Article

SAGE Open
January-March 2018: 1–9
© The Author(s) 2018
DOI: 10.1177/2158244018758147
journals.sagepub.com/home/sgo
SAGE

The Impact of Homogeneity of Answer Choices on Item Difficulty and Discrimination

Erkan Hasan Atalmis[1] and Neal Martin Kingston[2]

Abstract

This study explored the impact of homogeneity of answer choices on item difficulty and discrimination. Twenty-two matched pairs of elementary and secondary mathematics items were administered to randomly equivalent samples of students. Each item pair comparison was treated as a separate study with the set of effect sizes analyzed using meta-analysis and a moderator analysis. The results show that multiple-choice (MC) items with homogeneous answer choices tend to be easier than MC items with nonhomogeneous answer choices, but the magnitude was related to item content (algebra vs. geometry) and answer choice construction strategy. For algebra items, items with homogeneous answer choices are easier than those with nonhomogeneous answer choices. However, the difficulty of geometry items with homogeneous and nonhomogeneous is not statistically different. Taking into account answer choice construction strategy, the findings showed that items with homogeneous answer choices were easier than items with nonhomogeneous answer choices when different strategy was applied. However, the same construction strategy was applied; thus, the difficulty of items with homogeneous answer choices and nonhomogeneous answer choices was not statistically different. In addition, we found that item discrimination does not significantly change across MC items with homogeneous and nonhomogeneous answer choices.

Keywords

multiple-choice item, item-writing guidelines, homogeneity of answer choices, test validity, meta-analysis

Introduction

If construct-irrelevant artifacts of the test-development process interfere with the validity of inferences made from test scores, then the entire testing enterprise is at risk. However, relatively few empirical studies have examined the impact of item-writing guidelines on test performance.

Despite recent enthusiasm for technology-enhanced items, test construction with multiple-choice (MC) items remains popular due to short administration times, scoring objectivity, and the ability to make sound decisions with test results (Haladyna, Downing, & Rodriguez, 2002; McCoubrie, 2004). However, creating an MC item can be challenging. Previous studies have suggested that forming answer choices is a difficult part of the item-construction process (Haladyna & Downing, 1989; Hansen & Dexter, 1997), because each individual answer choice can potentially influence item quality.

Haladyna and his colleagues (2002) proposed 31 item-writing guidelines for creating high-quality MC items that contribute to overall test reliability and validity. The authors examined three commonly cited guidelines related to constructing answer choices: No.23 "Keep choices homogeneous," No.29 "Make all distracters plausible," and No.30 "Use typical errors of students."

However, application of one guideline can lead to a violation of another. For instance, answer choices based on common student errors may include choices that are not consistent with one another. In other words, constructing an MC item with plausible answer choices may harm their homogeneity in content and grammatical structure. Specifically, answer choices for math items seems homogeneous if similar number types, number of digits, or operations are used among answer options. Table 1 shows two examples of math items with plausible but not homogeneous answer choices. Choice D of Item 1 includes words, while the other answer choices contain only equations. Choice A of Item 2 is the only integer number, while the other choices are fractions.

Writing items under the constraints of these three guidelines requires more effort and time. However, the impact of homogeneity of answer choices on item difficulty (proportion of correct

[1]Kahramanmaras Sutcu Imam University, Turkey
[2]University of Kansas, Lawrence, USA

Corresponding Author:
Erkan Hasan Atalmis, Kahramanmaras Sutcu Imam University, Avsar Campus, Kahramanmaras 46100, Turkey.
Email: eatalmis@ksu.edu.tr

Table 1. Items with Plausible Answer Choices That Are Not Homogeneous.

Item 1	Item 2
Which is true between $\frac{1}{2}$ and $\frac{2}{3}$?	Which is $\frac{3}{2}\times\frac{4}{3}$?
A. $\frac{1}{2}>\frac{2}{3}$	A. 2
B. $\frac{1}{2}<\frac{2}{3}$	B. $\frac{9}{8}$
C. $\frac{1}{2}=\frac{2}{3}$	C. $\frac{12}{5}$
D. $\frac{1}{2}$ and $\frac{2}{3}$ cannot be compared	D. $\frac{17}{6}$
• Common errors of students • Plausible distractors • Answer choices are not homogeneous (Choice D is not parallel to other answer choices)	• Common errors of students • Plausible distractors • Answer choices are not homogeneous (Choice A is not parallel to other answer choices)

answer) and discrimination (correlation with the total test score) remains uncertain. The current study addresses this question.

Haladyna and Rodriguez (2013) suggested that answer choices that are nonhomogeneous in content and grammar can provide a cue to students about the correct answer. Cues can make items easier and influence item discrimination indices because students take advantage of such cues to choose correct answers.

Answer choice homogeneity is defined in different ways for MC items based on written statements and MC items based on mathematical expressions. For example, language similarity, content, and grammatical consistency can make word-based answer choices homogeneous. On the contrary, answer choices in mathematics items can be made homogeneous by using a consistent number of digits (three vs. two), type of number (integer vs. fraction), and format of answer choices (words vs. numbers). Although these differences may seem small, they may be a source of construct-irrelevant variance, caused, perhaps, by test wiseness or alternatively a student's suspicion that the item writer is trying to be tricky. This study examines the impact of homogeneity of answer choices only for mathematics MC items.

The few empirical test-development studies that have focused on answer choice similarity have had mixed results. Smith and Smith (1988) tested the impact of the similarity between correct choice (key) and incorrect student response for reading test items using different standard-setting methods, such as Angoff and Nedelsky methods. They found that answer choice similarity makes items more difficult using both methods. However, they did not adequately define how to classify answer choice similarity. Ascalon, Meyers, Davis, and Smits (2007) found parallel results. They focused on only MC items on a driver's license examination and classified answer choice similarity by comparing all distractors in an item with the correct answer based on distractor content, such as theme, words, and sentence length. On the contrary, Green (1984) found the opposite result using general information items.

Unlike previous studies that explored the impact of answer choice homogeneity on item difficulty for the items with word-based answer choices, the current study examines the impact of answer choice homogeneity on item difficulty and item discrimination of mathematics items with numerical answer choices. More specifically, we examine three research questions:

Research Question 1: Are item difficulty and item discrimination for items with homogeneous answer choices statistically different from item difficulty for items with nonhomogeneous answer choices?

Research Question 2: Are item difficulty and item discrimination for items with homogeneous answer choices statistically different from item difficulty for items with nonhomogeneous answer choices when controlling item content (algebra vs. geometry)?

Research Question 3: Are item difficulty and item discrimination for items with homogeneous answer choices statistically different from item difficulty for items with nonhomogeneous answer choices when controlling answer choice construction strategy?

Method

In this study, we replicated previous studies that compared the difficulty and discrimination of items with homogeneous answer choices and items with nonhomogeneous answer choices for mathematics items. We have also analyzed the interaction of these factors with item content and answer choices construction strategy.

Test Items

We selected 22 pairs of MC mathematics items from a state test forms in the United States developed for Grades 3 to 8

Table 2. Items with Homogeneous and Nonhomogeneous Answer Choices.

	Homogeneous	Nonhomogeneous
Item Pair 1 (algebra)	What is 2,508 ÷ 4? A. 102 B. 127 C. 602 D. 627	What is 1,620 ÷ 9? A. 68 B. 102 C. 120 D. 180
Item Pair 2 (algebra)	Solve for x: $\frac{3}{4}(x-2)=6$ A. $x=6$ B. $x=6\frac{1}{2}$ C. $x=9\frac{2}{3}$ D. $x=10$	Solve for x: $\frac{6}{5}(x+5)=15$ A. $x=7\frac{1}{2}$ B. $x=8\frac{1}{3}$ C. $x=8\frac{4}{5}$ D. $x=10$
Item Pair 3 (geometry)	Millie plans to paint a picture on a rectangular piece of paper. She has a piece of paper that measures 13 inches (in) by 17 in. Exactly how many square inches (in^2) of paper does Millie have? A. 30 in^2 B. 60 in^2 C. 221 in^2 D. 442 in^2	Tasha is covering a rectangular bulletin board with paper. The bulletin board is 14 feet (ft) long and 4 ft wide. Exactly how much paper does Tasha need to completely cover the bulletin board? A. 18 ft^2 B. 36 ft^2 C. 56 ft^2 D. 112 ft^2

and high school. We selected the pairs by identifying items that were based on the same specific learning standard, which is educational objectives students should possess at critical point of any course, within algebra or geometry and that had either homogeneous or nonhomogeneous answer choices. The coding of homogeneity and nonhomogeneity was determined by three judges who are researchers experienced at item-writing based on state standards and taxonomy. Each MC item had four answer choices: the key and three distractors. Although the items in each pair were written to the same specific learning standards and the stems were set up similarly, the item content was not identical as can be seen in the item pair examples in Table 2.

For Item Pair 1, the items on the right and the left side measure the same specific learning standard, which is "dividing a four-digit numbers by two-digit number." The answer choices of the item on the left are homogeneous because all choices have three-digit numbers. Moreover, the final digits in the choices are either 02 or 27. This also allows readers make a set of two similar choices: (102, 127) and (602, 627). It does not harm dissimilarity among answer choices because only one choice which is exactly different from others is not selected by take takers during the exam to provide cue. On the contrary, the answer choices of the item on the right are nonhomogeneous because Choice A contains a different number of digits than the other answer choices. It means that Choice A is exactly different from other in a particular way and student might tend to select this option as cue. This influences psychometric properties of items.

For Item Pair 2, both items measure the same specific learning standard, which is "multiply each term inside the brackets (both algebraic term and number) by fraction outside the brackets." The answer choices of the item on the left are homogeneous because two of the choices are integers and the other two choices are mixed fractions; integers and mixed fractions are equally represented in the answer choices. On the contrary, the answer choices of the item on the right side are nonhomogeneous because Choice D is a whole integer, while the other choices contain fractions.

For Item Pair 3, both items measure the same specific learning standard, which is "calculate the area of rectangle in real word problems." The answer choices of the item on the left are homogeneous because two of the choices contain two digits, whereas the other two choices contain three digits. The answer choices have two equal sets of similar choices in terms of number of digits. On the contrary, the answer choices of the item on the right side are nonhomogeneous because Choice D contains three digits, while the other choices contain two digits.

After we selected the item pairs, three judges determined whether answer choices in item pairs were constructed using "the same" or "different" strategies. Considering item pairs in Table 1, Item Pair 3 was coded as "using the same strategy" because each answer choice of items in Item Pair 3 was constructed using the same strategy. It means that Choice A stems from adding length and width, Choice B from the formula for perimeter, Choice C from area, and Choice D from area times 2.

However, Item Pair 1 was coded as "using different strategy" because not all answer choices of items in Item Pair 1 were constructed using the same strategy. For example, Choice B of homogeneous item (the left side) and Choice A of nonhomogeneous item (the right side) stem from dividing the last three digits of dividend by divisor while Choice C of homogeneous item and Choice B of nonhomogeneous item stem from the combination of dividing the first two digits, the third digit, and the last digit of dividend by divisor. However, Choice A of homogeneous item stems from dividing each digit of dividend by divisor while Choice C of nonhomogeneous item stem from the combination of dividing the first two digits of dividend by divisor and last two digits of dividend.

Item Pair 2 also was coded as "using different strategy" because not all answer choices of items in Item Pair 2 were constructed using the same strategy. The items are related to the distributive property of multiplication over addition/subtraction. Only Choice C of homogeneous item and nonhomogeneous item was constructed using the same strategy, which stems from multiplying the first term inside the brackets by fraction outside the brackets and adding/subtracting the second term inside the brackets and/from the fraction. For other answer choices of homogeneous and nonhomogeneous items, they were constructed using different strategy. For homogeneous item, Choice A stems from making sign error after multiplying each term inside the brackets by the function. Choice B of homogeneous item stems from multiplying the number on right side of equation by the function rather than multiplying each term inside the brackets by the function. For nonhomogeneous item, Choice B stems from multiplying each term inside the brackets. Choice D stems from equating the terms inside brackets to the number on right side of equation.

After answer choices construction strategy of all item pairs was examined, we calculated the item difficulty and discrimination for each item within the pairs by using classical test theory (CTT) approach. Item difficulty is calculated as the proportion of examinees answering the item correctly while item discrimination refers to the ability of an item to discriminate between students with high scores and low scores (Thorndike, 2005). We applied item-total correlation index to calculate item discrimination for each item in this study, which is one of most widely used method (Downing, 2005).

Figure 1 shows item difficulty and item discrimination index values for each item pairs. A total of seven item pairs had geometry content, whereas 14 item pairs had algebra content. In terms of answer options construction strategy, five item pairs were coded as "using the same strategy" while 17 item pairs were coded as "using the different strategy" (see Figure 1).

Participants

The items in each pair were administered to randomly equivalent, overlapping, and nonoverlapping samples of students selected from those participating in a state accountability assessment. Field test items were embedded in operational test forms and thus students were not aware that these items were field test items and did not count toward their scores.

To check the similarity of the random samples of students receiving each item within a pair, we took all of the demographic characteristics of the samples into account. For every one of the 22 item pairs, the samples of students differed by no more than .01 in the proportion of disability of students, .02 in the proportion of male students, .01 in the proportion of students from any racial group, and .01 in the proportion of students qualified for a free or reduced-price lunch. In addition, the difficulty of each test form was the same because mean total scores of the test forms were not more than .8, whereas standard deviation in the mean total scores was not more than .06. Table 3 shows item characteristics of each pair.

Meta-Analysis

Meta-analysis is a statistical analysis integrating and summarizing the results of individual quantitative studies on a particular topic (Glass, 1976). Meta-analysis allows researchers to compute effect sizes to combine mean and standard deviation values, p values, or correlation coefficients proposed by studies (Kulik & Kulik, 1989; Sen & Akbas, 2016). Moreover, moderator analysis can be integrated into meta-analysis for more precise estimate after studies are grouped based on moderator variables, such as age, gender, and subject area (Alpaslan, Yalvac, & Willson, 2017).

There are two approaches to compute effect size in meta-analysis: fixed-effects model and random-effects model. Under the fixed model, the same effect size is calculated for all studies and weighted based on the number of observations by that study (Borenstein, Hedges, Higgins, & Rothstein, 2012). Under random-effects model, effect size varies from study to study due to demographic characteristics of samples, such as differences in education level, age, and socioeconomic status (Cooper, 2010).

Item difficulty. Each item pair comparison was treated as a separate study in a random-effects meta-analysis. Thus, the difference in sample characteristics and item content across item pairs was accounted for by the random-effects error term. The software package Comprehensive Meta-Analysis was used to perform the analyses (Borenstein, Hedges, Higgins, & Rothstein, 2005). Table 4 presents the results of fixed- and random-effects models, which are based on different assumptions. True effect was the same across all item pairs in the fixed-effects model, whereas it varied from one item pair to another in the random-effects model (Borenstein et al., 2012).

The variation between item pairs, heterogeneity, was statistically significant, $Q(21) = 3{,}420.99$, $p < .001$. The corresponding I^2 was 99.39, which means that 99% of observed variance

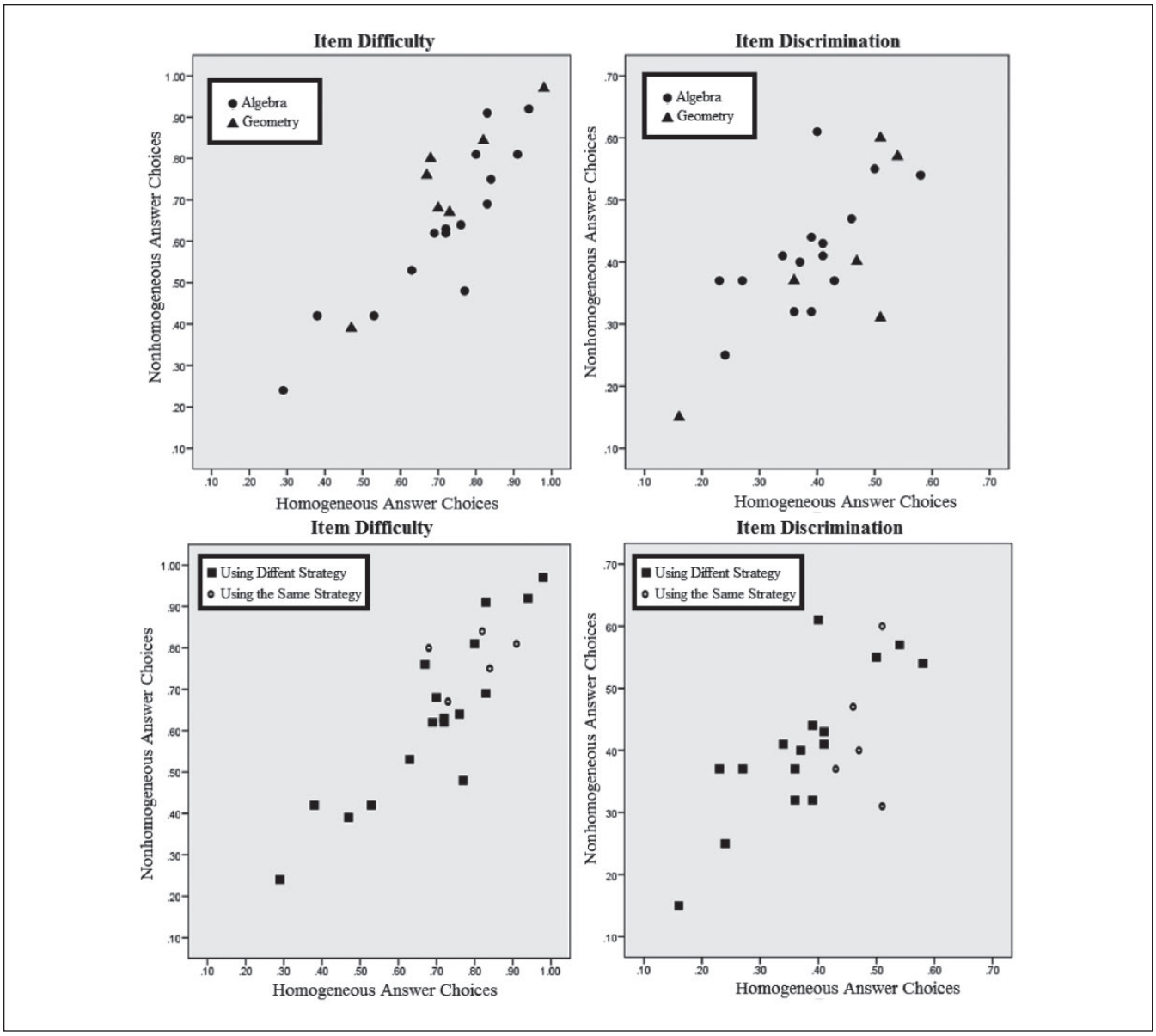

Figure 1. Item difficulty and discrimination for 22 item pairs.

in item pair effect sizes came from differences between item pairs that were not explained by sampling variability. The point estimates of the fixed-effects model and random-effects model in Table 4 show the difference in item difficulty between the items with homogeneous answer choices and the items with nonhomogeneous answer choices. Items with homogeneous answer choices were easier than items with nonhomogeneous answer choices with average effect sizes of .12 in fixed-effects model and .11 in random-effects model. The 95% confidence interval ranged from .11 to .13 in fixed-effects model and .02 to .20 in random-effects model.

We also conducted a mixed effects moderator analysis for the content area of the items (algebra vs. geometry) and answer choice construction strategy, as shown in Table 5.

According to Table 5, algebra items with homogeneous answer choices were easier than algebra items with nonhomogeneous answer choices, with an average effect size of .17, which is statistically significant ($z = 3.75$, $p < .001$). However, the difficulty of geometry items with homogeneous answer choices was not statistically different from the difficulty of geometry items with nonhomogeneous answer choices ($z = –0.31$, $p = .756$).

For items whose answer choices were constructed using different strategy in Table 5, items with homogeneous answer choices were easier than items with nonhomogeneous answer choices, with an average effect size of .12, which is statistically significant ($z = 2.46$, $p < .05$). However, for items whose answer choices constructed using the same

Table 3. Item Characteristics of Each Pair.

# of pair	Item type	Item characteristics		
		# of sample	Item difficulty	Item discrimination
1	Homogeneous	5,282	0.69	0.36
	Nonhomogeneous	5,266	0.62	0.32
2	Homogeneous	5,274	0.72	0.39
	Nonhomogeneous	5,266	0.62	0.32
3	Homogeneous	3,233	0.53	0.41
	Nonhomogeneous	3,205	0.42	0.41
4	Homogeneous	3,216	0.29	0.24
	Nonhomogeneous	3,217	0.24	0.25
5	Homogeneous	3,222	0.38	0.23
	Nonhomogeneous	3,195	0.42	0.37
6	Homogeneous	1,726	0.47	0.33
	Nonhomogeneous	1,705	0.39	0.3
7	Homogeneous	6,309	0.72	0.41
	Nonhomogeneous	6,262	0.63	0.43
8	Homogeneous	6,263	0.63	0.34
	Nonhomogeneous	6,263	0.53	0.41
9	Homogeneous	6,532	0.82	0.47
	Nonhomogeneous	6,532	0.84	0.4
10	Homogeneous	6,360	0.94	0.37
	Nonhomogeneous	6,360	0.92	0.4
11	Homogeneous	1,726	0.83	0.39
	Nonhomogeneous	1,726	0.91	0.44
12	Homogeneous	12,153	0.7	0.54
	Nonhomogeneous	12,153	0.68	0.57
13	Homogeneous	13,283	0.84	0.46
	Nonhomogeneous	13,283	0.75	0.47
14	Homogeneous	9,563	0.8	0.5
	Nonhomogeneous	9,563	0.81	0.55
15	Homogeneous	9,328	0.76	0.58
	Nonhomogeneous	9,328	0.64	0.54
16	Homogeneous	9,328	0.98	0.16
	Nonhomogeneous	9,328	0.97	0.15
17	Homogeneous	10,553	0.68	0.51
	Nonhomogeneous	10,553	0.8	0.31
18	Homogeneous	9,853	0.67	0.36
	Nonhomogeneous	9,853	0.76	0.37
19	Homogeneous	10,802	0.91	0.43
	Nonhomogeneous	10,802	0.81	0.37
20	Homogeneous	8,336	0.77	0.27
	Nonhomogeneous	8,336	0.48	0.37
21	Homogeneous	17,971	0.73	0.51
	Nonhomogeneous	17,971	0.67	0.6
22	Homogeneous	10,782	0.83	0.4
	Nonhomogeneous	10,782	0.69	0.61

strategy, the difficulty of items with homogeneous answer choices was not statistically different from the difficulty of items with nonhomogeneous answer choices (z = 0.65, p = .519).

Item discrimination. After we transformed the item discrimination indices for each of the 44 items from item-total correlations to Fisher z values, we generated fixed- and random-effects models for item discrimination values by using the software package Comprehensive Meta-Analysis (Borenstein et al., 2005). Table 6 shows the results of the fixed- and random-effects models.

The value of I^2 was 98.97, which means that 99% of the observed variance in item pair effect sizes came from differences between item pairs that were not explained by sampling variability. The variation between item pairs, heterogeneity, was statistically significant, $Q(21) = 2{,}033.11$, $p < .001$. The point estimate of the fixed-effects model and random-effects model showed that the difference in item discrimination between the items with homogeneous answer choices and the items with nonhomogeneous answer choices was not statistically different—fixed-effects model: $t(21) = -1.17$, $p = .22$; random-effects model: $t(21) = -0.08$, $p = .93$.

Results and Discussion

The purpose of this study was to explore the impact of the precise homogeneity of answer choices on item difficulty and discrimination of mathematics items. The findings showed that, overall, items with homogeneous answer choices were easier than items with nonhomogeneous answer choices, but the result depended on item content (algebra vs. geometry) and answer choice construction strategy. Algebra items with homogeneous answer choices were easier than algebra items with nonhomogeneous answer choices. However, the difference in difficulty of geometry items with homogeneous answer choices was not statistically significant than from the difficulty of geometry items with nonhomogeneous answer choices. Taking into consideration answer choice construction strategy, the findings showed that items with homogeneous answer choices were easier than items with nonhomogeneous answer choices when different strategy was used to construct answer choices of items with homogeneous and nonhomogeneous answer choices. On the contrary, when the answer choices of items were constructed using the same construction strategy, the difficulty of items with homogeneous answer choices and nonhomogeneous answer choices was not statistically different. Moreover, the very large I^2 statistic indicates that even when considering algebra versus geometry items, the source of most of the variation in difficulty across homogeneity conditions was undetermined. Also the impact of homogeneity of answer choices on discrimination was not statistically significant.

The results of this study provide empirical support to the growing body of test-development studies. Specifically, this study contributes to research on the impact of answer choice homogeneity on item psychometric characteristics. First, unlike past studies that focused on word-based items, the current study examined the impact of homogeneity of answer choices on psychometric characteristics of

Table 4. Fixed Effect and Random Effect of the Model for Item Difficulty.

	Effect size and 95% confidence interval				Test of null		Heterogeneity			
Model	# of studies	*M*	Lower limit	Upper limit	*z* value	*p* value	*Q*	*df* (*Q*)	Significance level	I^2
Fixed	22	0.12	0.11	0.13	35.33	.00	3,420.99	21	.00	99.39
Random	22	0.11	0.02	0.20	2.51	.01				

Table 5. Moderator Analysis for Content Area and Answer Choice Construction Strategy (Mixed-Effect Analysis).

		Effect size and 95% confidence interval				
	n	*M*	Lower limit	Upper limit	*z*	Significance level
Content area						
Algebra	15	0.17	0.08	0.26	3.75	.000
Geometry	7	−0.02	−0.14	0.11	−0.31	.756
Total	22	0.11	0.03	0.18	2.86	.004
Answer choice construction strategy						
Different strategy	17	0.12	0.03	0.22	2.46	.014
The same strategy	5	0.06	−0.13	0.26	0.65	.519
Total	22	0.11	0.02	0.2	2.48	.013

Table 6. Fixed Effect and Random Effect of the Model for Item Discrimination.

	Effect size and 95% confidence interval				Test of null		Heterogeneity			
Model	# of studies	*M*	Lower limit	Upper limit	*z* value	*p* value	*Q* value	*df* (*Q*)	Significance level	I^2
Fixed	22	0.00	−0.01	0.00	−1.17	.22	2,033.11	21.00	.00	98.97
Random	22	0.00	−0.07	0.06	−0.08	.93				

mathematics items. With regard to effect in item difficulty, our overall findings are consistent with Green (1984), yet inconsistent with Smith and Smith (1988) and Ascalon et al. (2007), in that on average the items with homogeneous answer choices were easier. However, when answer choices were constructed using the same strategy, it was found that the difficulty of items with homogeneous answer choices was not statistically different from the difficulty of items with nonhomogeneous answer choices, which is not consistent with the findings of Green (1984), Smith and Smith (1988), and Ascalon et al. (2007). One hypothesis that might explain these results is that the impact of option homogeneity depends on item content, because the discrepant results came from studies using reading test items, driver license items, and general information items. The common characteristic of such word-based items for examinees is that the probability of choosing correct answer using answer choices may be far higher compared with mathematics items because examinees may choose correct answer for a mathematics item only if they know the solution regardless of answer choices. Existing literature related to number of options on item psychometric properties may be considered to support this claim. For example, Rodriguez (2005) conducted a meta-analysis to determine the impact of on item difficulty and discrimination by examining 56 empirical studies, most of which contained social sciences and language arts items which are word-based items. The findings showed that items become easier when number of options decreased. However, a recent study conducted by Atalmış and Kingston (2017) designed MC mathematics items with four options and three options constructed using the same strategy and found that item difficulty was not statistically different across the items with four options and three options. These findings supported only if options of mathematics items were constructed using the same strategy, item difficulty may not be changed by homogeneity of answer choices or number of the options.

Another contribution of this study is that the current study examined the impact of answer choice homogeneity not only on item difficulty but also on item discrimination which differs from previous studies. The findings showed that item discrimination was not statistically influenced by mathematics items with homogeneous answer choices and nonhomogeneous answer choices even if different item contents and answer choice construction strategies were applied. A hypothesis that might explain these results is that items with homogeneous answer choices and nonhomogeneous answer

were constructed applying the same specific learning standards and the stems. Thus, when parallel mathematics item pairs are constructed, learning standards and stem similarity may be considered as higher priority than item content and answer choose construction strategy to keep item discrimination similarity between parallel items.

The last contribution of this study is the use of meta-analysis in unconventional contexts. Although meta-analysis is usually used for integrating and summarizing the results of individual quantitative studies on a particular topic, in this study, each item pair comparison was considered as a separate study to examine the impact of answer option homogeneity on item psychometric properties.

The results of this empirical study also apply to item writers and classroom teachers. Writing MC items with plausible distractors based on common student errors is always suggested (Haladyna et al., 2002). However, creating plausible distractors may affect answer choice homogeneity. Although each approach has merits, creating items with both plausible and homogeneous answer choices poses a challenge for item writers and classroom teachers. That is, item writers might need to spend significant extra time to construct fourth or fifth option which is plausible and homogeneous, and thus construct fewer items in a given amount of time. The results of the current study are informative for test creators on the usage of such approaches while designing a test.

Given the results of this study combines with the existing literature on item development, we recommend that the use of plausible answer options be given a higher priority than the use of homogeneous answer options in constructing MC mathematics items, but that both guidelines be considered.

There are some limitations of this study. First, there were only 22 pairs of items; thus, the number of items per grade level and the variety of types of nonhomogeneity were limited. Second, this study was conducted for only mathematics items. Most importantly, most of the variation in difficulty differences between the items with homogeneous and nonhomogeneous answer choices was not explained. We recommend additional exploration of item content and answer options construction strategy as moderator variables.

Authors' Note

An earlier version of this article was presented at the National Council on Measurement in Education (NCME) in Philadelphia, PA, USA, in 2014.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

Alpaslan, M. M., Yalvac, B., & Willson, V. (2017). A meta analytical review of the relationship between personal epistemology and self-regulated learning. *Turkish Journal of Education*, *6*, 48-67.

Ascalon, M. E., Meyers, L. S., Davis, B. W., & Smits, N. (2007). Distractor similarity and item-stem structure: Effects on item difficulty. *Applied Measurement in Education*, *20*, 153-170. doi:10.1080/08957340701301272

Atalmış, E. H., & Kingston, N. M. (2017). Three, four, and none of the above options in multiple-choice items. *Turkish Journal of Education*, *6*, 142-157. doi:10.19128/turje.333687

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. (2005). Comprehensive meta-analysis (Version 2.2.048) [Software]. Englewood, NJ: Biostat.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. (2012). *Introduction to meta-analysis*. Chichester, UK: John Wiley.

Cooper, H. (2010). *Research synthesis and meta-analysis*. Thousand Oaks, CA: SAGE.

Downing, S. M. (2005). The effects of violating standard item writing principles on tests and students: The consequences of using flawed test items on achievement examinations in medical education. *Advances in Health Sciences Education*, *10*(2), 133-143. doi:10.1007/s10459-004-4019-5

Glass, G. V. (1976). Primary, secondary and meta-analysis of research. *Educational Researcher*, *5*(10), 3-8.

Green, K. (1984). Effects of item characteristics on multiple-choice item difficulty. *Educational and Psychological Measurement*, *44*, 551-561. doi:10.1177/0013164484443002

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

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*, 309-334. doi:10.1207/S15324818AME1503_5

Haladyna, T. M., & Rodriguez, M. C. (2013). *Developing and validating test items*. New York, NY: Routledge.

Hansen, J. D., & Dexter, L. (1997). Quality multiple-choice test questions: Item-writing guidelines and an analysis of auditing test banks. *The Journal of Education for Business*, *73*, 94-97.

Kulik, J. A., & Kulik, C. L. C. (1989). The concept of meta-analysis. *International Journal of Educational Research*, *13*, 227-340.

McCoubrie, P. (2004). Improving the fairness of multiple-choice questions: A literature review. *Medical Teacher*, *26*, 709-712. doi:10.1080/01421590400013495

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

Sen, S., & Akbas, N. (2016). A study on multilevel meta-analysis methods. *Journal of Measurement and Evaluation in Education and Psychology*, *7*(1), 1-17.

Smith, R. L., & Smith, J. K. (1988). Differential use of item information by judges using Angoff and Nedelsky procedures. *Journal of Educational Measurement*, *25*, 259-274. doi:10.1111/j.1745-3984.1988.tb00307.x

Thorndike, R. M. (2005). *Measurement and evaluation in psychology and education* (7th ed.). Upper Saddle River, NJ: Pearson.

Author Biographies

Erkan Hasan Atalmis is an assistant professor at the Kahramanmaras Sutcu Imam University in Turkey. He earned PhD degree in Research, Evaluation, Measurement, and Statistics Program from University of Kansas. His research interest includes item and test development, type of grading system in educational assessment, and supplementary education.

Neal Kingston, PhD, came to the University of Kansas in 2006 and is a professor in the Research, Evaluation, Measurement, and Statistics Program and director of the Achievement and Assessment Institute. His research focuses on large-scale assessment, with particular emphasis on how it can better support student learning. He is the principal investigator/director or co-principal investigator of several large research projects, including Design and Development of a Dynamic Learning Maps Alternate Assessment, Kansas Assessment Program, Career Pathways Assessment System, and Development and Validation of Online Adaptive Reading Motivation Measures.