Psicología Educativa (2019) 25(1) 43-48

Psicología Educativa

http://journals.copmadrid.org/psed

Differential Length and Overlap with the Stem in Multiple-Choice Item Options: A Pilot Experiment

Giulia Casu[a] and Carmen García-García[b]

[a]University of Bologna, Italy; [b]Autonoma University of Madrid, Spain

ARTICLE INFO

Article history:
Received 21 July 2018
Accepted 22 October 2018

Keywords:
Item-writing guidelines
Multiple-choice items
Flawed items
Length of options
Lexical overlap
Item difficulty

ABSTRACT

Multiple-choice items are extensively used across different assessment contexts. A crucial requirement for ensuring their validity is their correct development, and a number of item-writing guidelines have been proposed that support item developers. This experimental pilot study aimed to investigate the effect of violating two item-writing guidelines: the differential length of the correct option compared to distractors and its lexical overlap with the stem. Standard and flawed items, respectively adhering to and deviating from guidelines, were randomly assigned to 55 college students and compared in their psychometric functioning. Results indicated that, in general, flawed items tended to be easier and less subject to random answers than standard ones, but significant differences were few. Discrepancies between standard and flawed subtests approached statistical significance with medium effect sizes. Although of interest, findings must be cautiously interpreted due to the small sample size. Implications for future research are discussed.

La longitud diferencial y el solapamiento con el enunciado en las opciones de ítems de opción múltiple: un experimento piloto

Palabras clave:
Directrices sobre redacción de ítems
Ítems de elección múltiple
Items con incumplimientos
Longitud de las opciones
Solapamiento léxico
Dificultad de los ítems

RESUMEN

Los ítems de opción múltiple son ampliamente utilizados en contextos de evaluación muy variados. Un requisito muy importante para garantizar su validez es su correcta redacción, y para ayudar a conseguirlo se han desarrollado una serie de directrices. El objetivo de este estudio piloto experimental fue investigar el efecto del incumplimiento de dos de estas reglas, más concretamente, la longitud diferencial de la opción correcta comparada con los distractores y su solapamiento léxico con el enunciado. Para ello, se asignó aleatoriamente a 55 estudiantes a las condiciones de responder a ítems que respetaban o que incumplían las mencionadas directrices y se compararon las propiedades psicométricas conseguidas por los ítems. Los resultados indican que, en general, los ítems con incumplimientos tendían a ser más fáciles y a recibir menos respuestas aleatorias; no obstante, había pocas diferencias significativas y el tamaño del efecto era medio. Aunque de interés, estos resultados deben ser interpretados con cautela debido al escaso tamaño muestral. Se comentan las implicaciones para futuras investigaciones.

The wide use of multiple-choice (MC) items across different evaluation contexts highlights the importance of their correct development and usage. With the objective of enhancing the validity of scores obtained from MC items, fundamental guidelines for MC item construction have been developed by different authors. Haladyna and Downing (1989a) settled the basis for MC item-writing by analyzing 46 textbooks and other sources, and proposed 43 consensual guidelines. The same authors also reviewed more than 90 studies to explore the validity of their recommendations and found that more than half of the guidelines had not been investigated at all (Haladyna & Downing, 1989b). In a replication of the latter review, Haladyna, Downing, and Rodriguez (2002) validated and reduced the original taxonomy of 43 item-writing rules to 31 guidelines, which have recently been reorganized and updated (Haladyna & Rodriguez, 2013). Other taxonomies for developing MC items were developed by Frey, Petersen, Edwards, Teramoto Pedrotti, and Peyton (2005) and Moreno, Martínez, and Muñiz (2006), which basically comprised the same advice as Haladyna et al.'s (2002). The latest pieces of advice for

Cite this article as: Casu, G. & García-García, C. (2019). Differential length and overlap with the stem in multiple-choice item options: A pilot experiment. *Psicología Educativa, 25,* 43-48. https://doi.org/10.5093/psed2018a20

Correspondence: giulia.casu3@unibo.it (G. Casu).

developing MC items were proposed by Moreno, Martínez, and Muñiz (2015), who drew up previous guidelines based on validity criteria, resulting in 9 general guidelines that summarize and subsume the previous ones by the same authors (Moreno et al., 2006) and by Haladyna et al. (2002).

Despite the availability of multiple guidelines, flawed MC items are commonly applied. For instance, MC items are frequently used that contain either no correct option or more than one correct option, excessive text in the stem, "all of the above" and "none of the above" options, distractors that appear poorly plausible or contain clues to the correct answer (Rodriguez, 1997). Violations of basic MC item-writing principles are also common in college entrance tests and exams (e.g., Atalmis, 2016; García, Ponsoda, & Sierra, 2011; Hijji, 2017). The importance of following the standard MC item-writing principles lies in that the usage of flawed MC questions negatively affects both tests and students (Haladyna & Rodriguez, 2013; Omer, Abdulrahim, & Albalawi, 2016; Pate & Caldwell, 2014), and introduces construct irrelevant variance (CIV) in the evaluation process (Downing, 2002, 2005). Indeed, CIV harms the evidence of validity of the assessment by interfering with the meaningful and exact interpretation of scores and negatively affecting the pass rate on the exam (Downing, 2002; Haladyna & Downing, 2004).

Authors agree that the standard MC item-writing guidelines have been built on consensus among item-writing experts rather than on the results of empirical studies. Thus, there is a need for further studies that finally validate or refute the proposed MC item-writing guidelines, and examine their impact on different psychometric indices (e.g., Haladyna & Downing, 1989a; Haladyna et al., 2002; Moreno et al., 2015). Specifically, Downing (2002) underlined the need for experimental studies in which standard and flawed (i.e., manipulated) items designed for assessing performance on a single domain are randomly assigned to examinees. However, to the authors' knowledge, research of this kind is still extremely rare (e.g., Caldwell & Pate, 2013).

Objective and Hypotheses

Following the claim by Downing (2002), this pilot study aimed to investigate whether the violation of two different guidelines for writing the options affected the psychometric functioning of MC items. The following MC item-writing guidelines by Haladyna and Rodriguez (2013) were considered: "Keep the length of options about equal" (guideline 20a) and "Avoid clang associations, options identical to or resembling words in the stem" (guideline 20c).

As to guideline 20a, all item-writing authors agree in that item options must be homogeneous in length (Albano & Rodriguez, 2018; Gierl, Bulut, Guo, & Zhang, 2017). However, a common mistake is that the correct option is longer than distractors (Omer et al., 2016; Rodriguez, 1997). Reviewing the MC item tests developed by college teachers in four different countries, Carter (1986) found that at least one item had a longer correct option in 86% of tests. Results of nonexperimental studies on the effects of violating this guideline are nonetheless still inconclusive. In investigating the possibility of predicting the difficulty of the Test of English as a Foreign Language (TOEFL), Freedle and Kostin (1993) found that the length of the incorrect options was negatively related to the difficulty index (i.e., proportion of subjects correctly answering an item), suggesting a detrimental effect of a greater number of words in the distractors. Similarly, a meta-analitic study by Rodriguez (1997) found that a longer correct option made items easier. Nevertheless, more recently, Martínez, Moreno, Martín, and Trigo (2009) found no difference in difficulty between standard items and items with differential length of one option compared to the rest, or between items with differential length in correct vs. incorrect options.

As to guideline 20c, previous observational studies analyzing the difficulty of TOEFL listening and reading comprehension items agree in that the lexical overlap with the stimulus or key text sentence makes the item easier when it occurs in the correct option, but makes the item more difficult when it occurs in a distractor (Freedle & Fellbaum, 1987; Freedle & Kostin, 1993, 1996; Ying-Hui, 2006).

These guidelines were selected for the present study as they have been mentioned among the most effective ones in providing cues to examinees (Morse, 1998). As suggested by Moreno et al. (2015), the extra attention paid on an option that stands out from the others for its length or wording might overlap with the response a person would give had such difference not existed. Moreover, compared to other item-writing rules, these are more suitable to an objective operationalization. The distinctness of an option relative to the others because of its length or overlap with elements in the stem can indeed be easily expressed in terms of number of words. Specifically, we explored whether a correct option that was either longer (guideline 20a) or more highly overlapped with the stem (guideline 20c), compared to distractors, affected item difficulty, defined as the proportion of examinees correctly answering an item, and proportion of random answers reported by examinees. Reported random answers were used as an index of perceived item difficulty. Assuming that random answers are given when a task is perceived to be difficult, perceived item difficulty was expected to be negatively, at least moderately associated with actual item difficulty (Bratfisch, Dornič, & Borg, 1972; Hambleton & Jirka, 2006; Wolf, Smith, & Birbaum, 1995).

Based on the above, the following experimental hypotheses were formulated. For guideline 20a, we hypothesized that: (1) items would be actually easier (i.e., have higher difficulty indices) when the correct option was longer than the distractors than when all options had approximately the same length; (2) items would be perceived as easier (i.e., the percentage of random answers reported by examinees would be lower) when the correct option was longer than the distractors than when all options had approximately the same length; and (3) items would be actually easier and perceived as easier (i.e., have higher difficulty indices and a lower percentage of random answers reported by examinees) as the visibility of the differential length (defined as a higher difference in the number of characters between the correct option and the longest distractor) increased. For guideline 20c, we hylothesized that: (1) items would be actually easier when the lexical overlap with the stem was higher in the correct option than when all options were approximately equally overlapped with the stem; (2) items would be perceived as easier when the lexical overlap with the stem was higher in the correct option than when all options were approximately equally overlapped with the stem; and (3) items would be actually easier and perceived as easier as the visibility of the lexical overlap between the correct option and the stem (defined as a higher difference in the number of words lexically overlapped with the stem between the correct option and the most overlapped distractor) increased.

Method

Participants

Fifty-five (65.5% females) Psychology students with a mean age of 23.7 years (SD = 6.8, range 19-52 years) participated in the study by completing one of four forms of a college MC test in basic Psychometrics. Forms A and B were completed by 15 (27.3%) and 14 (25.5%) examinees, respectively, whereas both Forms C and D were responded by 13 (23.6%) students.

Instruments

The questionnaire was composed by 18 MC items developed by methodology experts to assess student performance in basic

Psychometrics. Each item consisted of a question- or sentence-completion stem, followed by 3 vertically formatted options, only one of which was correct. Two versions of the questionnaire were designed. In the standard version, all items adhered to MC item-writing guidelines and showed adequate psychometric functioning as indicated by previous applications to large samples. In the flawed, manipulated version, half items contained violations of guideline 20a and the other half of guideline 20c. Violation of guideline 20a was defined as a surplus of 5 or more uninformative words (i.e., words with no information content, such as conjunctions, prepositions, pronouns, etc.) in the correct option compared to the longest distractor and was introduced in item # 1 to # 9. Violation of guideline 20c was defined as a surplus of 2 or more words lexically overlapped with the stem in the correct option compared to the most overlapped distractor and was introduced in items # 10 to # 18. Two independent methodology experts modified the standard items to introduce the guideline violation and ensured that each manipulated item included violation of one item-writing guideline only.

Design and Procedure

Four 18-item test forms were created to balance the levels (i.e., standard vs. flawed) of the independent variables (i.e., guideline 20a vs. guideline 20c) and control for order of presentation. Four items with violations of guideline 20a (items # 2, 3, 5, and 8) and four items with violations of guideline 20c (items # 11, 13, 16, and 17) were randomly assigned to Form A; the remaining manipulated items (items # 1, 4, 6, 7, and 9 for guideline 20a, and items # 10, 12, 14, 15, and 18 for guideline 20c) were assigned to Form B (Table 1). Form C and Form D contained the same items as Form A and Form B, respectively, but the order of presentation of the two halves of the questionnaire was reversed.

Subtests were composed by standard or flawed items, depending on the test form received. For example, as to guideline 20a, subtest 1 was formed by items # 2, 3, 5, and 8, that were flawed in forms A and C and standard in forms B and D; subtest 2 was composed by items # 1, 4, 6, 7, and 9 that were flawed in forms B and D and standard in forms A and C. For guideline 20c, subtest 3 included items # 11, 13, 16, and 17 that were flawed in forms A an C and standard in form B and D; finally, subtest 4 comprised items # 10, 12, 14, 15, and 18 that were flawed in forms B and D and standard in form A and C. Test forms were randomly assigned to students, ensuring that an approximately equal number of examinees received each form.

The questionnaire was applied to students during the last class of the Psychometrics course. Examinees were reassured that participation was voluntary and the questionnaire anonymous, and informed that the test score would neither be corrected for errors nor would in any way affect their Psychometrics course assessment mark. For each item, examinees were told that they could mark a cross in case of random answering or being insecure about the correct answer, as this would have helped the examiners to identify topics in need of reinforcement within the course program. MC item development was not a topic of the study program.

Table 1. Experimental Design

Form	Guideline 20a	Guideline 20c
A	2, 3, 5, 8 (subtest 1)	11, 13, 16, 17 (subtest 3)
B	1, 4, 6, 7, 9 (subtest 2)	10, 12, 14, 15, 18 (subtest 4)

Note. Flawed items included in each subtest are shown.

Data Analysis

To ensure the absence of order effects on examinees' performance, one-way analysis of variance (ANOVA) was performed to compare test scores between test forms.

Actual difficulty (i.e., proportion of correct answers) and perceived difficulty (i.e., percentage of random answers reported by examinees) were computed for each item, and their relationship was examined using bivariate correlations. The standard and flawed versions of each item were compared in their actual and perceived difficulty by performing a *z*-test.

For each subject, the proportion of correct answers and reported random answers in four different subtests was calculated. Mean proportion of correct answers and mean percentage of reported random answers in each subtest were compared between the two experimental conditions (i.e., standard vs. flawed) using ANOVA.

Finally, linear regression analyses were performed to investigate the influence of the visibility of the guideline violation on actual and perceived difficulty. For guideline 20a, the visibility of the violation was measured as the difference in the number of characters between the correct option and the longest distractor. For guideline 20c, the visibility of the violation was calculated as the difference in the number of words lexically overlapped with the stem between the correct option and the most overlapped distractor.

A power analysis indicated that, with α = .05 (two-tailed), at least 52 cases were needed to reach enough power (.80) to detect a large effect size. Evaluation of estimates was based on both statistical significance (significance level set at $p \leq .05$) and effect-size measures, Cohen's *d* of 0.20 being considered small, 0.50 medium, and 0.80 large, and R^2 and η^2 of .01 being considered small, .09 medium, and .25 large (Cohen, 1988). Statistical analyses were performed with IBM SPSS 22 (SPSS Inc., Chicago, IL). Power analysis was performed with G*Power 3.1 (Faul, Erdfelder, Lang, & Buchner, 2007).

Results

Order Effects

No significant difference was found between test forms in total test scores, indicating the absence of order effects, $F(3, 51) = 1.43$, $p = .25$, $\eta^2 = .08$ (Table 2). The higher mean score in Form B might be attributable to the lower mean percentage of reported random answers in items that make up this form (Table 3 and Table 4).

Table 2. Mean Scores across Test Forms

Form	*n*	*M* (*SD*)
A	15	9.20 (2.88)
B	14	11.29 (2.81)
C	13	9.62 (3.57)
D	13	9.15 (3.29)

Guideline 20a

Actual and perceived difficulty were moderately, although non-significantly, negatively correlated ($r = -.41$, $p = .10$).

As shown in Table 3, the mean (actual) difficulty index was slightly higher for items with differential length of the correct option than for items with approximately same length options. As to individual items, six out of nine items were easier in their flawed version; however, the proportion of correct answers was significantly, moderately higher in the flawed than in the standard version for three items only.

The mean proportion of random answers reported by examinees was almost equal between standard and flawed items. As to individual items, there was no clear pattern between item manipulation and perceived item difficulty.

For both subtests 1 and 2, results of ANOVAs indicated that the difference in mean percentage of correct answers between the standard and flawed versions approached statistical significance

Table 3. Item Descriptive Statistics by Experimental Condition for Guideline 20a

	Actual difficulty[1]						Perceived difficulty[2]			
	Standard		Flawed				Standard	Flawed		
Item	*n*		*n*		*z*	*r*			*z*	*r*
1	27	.48	26	.81	-2.51*	.34	63%	35%	2.04*	.27
2	27	.85	28	.82	0.30	.04	11%	25%	-1.35	.18
3	27	.56	28	.54	0.15	.02	30%	43%	-1.00	.14
4	28	.61	26	.50	0.81	.11	68%	73%	-0.40	.05
5	26	.35	28	.46	-0.82	.11	42%	71%	-2.15*	.29
6	27	.14	27	.37	-1.94*	.26	50%	41%	0.66	.09
7	28	.50	26	.62	-0.89	.12	46%	19%	2.11*	.28
8	27	.22	28	.64	-3.14*	.42	37%	29%	0.63	.09
9	28	.46	27	.52	-0.45	.06	29%	37%	-0.63	.09
M (*SD*)		.46 (.21)		.59 (.15)			42% (0.18)	41% (0.19)		

Note. [1]Proportion of correct answers. [2]Percentage of reported random answers.
*$p \leq .05$.

(Table 5). In both cases, the flawed version tended to have a significantly higher percentage of correct answers, compared to the standard one, with a medium effect size.

Finally, the visibility of the violation had no significant effects on actual difficulty, although the strength of the association was moderate, $\beta = .33$, $R^2 = .11$, $F(1, 16) = 1.89$, $p = .19$, or on perceived difficulty, $\beta = -.02$, adjusted $R^2 = .00$, $F(1, 16) = .010$, $p = .93$.

Guideline 20c

Actual and perceived difficulty were moderately, although non-significantly, negatively correlated ($r = -.30$, $p = .23$).

Items in which the lexical overlap with the stem was higher in the correct option compared to distractors had a slightly higher mean (actual) difficulty index than standard items, in which all options were approximately equally overlapped with the stem (Table 4). As to individual items, seven out of nine items were easier in their flawed version; nevertheless, differences in item difficulty were negligible, and reached statistical significance in one item only, with a medium effect size.

The mean proportion of reported random answers was slightly higher in standard items. Nevertheless, there was no clear relation between item manipulation and perceived item difficulty.

For subtest 4, results of ANOVAs indicated that the flawed version tended to have a significantly higher percentage of correct answers, compared to the standard one, with this difference approaching statistical significance and being moderate in magnitude (Table 5). No significant effect of guideline manipulation was instead found for subtest 3.

Finally, the visibility of the violation had no significant effects on actual difficulty, although the association was moderate in strength, $\beta = .41$, $R^2 = .17$, $F(1, 16) = 3.16$, $p = .09$. Instead, a higher difference in the number of words lexically overlapped with the stem between the correct option and the most overlapped distractor was associated with a significantly lower perceived difficulty, with a large effect size, $\beta = -.50$, $R^2 = .25$, $F(1, 16) = 5.19$, $p = .04$.

Discussion

This pilot study aimed to investigate whether the violation of two guidelines by Haladyna and Rodriguez (2013) for writing MC item options affected item psychometric characteristics. Actual difficulty (i.e., proportion of examinees correctly answering the item) and perceived difficulty (defined as the percentage of reported random answers) of a set of standard items (i.e., with no violations of any item-writing guideline) were compared to those of a flawed version of the same items (i.e., with violation of one of the two guidelines). The following guidelines were considered: "Keep the length of options about equal" (guideline 20a), and "Avoid clang associations, options identical to or resembling words in the stem" (guideline 20c). For guideline 20a, flawed items had differential length of the correct option, defined as a surplus of 5 or more words relative to the longest distractor. As to guideline 20c, lexical overlap between the correct option and the item stem was introduced in flawed items, and

Table 4. Item Descriptive Statistics by Experimental Condition for Guideline 20c

	Actual difficulty[1]						Perceived difficulty[2]			
	Standard		Flawed				Standard	Flawed		
Item	*n*		*n*		*z*	*r*			*z*	*r*
10	28	.50	27	.70	-1.51	.20	43%	11%	2.66*	.36
11	27	.48	28	.54	-0.45	.06	26%	25%	0.09	.01
12	28	.50	27	.52	-0.15	.02	46%	33%	0.99	.13
13	27	.59	28	.43	1.19	.16	22%	39%	-1.37	.18
14	28	.32	27	.70	-2.82*	.38	50%	52%	-0.15	.02
15	28	.64	27	.74	-0.80	.11	39%	33%	0.46	.06
16	27	.48	28	.50	-0.15	.02	59%	54%	0.37	.05
17	27	.78	28	.75	0.26	.04	33%	46%	-0.99	.13
18	28	.57	26	.58	-0.07	.01	43%	31%	0.91	.12
M (*SD*)		.54 (.13)		.61 (.12)			40% (.12)	36% (.14)		

Note. [1]Proportion of correct answers. [2]Percentage of reported random answers.
*$p \leq .05$.

Table 5. Subtest Comparisons for Proportion of Correct Answers

Guideline 20a						
	Standard		Flawed			
Subtest	*n*	*M* (*SD*)	*n*	*M* (*SD*)	*F*(1, 53)	*d*
1 (items # 2, 3, 5, 8)	27	.49 (.23)	28	.62 (.26)	3.93, $p = .053$ $\eta^2 = .07$	0.54
2 (items # 1, 4, 6, 7, 9)	28	.44 (.20)	27	.55 (.22)	3.84, $p = .055$ $\eta^2 = .07$	0.53
Guideline 20c						
	Standard		Flawed			
Subtest	*n*	*M* (*SD*)	*n*	*M* (*SD*)	*F*(1, 53)	
3 (items # 11, 13, 16, 17)	27	.58 (.29)	28	.55 (.30)	0.14, $p = .71$ $\eta^2 = .003$	0.10
4 (items # 10, 12, 14, 15, 18)	28	.51 (.26)	27	.64 (.25)	3.92, $p = .053$ $\eta^2 = .07$	0.52

defined as a surplus of 2 or more words lexically overlapped with the stem compared to the most overlapped distractor.

Results indicated that, for guideline 20a, the proportion of correct answers was in general higher for flawed than for standard items. However, this difference was statistically significant for three out of nine items only, with medium effect sizes. When considering comparable subtests, the proportion of correct answers was higher in subtests with differential length of the correct option, with these differences approaching statistical significance and being moderate in magnitude. Thus, altogether, flawed items tended to be easier than standard ones, as hypothesized. It must be nonetheless acknowledged that the differences in actual difficulty between standard and flawed items were not systematic. This partly contradicts meta-analytic findings by Rodriguez (1997), who concluded that a higher length of the correct option relative to distractors makes items easier, but is in line with more recent findings by Martinez et al. (2009), who found no differences between standard and flawed items. However, it must be noted that Martinez et al. (2009) included in their analyses also distractors with differential length compared to the rest of options. Most of all, these studies did not consider standard and flawed versions of a same item, which makes comparisons with previous evidence difficult to carry out.

With respect to guideline 20c, the mean proportion of correct answers was slightly higher for flawed items compared to standard ones, suggesting that flawed items tended to be easier. Nevertheless, differences in actual item difficulty between standard and flawed items were minor, and reached the statistical significance in one out of nine items only, with a medium effect size. In addition, for this guideline one subtest only showed an almost significantly higher proportion of correct answers in its flawed version than in its standard version. Thus, we found no strong evidence in support of initial hypotheses that items containing lexical overlap between the correct option and the stem would be easier than the corresponding standard items. This contradicts previous findings that the lexical overlap with the stem makes the item easier when it is in the correct option (Freedle & Fellbaum, 1987; Freedle & Kostin, 1993, 1996; Ying-Hui, 2006). Again, it must be nonetheless noted that these studies did not have an experimental design, which makes comparisons with our findings problematic.

With respect to perceived item difficulty, for both guidelines this index was negatively, moderately correlated with actual item difficulty, as expected. However, the mean proportion of random answers reported by examinees was almost equal between standard and flawed items for guideline 20a, while it was slightly lower for flawed items in case of violation of guideline 20c. In general, no clear pattern between item manipulation and perceived item difficulty could be observed for either of the two guidelines, which does not support initial hypotheses.

Finally, we xamined whether the visibility of the guideline violation was associated with item psychometric characteristics. For guideline 20a, the visibility of the differential length of the correct option compared to distractors was unrelated to actual and perceived item difficulty, in contrast with hypotheses. For guideline 20c, a greater visibility of the lexical overlap between the correct option and the item stem was associated with a lower percentage of random answers reported by examinees. This was partly in line with hypotheses and in line with findings that the visibility of guideline 20c violation was a significant predictor of a lower item difficulty (Freedle & Kostin, 1993, 1996; Ying-Hui, 2006).

To our knowledge, this pilot study was one of the first to use an experimental design to test the validity of MC item-writing guidelines. Altogether, findings suggest a tendency for MC items violating guidelines 20a and 20c to be easier and less subject to random answers than items with no violation of any guideline. Nevertheless, it must be noted that significant individual differences between standard and flawed items were only few in number, and discrepancies between flawed and standard subtests only approached statistical significance. This might be attributable to the small sample size, which limited statistical power (post-hoc achieved power was .44 in the present study), as the estimated effect sizes were medium according to Cohen's criteria (Cohen, 1988). On the other hand, the true difference in psychometric functioning between standard and flawed items might be a null one, which needs to be addressed in future studies using larger samples.

The limited sample size also precluded a comparison of standard and flawed items in their ability to discriminate between higher- and low-performing students, as each test form was completed by too few examinees. In addition to the small sample size, other limitations of this pilot study include that our findings refer to the performance of college students in a very specific domain. Moreover, examinees were informed that test score would not influence their final mark, which may have affected their motivation and performance, with potential consequences on results.

In conclusion, although of interest, results of the present pilot study must be cautiously interpreted. Further studies on larger samples are required to definitely assess whether the violation of the examined guidelines is associated with lower item difficulty and less random answers, and to explore the effects of these violations on item discrimination. Future research should also investigate how the application of flawed items contributes to variance irrelevant to the construct being measured. Indeed, test-wiseness is an individual's ability to take advantage of test characteristics and format that is unrelated to his/her knowledge or ability level; thus, examinees' experience with MC items or their differential skill in taking tests might represent important sources of variance in test scores (Milman, Bishop, & Ebel, 1965; Papenberg,

Willing, & Musch, 2017; Thorndike, 1951). Another factor potentially affecting MC test performance is an examinee's cognitive style, especially in case of flawed MC items (Armstrong, 1993). Studies specifically designed to examine the effect of the above mentioned variables are therefore encouraged to increase our understanding of how examinees interface with flawed items.

Conflict of Interest

The authors of this article declare no conflict of interest.

References

Albano, A. D., & Rodriguez, M. C. (2018). Item development research and practice. In S. Elliott, R. Kettler, P. Beddow, & A. Kurtz (Eds.), *Handbook of accessible instruction and testing practices* (2nd ed.) (pp. 181-198). Cham, Switzerland: Springer.

Armstrong, A. M. (1993). Cognitive-style differences in testing situations. *Educational Measurement: Issues and Practice, 12*(3), 17-22. https://doi.org/10.1111/j.1745-3992.1993.tb00538.x

Atalmis, E. H. (2016). Do the guideline violations influence test difficulty of high-stake test? An investigation on university entrance examination in Turkey. *Journal of Education and Training Studies, 4*(10), 1-7.https://doi.org/10.11114/jets.v4i10.1738

Bratfisch , O., Dornič , S., & Borg , C. (1972). *Perceived difficulty of items in a test of reasoning ability*. Institute of Applied Psychology, University of Stockholm. Stockholm.

Caldwell, D. J., & Pate, A. N. (2013). Effects of question formats on student and item performance. *American Journal of Pharmaceutical Education, 77*(4), 71. https://doi.org/10.5688/ajpe77471

Carter, K. (1986). Test-wiseness for teachers and students. *Educational Measurement: Issues and Practice, 5*(4), 20-23. https://doi.org/10.1111/j.1745-3992.1986.tb00495.x

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Downing, S. M. (2002). Construct-irrelevant variance and flawed test questions: Do multiple-choice item writing principles make any difference? *Academic Medicine, 77*(Suppl. 10), S103–S104. https://doi.org/10.1097/00001888-200210001-00032

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*, 133-143. https://doi.org/10.1007/s10459-004-4019-5

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral and biomedical sciences. *Behavior Research Methods, 39*, 175-191. https://doi.org/10.3758/BF03193146

Freedle, R., & Fellbaum, C. (1987). An exploratory study of the relative difficulty of TOEFL'S listening comprehension items. In R. Freedle & R. Duran (Eds.), *Cognitive and linguistic analyses of test performance* (pp. 162-192). Norwood, NJ: Ab1ex.

Freedle, R. O., & Kostin, I. W. (1993). *The prediction of TOEFL reading comprehension item difficulty for expository prose passages for three item types - main idea, inference, and supporting idea items* (TOEFL Research Report RR-93-13). Princeton, NJ: Educational Testing Service.

Freedle, R., & Kostin, I. (1996). *The prediction of TOEFL listening comprehension item difficulty for minitalk passages: Implications for construct validity* (TOEFL Research Report RR-96-29). Princeton, NJ: Educational Testing Service.

Frey, B. B., Petersen, S., Edwards, L. M., Teramoto Pedrotti, J., & Peyton, V. (2005). Item-writing rules: Collective wisdom. *Teaching and Teacher Education, 21*, 357-364. https://doi.org/10.1016/j.tate.2005.01.008

García, C., Ponsoda, V., & Sierra, A. (2011). Prediction of item psychometric indices from item characteristics automatically extracted from the stem and option text. *International Journal of Continuing Engineering Education and Life-Long Learning, 21*, 210-221. https://doi.org/10.1504/IJCEELL.2011.040199

Gierl, M. J., Bulut, O., Guo, Q., & Zhang, X. (2017). Developing, analyzing, and using distractors for multiple-choice tests in education: A comprehensive review. *Review of Educational Research, 87*, 1082-1116. https://doi.org/10.3102/0034654317726529

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

Haladyna, T. M., & Downing, S. M. (1989b). The 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. (2004). Construct-irrelevant variance in highstakes testing. *Educational Measurement: Issues and Practice, 23*(1), 17-27. https://doi.org/10.1111/j.1745-3992.2004.tb00149.x

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines. *Applied Measurement in Education, 15*, 309-334. https://doi.org/10.1207/S15324818AME1503_5

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

Hambleton, R. K., & Jirka, S. J. (2006). Anchor-based methods for judgmentally estimating item statistics. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 399-420). Mahwah, NJ: Erlbaum.

Hijji, B. M. (2017). Flaws of multiple choice questions in teacher-constructed nursing examinations: A pilot descriptive study. *Journal of Nursing Education, 56*, 490-496. https://doi.org/10.3928/01484834-20170712-08

Martínez, R. J., Moreno, R., Martín, I., & Trigo, M. E. (2009). Evaluation of five guidelines for option development in multiple-choice item-writing. *Psicothema, 21*, 326-330.

Milman, J., Bishop, C., & Ebel, R. (1965). An analysis of test-wiseness. *Educational and Psychological Measurement, 25*, 707-726. https://doi.org/10.1177/001316446502500304

Moreno, R., Martínez, R. J., & Muñiz, J. (2006). New guidelines for developing multiple-choice items. *Methodology, 2*, 65-72. https://doi.org/10.1027/1614-2241.2.2.65

Moreno, R., Martínez, R. J., & Muñiz, J. (2015). Guidelines based on validity criteria for the development of multiple choice items. *Psicothema, 27*, 388-394. https://doi.org/10.7334/psicothema2015.110

Morse, D. T. (1998). The relative difficulty of selected test-wiseness skills among college students. *Educational and Psychological Measurement, 58*, 399-408. https://doi.org/10.1177/0013164498058003003

Omer, A. A., Abdulrahim, M. E., & Albalawi, I. A. (2016). Flawed multiple-choice questions put on the scale: What is their impact on students' achievement in a final undergraduate surgical examination? *Journal of Health Specialties, 4*, 270-275. https://doi.org/10.4103/2468-6360.191908

Papenberg, M., Willing, S., & Musch, J. (2017). Sequentially presented response options prevent the use of testwiseness cues in multiple-choice testing. *Psychological Test and Assessment Modeling, 59*, 245-266.

Pate, A., & Caldwell, D. J. (2014). Effects of multiple-choice item-writing guideline utilization on item and student performance. *Currents in Pharmacy Teaching and Learning, 6*, 130-134. https://doi.org/10.1016/j.cptl.2013.09.003

Rodriguez, M. C. (1997, April). *The art & science of item writing: A meta-analysis of multiple choice item format effects*. Paper presented at the Annual Meeting of the American Educational Research Association. Chicago, IL. Abstract retrieved from http://edmeasurement.net/research/artandscience.pdf

Thorndike, R. (1951). Reliability. In E. F. Lindquist (Ed.), *Educational measurement* (pp. 560-620). Washington, DC: American Council on Education.

Wolf, L. F., Smith, J. K., & Birnbaum, M. E. (1995). Consequence of performance, test, motivation, and mentally taxing items. *Applied Measurement in Education, 8*, 341-351. https://doi.org/10.1207/s15324818ame0804_4

Ying-Hui, H. (2006). An investigation into the task features affecting EFL listening comprehension test performance. *Asian EFL Journal Quarterly, 8*(2), 33-54.