make recommendations, so that agreement can be reached among content experts and psychometricians; the third step is to simulate new item scores for the removed options and response selections. The last step evaluates the impact of these actions on the test.

Table 2 Summary Statistics of the Answer Key and the Four Least Attractive Options in an MC.1 Item

Option	Q	Key	R	S	T
Freq.O	.00	.79	.01	.00	.00
Score.O	5.40	14.27	10.10	6.81	7.10
Poly.O	−.66	.62	−.24	−.56	−.53

Note. MC.1 = traditional multiple-choice items that ask test takers to select one option, Freq.O = observed relative frequency of an option; Poly.O = the largest polyserial coefficient; Score.O = the average total test score of the subgroup that chose the particular option.

Flagging Options

Nonfunctioning options are the least popular options and the options with the smallest discrimination power. Previous studies (Lord, 1980; Rodriguez, 2005) have shown that removing options with the smallest discrimination power is preferable in terms of maintaining a high test internal consistency reliability. To estimate the discrimination power of an option, we can use the polyserial coefficient that measures the association between the option and the total test score (Drasgow, 1986) or calibrate item discrimination parameters using the nominal response model (NRM; Bock, 1972).

However, because of the large numbers of options and response selections and thus sparse data for these response selections in our sample, NRM calibration is not feasible, and the polyserial coefficient is not reliable for options that have very small frequencies. Therefore we used the least popular option as the criterion to identify nonfunctioning options for each item. For this, we computed option summaries for each item: the observed relative frequency of an option (Freq.O), the average total test score of the subgroup that chose the particular option (Score.O), and the polyserial coefficient of the option (Poly.O).

Table 2 shows statistical summaries of an example MC.1 item with more than eight options. The rows of Table 2 show Freq.O, Score.O, and Poly.O, respectively. For this item, the key has the highest subgroup average total score (Score.O = 14.27) and the largest polyserial coefficient (Poly.O = .62). Most test takers answered the item correctly (Freq.O = .79). On the other hand, Options Q, R, S, and T are the least attractive options, with less than 1% selection rates. These four options are flagged as nonfunctioning options for this MC.1 item.

Figure 1 shows the option characteristic curves (OCCs) of this MC.1 item. The OCC is the estimated conditional probability of selecting the option for a given total score. The first panel shows the OCC of the key for this item, where the *y*-axis stands for the conditional probability and the *x*-axis stands for the total test score. The OCC curve is estimated by the smoothing spline method (Hastie & Tibshirani, 1990), and the band around the OCC is the estimation error band. The plot again shows that the key has a monotonically increasing OCC. On the other hand, Options Q, R, S, and T have few data with flattened OCCs. R codes for producing such OCCs are presented in Appendix A.

Besides the OCCs of each option/response selection, the option response time plots can be constructed similarly to the OCC plots; inspection of these response time plots (refer to Figure 1B in the appendix) can help providing information

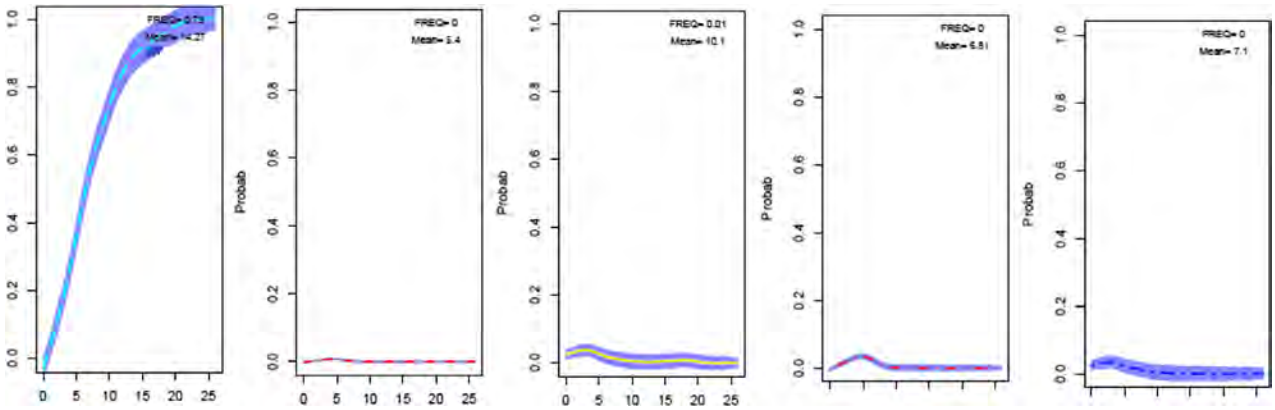

Figure 1 Option characteristic curves of an MC.1 item for the key (first panel) and the least attractive options Q, R, S, and T (penultimate panel). MC.1 = traditional multiple-choice items that ask test takers to select one option.

Table 3 An MC.m Item and Its 20 Most Popular Response Patterns (From P.1 to P.20)

	P.1	P.2	P.3	P.4	P.5	Omit	P.7	P.8	P.9	P.10
Freq.O	.50	.12	.12	.08	.02	.02	.01	.01	.01	.01
Score.O	14.94	11.77	11.24	15	12.5	3.04	11.23	9.67	13.45	13.32
Poly.O	.47	−.15	−.22	.21	−.04	−.86	−.15	−.27	.03	.02
	P.11	*P*.12	*P*.13	*P*.14	*P*.15	*P*.16	*P*.17	*P*.18	*P*.19	*P*.20
Freq.O	.01	.00	.00	.00	.00	.00	.00	.00	.00	.00
Score.O	8.08	9.13	8.74	10.72	6.61	6.91	5.27	12.11	6.94	9.65
Poly.O	−.37	−.28	−.32	−.16	−.49	−.45	−.58	−.06	−.43	−.23

Note. Freq.O = observed relative frequency of an option; Poly.O = the largest polyserial coefficient; Score.O = the average total test score of the subgroup that chose the particular option; MC.m = multiple-choice items with multiple selections.

about how much time different score subgroups took to choose each option/response selection (refer to Figure B1). Overall, higher score subgroups took less time to make their selections.

As an example of MC.m items, Table 3 and Figure 2 show the summaries and OCCs of such an MC.m item. This item also has more than eight options, and it asks test takers to select all options that apply. In the data, we observed 234 different response selections/patterns. Note that different response selections/combinations of the same options are treated as the same. For example, "1,2,3," "2,3,1," and "3,2,1" are recoded as the same pattern. Table 3 gives the 20 most popular response patterns (from P.1 to P.20) in descending order of relative frequency. Response pattern P.1 is the intended key and also functioned statistically as the key because about half of the test takers chose this pattern (Freq.O = .50) and because it has the second highest subgroup score (Score.O = 14.94) and the largest polyserial coefficient (Poly.O = .47) among the 20 most popular response patterns. Note that the sixth most popular behavior of the sample is making no selection, which is labeled as "Omit." The response pattern P.4 has the highest subgroup score (Score.O = 15.00), with $n = 584$ test takers (Freq.O = 8%); in addition, the polyserial coefficient is .21 for this response selection. Therefore this response selection seems to be a candidate for partial credit for this complex item (Frary, 1989; Lord, 1980; Rodriguez, 2005).

We also observed that Options 3, 6, 10, and 12 rarely appeared in the 234 different response selections, and they are flagged as nonfunctioning options for this MC.m item. Figure 2 implies the same message as that in Table 3.

Inspection of the option/response selection plots of response time (refer to Figure B2) shows that the four most popular response patterns have similar response time patterns: higher score groups took less time to select the specific response pattern.

Decision Rules

After examining options and response selections for all items on the test, the flagged options were presented to content experts for discussion. On the basis of their feedback, a maximum number of options was decided so that the simulated revised test would not be too different from the original one and removal of the nonfunctioning options would not have a significant detrimental impact on item integrity. For the partial credit assignment to a response selection, we decided to apply the following rules, all of which must be satisfied before simulating partial credit:

1. This response selection contains the key (or a component of the key) and another option that is irrelevant or does not conflict with the key.
2. The total score mean of the subgroup who had this pattern is close to that of the subgroup who selected the key.
3. This response selection has a polyserial coefficient larger than .10.

Note that the second and third decision rules will result in reliability increases, as only response selections that are positively related to the total score criterion are given partial credit.

Simulation of Responses

Once options are identified as nonfunctioning, they are removed from the items. Responses of test takers whose chosen options were removed have to be reassigned to create new data sets. In a previous simulation study, Guo et al. (2018) used two schemes in the assignment to represent two extreme cases. One was the random assignment, which assumed that

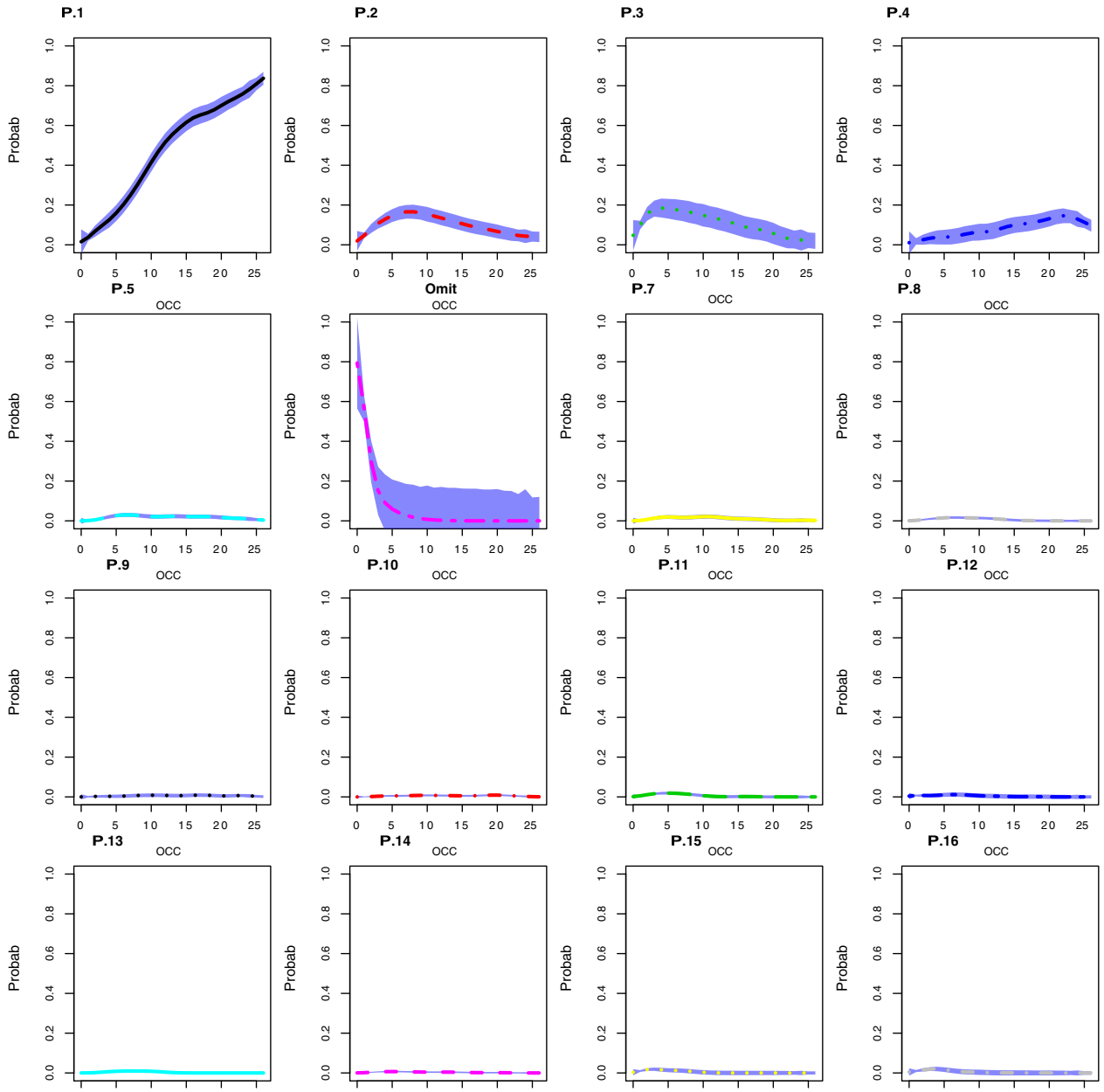

Figure 2 The 16 most frequent response selections for an MC.m item. MC.m = multiple-choice items with multiple selections.

test takers would randomly guess among the remaining options if the options they would have chosen were unavailable. Hence the newly created data may contain more noise than the original data. The other scheme was the conditional probability assignment that used the probabilities of choosing the remaining options conditional on the test taker's ability. The conditional probability assignment is an ideal situation for response assignment because it assumes that test takers whose chosen options were removed would behave in the same way as those who had similar ability and had chosen the remaining options. The two simulation schemes of response assignments may provide reasonable coverage of test takers' possible response behaviors when their preferred options are not available (Guo et al., 2018; Haladyna & Downing, 1989a; Rodriguez, 2005).

However, again because of the large number of response selections for the applicable items, it was not feasible to use the conditional probability assignment in our simulations. In addition, the flagged options or response selections containing the flagged options seemed to be irrelevant to the related questions. Therefore, in the following analysis, we use the random assignment scheme to produce new data sets.

Evaluation

To evaluate the combined impact of reduced numbers of options and partial credit, item summaries (item difficulty and item polyserial correlation) and test reliability of the new data sets are compared to the original data. It is expected that item difficulty will decrease as the number of options decreases, but the test reliability may increase with partial credit assignments.

Results

Each option or response selection for the 26 items on the test was examined in terms of summary statistics and OCC plots, to flag options that were not functioning and response selections that were potential candidates for partial credit. The flagged options and response selections were presented to content experts. On the basis of their feedback, we focused on items with more than eight options and reduced the number of options to eight so that the simulated revised test would not be too different from the original one. All the flagged nonfunctioning options were investigated by content experts, and their removals were confirmed not to have significant detrimental impact on item integrity.

Discussions among content experts and psychometricians resulted in removing 13 options in four items and assigning partial credit to five response selections in three items. Because nonfunctioning options were the least popular options, the percentages of test takers whose chosen options were removed were small, ranging from .80% to 3.47%. Item response time differences between test takers whose chosen options/response selections were removed and those whose were kept were not large, even though the two sample *t*-tests were statistically significant (refer to Table B1).

We tried a partial credit of 1/3 point and 1/2 point, respectively, as suggested by content experts, who were more comfortable with 1/3 point than with 1/2 point. Note that the reliability coefficient estimate was .783 for the operational data. Simply rescoring the test with partial credit brought the test reliability up to .793 for a partial credit of 1/3 point and to .796 for a partial credit of 1/2 point. Assigning partial credit alone also slightly reduced the standard error of measurement (SEM).

To evaluate impact on the test by simultaneously removing nonfunctioning options and assigning partial credit on the studied test, we simulated the replacement of removed nonfunctioning responses by remaining options where applicable. In the simulation, for MC.1 items with eight remaining choices after the removal of nonfunctioning options, we used the chance rate of 1/8 to randomly assign a correct response to the test taker if his or her option was removed. For MC.m items, we used a chance rate of $1/K$ to randomly assign full credit (correct response) to the test taker and a chance rate of k/K to randomly assign partial credit, where K is the number of the most popular response selections (in our data, $K = 20$) for those MC.m items and k is the number of response selections that will receive partial credit.

Table 4 shows the test and item statistics of the simulated data in 100 replications. As expected, with partial credit and removal of nonfunctioning options, the test internal consistency reliability coefficient estimate increased (from .78 to .79 or .80), and the increment was larger when the partial credit value was 1/2. The SEM of the simulated data was nearly the same. The test total score increased as well (from 13.04 to 13.4 or 13.33), and the score standard deviation slightly increased (from 5.07 to 5.14 or 5.18). Note that in operational settings, the revised test score should be equated to the original test score to ensure comparability (Kolen & Brennan, 2004).

Table 4 also shows that removing nonfunctioning options alone (i.e., partial credit = 0) had minimal impact on items and the test; that is, item difficulty, polyserials, and test reliability are about the same as for the original test. However, with partial credit assignment, item difficulty decreased and item discrimination power increased for those affected items, except for one item: Item 19. Note that even though the polyserial coefficient of this item became slightly smaller with partial credit assignment, the test reliability increased. Without partial credit on this item, on the other hand, the polyserial coefficient of Item 19 maintained the same, but the test reliability was smaller. For items with removed options, their summary statistics are about the same, and variation of summary statistics in the 100 replications is negligible, both because of the small percentages of removed options/response selections.

Overall, as evident in previous studies, when the number of options decreases, the item might become easier because of a larger random guessing chance, and the test scores may increase (Rodriguez et al., 2014). For the same reason, item discrimination and test reliability may decrease. However, the decrement in the item discrimination power and the test reliability is counterbalanced by the partial credit assignment. In fact, the simulated tests had slightly higher reliability because of the partial credit assignment and the small portions of removed options.

Table 4 Summary Statistics Comparison Between the Original Data and the Simulated Data

	Original		Partial credit = 0				Partial credit = 1/3				Partial credit = 1/2			
Test	Rel	SEM	Rel		SEM		Rel		SEM		Rel		SEM	
			Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*
	.78	2.26	.78	.0001	2.26	.0001	.79	.0001	2.25	.0001	.80	.0001	2.25	.0001
Test	Score	*SD*	Score mean		Score *SD*		Score mean		Score *SD*		Score mean		Score *SD*	
			Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*
	13.04	5.07	13.04	.0008	5.06	.0009	13.24	.0008	5.14	.0009	13.33	.0008	5.18	.0009
			Score		Poly		Score		Poly		Score		Poly	
Item	Score	Poly	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*
1*	.81	.62	.81	.0005	.62	.0016	.81	.0005	.62	.0016	.81	.0005	.62	.0016
2*	.90	.57	.90	.0004	.57	.0015	.90	.0003	.57	.0017	.90	.0003	.57	.0014
3	.70	.54	.70	.0000	.54	.0003	.70	.0000	.54	.0002	.70	.0000	.54	.0002
4*#	.50	.47	.50	.0005	.46	.0014	.53	.0005	.49	.0014	.55	.0005	.49	.0015
5	.56	.33	.56	.0000	.33	.0002	.56	.0000	.33	.0002	.56	.0000	.33	.0003
6	.32	.31	.32	.0000	.31	.0002	.32	.0000	.31	.0002	.32	.0000	.31	.0002
7	.57	.54	.57	.0000	.54	.0002	.57	.0000	.54	.0002	.57	.0000	.54	.0002
8	.53	.58	.53	.0000	.58	.0002	.53	.0000	.58	.0002	.53	.0000	.58	.0002
9	.54	.51	.54	.0000	.51	.0002	.54	.0000	.51	.0002	.54	.0000	.51	.0002
10	.63	.54	.63	.0000	.54	.0002	.63	.0000	.54	.0003	.63	.0000	.54	.0002
11	.57	.51	.57	.0000	.51	.0002	.57	.0000	.51	.0002	.57	.0000	.51	.0002
12	.47	.48	.47	.0000	.48	.0002	.47	.0000	.48	.0002	.47	.0000	.48	.0002
13	.56	.56	.56	.0000	.56	.0002	.56	.0000	.56	.0002	.56	.0000	.55	.0002
14*	.40	.65	.40	.0003	.64	.0009	.40	.0004	.64	.0009	.40	.0002	.64	.0007
15	.66	.59	.66	.0000	.59	.0003	.66	.0000	.59	.0002	.66	.0000	.58	.0002
16	.51	.41	.51	.0000	.41	.0002	.51	.0000	.41	.0003	.51	.0000	.41	.0002
17	.61	.57	.61	.0000	.57	.0002	.61	.0000	.57	.0003	.61	.0000	.57	.0002
18	.50	.57	.50	.0000	.57	.0003	.50	.0000	.57	.0002	.50	.0000	.57	.0002
19#	.18	.68	.18	.0000	.68	.0002	.28	.0000	.65	.0002	.33	.0000	.66	.0002
20	.54	.49	.54	.0000	.49	.0002	.54	.0000	.49	.0002	.54	.0000	.49	.0002
21	.52	.52	.52	.0000	.52	.0002	.52	.0000	.52	.0002	.52	.0000	.53	.0002
22	.48	.52	.48	.0000	.52	.0002	.48	.0000	.52	.0002	.48	.0000	.52	.0002
23	.39	.64	.39	.0000	.64	.0002	.39	.0000	.64	.0002	.39	.0000	.64	.0002
24#	.18	.52	.18	.0000	.52	.0003	.25	.0000	.62	.0002	.28	.0000	.62	.0002
25	.40	.17	.40	.0000	.17	.0002	.40	.0000	.16	.0003	.40	.0000	.16	.0003
26	.28	.24	.28	.0000	.24	.0003	.28	.0000	.23	.0003	.28	.0000	.23	.0003

Note.* = items with removed options; # = items with partial credit; Poly = polyserial coefficient; REL = reliability; SEM = standard error of measurement.

Note that the Spearman – Brown formula can be used to predict the reliability of a lengthened or shortened test (Haertel, 2006, p. 77):

$$\rho = \frac{(1/x)\,\rho_n}{1 + (1/x - 1)\,\rho_n},$$

where ρ and ρ_n are the reliabilities of the original test and the modified test, respectively, and x is the factor by which the test is lengthened or shortened. For the current 26-item test, the original test reliability is $\rho = .7826$, and the test needs to be $x = (1/\rho - 1)/(1/\rho_n - 1) = 1.081$ times longer to reach the increased reliability of $\rho_n = .7955$. This indicates that by simultaneously assigning partial credits and removing nonfunctioning options, the impact on test reliability is equivalent to adding two more items to the test.

Discussion

With advances in educational assessment, researchers and test developers have been developing new MC item types to measure complex skills required in the digital age. For these new MC item types, it is very challenging for test developers to predetermine how many options are appropriate and to preassign partial credit relevant to test construct so that the

test is efficient, is reliable, and possesses the best possible psychometric properties. In this study, using a real data set as an illustration, we presented a step-by-step statistical approach to analyzing new MC item types to improve the test psychometric properties without sacrificing content integrity. Our approach is guided by program-specific data, psychometric principles, and content experts' knowledge. The analysis and simulation methods can help psychometricians to evaluate item and test performance, and the results can assist test developers in creating high-quality new item types and revising the test design in new form development.

As shown in our analyses, large numbers of options may be of limited value because many of them are nonfunctioning and are selected by few test takers. Removal of those options has limited or no impact on the psychometric properties of the item and the test, though it might impact an item's verisimilitude. The elimination or avoidance of nonfunctioning options may also help to make test items more accessible, particularly by reducing the required amount of reading and eliminating potential sources of confusion (Rodriguez, 2005; Schneid et al., 2014).

Furthermore, partial credit for these new MC item types helped to extract more accurate information from the complex response selections, and therefore it may increase test internal consistency without lengthening a conventional MC.1 test, as shown in our real data illustration. Partial credit may also encourage test takers' engagement with the test and help to provide feedback to enhance learning for some formative assessments (Frary, 1989).

Nevertheless, our study has several limitations. First, because the data were collected from an operational testing program, we decided to use eight as the maximum number of options on the test so that the research results can guide new form development and the new forms will have properties that are relatively similar to the original one. For a testing program that does not need to maintain scale, different numbers of options should be experimented with, in line with the proposed procedures in this study. Other limitations include the nonparametric approach in estimating OCCs and response time; that is, we could not use a parametric item response model for item analysis because of the restriction imposed by the large numbers of response selections for the MC.m items — and for the same reason, use of item response time was limited in our study. A modified NRM approach as in Haberman and Lee (2017) could be explored, in which distractor selection and proficiency are conditionally independent given that the test taker's selection was incorrect, even though this approach does not result in the test taker gaining more credit with fewer options. Further studies may also investigate parametric modeling of response selections and response time with appropriate data.

Both processes suggested here have implications for test development. First, MC items are generally written with the intention of including some very plausible distractors to increase discrimination power: More proficient examinees are expected to be more likely than less proficient ones to be able to distinguish an attractive, but incorrect, option/choice or combination of options/choices (patterns) from a correct one. A partial credit model changes the functionality of these distractors. Test developers expecting that such an option might receive partial credit would need to develop a different strategy in crafting distractors. Second, it is informative to simulate posttest changes to test takers' responses by removing options and reassigning responses among the remaining options, and the practical use of the simulation data lies in the development of some guidelines for appropriate numbers of options/choices in newly written or revised items. Items could be crafted so as to have less need for greater numbers of options. For example, instead of asking test takers to consider all of the sentences in a long passage, they could be asked to consider some plausibly selected subset of the sentences.

Overall, compared to previous studies that mostly used empirical data collected from real test administrations/field trials of revised tests, our simulation approach is obviously economic, efficient, and informative in guiding content experts in developing and redesigning new item types and new assessments. In addition, because our proposed item analysis approach relies on raw item responses and response times, it can be immediately applied to other new item types. For example, the technology-enhanced and interactive item types require test takers to use drag-and-drop or slider controls to form a response (National Center for Education Statistics, 2019); the sequences of actions can be treated as raw responses and analyzed accordingly. Results may inform test developers how well the items functioned and help educators to understand how well students reacted to such items. In a case where scoring such items is in question, our analysis approach is informative as well when assigning partial credit to certain response sequences.

Our study is intended to develop a generic statistical approach for analyzing new item types and providing useful information to test developers and testing programs for decision-making to improve the assessment. Implementing changes to items and tests in operational testing programs requires a joint effort between all aspects of the programs, not only between statistics and psychometrics.

Acknowledgments

We thank Matt Johnson for his review of this study and Shelby Haberman, Rick Morgan, and Sooyeon Kim for their comments on the initial draft.

References

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, *37*, 29 – 51. https://doi.org/10.1007/BF02291411

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

Drasgow, F. (1986). Polychoric and polyserial correlations. In N. Johnson & S. Kotz (Eds.), *Encyclopedia of statistical sciences* (Vol. *7*, pp. 68 – 74). John Wiley.

Frary, R. (1989). Partial-credit scoring methods for multiple-choice tests. *Applied Measurement in Education*, *2*(1), 79 – 96. https://doi.org/10.1207/s15324818ame0201_5

Goldhammer, F., Martens, T., & Lüdtke, O. (2017). Conditioning factors of test-taking engagement in PIAAC: An exploratory IRT modelling approach considering person and item characteristics. *Large-Scale Assessments in Education*, *5*(1), Article 18. https://doi.org/10.1186/s40536-017-0051-9

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., 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). Educational Testing Service. https://doi.org/10.1002/ets2.12209

Haberman, S., & Lee, Y.-S. (2017). *A statistical procedure for testing unusually frequent exactly matching responses and nearly matching responses* (Research Report No. RR-17-23). Educational Testing Service. https://doi.org/10.1002/ets2.12150

Haertel, E. H. (2006). Reliability. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 65 – 110). American Council on Education/Praeger.

Haladyna, T. M., & Downing, S. M. (1988, April 8 – 12). *Functional distractors: Implications for test-item writing and test design* [Paper presentation]. Annual meeting of the American Educational Research Association, New Orleans, LA, United States.

Haladyna, T. M., & Downing, S. M. (1989a). A taxonomy of multiple-choice item writing rules. *Applied Measurement in Education*, *2*(1), 37 – 50. https://doi.org/10.1207/s15324818ame0201_3

Haladyna, T. M., & Downing, S. M. (1989b). Validity of a taxonomy of multiple-choice item writing rules. *Applied Measurement in Education*, *2*(1), 51 – 78. https://doi.org/10.1207/s15324818ame0201_4

Halpern, D. F. (2010). *Halpern critical thinking assessment manual*. Schuhfried.

Hastie, T. J., & Tibshirani, R. J. (1990). *Generalized additive models*. Chapman and Hall.

Kolen, M., & Brennan, R. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). Springer. https://doi.org/10.1007/978-1-4757-4310-4

Ku, K. Y. L. (2009). Assessing students' critical thinking performance: Urging for measurements using multi-response format. *Thinking Skills and Creativity*, *4*, 70 – 76. https://doi.org/10.1016/j.tsc.2009.02.001

Liu, O. L., Frankel, L., & Roohr, K. C. (2014). *Assessing critical thinking in higher education: Current state and directions for next-generation assessment* (Research Report No. RR-14-10). Educational Testing Service. https://doi.org/10.1002/ets2.12009

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

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

R Core Team. (2019). *R: A language and environment for statistical computing* [Computer software]. R Foundation for Statistical Computing. http://www.Rproject.org/

Rios, J. A., Guo, H., Mao, L., & Liu, O. L. (2017). Evaluating the impact of careless responding on aggregated-scores: To filter unmotivated examinees or not? *International Journal of Testing*, *17*(1), 74 – 104. https://doi.org/10.1080/15305058.2016.1231193

Rodriguez, M. C. (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

Rodriguez, M. C., Kettler, R. J., & Elliott, S. N. (2014). Distractor functioning in modified items for test accessibility. *SAGE Open*, *4*(4), 215824401455358. https://doi.org/10.1177/2158244014553586

Schneid, S. D., Armour, C., Park, Y. S., Rudkowsky, R., & Bordage, G. (2014). Reducing the number of options on multiple-choice questions: Response time, psychometrics and standard setting. *Medical Education*, *48*(10), 1020 – 1027. https://doi.org/10.1111/medu.12525

Thissen, D. (1976). Information in wrong responses to the Raven progressive matrices. *Journal of Educational Measurement*, *13*(3), 201 – 214. https://doi.org/10.1111/j.1745-3984.1976.tb00011.x

Thissen, D., Steinberg, L., & Fitzpatrick, A. R. (1989). Multiple-choice models: The distractors are also part of the item. *Journal of Educational Measurement*, *26*(2), 161 – 176. https://doi.org/10.1111/j.1745-3984.1989.tb00326.x

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

Appendix A

R Codes for Generating Item Plots

The following codes help to generate the OCC plots for item *j*. This item has 12 options that are numerically coded as $k = 1, 2, \ldots, 12$, respectively, in the raw item responses.

```
library(gplots);

library(polycor)

#J:   the number of items on the test
#crit: sum score vector of N test takers
#data: item raw response matrx of NxJ before scoring items, where N is the num-
ber
#    of test takers, and J is the number of items
#j:  the studied item
#k:  numerical value of the kth option for Item j

##### OCC functions#####################################################
 OCCfun<-function(data, crit, j, k)
 {
  sx=crit
  ir=(data[,j]==k) #response k is assigned 1, ow it is zero
  freq=round(sum(ir, na.rm=TRUE)/length(data[,1]),2) # frequency of response k
  m=round(mean(sx[ir], na.rm=TRUE),2) #group mean score with response k
  rs="NA" #compute item polyserial coefficient
  if (sum(ir, na.rm=TRUE)>0)
     {rs=round(polyserial(sx,ir, ML=TRUE, std.err=TRUE)$rho,2)}
  ###spline smoothing & error band####
  DD<-cbind(sx, ir); DD=na.omit(DD); sx<-DD[,1]; ir<-DD[,2]
  fit <- smooth.spline(sx, ir, df = 10)
  res <- (fit$yin - fit$y)/(1-fit$lev); sigma <- sqrt(var(res));
  SE=sigma*sqrt(fit$lev)
  list(P.sp=fit$y, xx=fit$x, SE=SE, FF=freq, M=m, RR=rs)
 }
##### OCC plots#########################################################
 PlotOCCfun<-function(data, crit, j, k, J)
 {
  OCC<-OCCfun(data, crit, j, k) #OCC functions
  plot(OCC$xx, OCC$P.sp, type="n", ylim=c(0,1), xlim=c(0, J), ylab="Probab",
               xlab="OCC", main=paste("item", j, " Option=", k))
  bluetrans <- rgb(0, 0, 250, 120, maxColorValue=255)
```

```
  polygon(c(OCC$xx, rev(OCC$xx)), c(-2*OCC$SE+OCC$P.sp, rev(2*OCC$SE+OCC$P.sp)),
                          col = bluetrans, border = NA)
  lines(OCC$xx, OCC$P.sp, col=k, lty=k, lwd=2)
  ####summary stats
  ff=OCC$F
  mm=OCC$M
  rr=OCC$RR
   text(J/2,1, c(paste("FREQ=",ff)), cex=.8)
   text(J/2,.9, c(paste("Mean=",mm)), cex=.8)
   text(J/2,.8, c(paste("R-bis=",rr)), cex=.8)
   }
###assign values to crit, data=data.frame(data), j and J;######################
par(mfrow=c(3,4), mai=c(.1,.51,.51,.1))
 for (k in 1:12)
 {
  PlotOCCfun(data,crit, j, k, J)
 }
```

Appendix B

Response Time Information

Table B1 Proportion and response time of removed responses

Item	Percentage (%)	Removed log (RT) mean	Kept log (RT) mean	*t*-test *p*-value
1	1.58	4.58	4.76	.048
2	.81	4.19	3.91	.023
4	3.47	3.99	4.15	.004
14	.78	3.68	4.01	.006

Note. Logarithm of response time was used to approximate normality. RT = response time.

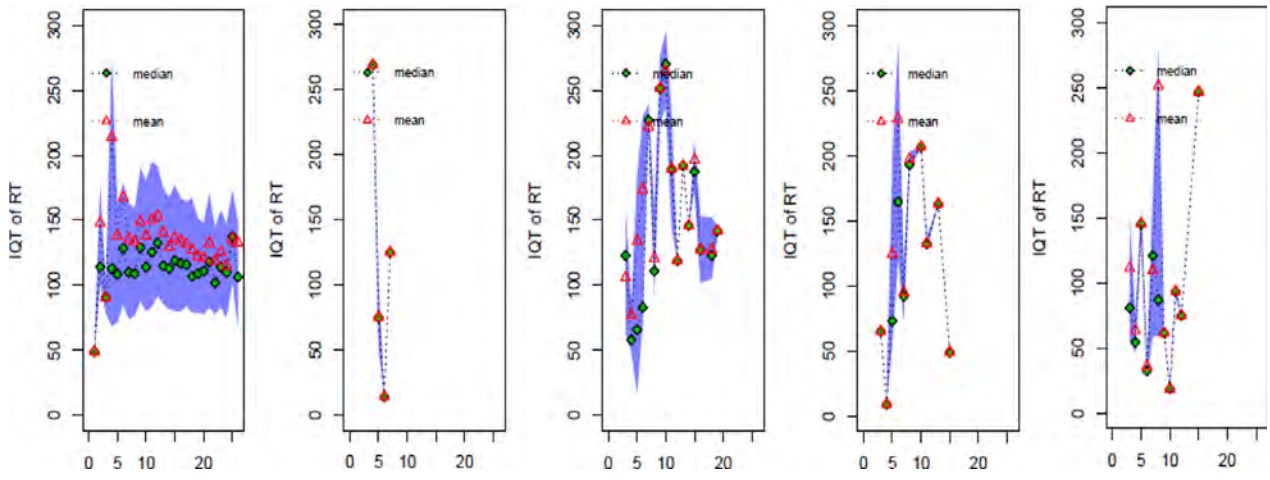

Figure B1 Conditional item response time for each option of the MC.1 item in Figure 1. In each panel, the *x*-axis stands for the total test score, and the *y*-axis stands for the response time. The triangle (diamond) is response time mean (median) at each score point, and the shaded area is the interquartile range of response time at that score point. The panels display the response time for the key, and the removed options are Q, R, S, and T, respectively. MC.m = multiple-choice items with multiple selections.

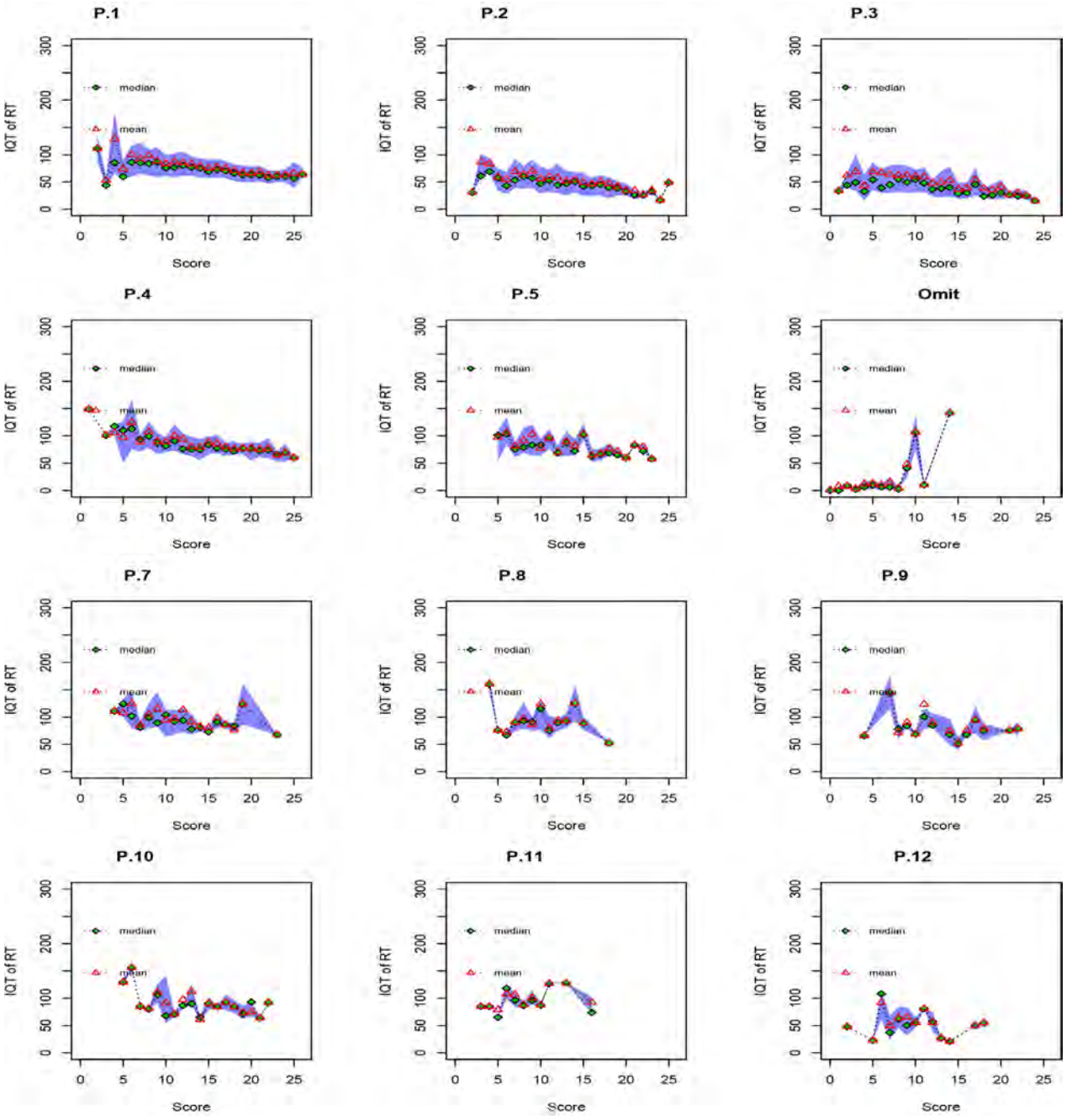

Figure B2 Conditional item response time for the 12 most popular response selections of the MC.m item in Figure 2. In each panel, the *x*-axis stands for the total test score, and the *y*-axis stands for the response time. The triangle (diamond) is response time mean (median) at each score point, and the shaded area is the interquartile range of response time at that score point. MC.m = multiple-choice items with multiple selections.

Suggested citation:

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

Action Editor: Shelby Haberman

Reviewers: Sooyeon Kim and Rick Morgan