Psychological Test and Assessment Modeling, Volume 54, 2012 (1), 3-19

An analysis of the discrete-option multiple-choice item type

Neal M. Kingston[1]*, Gail C. Tiemann*[2]*,*
Harold L. Miller Jr. [3] *& David Foster*[4]

Abstract

The discrete-option multiple-choice (DOMC) item type was developed to curtail cheating and reduce the impact of testwiseness, but to date there has been only one published study of its statistical characteristics, and that was based on a relatively small sample. This study was implemented to investigate the psychometric properties of the DOMC item type and systematically compare it with the traditional multiple-choice (MC) item type. Test forms written to measure high school-level mathematics were administered to 802 students from two large universities. Results showed that across all forms, MC items were consistently easier than DOMC items. Item discriminations between DOMC and MC items varied randomly, with neither performing consistently better than the other. Results of a confirmatory factor analysis was consistent with a single factor across the two item types.

Key words: computer-based testing, innovative item types, discrete-option multiple-choice, multiple-choice, testwiseness

[1] *Correspondence concerning this article should be addressed to:* Neal Kingston, PhD, Center for Educational Testing and Evaluation, Joseph R. Pearson Hall, 1122 West Campus Road, Room 738, Lawrence, Kansas, 66045; email: nkingsto@ku.edu

[2] University of Kansas

[3] Brigham Young University

[4] Kryterion, Incorporated

Introduction

The multiple-choice (MC) test-item type has a long history. Over the last 90 years, empirical research has established this type of selected-response item as an efficient and effective method of measuring cognitive ability (Downing, 2006). From its historical roots in sorting and classifying World War I soldiers to its modern-day use in accountability and professional licensure testing, the MC item has shown flexibility and versatility in a variety of testing situations.

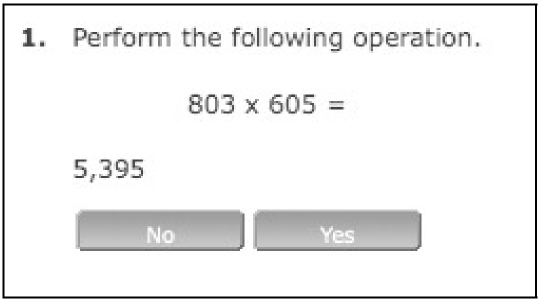
Throughout this course of development and use, many have described the strengths and weaknesses of the MC item. Tanner (2003) noted that because of the relatively short amount of time needed for students to take an MC item, its use allows for a thorough and representative sampling of the target domain, thus reducing the threat of construct underrepresentation. In addition, because scoring is straightforward, machine or computer scoring is efficient, objective, and reliable (Downing, 2006). Overall, because of their strong measurement properties, evidence to support the validity of scores derived from MC tests is more readily assembled and evaluated compared with constructed-response items (Downing, 2006).

MC items are not without their limitations, however. There is some criticism that MC items are not consistent with situations students face in the real world (Tanner, 2003) and that a narrowing of the curriculum can occur when MC-based accountability tests focus on low-level content (Boyd, 2008; Nichols & Berliner, 2007). Others, however, demonstrate that MC questions can indeed be written to reflect higher-order thinking (Nitko, 2004; Taylor & Smith, 2009); thus the reflection of low-level content can be the fault of the item writer (Downing, 2006). Indeed, adhering to item-writing rules can reduce a number of possible flaws in an MC item (Frey, Petersen, Edwards, Pedrotti, & Peyton, 2005).

In addition to content-related flaws, performance on MC items can be affected by *testwiseness*, or the ability to determine a correct answer without actually grasping the related content. Although MC items certainly allow for random guessing, test takers can also use partial knowledge to eliminate incorrect answers (Downing, 2006). Indeed, numerous resources and "how-tos" exist to teach test takers how to "guess intelligently" in order to increase the probability of answering an item correctly (Blackey, 2009). The ability to change answer choices is also a component of testwiseness. Geiger (1997) studied answer changing with undergraduate psychology students, finding that for every one point lost in changing an item response, three points were gained. Geiger's results replicated Benjamin, Cavell, and Shallenberger's (1984) meta-analysis, which found a two- to three-point increase for every point lost in answer-changing behavior. Overall, numerous empirical studies have concluded that testwiseness adds variance to scores, with many finding a positive relationship between testwiseness and test performance. Geiger's study, in particular, correlated scores on a measure of testwiseness with test performance and concluded that a moderate portion of test variance could be contributed to testwiseness. Clearly, inflated scores resulting from testwiseness detract from score validity by increasing construct-irrelevant variance.

The discrete-option multiple-choice item type

Variations of the traditional MC item have been created in an effort to reduce these limitations and enhance construct representation (Downing, 2006; Sireci & Zenisky, 2006). One such item type is the computer-based discrete-option multiple-choice (DOMC) item (Foster & Miller, 2009). The DOMC is similar to the traditional MC item in that a stem is offered, and students must choose from a limited number of possible responses. However, with the DOMC item, test takers are randomly presented with one answer option on the screen at a time (see Figure 1).



1. Perform the following operation.

803 x 605 =

5,395

Figure 1:
Screen shot of discrete-option multiple-choice item.

After reviewing the response option, test takers decide whether the option is correct by selecting “yes” or “no.” The item is scored correct if the test taker answers “yes” to the correct response. A test taker can answer incorrectly by either selecting “no” when the correct option is presented or by selecting “yes” when an incorrect answer is presented. The item is “complete” once the computer determines that the student has incorrectly or correctly answered the item. However, once the test taker has selected “yes” or “no” to the correct response, the system may present another, unscored option. This helps prevent the correct answer from being cued. Note that test takers are not allowed to go back and review previously shown responses or change their answers.

While the DOMC item type cannot control for poorly created or worded items, Foster and Miller (2009) hypothesized that the new item type could help control for testwiseness. As mentioned earlier, one of the key limitations of the traditional MC item is the ability to eliminate answering based on partial knowledge or changing answers. In a series of small empirical studies using psychology content and administered to undergraduate psychology students, Foster and Miller compared the psychometric characteristics of the DOMC item with the traditional MC item. They found that most DOMC items were significantly more difficult than MC, though some were about the same or easier. The difference in the mean *p* values was 10.7 %. Though they expected point-biserial correlations to be higher for the tests using DOMC items, 40 % of their DOMC items had higher point-biserial correlations than their MC counterparts. Related to item analysis, the authors concluded that the formats behaved differently based on the content of the question. Additional results indicated that test takers took about 10 % less time to answer the DOMC items, suggesting that considering fewer response options contributed to the reduction. Survey results indicated that half of the students in the sample strongly or moderately preferred the MC item over the DOMC item. Overall, the authors concluded that based on their small study the DOMC item had the potential to reduce the effects of testwiseness, improve test security, and improve the fairness of an assessment by reducing construct-irrelevant variance. However, these initial studies used small sample sizes at a single university, thus limiting generalizability.

The purpose of the current study was to expand on Foster and Miller's (2009) initial effort to test the psychometric differences between the DOMC and the traditional MC item through a larger-scale, more closely controlled study. In this study, the following comparisons between DOMC and MC items were made: frequency distribution, mean score, reliability, factor structure, and differential item functioning.

Method

Four parallel item sets were created from items written to measure high school-level mathematics. Test specifications were created with item categories based on high-school standards and benchmarks described in *Content Knowledge: A Compendium of Standards and Benchmarks for K-12 Education*" by Kendall and Marzano (2000). Test specifications are described in Table 1.

Items were written to match 10 item-description categories (see Table 1), and one item from each category was placed in each of four 10-item sets. The sets were designed to be parallel in terms of content and judgment of likely difficulty. The sets were placed on four test forms using varying modes of administration, DOMC and MC, in order to systematically compare item performance. Thus, the test forms represented four possible test-administration conditions.

Table 2 presents an overview of the test design. DOMC indicates that the set was administered in the discrete-option multiple-choice format. MC indicates the set was administered in the traditional multiple-choice format.

Item sets 1-4 each consisted of 10 items, for a total of 40 unique items across all forms. The item sets each reflected the same item descriptions and presented items in the order of item descriptions shown in Table 1. For all test forms, items in sets 1 and 2 were administered first (as items 1-10 and 11-20, respectively) in order to (a) allow linking with section scores of sets 3 and 4, and (b) provide the test taker with some familiarization with the item types and the test-delivery system. Forms A1 and A2 (as well as B1 and B2) differed only in that the position of sets 3 and 4 on the forms were reversed in order to control for possible item-type order effects. Items in set 3 were presented as DOMC items on the A forms and as MC items on the B forms. Items in set 4 were presented as MC items on the A forms and DOMC items on the B forms.

Table 1:
Test specifications

Benchmark/Standard	Item description
Uses basic and advanced procedures while performing the processes of computation	A. Multiply two three-digit numbers where the first digit is greater than 5 and the second digit is 0. B. Solve a numeric equation requiring knowledge of order of operations. Problem will include one of each of the following operations: parentheses, exponentiation, multiplication, division, and subtraction, and will be written on one horizontal line. C. Reduce an equation to simplest form. Equation will include two variables and two constants represented algebraically. D. Solve a simple permutation word problem.
Understands and applies basic and applied properties of the concepts of measurement	E. Determine the relationship between area and volume. F. Solve a simple word problem using velocity.
Understands and applies basic and applied properties of probability	G. Determine probability using a counting procedure, enumerating equally likely events and dividing events of interest by total events. H. Determine conditional probabilities from a three-by-three contingency table.
Understands and applies basic and applied properties of functions and algebra	I. Understand the meaning of slope and intercepts in linear functions. J. Solve a system of linear equations with two variables.

Table 2:
Test-form design

Form	Items 1-10	Items 11-20	Items 21- 30	Items 31-40
A1	DOMC item set 1	MC item set 2	DOMC item set 3	MC item set 4
A2	DOMC item set 1	MC item set 2	MC item set 4	DOMC item set 3
B1	DOMC item set 1	MC item set 2	DOMC item set 4	MC item set 3
B2	DOMC item set 1	MC item set 2	MC item set 3	DOMC item set 4

Sample

Undergraduate and graduate students from Brigham Young University and the University of Kansas participated in the study during the spring and fall of 2009. Students likely to exhibit advanced knowledge of mathematics, including mathematics, mathematics education, engineering, computer science, accounting, and economics majors were excluded from the study, as we believed the content would be too easy and thus provide no differentiation across item types. Students were paid $5 to participate.

The final sample consisted of 802 individuals. Fifty-one percent were female, 38 percent were male, and 12 percent did not indicate a gender. Ten percent indicated that they were non-resident aliens. One percent were of unknown race and 3 % described themselves as Hispanic of any race. Sixty-four percent were white, 14 % were Asian, 4 % were of two or more races, and 1 % were Black or African American. Less than 1 % were American Indian or Alaska Native, or Native Hawaiian or other Pacific Islander. Fourteen percent did not indicate a race.

Data collection

Assessments were delivered online using Webassessor™ test-administration software, which offers DOMC functionality. Prior to data collection, unidentifiable test codes were generated and assigned to the four research conditions: test forms A1, A2, B1, and B2, as described above. The list of test codes was then compiled and ordered randomly. During data collection, examinees were given the first unused test code from the list, which also served as a login to the Webassessor™ platform. While random assignment of examinees to test forms was a goal of the project, an error in test-code assignment prevented this from occurring until about halfway through data collection. Once the error was discovered, test codes were regenerated to balance the delivery of the remaining forms and then randomly assigned to further examinees. All test delivery was proctored by trained research assistants. For this study, DOMC answer options were delivered in sequential order; participants therefore experienced all response options in the same manner. Since the study included strong covariates, the lack of complete randomization was not considered a significant flaw, and all data were used for the analyses presented herein.

Procedures

Raw frequencies were determined for all possible item-set scores and compared between MC items and their DOMC counterparts in item set 3 and again for item set 4. Classical item statistics were calculated for items and item sets across forms, including item difficulties, point-biserial correlations, internal consistencies, and intercorrelations. Item difficulties were normalized for analyses involving correlation of DOMC and MC item sets. Coefficient alpha was estimated for each item set under each administration condition. Multivariate analysis of covariance (MANCOVA) was used to statistically compare raw means for MC and DOMC items offered in set 3 and in set 4. Scores on sets 1 and 2 served as the covariates, since these item sets were offered in the same format and position across all forms. Confirmatory factor analysis was used to compare three different models of test structure. The Mantel-Haenszel statistic was used to assess differential item functioning between male and female participants.

Results

Item-set statistics

Table 3 lists and compares the score frequencies and cumulative-score percentages for item set 3 when offered as DOMC or as MC in test position 3 (i.e., forms A1 and B2). Table 4 offers the same information for item set 4 when offered in position 3 (i.e., forms A2 and B1). Lack of consistent random assignment resulted in fewer examinees taking forms A2 and B2. However, the cumulative percentages for both sets 3 and 4 show that lower scores were more frequent when the set was offered as DOMC items than when offered as MC items.

Means for DOMC items in set 1 and MC items in set 2 are summarized in Table 5. Across all forms, items offered as MC had higher *p* values than those offered as DOMC. Item-set scores across the forms varied by 1.67 standard deviations for the MC items and by 2.01 standard deviations for DOMC items.

Table 3:
Raw frequencies and cumulative score percentages for item set 3 in test position 3

Item set 3 – Raw frequencies (Groups are not random)				
	Count		Cum. Percent	
Score	DOMC	MC	DOMC	MC
0	0	0	0.0%	0.0%
1	7	2	1.3%	0.7%
2	8	2	2.8%	1.5%
3	28	6	8.1%	3.7%
4	62	17	19.8%	9.9%
5	85	19	35.8%	16.9%
6	96	25	54.0%	26.1%
7	89	65	70.8%	50.0%
8	82	73	86.2%	76.8%
9	61	44	97.7%	93.0%
10	12	19		
n	530	272		
Mean	6.23	7.21		
SD	1.95	1.80		

Table 4:
Raw frequencies and cumulative score percentages for item set 4

Item set 4 – Raw frequencies (Groups are not random)				
	Count		Cum. Percent	
Score	DOMC	MC	DOMC	MC
0	2	0	0.0%	0.0%
1	4	0	1.3%	0.7%
2	11	3	2.8%	1.5%
3	18	14	8.1%	3.7%
4	32	30	19.8%	9.9%
5	42	44	35.8%	16.9%
6	46	60	54.0%	26.1%
7	45	117	70.8%	50.0%
8	43	115	86.2%	76.8%
9	23	84	97.7%	93.0%
10	6	63		
n	530	530		
Mean	5.96	7.21		
SD	2.10	1.83		

Table 5:
Item set 1 and item set 2 means across forms

Group taking form		Raw mean		95% Confidence interval	
	n	DOMC Set 1	MC Set 2	DOMC Set 1	MC Set 2
A1	249	6.2	7.0	[5.95, 6.45]	[6.79, 7.21]
A2	281	6.5	7.2	[6.26, 6.74]	[7.00, 7.40]
B1	141	6.1	6.9	[5.77, 6.43]	[6.62, 7.18]
B2	131	6.0	6.8	[5.66, 6.34]	[6.51, 7.09]

Multivariate analysis of covariance

In order to rule out an order effect between forms A1 and A2, as well as B1 and B2, MANCOVA was used with the test form as the independent variable, the scores on sets 3 and 4 as dependent variables, and the scores on sets 1 and 2 as covariates. Significant differences among the test forms and the set scores were found, Wilks's Λ = .71, *F* (6,

1590) = 50.36, p = .000. Analyses of covariance (ANCOVA) for each dependent variable were conducted as follow-up tests to the first MANCOVA. The ANCOVAs for sets 3 and 4 were significant, both at the p = .000 level. Post hoc analyses for the item-set variables consisted of pairwise comparisons to find which test forms were different from each other. Results indicated that, for both set 3 and set 4, form A1 was not significantly different from form A2, nor was form B1 significantly different from B2. Thus, no order effect was indicated. Based on these results, scores from forms A1 and A2, as well as B1 and B2, were dummy coded as simply group A and group B for subsequent analyses.

A one-way MANCOVA was then conducted to determine the effect of the item type, DOMC or MC, on the dependent variables, set 3 and set 4 scores. Scores for sets 1 and 2 again served as covariates. Significant differences were found among the two groups and two sets, Wilks's Λ = .71, F (2, 797) = 163, p = .000. The multivariate η^2 based on Wilks's Λ was .29.

ANCOVAs on each dependent variable were conducted as follow-up tests to the MANCOVA. The ANCOVA for set 3 scores was significant, F (1,801) = 108.98, p = .000, η^2 = .12. The ANCOVA for set 4 scores was also significant, F (1,801) = 107.32, p = .000, η^2 = .12. Set 4 offered items in MC format on the group A forms and DOMC on the group B forms. Table 6 presents the means, adjusted means, standard errors, and adjusted standard errors of estimate for sets 3 and 4.

Table 6:
Item-set statistics: MANCOVA

	Set 3		Set 4	
	DOMC	MC	DOMC	MC
Raw Means	6.23	7.21	5.96	7.31
Adjusted Means	6.17	7.34	6.10	7.24
Raw SEE	.08	.11	.13	.08
Adjusted SEE	.07	.09	.09	.06

Internal consistency and intercorrelations

Coefficient alpha was computed as an internal estimate of reliability. Results for each item type by item set are listed along the diagonals in Table 7. Correlations between the scores from all the item sets are described in the lower left quadrant of the table, with correlations corrected for attenuation due to the unreliability of each of the paired scores (Haertel, 2006, equation 42) listed in the upper-right quadrant. Note that the table does not include a DOMC 2 entry because item set two was only administered in multiple-choice format. Similarly there is no MC 1 item set as item set one was only administered in DOMC format. Because of the high number of attenuated correlations, the bulk of the data suggest the DOMC and MC items are measuring the same construct.

Table 7:
Reliabilities and intercorrelations

	DOMC 1	DOMC 3	DOMC 4	MC 2	MC 3	MC 4
DOMC 1	**.56**	>1.00	.96	>1.00	.92	>1.00
DOMC 3	.61	**.54**	--	.99	--	>1.00
DOMC 4	.55	--	**.59**	>1.00	>1.00	--
MC 2	.55	.50	.56	**.47**	>1.00	>1.00
MC 3	.49	--	.59	.50	**.51**	--
MC 4	.61	.61	--	.54	--	**.57**

Classical item statistics

Again, each item in set 3 and set 4 was offered in the DOMC format to some students and in the MC format to others, depending on test-form assignment. In Figure 2, normalized *p* values (Henryssen, 1971, p. 139) for DOMC items are plotted against their counterpart MC item. Normalized p is the z-score that corresponds to the proportion of the distribution. For example a z-score of 0 corresponds to a p of .5 and a z-score of 1 corresponds to a p of .84. The normalized p is used because it has better statistical characteristics than p (Henryssen, 1971, p. 139).

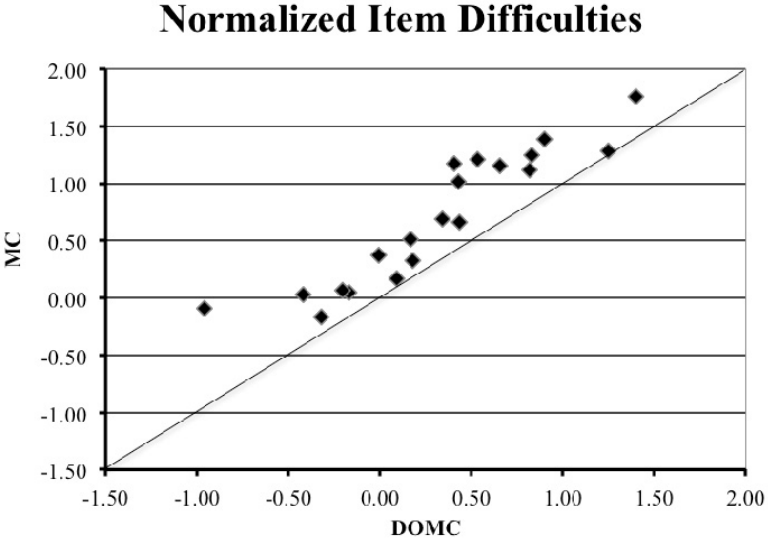


Figure 2:
Normalized item difficulties.

The plot in Figure 2 demonstrates that, at the item level, each item was easier when offered as MC than when offered as DOMC and that the correlation between item difficulties was high (.93). One item was considerably less difficult when offered as MC, with a *p* value of 0.17 as DOMC and 0.46 as MC (normalized *p* values of -0.96 and -0.09, respectively). This item was from the "simple combination" content category.

In Figure 3, biserial correlations for DOMC items were plotted against their counterpart MC items. Biserial correlations were calculated using the 10 item section total as the criterion. While correlations were not part-whole corrected, this was equally true for both item formats and thus not expected to influence the relationship between the two. Results showed a modest correlation between biserial correlations (.36) and that some items were better discriminating as DOMC and others as MC. There appeared to be no relationship between item content and whether biserial correlations were higher for one item type or the other. Six items discriminated almost identically as DOMC and MC (though slightly favoring DOMC). Seven items discriminated less well as MC than as DOMC. Two items discriminated considerably less well as DOMC than as MC, with the remaining four items discriminating somewhat less well. Examination of item content provided no indication as to why the two items discriminated considerably less well.

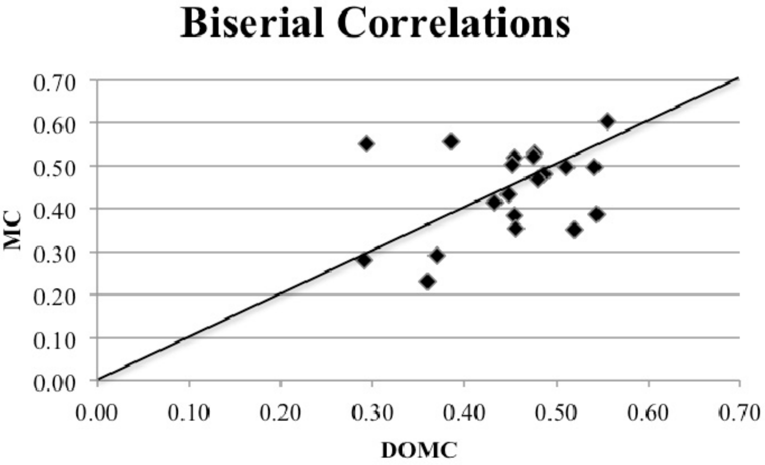


Figure 3:
Biserial correlations.

Confirmatory factor analysis

A confirmatory factor analysis was performed, comparing three different measurement models for the combined data set of responses to DOMC and MC items. An item-parcel approach was used in order to minimize the impact of statistical artifacts (Cattell &

Burdsal, 1975). Set 1 and set 2, as described earlier, were highly parallel in content, with one administered as DOMC and the other as MC. The two sets were each divided into parcels of three, three, and four items, with the content of the parcels from set 1 parallel to the content from the corresponding parcel in set 2.

The first of the three models posited a single factor underlying all six parcels. Based on problems with model fit, a second approach used a single factor with pairwise fixed-error variances. This reflected the high degree of parallelism in the content of the items in the paired parcels. The third model used two factors, one for each item type. Figures 4, 5, and 6 show each of these models. SAS Proc Calis was used to fit all models. Table 8 presents the results of these analyses.

The highlighted cells indicate the model that had the best fit to the data for the criterion indicated in that row. For four of the five criteria, the one-factor model with paired-error variances showed the best fit. For *p* influenced by sample size, the two-factor model showed a marginally better fit than the one-factor model with paired-error variances. Based on these results, we believe a one-factor model (with paired-error variances) is sufficient to explain the observed data.

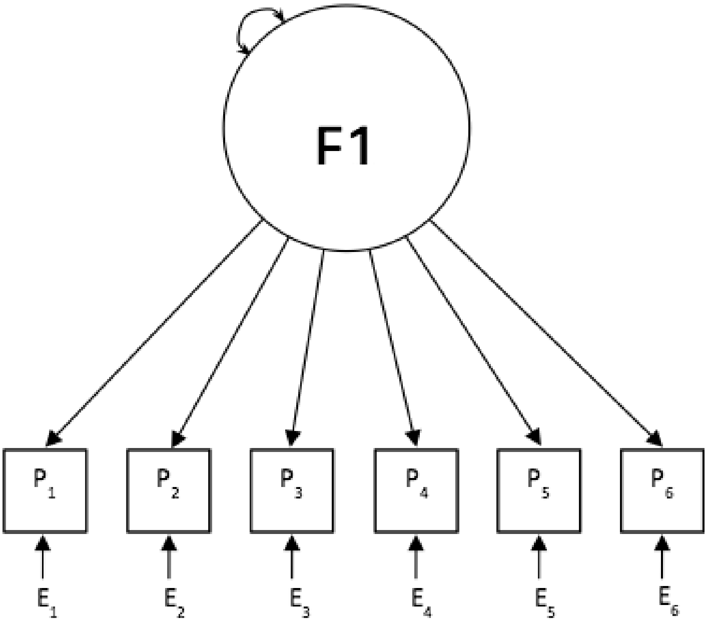


Figure 4:
Single-factor model.

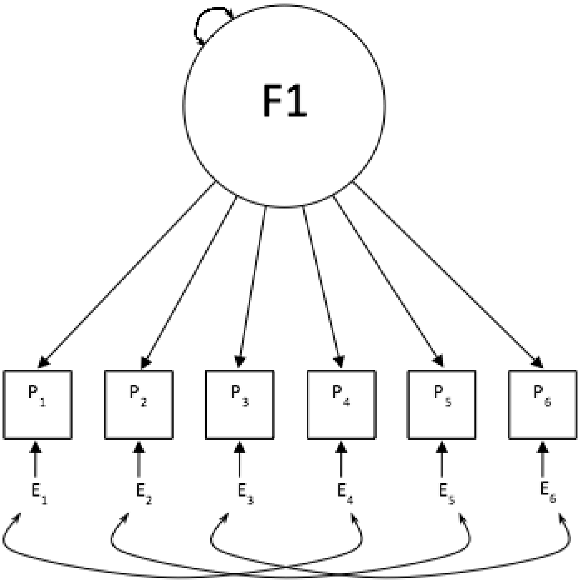


Figure 5:
Single-factor model with pairwise fixed error variances.

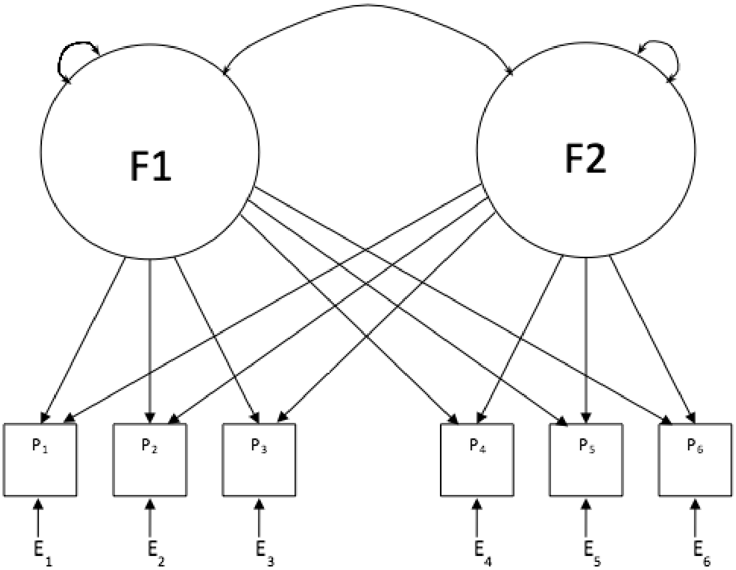


Figure 6:
Two-factor model.

Table 8:
Confirmatory factor analysis results

Statistic	One Factor	One Factor With Pairwise Fixed-Error Variances	Two Factors
Chi-square	51.20	12.90	5.81
df	9	6	2
p (influenced by sample size)	>.0001	.04	.05
Chi-square/df	5.70	2.20	2.90
RMSEA	.08	.04	.05
Bentler's comparative fit index (should exceed .90)	.93	.99	.95
Bentler & Bonett's non-normed index	.88	.97	.95

Differential item functioning

Using DIFAS (Penfield, 2005), differential item functioning (DIF) analysis was performed comparing males with females using data from sets 1 and 2, which were administered to all examinees. Because of insufficient numbers of students for other demographic characteristics, this was the only comparison made. Two different sets of decision rules were applied. The combined decision rule (Penfield, 2003) flags an item if either the Mantel-Haenszel chi-square or the Breslow-Day chi-square statistic is significant at $p < 0.025$. The second set of rules was based on the ETS categorization procedure (Zieky, 1993), which uses a combination of effect size and statistical significance.

Under the combined decision rule, three DOMC items and three MC items were flagged. Females performed better on two of the three DOMC items, as well as on two of the three MC items. Using the ETS categorization procedure, no items showed large levels of DIF, whereas five items showed moderate levels of DIF. One was a DOMC item that had also been flagged by the combined decision rule and favored females. Four were MC items, three of which had been flagged by the combined decision rule. Two of these four items favored females, and two favored males.

Discussion

The goal of this study was to explore the psychometric differences between the DOMC and traditional MC item types through a larger, more controlled study than had been attempted before. DOMC-item performance was directly compared to MC-item performance through assignment of test takers to systematically varied test forms. Results showed that, across all forms, MC items were consistently easier than DOMC items. Mean scores from DOMC item sets were also lower than for MC sets, which could possibly be attributed to a reduced impact of testwiseness in responding to DOMC items or alternatively to the fact that examinees cannot revisit options once they make a choice to select or not select an option. Item discriminations between DOMC and MC items varied, with neither performing consistently better than the other. Additionally, variation among reliabilities of the DOMC and MC sets showed no consistent pattern. Thus the results were consistent with those from the prior, limited study of the DOMC item type.

Confirmatory factor analysis provided no evidence that MC and DOMC are measuring different constructs. This result calls to question whether the two item types are differentially impacted by testwiseness, although it is possible that testwiseness effects are too highly correlated with achievement to be detected with confirmatory factor analysis.

Based on the results of this study, there appear to be no psychometric reasons for excluding DOMC items from testing programs. Note, however, that correlational analyses (such as the confirmatory factor analysis performed in this study) are not always sensitive to group differences, and DIF analyses were limited to male–female. The generalizability of the results of this study is limited by the choice of population (college students) and subject matter (mathematics). However, the results are consistent with those from previous studies in which the subject matter was psychology.

The DOMC format provides an easy way to increase item difficulty without changing the constructs measured. Additional approaches may be necessary to understand whether the DOMC and MC formats tap into different constructs. The analyses provided in this study do not shed light on whether the use of DOMC items can reduce cheating. Future research will look at this issue.

Authors' note

Neal M. Kingston, Department of Psychology and Research in Education, Center for Educational Testing and Evaluation, University of Kansas; Gail C. Tiemann, Center for Educational Testing and Evaluation, University of Kansas; Harold L. Miller, Jr., Department of Psychology, Brigham Young University; David Foster, Kryterion, Incorporated, Lindon, Utah.

This research was supported in part by the Center for Educational Testing and Evaluation; and the Brigham Young University Office of Undergraduate Education. The authors would like to thank Kryterion, Incorporated for use of the Webassessor™ testing software during this study. We also thank several graduate and undergraduate research assistants for their help in conducting the testing sessions.

References

Benjamin, L. T., Jr., Cavell, T. A., & Shallenberger, W. R., III (1984). Staying with initial answers on objective tests: Is it a myth? *Teaching of Psychology, 11*(3), 133-141.

Blackey, R. (2009). So many choices, so little time: Strategies for understanding and taking multiple-choice exams in history. *The History Teacher, 43*(1), 53-66.

Boyd, B. T. (2008). Effects of state tests on classroom test items in mathematics. *School Science and Mathematics, 108*(6), 251-262.

Cattell & Burdsal (1975). The radial parcel double factoring design: A solution to the item-vs-parcel controversy. *Multivariate Behavioral Research*. Vol 10(2), Apr 1975, pp. 165-179

Downing, S. M. (2006). Selected-response item formats in test development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 3-26). Mahwah, NJ: Erlbaum.

Foster, D., & Miller, H. (2009). A new format for multiple-choice testing: Discrete-option multiple-choice. Results from early studies. *Psychology Science Quarterly, 51*(4), 355-369.

Frey, B. B., Petersen, S., Edwards, L. M., Pedrotti, J. T., & Peyton, V. (2005). Item-writing rules: Collective wisdom. *Teaching and Teacher Education: An International Journal of Research and Studies, 21*(4), 357-364.

Geiger, M. A. (1997). An examination of the relationship between answer changing, testwiseness, and examination performance. *Journal of Experimental Education, 66*(1), 49-60.

Haertel, E H. (2006). Reliability. In Robert Brennan (Ed.), *Educational Measurement*. Westport, CT: Praeger Publishers.

Henryssen, S. (1971). Gathering, analyzing, and using data on test items. In R. L. Thorndike (ed.), *Educational measurement*, 2nd Ed, (pp. 130-159). Washington, DC: American Council on Education.

Kendall, J. S., & Marzano, R. J. (2000). *Content knowledge: A compendium of standards and benchmarks for K-12 education* (3rd ed.). Aurora, CO/Alexandria, VA: McREL/ Association for Supervision and Curriculum Development.

Nichols, S. L., & Berliner, D. C. (2007). *Collateral damage: How high-stakes testing corrupts America's schools*. Cambridge, MA: Harvard Education Press.

Nitko, A. J. (2004). *Educational assessment of students* (4th ed.). Upper Saddle River, NJ: Merrill.

Penfield, R. D. (2003). Applying the Breslow-Day test of trend in odds ratio heterogeneity to the analysis of nonuniform DIF. *Alberta Journal of Educational Research, 49*, 231-243.

Penfield, R. D. (2005). DIFAS: Differential item functioning analysis system. *Applied Psychological Measurement, 29*, 150-151.

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

Tanner, D. E. (2003). Multiple-choice items: Pariah, panacea, or neither of the above? *American Secondary Education, 31*(2), 27-36.

Taylor, M., & Smith, S. (2009). How do you know if they're getting it? Writing assessment items that reveal student understanding. *Science Scope, 32*(5), 60-64.

Zieky, M. (1993). Practical questions in the use of DIF statistics in test development. In P.W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 337-364). Hillsdale, NJ: Lawrence Erlbaum.