Measuring the Power of Learning.™

Research Report
ETS RR–18-22

A Simulation-Based Method for Finding the Optimal Number of Options for Multiple-Choice Items on a Test

Hongwen Guo

Jiyun Zu

Patrick Kyllonen

December 2018

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

A Simulation-Based Method for Finding the Optimal Number of Options for Multiple-Choice Items on a Test

Hongwen Guo, Jiyun Zu, & Patrick Kyllonen

Educational Testing Service, Princeton, NJ

For a multiple-choice test under development or redesign, it is important to choose the optimal number of options per item so that the test possesses the desired psychometric properties. On the basis of available data for a multiple-choice assessment with 8 options, we evaluated the effects of changing the number of options on test properties (difficulty, reliability, and score comparability) using simulation. Using 2 criteria (low frequency and poor discrimination) to remove nonfunctioning options and 2 schemes (random and educated guessing) to model hypothetical response behavior for the removed options, we found that decreasing the number of options (from 8) created an easier test form but that a test form with reduced options could be more reliable if low-discriminating options were removed and an educated guessing strategy were assumed. We present a rationale for the optimal number of options for this test being approximately 5, which would result in a shorter test while preserving its psychometric quality. Simulation methods discussed in this report could be applied to any test to compare the effects of changing the number of options.

Keywords Test reliability; nonfunctioning options; item discrimination

doi:10.1002/ets2.12209

What is the optimal number of response options for items in a multiple-choice test? In practice, many multiple-choice tests have four or five alternative choices per item, particularly in standardized testing. Some tests have even more options per item. New assessments are frequently being developed and piloted, and existing tests are undergoing redesigns periodically, which may involve changing the number of response options. The new *SAT*® test (launched March 2016), for example, reduces the number of response options per item from five to four (Princeton Review, 2016). For a test being developed or being redesigned, how can we make decisions on the optimal number of options? On the basis of available data, can we predict how psychometric properties of the test, such as item difficulty, test reliability, and score comparability, will be affected if the number of response options changes?

There is an extensive literature on the topic of how the number of response options affects test quality. Grier (1975) showed that for maximizing reliability, having three options was better for tests with at least 18 items, and having two options was better for tests with fewer items. Lord (1980) proposed *proportionality* as a guide to determine the optimal number of items. Proportionality assumes that total testing time (T) is proportional to the product of the number of items (n) and the number of options (a) per item, that is, $T \propto na$. Using several approaches, Lord found that for most of the ability range, the optimal number of response options was three, assuming a fixed testing time (i.e., fewer response options per item enables faster per item responding and thereby allows for more items per test). However, for test takers with high ability, the optimal number of options was two, enabling longer tests, and for test takers with low ability, the optimal number of options was five, with shorter tests. Employing an item response theory (IRT) framework, he explained, "The effect of decreasing the number of choices per item while lengthening the test proportionately is to increase the efficiency of the test for high-level examinees and to decrease its efficiency for low-level examinees" (p. 110).

Budescu and Nevo (1985) examined the validity of Lord's (1980) proportionality assumption in three tests administered with two, three, four, and five options per item and found that their empirical results did not support it. Furthermore, the data indicated that the method by which options were deleted could affect test information. In his meta-analysis, Rodriguez (2005) found that reducing the number of options (but keeping the number of items constant) generally reduced item difficulty, item discrimination, and test reliability, unless the options deleted were ineffective ones. On the basis of empirical studies in the literature, Rodriguez argued that examinees were unlikely to engage in blind guessing

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

but rather were more likely to use educated guessing, essentially reducing the four- or five-option item to a three- or two-option item. He concluded that three options were optimal for multiple-choice items in most settings. He found that changing from five or four options to three had little to no effect on multiple-choice item difficulty and discrimination and test score reliability, on average.

Studies have also explored the effects of changing the number of response options on item quality empirically (Baghaeri & Amrahi, 2011; Rodriguez, Kettler, & Elliott, 2014; Schneid, Armour, Park, Rudkowsky, & Bordage, 2014). Using a common-item equating design approach, Baghaeri and Amrahi (2011) found that, with one exception, no significant change was observed in item difficulties, item fit statistics, and reliabilities across three test forms with different numbers of options. The exception concerned the discrimination power of distractors, which was found to be inversely affected by the number of options per item (i.e., a higher number of options lowered the discrimination power). Schneid et al. (2014) administered two versions (three vs. four or five options per item) of a computerized exam. They found that using three-option multiple-choice questions might strengthen validity by allowing additional questions to be tested in a fixed amount of testing time with no deleterious effect on the reliability of the test scores.

Most measurement textbook authors have suggested writing as many options as feasible (Haladyna & Downing, 1989a; Rodriguez, 2005) or developing as many functional distractors as feasible (Haladyna & Downing, 1989b). Haladyna and Downing (1988) defined the functional distractors to be options that have either (a) a significant negative point-biserial correlation with the total test score, (b) a negatively sloping item characteristic curve, or (c) a frequency of response greater than 5% for the total group. They also found that the number of functional distractors per item was unrelated to item difficulty and positively related to item discrimination.

Multiple-choice questions often contain several options that examinees rarely or never select (Schneid et al., 2014). Such nonfunctional options might not be plausible to even minimally competent examinees and therefore result in increased time spent on each item without making any contribution to item discrimination. For newly generated multiple-choice questions, test writers intentionally avoid three-option questions by discarding any questions that have only three plausible options or by adding fillers such as "all of the above" and "none of the above." These item-writing practices are undesirable because they introduce construct-irrelevant variance into the assessment (some examinees may be relatively more or less prone to select "all of the above" or "none of the above" independently of item content). Rodriguez et al. (2014) used a large multistate data set to measure the impact on distractor functioning. Distractor functioning was neither systematically improved nor weakened by reducing the number of options. The elimination of nonfunctioning options promoted several goals in making test items more accessible to all students, particularly by reducing the per-item testing time, reducing the required amount of reading, and eliminating potential sources of confusion (Rodriguez, 2005).

As is evident in most of these studies, reducing the number of item options from the standard four or five or more can affect item difficulty, item discrimination, and test reliability to various degrees. In some cases, it may be possible to conduct a field trial in which the number of different options is varied across test forms, and analysis can quantify the effects of different numbers of options on test quality. However, in other cases, such a field trial may be impractical, and test developers may instead be interested in exploring the effects of number of options on test quality based on existing data from a given form or set of test forms. The question would be, Can the existing data shed light on the effects of changing the number of options for the test, to preserve (while reducing testing time) or enhance (while keeping testing time constant) its psychometric properties?

Studies suggest that eliminating options by dropping those shown by item analysis to be least discriminating or not selected is a desirable practical procedure that should yield better results than simply eliminating distractors at random (Lord, 1980; Rodriguez et al., 2014; Williams & Ebel, 1957). The purpose of our study was to propose a simulation-based method to evaluate the possible impact of reducing the number of response options and to provide guidelines for further investigating the optimal number of response options per item for a multiple-choice test. To evaluate the hypothesis that changing the number of response options can affect item quality, we analyze existing data sets from an inductive reasoning test and simulate the systematic removal of response options based on two different strategies: (a) option response frequency (the *frequency* criterion) and (b) option discrimination power (the *discrimination* criterion). After eliminating options, we simulate examinee behavior by reassigning examinee responses, either (a) randomly or (b) probabilistically, conditioned on ability (or total test scores).

To evaluate the impact of reduced number of response options per item on the test, we compute changes in psychometric properties, specifically, item difficulty, item discrimination power, test reliability, and test scores obtained from

different numbers of response options. We hypothesize that excluding nonfunctioning options using either a frequency or discrimination criterion should generally result in better psychometric properties, particularly when we assume that examinees who selected a deleted option would have selected a second-choice option based on their ability rather than randomly.

Method

Our general method was to (a) take an existing data set that included response information for a 33-item test with eight response options per item, (b) remove nonfunctioning options and reassigning responses using other options, then (c) conduct various analyses to examine the psychometric properties of versions of the test that varied on the number of response options per item.

Data Set

Responses were from a sets-and-matrices test battery designed to measure general reasoning ability. Carroll (1993) identified letter sets and matrices to be among the best and most characteristic tests of inductive reasoning. Data were collected via the Qualtrics Research Suite survey platform. Data were collected to educational attainment specifications: equal numbers of (a) third- and fourth-year college students, (b) college graduates, (c) master's program students, (d) master's degree holders, and (e) PhD holders. There were no specified targets for men, women, or members of different racial/ethnic subgroups, but the recruiting strategy was designed to achieve gender and racial/ethnic and age diversity. All participants were U.S. citizens.

The data set we used for illustration purposes contained item responses from 485 test takers (42% male, 57% female). The test had 33 items, and each item had eight response options. An example item is shown in Figure 1.

Identifying and Removing Nonfunctioning Options

Following Rodriguez (2005), we used two criteria to identify nonfunctioning options. The *frequency criterion* treats options with the lowest selection frequencies as the nonfunctioning options for an item. This criterion can be used for both large- and small-sample data sets.

The *discrimination criterion* treats options with the lowest correlation with overall trait level (i.e., the factor measured by the test) as the nonfunctioning options for an item. There are two different approaches for estimating correlation with trait level. For large-sample size data sets, IRT models like the nominal response model (NRM; Bock, 1972) can

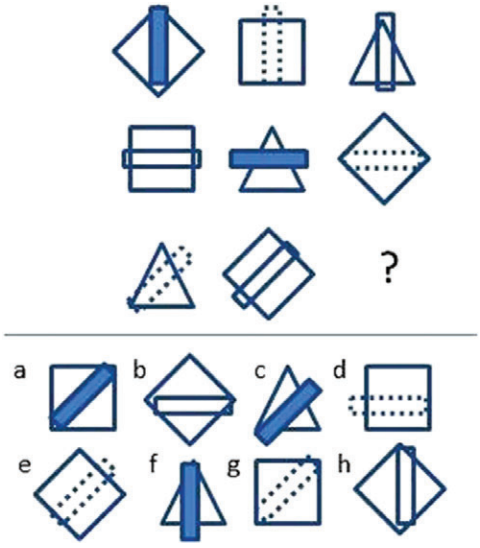

Figure 1 Example of a typical matrix problem: Choose the figure, a – h, that can replace the question mark (?).

be used to calibrate item discrimination parameters. For small-sample data sets, it is not appropriate to use IRT models; instead, the item polyserial correlation (Drasgow, 1986) for each option can be used as the option discrimination power index.[1] A discrimination index that is well above (or below) zero indicates high discrimination power for differentiating examinees with high and low ability. We treated the options with the lowest discrimination indices in absolute value[2] as the nonfunctioning options for an item; that is, when the discrimination index of an option was close to zero, we removed this option. This is a reasonable procedure, because an option with a low-discrimination index in absolute value does not discriminate examinees with high and low ability (Lord, 1980; Rodriguez, 2005). Once the distractors were identified to be nonfunctioning, they were removed from the item one by one.

Response Reassignment

Responses of test takers whose chosen options are removed using one of the strategies described (low frequency, poor discrimination) were reassigned to a different option. This simulates a data set in which the removed option is not present. Two schemes were used for reassignment.

In *random reassignment*, the remaining options were equally likely to be assigned as the new response. Random reassignment assumes that if an option a test taker would have chosen is not available, the test taker would have randomly guessed among available options. The justification here is that we have no information as to the respondent's second choice, and given that what the respondent believed to be the correct answer (based on his or her response) is not the correct answer, it may be at least plausible to think that the respondent was essentially guessing as to his or her first choice. A random reassignment scheme results in the newly created data set containing more noise than the original one, due to the pure random guessing assumption. As the number of options decreases, we lose more information in the data, and consequently, the score is expected to become less reliable.

Educated guessing reassignment is a conditional probability assignment scheme. For small-sample data sets, the probability of a particular option being reassigned to a test taker is proportional to its conditional probability given the test taker's total score. Because of possible sparse data at some score points, a kernel smoothing method (Ramsay, 1991) is used to produce the conditional probability of an option. More specifically, let $Pjm(x)$ be the conditional probability of choosing option m of Item j, given the total score x, and let $K(x)$ be a kernel function, a nonnegative and continuous function integrated to 1, such as the normal density function and the density function of a uniform distribution. The kernel estimation of $Pjm(x)$ is

$$\widehat{P}jm\,(x) = \sum_{i=1}^{N} w_i\,(x)\; Y_{jm}^{(i)},$$

where

$$wi\,(x) = \frac{K\left(\frac{x-Xi}{h}\right)}{\sum_{j=1}^{N} K\left(\frac{x-Xj}{h}\right)},$$

and where $Y_{jm}^{(i)} = 1$ if the ith examinee's response to Item j is option m and $Y_{jm}^{(i)} = 0$ otherwise, X_i is the ith examinee's total test score, and h is the bandwidth. In this report, we used the normal density function as the kernel function for smoothing. A curve that describes the conditional probability of one option against the total score is called the option characteristic curve (OCC). Educated guessing reassignment is an ideal situation of response assignment because we assume that test takers whose chosen options are removed behave in the same way as did others of the same ability level who have chosen the remaining options. For example, if Option 1 is removed from Item j, which has eight options, then the response of the test taker with a total score of x who chose Option 1 will be assigned an option of 2, 3, … , or 8, randomly sampled from a multinomial distribution with the probabilities being

$$\left(\widehat{P}j2\,(x)\,/T, \widehat{P}j3\,(x)\,/T,\ \ldots\ , \widehat{P}j8\,(x)\,/T\right),$$

where

$$T = \sum_{m=2}^{8} \widehat{P}_{jm}\,(x)\,.$$

For data sets with sample sizes large enough to run NRM, the conditional probability is the item response function of an option given a test taker's ability.

Under educated guessing reassignment, the newly created data may contain as much information as the original data. As the number of options decreases, for a fixed-length test, we would expect that the score reliability under educated guessing reassignment would remain more stable than that under random reassignment.

Although the two reassignment schemes, random and educated guessing, are quite different, and each represents an ideal case, they both can be justified as at least somewhat realistic. For example, random guessing has been cited as perhaps the greatest risk preventing more common use of fewer options (Rodriguez, 2005) and is essentially an assumption of classical test theory. On the other hand, experimental evidence has suggested that test takers make educated guesses when they can and that the effects of guessing are negligible across subject areas and age groups (Haladyna & Downing, 1989b; Rodriguez, 2005). Therefore our two proposed response reassignment simulation schemes may provide reasonable coverage of possible realistic empirical scenarios and useful end points for a sensitivity analysis.

Analysis

To evaluate test changes when the number of options was reduced, summaries of item difficulty and item polyserial correlation of the new, simulated data sets were compared to the original data set. Paired *t*-tests were used to compare means. We expected that item difficulty would decrease as the number of options decreased.

What matters most are test reliability and score quality. We compared test reliability using Cronbach's alpha as well as the standard error of measurement for the different data sets. Because of the change of the number of options, variability of test scores change too. Therefore, we also calculate the corrected reliability for restriction in range (CRT α; Haertel, 2006, p. 84). In addition, because of the change in test difficulty, the total scores on the tests with different numbers of options were not comparable. Note that most testing programs usually maintain a scale for score uses, and equating results are informative for observing the magnitude of score changes and whether a new scale would be needed if the change were too large. Hence, we used equating procedures, particularly, the single-group equipercentile equating method (Kolen & Brennan, 2004), to show how scores on simulated forms compared to each other. The changes in individual equated scores were also compared with respect to the conditional standard error of measurement (CSEM).

There are many methods to compute CSEM (e.g., see Woodruff, 1990). For simplicity, we used the binomial error model proposed by Lord (1984). If each examinee is given a different random sample of J items, then an estimate of the squared CSEM for a randomly selected examinee with the observed score X is (Lord, 1984)

$$\mathrm{CSEM}^2(X) = \frac{X(J-X)}{(J-1)} \times (1-K),$$

where

$$K = \frac{J(J-1)s_p^2}{\overline{x}\left(J-\overline{x}\right) - s_x^2 - Js_p^2},$$

where $\overline{x}$ and s_x^2 are the sample mean and variance of the number right scores, and s_p^2 is the variance of item difficulties in terms of P^+, the proportion of correct answers. This method was justified primarily by empirical validation. It has the advantage of requiring neither large samples nor complicated calculations. The comparison of equated scores with each other and with respect to CSEM may provide a guideline for choosing the optimal number of options for a particular test.

Simulating Data

We first used the frequency criterion to remove options with the smallest frequencies, to simulate a new data set. For example, for Item 1 (refer to the upper portion of Table 1), the key was E, and the percentages of examinees choosing options A, B, C, D, F, G, and H were .06, .07, .03, .01, .05, .02, and .10, respectively. To construct a new data set, N7, with seven options for Item 1, we removed the least popular option D; we reassigned new responses to test takers who chose D, and the new responses were chosen from one of the remaining options A, B, C, F, G, H, and E with equal weight (*random guessing* reassignment) or with unequal weight (*educated guessing* reassignment). More specifically, the R function *Sample*( ) is used to generate a response without replacement (R Core Team, 2014). We simulated item responses for other items in the same way to create the N7 data set.

Table 1 Summary Statistics for Item 1, Whose Key Was E, Using Data Sets N8, N5, and N2

Option	A	B	C	D	E	F	G	H
Data set N8								
Freq.	.06	.07	.03	.01	.66	.05	.02	.10
Mean	8.04	9.44	8.53	5.67	12.13	8.71	6.44	9.16
Polyserial	−.34	−.19	−.24	−.53	.44	−.24	−.50	−.22
Data set N5								
Freq.	.07	.08	NA	NA	.67	.06	NA	.11
Mean	11.31	12.34	NA	NA	14.89	11.3	NA	11.79
Polyserial	−.32	−.20	NA	NA	.43	−.30	NA	−.22
Data set N2								
Freq.	NA	NA	NA	NA	.76	NA	NA	.24
Mean	NA	NA	NA	NA	20.06	NA	NA	17.96
Polyserial	NA	NA	NA	NA	.35	NA	NA	−.36

Note. NA in frequency, mean, or polyserial correlation indicates that the corresponding option was removed from the data.

To construct data set N6, with six options per item, for Item 1, we removed the two least frequently chosen options, D and C. We then reassigned new responses to test takers who chose C or D, and the new responses were chosen from one of the remaining options A, B, E, F, or G, with equal weight (random guessing reassignment) or with unequal weight (educated guessing reassignment). Other item responses were simulated similarly.

Datasets N5, N4, and N3 were simulated in the same fashion. In data set N2, Item 1 had only two options (Options E and H; note that the key, Option E, is always reserved in all the new data sets); we assigned new responses to test takers who chose options other than E and H, and the new responses were either E or H with equal weight (random guessing) or with unequal weight (educated guessing). Thus we created two sets (one set under random guessing, the other under educated guessing) of six new data sets N7, N6, N5, N4, N3, and N2, based on the original data (denoted as N8).

For the discrimination criterion, because of limited sample sizes, we used the polyserial correlation approach. Thus the two sets (again, one set under random guessing, the other under educated guessing) of simulated data sets (N7d, N6d, N5d, N4d, N3d, and N2d) removed options with the smallest polyserial correlations in absolute value. The response assignment was generated in the same way as described in the preceding section.

In total, from the original data set, we simulated 2 (low frequency vs. low discrimination) × 2 (random vs. educated guessing) × 6 (number of response options 2, 3, 4, 5, 6, and 7) = 24 data sets. All the simulations and analyses were carried out using R (R Core Team, 2014).

Results

We present results separately by how response options were removed (low frequency vs. poor discrimination). Results within a response-option-removal strategy are presented for both reassignment strategies, random and educated guessing. Results are shown both graphically and in tables.

Using the Low-Frequency Criterion to Simulate the Removal of Options

The items on the studied test were expected to provide forms that were difficult (to discriminate examinees with high ability) and reliable. The test developers were aware of the value of attractive distractors in creating difficult items. Therefore we were not expecting to find many nonfunctioning distractors. Nevertheless, the empirical item plots showed that many of the options of the items were nonfunctioning; that is, test takers did not select them or they did not discriminate between test takers of high and low ability. This meant that removing nonfunctioning distractors would be a viable procedure for evaluating the effects of the number of options in this study given this data set.

Assuming Random Guessing to Generate a New Selected Option

Under random guessing reassignment, the newly created data should contain more noise than the original data, and therefore we would expect that tests become less reliable (such as lower average item-total correlation and greater standard

error of measurement) as the number of options decreases. Using the low-frequency criterion to remove options and assuming random guessing, Figure 2 shows the OCC curves of Item 1 (an easy item with $P^+ = .66$) whose key happened to be the most popular option, E. The top, middle, and lower panels of Figure 2 show OCC curves for the same item with N8 (the original data), N5 (with the three least selected options removed), and N2 (with the six least selected options removed), respectively. Table 1 shows option selection statistics for Item 1 using N8, N5, and N2. For N8 in Table 1, the "Freq." row shows that 66% of the test takers chose the key, Option E; less than 10% chose H, A, or B; and less than 5% chose C, D, or G. The row labeled "Mean" of Table 1 corresponds to score mean for each option group. The group who chose the key had the highest score mean. The row labeled "Polyserial" corresponds to polyserial correlation coefficient (Drasgow, 1986) of each option. The key had the highest and positive discrimination power. The distractors all showed negative polyserial correlation coefficients, and the values were not close to zero. However, in view of the small frequencies for some options, the negative coefficients may not have been reliable.

Table 1 shows information for the same item when the three least selected (lowest frequency) options were removed in N5, the simulated data with five options. Table 1 shows that removing three options had limited impact on the key OCC or discrimination power. Table 1 also displays item statistics with two options based on N2. In this case, some impact of the number of options could be observed in item discrimination and OCCs.

Figure 3 presents the same information for Item 14, a difficult item (with $P^+ = .08$) whose key, B, was the second least popular option. For Item 14, item characteristic changes from eight options to five options is limited, but it is significant from eight options to two options.[3] Table 2 shows that Options C and D had little discrimination power in the original data, nor did Options A, F, and G.

All item plots show that when the number of options decreased, the discrimination power of the key and that of the distractors, in terms of the polyserial coefficient (Drasgow, 1986), decreased. Particularly, the value of the distractor discrimination became more negative.

Assuming Educated Guessing to Generate a New Selected Option

Under the educated guessing assumption, test takers whose chosen options were removed (in this case, owing to low frequency) are assumed to behave in the same way as would be expected of others of their ability level in choosing responses. Hence we would expect that as the number of options decreases, the simulated responses should contain nearly as much information and be nearly as reliable as the original data. Table 3 shows that under the educated guessing assumption, item difficulty and item discrimination decreased as the number of options decreased. However, while form difficulty changed to a similar degree, discrimination power decreased faster under random guessing than under educated guessing.

Table 4 shows that for both random and educated guessing, as the number of options decreased, the mean number correct increased, measurement error increased, and reliability decreased. All changes in score means were statistically significant compared to the original data based on the paired t-tests. The change in score mean, score variance, and test reliability was larger in magnitude under random guessing than under educated guessing, as we expected.

However, because of the change in test difficulty due to changes in the number of options, number-correct scores on forms with different numbers of options are not strictly comparable. Hence we used equating (single-group equipercentile equating; Kolen & Brennan, 2004) to show how scores on simulated forms compared to each other. We used the R package equate (Albano, 2016) in conducting equating. Because of sparse data, the loglinear presmoothing method (Holland & Thayer, 2000) was used with three moments preserved. The changes in individual equated scores were also compared with respect to the CSEM. We observed that the test taker's score was boosted when the number of options decreased. Figure 4 shows the difference in number correct on the original form (N8) and an equated form (N2 – N7) as a function of number correct on the original form (N8). For example, a score of 15 on N8 corresponds to a score of 15.1 on N7, 15.2 on N6, 15.8 on N5, 16.6 on N4, 18.1 on N3, and 23.2 on N2, respectively. This was done for both random and educated guessing. Compared to educated guessing, the equated scores using random guessing were higher for the lower score range (particularly N2 and N3). Nevertheless, under both schemes, the differences of equated scores were within a CSEM for N7, N6, and N5.

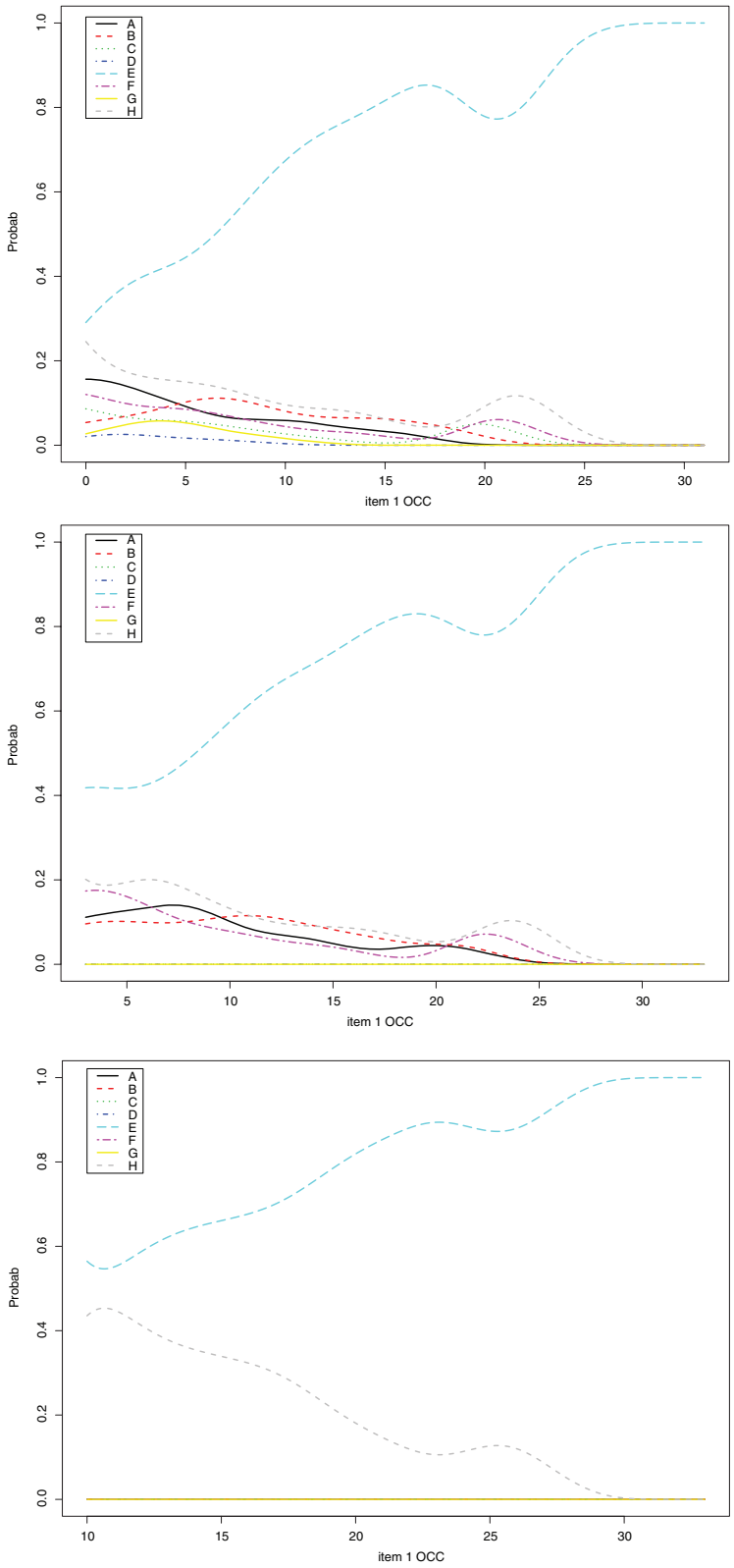

Figure 2 Option characteristic curve plots of Item 1, for the original data with eight (top), five (middle), or two (bottom) options, after low-frequency options were removed, assuming random guessing among remaining options.

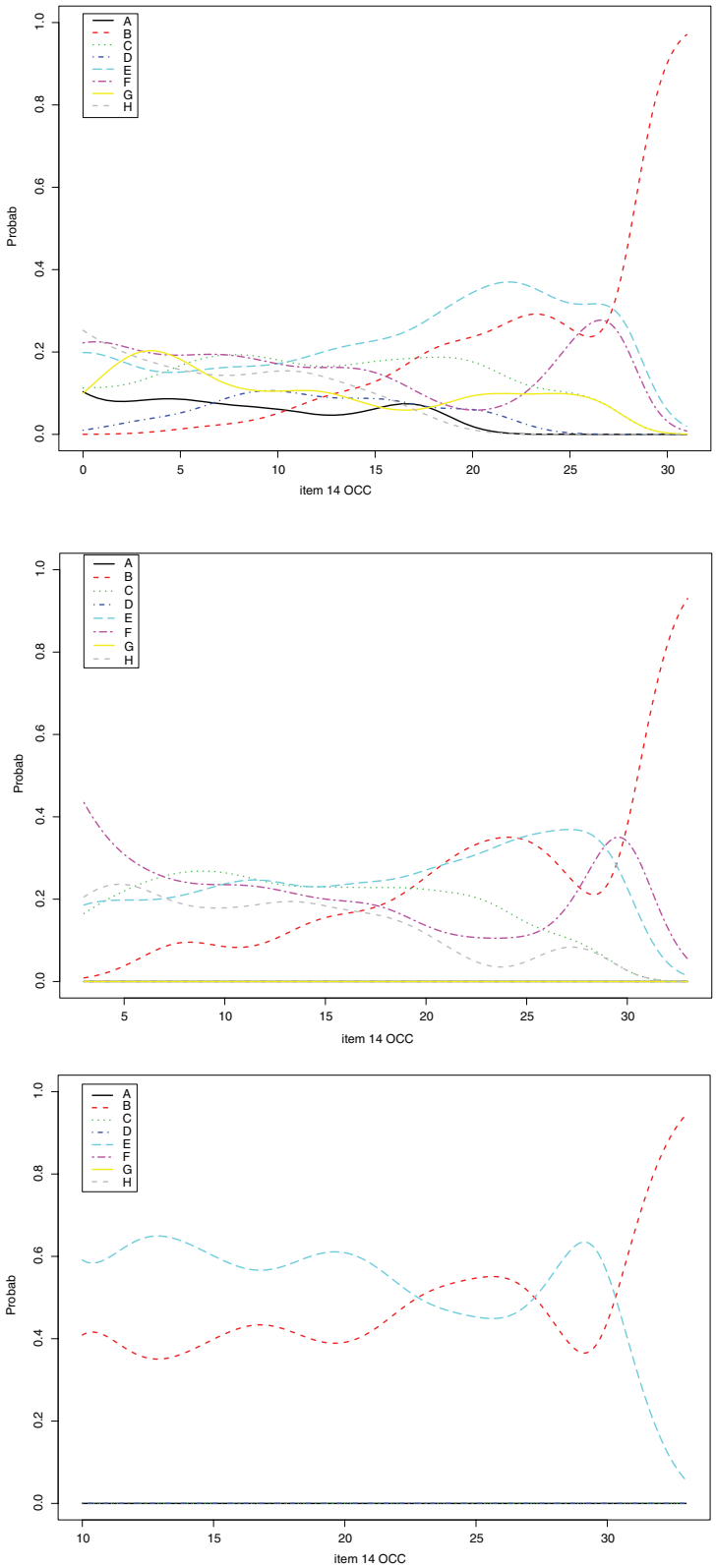

Figure 3 Option characteristic curves plots of Item 14 for the original data with eight options (top), five (middle), or two (bottom) options, after low-frequency options were removed, assuming random guessing among remaining options.

Table 2 Summary Statistics for Item 14, Whose Key Was B, Using Data Sets N8, N5, and N2

Option	A	B	C	D	E	F	G	H
Data set N8								
Freq.	.06	.08	.17	.08	.20	.16	.11	.13
Mean	9.50	15.95	10.90	11.00	12.12	10.04	9.63	9.05
Polyserial	−.16	.47	−.01	.00	.16	−.13	−.17	−.27
Data set N5								
Freq.	NA	.14	.22	NA	.24	.21	NA	.19
Mean	NA	16.74	13.49	NA	14.59	12.82	NA	12.35
Polyserial	NA	.36	−.06	NA	.12	−.16	NA	−.24
Data set N2								
Freq.	NA	.48	NA	NA	.52	NA	NA	NA
Mean	NA	20.12	NA	NA	19.02	NA	NA	NA
Polyserial	NA	.19	NA	NA	−.19	NA	NA	NA

Note. NA in frequency, mean, or polyserial correlation indicates that the corresponding option was removed from the data.

Table 3 Means and Standard Deviations of Probability Correct and Item-Total Polyserial Correlations of Original and Simulated Test Forms Using a Frequency Criterion to Remove Nonfunctioning Options

	Random guessing		Educated guessing	
Form	M_p (SD_p)	M_r (SD_r)	M_p (SD_p)	M_r (SD_r)
Original data (N8)	.39 (.22)	.48 (.12)	.39 (.22)	.48 (.12)
N7	.40 (.22)	.46 (.12)	.41 (.23)	.47 (.11)
N6	.40 (.22)	.45 (.12)	.42 (.23)	.46 (.12)
N5	.42 (.21)	.44 (.11)	.43 (.23)	.45 (.11)
N4	.44 (.21)	.40 (.11)	.46 (.22)	.44 (.10)
N3	.49 (.19)	.36 (.09)	.50 (.22)	.42 (.09)
N2	.59 (.15)	.31 (.11)	.59 (.20)	.41 (.10)

Note. All mean difficulty values (p) and mean polyserial correlations (r) were significantly different from those of the original data, based on paired t-tests.

Table 4 Means, Standard Deviations, Alphas, Corrected Alphas for Range Restriction, and Standard Errors of Measurement of Number-Correct Scores of Original and Simulated Test Forms Using a Frequency Criterion to Remove Nonfunctioning Options

	Random guessing					Educated guessing				
Form	M_{nc}	SD_{nc}	α_{nc}	CRT α_{nc}	SEM_{nc}	M_{nc}	SD_{nc}	α_{nc}	CRT α_{nc}	SEM_{nc}
Original data (N8)	12.89	5.06	.78		2.38	12.89	5.06	.78		2.38
N7	13.18	4.96	.76	.77	2.41	13.49	5.06	.78	.78	2.39
N6	13.36	4.86	.75	.75	2.43	13.8	4.99	.77	.77	2.41
N5	13.82	4.80	.74	.74	2.47	14.24	4.94	.76	.76	2.43
N4	14.67	4.48	.68	.69	2.53	15.06	4.81	.74	.74	2.46
N3	16.16	4.15	.60	.62	2.61	16.39	4.79	.73	.74	2.5
N2	19.61	3.60	.46	.48	2.65	19.61	4.44	.68	.69	2.5

Note. All score means were significantly different from those of the original data, based on the paired t-tests. CRT α_{nc} = corrected alpha coefficient (Haertel, 2006, p. 84). nc = number correct. SEM = standard error of measurement.

Results Using the Low-Discrimination Criterion to Remove Options

In this section, we present results obtained when options were removed by the discrimination criterion. Results on N2 are not presented because of data sparseness for some item scores.

Table 5 shows that when options were eliminated based on low discrimination (rather than low frequency), roughly the same results occurred for random guessing (decreased item difficulty and discrimination). However, for educated guessing, the results were quite different. Item difficulty decreased more dramatically, and item discrimination actually increased. Subsequently, mean number correct increased, particularly under educated guessing (see Table 6).

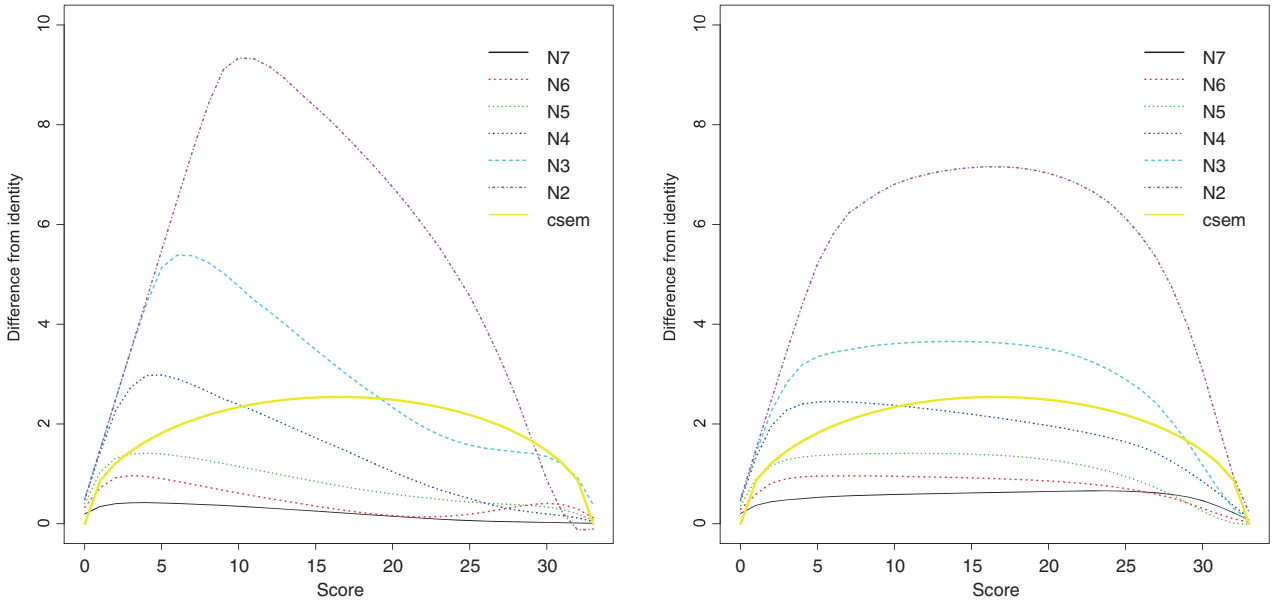

Figure 4 The equating function differences from the identity line under random guessing (left) and educated guessing (right). Options removed by low frequency.

Table 5 Means and Standard Deviations of Probability Correct and Item-Total Polyserial Correlations of Original and Simulated Test Forms Using the Low-Discrimination Criterion to Remove Nonfunctioning Options

	Random guessing		Educated guessing	
Form	M_p (SD_p)	M_r (SD_r)	M_p (SD_p)	M_r (SD_r)
Original data (N8)	.39 (.22)	.48 (.12)	.39 (.22)	.48 (.12)
N7d	.41 (.22)	.47 (.11)	.46 (.22)	.51 (.13)
N6d	.44 (.21)	.45 (.11)	.52 (.22)	.54 (.13)
N5d	.46 (.20)	.44 (.11)	.59 (.21)	.58 (.12)
N4d	.50 (.19)	.42 (.11)	.66 (.22)	.60 (.12)
N3d	.56 (.17)	.40 (.10)	.74 (.19)	.59 (.11)

Note. All mean difficulty values (p) and mean polyserial correlations (r) were significantly different from those of the original data, based on paired t-tests.

All of the changes in score means were statistically significant compared to the original data based on the paired t-tests.

Unlike results found using the frequency criterion to remove options, compared to educated guessing, the changes in score mean, score variance, and test reliability were larger in magnitude under educated guessing. Test reliability decreased as a result of removing options under random guessing, but reliability increased under educated guessing.

Figure 5 shows the difference between equated scores (on the y-axis) of simulated forms (N7 – N2) and the score on the original form (N8). For example, in Figure 5, a score of 15 on N8 corresponds to a score of 15.8 on N7, 16.5 on N6, 17.2 on N5, 18.8 on N4, and 20.8 on N3, respectively, under random guessing. Compared to random guessing, the equated scores using educated guessing were much higher, particularly for the upper score range. Under random guessing, the differences in equated scores were within CSEM for N7, N6, and N5, but under educated guessing, none of the equated scores were within the CSEM of the original scores on N8.

In summary, item difficulty, item discrimination, and test reliability tended to decrease as the number of options decreased, except for the case in which options were removed based on poor discrimination and assuming an educated guessing strategy, in which case item discrimination and test reliability increased.

Table 6 Means, Standard Deviations, Alphas, and Standard Errors of Measurement of Number-Correct Scores of Original and Simulated Test Forms Using the Low-Discrimination Criterion to Remove Nonfunctioning Options

	Random guessing					Educated guessing				
Form	M_{nc}	SD_{nc}	α_{nc}	CRT α_{nc}	SEM_{nc}	M_{nc}	SD_{nc}	α_{nc}	CRT α_{nc}	SEM_{nc}
Original data (N8)	12.89	5.06	.78		2.38	12.89	5.06	.78		2.38
N7d	13.68	5.05	.77	.77	2.43	15.08	5.55	.81	.81	2.42
N6d	14.36	5.00	.75	.76	2.48	17.18	5.97	.84	.84	2.4
N5d	15.30	4.92	.73	.74	2.54	19.47	6.41	.87	.87	2.33
N4d	16.58	4.85	.72	.72	2.58	21.67	6.22	.87	.87	2.22
N3d	18.54	4.60	.68	.69	2.61	24.34	5.64	.86	.86	2.09

Note. All score means were significantly different from those of the original data, based on the paired *t*-tests. CRT α_{nc} = corrected alpha coefficient (Haertel, 2006, p. 84). nc = number correct. SEM = standard error of measurement.

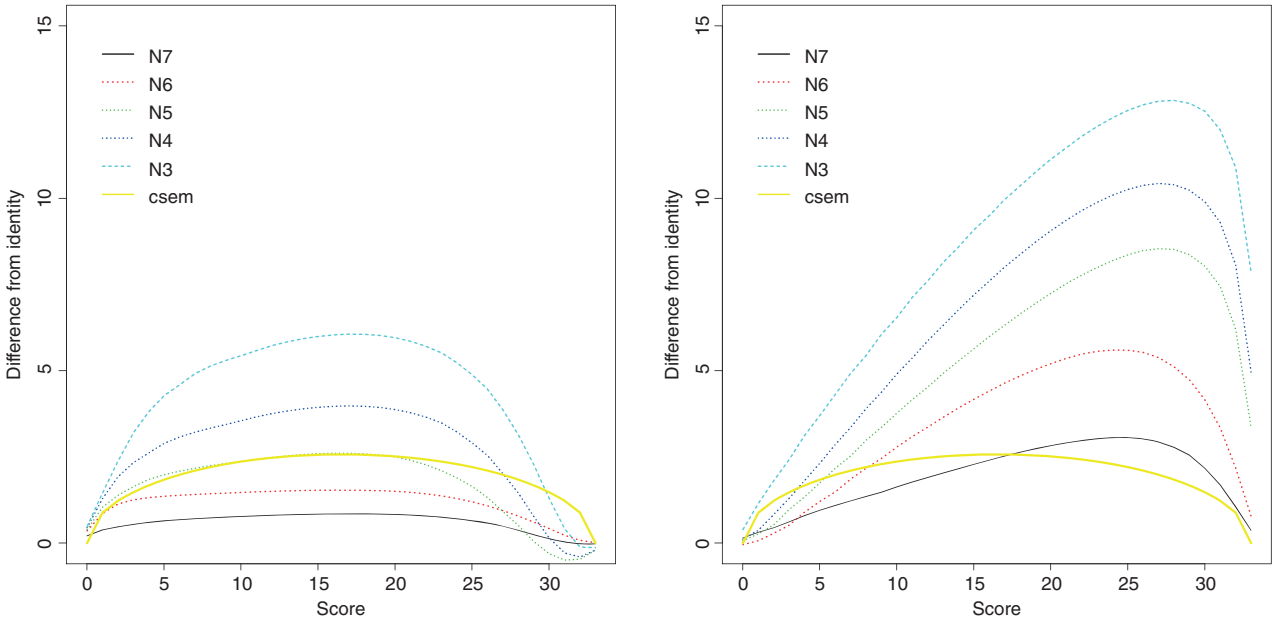

Figure 5 The equating function differences from the identity line under (left) random guessing and (right) educated guessing. Options removed by low discrimination.

Discussion

Using a real data set, we presented a simulation-based method to provide reasonable coverage of the possible impact of reducing the number of options on multiple-choice tests. The following summarizes our findings.

Impact on Form Difficulty

As the number of options decreased, items became easier (and slightly less variable). This was true regardless of whether options were removed based on low frequency or poor discrimination. However, the effect was especially pronounced when options were removed based on low discrimination and assuming that respondents engaged in an educated guessing strategy, in which case forms with fewer options were much easier.

Impact on Reliability

As the number of options decreased, reliability tended to decrease (and measurement error increased), especially if random guessing was deployed and low frequency was used to remove options. An exception was found for educated guessing

Table 7 Number of Additional Items for N5 to Reach the Original Reliability, .78

	Remove by low frequency		Remove by low discrimination	
	Random guessing	Educated guessing	Random guessing	Educated guessing
N5 reliability	.74	.76	.73	.87
Additional number of items	8	4	10	−15

after low-discrimination options were removed: Only under this condition did reliability actually increase as options were removed, at least to a point.

Impact on Scores

The impact of the number of options on number-correct scores could be substantial, particularly when options were removed by low discrimination. For example, from the original test with eight options to a form with only five (N5), number correct increased by 0.92 (using the frequency criterion and random guessing; Table 4), 1.34 (using the frequency criterion and educated guessing; Table 4), 2.31 (using the discrimination criterion and random guessing; Table 6), and 6.48 (using the discrimination criterion and educated guessing; Table 6), and the latter two values are comparable to or larger than SEM. Note that because the majority of keys are the most popular options on the test, the mean score under educated guessing is higher than that under random guessing.

On the basis of the results, number-correct scores obtained from the original and the simulated tests with reduced options were not comparable. Equating was used to compare scores. Equated scores of the simulated test forms increased as the number of options decreased. Except for educated guessing following removal of low-discrimination options, the differences between the original and equated scores for N5, N6, or N7 were within the CSEM in the whole score range, but reducing the number of options further, to N4, and especially to N2 and N3, led to scores being larger than the CSEM in most of the score range. This suggests that to maintain psychometric properties, such as test difficulty, test reliability, and test equating, but at the same time to be able to reduce testing time, it seems that five (N5) may be the an appropriate number of options per item for this test.

Impact on Test Length

Although we do not have data on response time under reduced options, it seems reasonable that reducing options would result in lower per item response time, which means that more time during the testing session could be used for additional items. According to the Spearman-Brown formula (Haertel, 2006, p. 77), for the current 33-item test, one more item may increase the reliability to

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

where k represents the factor by which the test lengthened, ρ is the estimated new reliability, and ρ_x is the original reliability. By the preceding formula, for five-option items on the test to reach the original reliability (.78) of the eight-option test, we would need to add four, eight, or 10 new items to the test under the worst case. Note that using the discrimination criterion and the conditional probability assignment of responses, the reliability was .87, well above .78 (Table 7).

How Should Options Be Removed?

Comparing the two criteria for removing nonfunctioning options, eliminating options that are selected infrequently has a smaller impact on score means because the number of affected test takers is smaller. However, item discrimination power determines test reliability, and options with the lowest frequencies may not have the lowest discrimination power. Therefore test reliability may decrease to a greater degree when the frequency criterion is used to define nonfunctioning options. If maintaining test reliability is the ultimate goal, the discrimination criterion is recommended for removing nonfunctioning options. This practice was also recommended in previous studies (Lord, 1980; Rodriguez et al., 2014; Williams & Ebel, 1957).

Table 8 Means and Standard Deviations of Probability Correct and Item-Total Polyserial Corrections of Original and Simulated Test Forms Using the Low-Discrimination Criterion to Remove Nonfunctioning Options

Form	M_p	SD_p	M_r	SD_r
N8	.39	.22	.48	.12
N7m	.43	.21	.48	.11
N6m	.46	.21	.47	.11
N5m	.49	.20	.48	.10
N4m	.50	.21	.49	.09
N3m	.58	.18	.45	.09

Realism of the Simulation's Assumptions on Respondent Behavior

Random guessing assumes that test takers randomly guess among the remaining options when their chosen options are removed. This introduces more noise, so item discrimination power and test reliability decrease more compared to educated guessing as the number of options decreases.

Educated guessing assumes that test takers whose chosen options are removed will behave in the same way as others (of the same ability) whose chosen options are functioning. This leads to a larger score mean increase (compared to random guessing) as options are removed, and if options are removed based on the discrimination criterion, then the test actually gets more reliable, and item discrimination power is increased, as the number of options decreases.

Both hypothetical behaviors may not be realistic, and researchers cannot control test takers' response patterns. It is challenging to understand and model test takers' real behavior on multiple-choice tests when the number of options is reduced.

Lord (1980) showed theoretically that the three-option item provided the most information curves at the midrange of the score scale, whereas the two-option multiple-choice item provided the most information for high-scoring examinees and the four- and five-option formats provided the most information for examinees who tended toward low scores. Therefore high-scoring students may be less inclined to random guess, thereby not needing as many options as students with lower scores, who are more inclined to random guess (Haladyna & Downing, 1989b; Levine & Drasgow, 1983). Therefore the educated guessing strategy assumes that, given their true ability (approximated by their scores on the original test), examinees' responses to items are based on their ability regardless of whether their chosen options are removed. Empirical studies reported an increase or no difference in item discrimination and in reliability (Haladyna, Downing, & Rodriguez, 2002, and reference therein) with fewer options. Hence, in our study, we hypothetically assumed two extreme response strategies: random guessing and educated guessing.

Some studies (Attali & Bar-Hillel, 2003) showed that some examinees usually chose Option C when guessing because they believed that Option C is very likely to be the correct answer. How would the test property change if the test takers use different guessing strategies? To investigate this issue, we experimented with one more set of simulations with a mixed guessing strategy that is a combination of random guessing, educated guessing, and selecting Option C only. Specifically, we removed nonfunctional options by the discrimination criterion first. Next, to reassign responses, for each item on the test, one-third of the test takers were randomly selected and assumed to use an educated guessing strategy, another one-third were randomly selected and assumed to use a random guessing strategy, and the remaining one-third of test takers were assumed to choose Option C all the time. This way, each test taker was likely to use different guessing strategies on different items. The results are shown in Tables 8 and 9.

Under the mixed guessing strategy (refer to Table 8), item difficulty decreases as the number of options becomes smaller, similar to what is shown in Table 5. As expected, item difficulty and discrimination power are between the random guessing and educated guessing results.

Similar to what is shown in Table 6, we observe in Table 9 that, under the mixed guessing strategy, test forms became easier (scores increase) as the number of options decreases. The score means are between those obtained by random and educated guessing. The test reliability, however, is rather stable and comparable to that of the original data.

The simulation results under mixed guessing strategy lend more support to the claim that our simulation-based method may provide reasonable coverage and approximation of empirical data should the tests with reduced numbers of options be administered. Note that decisions on the optimal number of options should not be based solely on the simulation results;

Table 9 Means, Standard Deviations, Alphas, and Standard Errors of Measurement of Number-Correct Scores of Original and Simulated Test Forms Using the Low-Discrimination Criterion to Remove Nonfunctioning Options

Form	*M*	*SD*	Alpha	SEM
N8	12.89	5.06	.78	2.38
N7 m	14.20	5.23	.78	2.45
N6 m	15.15	5.29	.78	2.48
N5 m	16.32	5.49	.79	2.51
N4 m	16.39	5.51	.80	2.48
N3 m	19.30	5.03	.75	2.53

rather, the simulation results should be treated as the first step in guiding test developers, researchers, and practitioners in choosing appropriate numbers of options for further studies on the tests under development or redesign.

Implementation

The value of the simulation-based method resides in the numerical comparison of the impact of the number of options per test item; changing the number of options may have implications for cost efficiency and labor savings and may thereby be useful and informative in guiding practitioners. For instance, for the eight-choice test we illustrated, we compared results from seven simulated forms with numbers of options per item being two, three, four, five, six, and seven, respectively. This can be easily implemented in simulations but not in practice. Results from the two extreme response assignments showed that reducing the number of options from eight to five may not cause dramatic damage to psychometric properties of the test, in terms of test difficulty, test reliability, and score compatibility.

As Haladyna and Downing (1988) suggested, item writers can produce as many options as possible for an item in the beginning stages of test development. It is not possible to administer all possible numbers of options for a test to decide what the optimal number is. Therefore the results from our simulation-based method can guide researchers, test developers, and practitioners in choosing limited numbers of options in the second round of field trials so that empirical data can be collected. If, in addition, response time can be collected for the test, the response time difference between the original test and the tests with reduced numbers of options can further guide us on how many new items we might appropriately add to the test, if necessary.

Limitations

One limitation of our study is that the assessment we used for illustration purposes has eight options per item. While it is typical for a test that measures inductive reasoning to have eight options, it is unusual for educational assessments. Another limitation is that simulations were conducted only once under each condition of the 2 (criteria) × 2 (schemes) × 6 (number of options) = 24 design conditions in this study. Researchers can replicate the simulations and analyses as many times as necessary to reach stable results. In addition, one can adopt the asymptotic results (Guo, 2016) to predict score distributions when the number of options is reduced and when the random assignment of response scheme is assumed.

Acknowledgments

The authors would like to thank Gautam Puhan, Sandip Sinharay, Sooyeon Kim, and several anonymous reviewers for their comments on the report. This publication is based on work supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA) SHARP program, via contract 2015-14120200002-002, and is subject to the Rights in Data – General Clause 52.227-14, Alt. IV (DEC 2007). The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, expressed or implied, of ODNI, IARPA, or the U.S. government or of any of the authors' host affiliations. The U.S. government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotations therein.

Notes

1 As a reviewer pointed out, the classic point-total correlation for distractors is biased because it often simply codes the studied distractor as 1 and the others as 0. That is, the 0 code includes both test takers who may have selected the correct responses and those who selected a distractor. We ran additional analyses to investigate the differences between the classic point-total correlation and an alternate one in which the code 0 included only the test takers who selected the correct option. We found that the ranking differences of the point-total correlation between the two calculation schemes were mostly 0 or at most 1, which confirmed the reviewer's comment that the decisions made in the two calculation schemes were relatively similar.

2 In practice, an item may have a distractor that has a nonnegligible positive discrimination value and is competitive for the key. That is, the item potentially could have double keys. Such a dysfunctioning distractor should be removed or revised before identifying nonfunctioning options. In our data, we did not see double-key items, and all distractors had negative or near-zero correlation coefficients with the total scores.

3 As a reviewer pointed out, the probability of choosing either option is expected to be approximately .5 for examinees with low ability. In the simulated data N2, few test takers scored below 15. Because of the sparse data and the smoothing method we used to draw the curve, one simulation of responses may cause probability to be diverted from .5 at the lower end of the scores.

References

Albano, A. (2016). Equate: An R package for observed-score linking and equating. *Journal of Statistical Software, 74*(8), 1–36. https://doi.org/10.18637/jss.v074.i08

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*, 109–128. https://doi.org/10.1111/j.1745-3984.2003.tb01099.x

Baghaeri, P., & Amrahi, N. (2011). The effects of the number of options on the psychometric characteristics of multiple choice items. *Psychological Test and Assessment Modeling, 53*, 192–211.

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*, 183–196. https://doi.org/10.1111/j.1745-3984.1985.tb01057.x

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor analytic studies*. New York, NY: Cambridge University Press.

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

Grier, J. (1975). The number of alternatives for optimum test reliability. *Journal of Educational Measurement, 12*, 109–112. https://doi.org/10.1111/j.1745-3984.1975.tb01013.x

Guo, H. (2016). Predicting rights-only score distributions from data collected under formula score instructions. *Psychometrika, 82*, 1–16. https://doi.org/10.1007/s11336-016-9550-9

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

Haladyna, T. M., & Downing, S. M. (1988, April). *Functional distractors: Implications for test-item writing and test design*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

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

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education, 15*, 309–333. https://doi.org/10.1207/S15324818AME1503_5

Holland, P. W., & Thayer, D. T. (2000). Univariate and bivariate loglinear models for discrete test score distributions. *Journal of Educational and Behavioral Statistics, 25*, 133–183. https://doi.org/10.3102/10769986025002133

Kolen, M., & Brennan, R. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York, NY: Springer.

Levine, M. V., & Drasgow, F. (1983). The relation between incorrect option choice and estimated ability. *Educational and Psychological Measurement, 43,* 675–685. https://doi.org/10.1177/001316448304300301

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

Lord, F. (1984). Standard errors of measurement at different ability levels. *Journal of Educational Measurement, 21*, 239–243. https://doi.org/10.1111/j.1745-3984.1984.tb01031.x

Princeton Review. (2016). *The new SAT*. Retrieved from http://www.princetonreview.com/college/sat-changes

Ramsay, J. (1991). Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika, 56*, 611–630. https://doi.org/10.1007/BF02294494

R Core Team. (2014). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.Rproject.org/

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*, 3 – 13. https://doi.org/10.1111/j.1745-3992.2005.00006.x

Rodriguez, M., Kettler, R., & Elliott, S. (2014, October 7). Distractor functioning in modified items for test accessibility. *Sage Open*, *4*(4). https://doi.org/10.1177/2158244014553586

Schneid, S., 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*, 1020 – 1027. https://doi.org/10.1111/medu.12525

Williams, B. & Ebel, R. (1957). The effect of varying the number of alternatives per item on multiple-choice vocabulary test items. *The yearbook of National Council on Measurements Used in Education* (no. 14, pp. 63 – 65). Ames, IA: National Council on Measurements Used in Education.

Woodruff, D. (1990). Conditional standard error of measurement in prediction. *Journal of Educational Measurement, 27*, 191 – 208. https://doi.org/10.1111/j.1745-3984.1990.tb00743.x

Suggested citation:

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

Action Editor: Gautam Puhan

Reviewers: Sandip Sinharay and Sooyeon Kim