Journal of Education, Teaching and Social Studies
ISSN 2642-2336 (Print) ISSN 2642-2328 (Online)
Vol. 2, No. 4, 2020
www.scholink.org/ojs/index.php/jetss

Original Paper

The Effective Use of Negative Stems and “All of the Above” in Multiple-Choice Tests in College Courses

Michael Joseph Wise[1*]

[1] Department of Biology, Roanoke College, 221 College Lane, Salem, Virginia 24153, USA

[*] Michael Joseph Wise, E-mail: wise@roanoke.edu

ORCID: https://orcid.org/0000-0003-0091-303X

Received: September 8, 2020 Accepted: September 28, 2020 Online Published: October 2, 2020

doi:10.22158/jetss.v2n4p47 URL: http://dx.doi.org/10.22158/jetss.v2n4p47

Abstract

Over the past few decades, test-writing experts have converged on a set of best-practice guidelines for constructing multiple-choice (MC) items. Despite broad acceptance, some guidelines are supported by scant or inconsistent empirical evidence. This study focused on two of the most-commonly violated of these guidelines: the use of negatively oriented stems (e.g., those using the qualifiers “not” or “except”) and the use of “all of the above” (AOTA) as a response option. Specifically, I analysed the psychometric qualities of 545 MC items from science courses that I taught at a liberal arts college. In this dataset, items with negatively oriented stems did not differ in difficulty or discriminability from questions with positively oriented stems. Similarly, items with AOTA as a response option did not differ in difficulty or discriminability from those without AOTA as an option. Items that used AOTA as a distractor were significantly more difficult, and slightly more discriminating, than were items that used AOTA as the key. Although they must be written with extra attention to detail, this study suggests that MC items with negative stems or AOTA as a response option can be effectively employed for assessment of content mastery in a classroom setting.

Keywords

all of the above, difficulty, discriminability, multiple-choice items, negatively worded stems

1. Introduction

The multiple-choice question (i.e., MC item) is a valuable tool for testing students’ mastery of constructs in an efficient, reliable, and objective manner. Although MC items had their origin over a century ago (Kelly, 1916; Chapman & Toops, 1919), only in recent years have researchers and pedagogists begun to converge on a set of best-practice guidelines for the writing of MC items (Aiken, 1987; Haladyna & Downing, 1989a, 1989b; Haladyna, 1997; Frey, Petersen, Edwards, Pedrotti, & Peyton, 2005; Moreno, Martínez, & Muñiz, 2006; Haladyna & Rodriguez, 2013; Towns, 2014; Rodriguez & Albano, 2017). Although many of these guidelines are commonsense rules of thumb, the rationale for other guidelines is more contentious, and these guidelines continue to be refined by new experimental and observational studies. The current study focuses on two categories of MC items for which opinions are mixed: 1) items with negatively worded stems, and 2) items with "all of the above" as a response option. Although these two types of items are popular with many test-writers, others condemn them as among the most insidious of item-writing flaws (Tarrant, Knierim, Hayes, & Ware, 2006; Abdulghani et al., 2015; Ibbett & Wheldon, 2016; Rush, Rankin, & White, 2016).

The standard format for a multiple-choice (MC) item is a short question or statement (the "stem"), followed by a list of several response options, including one correct option (the "key") and multiple incorrect "distractors." A negatively worded stem in an MC item contains such qualifier words as "not," "false," or "except," as in, "Which of the following is *not*… ?" The main criticism against negative stems is that they increase the complexity of the cognitive task of interpreting the item, independent of the students' knowledge of the concept being tested by the item (Dudycha & Carpenter, 1973; Cassels & Johnstone, 1984; Violato & Harasym, 1987; Harasym, Price, Brant, Violato, & Lorscheider, 1992; Tamir, 1993; Tarrant et al., 2006). A student rushed for time may overlook the negation in the stem and may select the first response option that he or she recognizes to be a true statement—not realizing that the statement of truth is a distractor, and the key to the question is a false statement. This phenomenon would cause items with negative stems to be more difficult than items testing the same construct with positively worded stems. Moreover, because students with mastery of the construct being tested will be at risk of misreading the negative stem, such items would be expected to have lower discriminability than items with positive stems (Chiavaroli, 2017).

The primary concern over the use of an "all of the above" (AOTA) response option is that such items are prone to cueing that enables test-wise students with incomplete mastery to answer the items correctly (Haladyna & Rodriguez, 2013; Ibbett & Wheldon, 2016; Rodriguez & Albano, 2017). For example, consider a four-response MC item with AOTA (the fourth option) as the key. A student with partial knowledge of the construct may recognize that two of the options are correct, and thus know that AOTA must be the key, even if he or she has no knowledge of the content in the third response option. Alternatively, a student may recognize that one of the response options is incorrect, and thus he or she can eliminate AOTA from consideration as well. Another problem arises due to the tendency of many test-writers to include AOTA as a response option predominantly as the key (Hansen & Lee, 1997; Poundstone, 2014). If students know that AOTA is rarely wrong, then they can answer items correctly by guessing AOTA, even without any knowledge of the construct being tested by the item. The flip side of this problem is when test-writers include AOTA as a response option because it is simpler than devising another plausible distractor. Test-wise students may recognize that AOTA is just an afterthought, and they can focus only on the other response options. The upshot of each of these potential pitfalls is that MC items with AOTA as a response option would tend to be less difficult. Moreover, susceptibility to guessing would make items with AOTA, either as a distractor or as a key, less discriminating, which would lead to lower reliability of MC tests (Tollefson, 1987; Moncada & Moncada, 2010).

Given these potential pitfalls, why do so many test-writers still include MC items with negative stems and AOTA response options? One reason is that writing plausible distractors is a difficult, time-consuming task (Aiken, 1987; Hansen & Lee, 1997; Chéron, Ademi, Kraft, & Löffler-Stastka, 2016; Chiavaroli, 2017). Consider that for each five-response item, a test-writer needs to devise four responses that will sound reasonable to students who have incomplete knowledge of the construct. However, if the item is written with a negative stem, then the test-writer has to come up with only one distractor that is a reasonable-sounding false statement. The rest of the options will be true statements, which take less creativity and effort to devise. If a test-writer is having difficulty coming up with even a single plausible false statement about a construct, then using AOTA as the key automatically solves that problem. Finally, if a test-writer feels a need to add just one more distractor to an item, then AOTA can always be put at the bottom of the list of response options.

Instead of the easy route of blaming the shortcomings of test-writers, one should also consider the possibility that some constructs can be tested more effectively with questions that have more true options than false options. Such constructs may be more common in some fields of study than others, and the disciplines of biology offer many examples. Consider, for instance, MC items that address the following constructs: elements that make up a macromolecule, organs that make up an organ system, symptoms in common to a disease, taxonomic classes within a single phylum, species that share a type of biome, etc. Given that items with negative stems or with AOTA as a response option arise organically to test-makers, it does not make sense to avoid using these strategies outright. Whether these are appropriate variations of MC items—as opposed to flaws in item writing—is an empirical question that can best be addressed by analyses from a diversity of testing situations.

The goal of the current study was to analyze MC items written for and employed in classroom tests—rather than items that were artificially altered to test specific hypotheses—as the former are more likely to represent wording and formatting choices that a teacher would actually use in a test. Specifically, I calculated difficulties and discrimination indices for 545 MC items that I have used in college-level biology and environmental science classes. These parameters were used to address the main concerns that have led researchers to regard negative stems and AOTA options as writing flaws:

1) Are MC items with negative stems more difficult than those with positive stems? 2) Are MC items with negative stems less discriminating than those with positive stems? 3) Are MC items with AOTA as a response option less difficult than those without AOTA? 4) Are MC items with AOTA as a response option less discriminating than those without AOTA? In addition, I asked whether the difficulty or discriminability differed between items in which AOTA was the key and items in which AOTA was a distractor.

2. Methods

2.1 The Data Set

The MC items analyzed for this study were written for tests in six biology and environmental science courses that I taught at Roanoke College between 2012 and 2019 (Table 1). The dataset consists of 545 five-response MC items from 14 tests, each of which was taken by 23-51 students (Table 1). The stems of 111 MC items were negatively worded, and 434 were positively worded. The negating word ("not", "except", or "false") was emphasized by boldface, italics, and/or underlines in 100 of these 111 items. Fifty of the MC items included AOTA as a response option—10 as the key and 40 as a distractor.

Table 1. Courses Included in this Study, with the Number of Students Taking Each Exam, and the Total Number of Multiple-choice Items on the Exams per Course

Course	Year	Students	MC items	Exams included
BIOL 120: Principles of Biology	2013	48	42	Final
BIOL 125: Biodiversity	2015	23-25	74	Four hourly
BIOL 180: Exploring Biological Diversity	2018-9	31-35	151	Five hourly & final
BIOL 205: General Ecology	2012	34	57	Final
BIOL 205: General Ecology	2015	51 & 44	72	Midterm & final
INQ 251: Bugs in the System	2016	23	50	Final
ENSC 101: Intro to Environmental Science	2019	33-34	173	Two hourly & final

2.2 Psychometric Characteristics

I calculated two fundamental psychometric variables for each MC item: difficulty and discriminability. The difficulty of a test item is traditionally quantified as the proportion (p) of test-takers who answered the item correctly. That is, the lower the value of p, the greater the difficulty.

I used the point-biserial method to quantify a discrimination index (*DI*) for each item. The *DI* is equivalent to a Pearson product-moment correlation between whether each student answered the question correctly (1) or incorrectly (0) and the student's total score on all the MC items of a test. A sufficiently discriminating item will have a *DI* greater than 0.2, while an item with a *DI* greater than 0.4 is considered to be excellent (Towns, 2014; Chiavaroli, 2017).

2.3 Statistical Analysis

To assess the influence of the stem orientation (negative or positive) and the presence of AOTA as a response option on the difficult and discriminability of the MC items, I ran two analyses of variance (ANOVAs)—one with difficulty (p) as the response variable and one with *DI* as the response variable. Both ANOVAs included two explanatory factors, which were treated as dummy (i.e., indicator) variables: 1) negative stem orientation (or not), and 2) inclusion of AOTA (or not). To compare the functioning of the AOTA response option when used as a key versus a distractor, I performed two t-tests—one with p and one with *DI* as the response variable. Finally, to assess whether highlighting the negating word in a negative stem affected an item's functioning, I performed two additional t-tests—one with *p* and one with *DI* as the response variable. All statistical analyses in this study were performed using JMP$_{IN}$ v. 4.0.4 (SAS Institute Inc., Cary, North Carolina, USA).

3. Results

3.1 Item Difficulty

The mean (±1 SD) difficulty of the 545 MC items was 0.67 ±0.22. The orientation of the stems had no effect on the items' difficulties ($F_{1,542}$ = 0.0018, *P* = 0.96; Table 2A), as the mean difficulty (±1 SEM) for items with positive and negative stems was 0.67 ±0.01 and 0.66 ±0.02, respectively (Figure 1A). In the 111 items with negative stems, highlighting the negative term did not significantly affect the difficulty (t_{109} = 1.263; *P* = 0.21), though the trend was that those with the negative term highlighted were actually more difficult (*p* = 0.65 ±0.01; N = 100) than those without highlighting (*p* = 0.74 ±0.07; N = 11).

Table 2. Summary of Results of Analyses of Variance (ANOVAs) of the Effects of Test Item Type on the Difficulty and Discrimination Index of the Items

Source of variation	Df	Mean square	*F*-ratio	*P*-value
A. Difficulty				
Stem orientation	1	0.000086	0.0018	0.97
AOTA	1	0.068501	1.4019	0.24
Error	542	0.048862		
B. Discrimination index				
Stem orientation	1	0.031189	0.8347	0.36
AOTA	1	0.020220	0.5411	0.46
Error	542	0.037365		

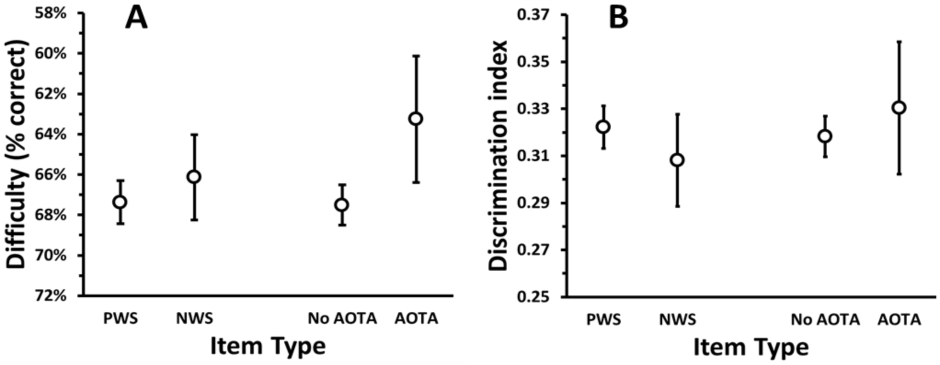

Figure 1. Comparisons of A. Difficulty (% Correct) and B. Discriminability (*DI*) of Different Formats of Multiple-choice Items. Points and Bars Represent Means ±1 SEM. PWS = Positively Worded Stem; NWS = Negatively Worded Stem; AOTA = All of the Above

Including "all of the above" (AOTA) as a response item did not significantly affect the difficulty of the items ($F_{1,542}$ = 1.4019; P = 0.24; Table 2A), as the mean difficulties (± 1 SEM) for items with AOTA and without AOTA were 0.68 ± 0.03 and 0.63 ± 0.01, respectively (Figure 1A). Among the 50 items with AOTA as a response option, the items were significantly easier (t_{48} = 1.263; P = 0.04; Figure 2A) when AOTA was used a key (p = 0.76 ± 0.07; N = 10) than when AOTA was a distractor (p = 0.60 ± 0.03; N = 40).

3.2 Item Discriminability

The mean (± 1 SD) discrimination index of the 545 MC items was 0.32 ± 0.19. The orientation of the stems had no effect on the discriminability ($F_{1,542}$ = 0.8347; P = 0.36; Table 2B), as the mean *DI* (± 1 SEM) for items with positive and negative stems was 0.32 ± 0.01 and 0.31 ± 0.02, respectively (Figure 1B). As with item difficulty, highlighting the negative term did not significantly affect the discriminability (t_{109} = 0.830; P = 0.41). Even so, items with the negating word highlighted showed a non-significant trend in the direction of being more discriminating (*DI* = 0.31 ± 0.02; N = 100) than those without highlighting (*DI* = 0.26 ± 0.06; N = 11).

Including AOTA as a response item did not significantly affect the discriminability of the items ($F_{1,542}$ = 0.5411; P = 0.46; Table 2B), as the mean *DI's* (± 1 SEM) for items with AOTA and without AOTA were 0.33 ± 0.03 and 0.32 ± 0.01, respectively (Figure 1B). Among the 50 items with AOTA as a response option, the items were slightly, but not significantly, more discriminating (t_{48} = 0.952; P = 0.35) when AOTA was used as a distractor (*DI* = 0.34 ± 0.03; N = 40) than when AOTA was the key (*DI* = 0.28 ± 0.06; N = 10; Figure 2B).

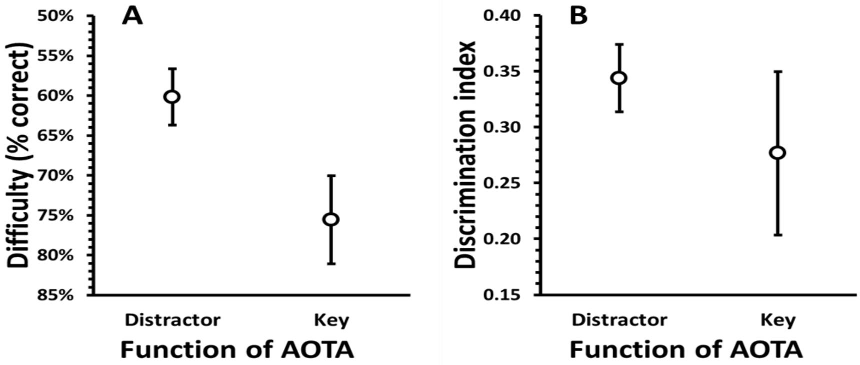

Figure 2. Comparisons of A. Difficulty (% Correct) and B. Discriminability (*DI*) of Multiple-choice Items when All of the Above (AOTA) is Included as a Distractor or as the Key. Points and Bars Represent Means ± 1 SEM

4. Discussion

4.1 Negatively Worded Stems

Although opposition to the use of negative stems in MC items is not universal, the majority of authors who have expressed an opinion on the orientation of stems suggest that test-writers should avoid negative orientations (Hansen & Lee, 1997; Haladyna, 2004; Towns, 2014; Rush et al., 2016; Chiavaroli, 2017). The main concern is that a more cumbersome logical structure of negative stems will inflate the difficulty of the items in a way that decouples the understanding of the intent of the item from a student's mastery of the construct being tested (Haladyna & Rodriguez, 2013; Rodriguez & Albano, 2017). As a consequence, MC items with negative stems would be less discriminating than those with positive stems. The results of the current study, however, provide no evidence to justify these concerns. In fact, the mean difficulties of the items with negative and positive stems were nearly identical (at 0.67 and 0.66, respectively). Moreover, items with negative stems were almost exactly the same in discriminability as items with positive stems (mean *DI* = 0.32 and 0.33, respectively). In no practical way, therefore, did the MC items with negative stems show symptoms of being flawed.

The current study's findings are consistent with the relatively small number of other empirical studies that have addressed the potential impact of stem orientation on the psychometric characteristics of MC questions (Violato & Harasym, 1987; Haladyna, Downing, & Rodriguez, 2002). While a minority of empirical evidence suggests that MC items with negative stems can be slightly more difficult than those with positive stems (Dudycha & Carpenter, 1973; Cassels & Johnstone, 1984; Haladyna & Downing, 1989b; Rodriguez, 1997), there appear to be no published empirical studies showing that the discrimination index differed between items with negatively and positively oriented stems.

Finally, it has often been suggested that highlighting the negative phrasing in the stem can act as an aid to reduce the chance that students will misread an item (Hansen & Lee, 1997; Haladyna et al., 2002; Towns, 2014). This suggestion is certainly logical, but there is a dearth of evidence to support the assumption that it is helpful. Surprisingly, in the current study, students gave incorrect answers to the items with highlighted negating words 9% more frequently than they did when the negating word was not highlighted. Perhaps students were actually more likely to skip over the negating word when it was written in a different style than the rest of the words in the stem. Nevertheless, the discrimination index was 20% greater when the negating word was highlighted. It should be kept in mind that the study was not designed to test the impact of this highlighting, and thus its power to make statistical inferences regarding highlighting is low. Because my goal in writing the items was to make them as clear as possible, I ended up highlighting the negating word in 90% of the items with negative stems. Perhaps I was subconsciously more likely to neglect to highlight the negating word when the item seemed particularly simple. Whatever the cause, the surprising trend in this study argues for additional studies on the practice of highlighting the negating terms in MC items with negative stems.

4.2 All of the Above

Some critics have claimed that using AOTA (or none of the above) as a response option can increase the difficulty of an MC item due to the relatively more complicated wording needed to set up a list of response options that are all correct (or all incorrect) (Frey et al., 2005; Rush et al., 2016). More commonly, critics worry that using AOTA will decrease the difficulty due to cues that will increase the susceptibility of MC items to guessing by test-wise students (Hansen & Lee, 1997; Haladyna &

Rodriguez, 2013; Ibbett & Wheldon, 2016; Rodriguez & Albano, 2017). Either of these effects would also be detrimental to the discriminability of MC items. In the current study, items with AOTA as a response option were answered correctly 4.2% more frequently than were items without AOTA. Although this difference in difficulty was not statistically significant, it is consistent with the expectation of AOTA questions being susceptible to cueing. These items were significantly less difficult when AOTA was the key rather than a distractor. That result suggests that students with partial knowledge of the construct (e.g., recognizing that two of the response options are correct) are able to conclude that AOTA must be the key.

Importantly, any decrease in the difficulty of items with an AOTA option did not translate to a difference in discriminability in this study: MC items with AOTA had a mean *DI* of 0.33, and items without AOTA had a mean of 0.32. Items in which AOTA was a distractor rather than the key were slightly, but not significantly, more discriminating (*DI* = 0.34 and 0.28, respectively). Despite the amount of criticism about using AOTA in MC items, there are surprisingly few studies that have found the presence of AOTA to have a negative effect on discriminability (Dudycha & Carpenter, 1973; Mueller, 1975). Along with the bulk of previous evidence, the results of this current study argue for the utility of using AOTA as a response option in MC items. Certainly, the current study provides no evidence that the use of AOTA should be considered a flaw.

In order to minimize guessing by test-wise students, the frequency with which AOTA or NOTA (none of the above) should be the key (rather than a distractor) is 1/*k*, where *k* is the number of response options per item (Frary, 1991; DiBattista, Sinnige-Egger, & Fortuna, 2014). That is, students should not have reason to believe that AOTA has a greater or lesser chance of being the key than any other response option. Adherence to these ideal frequencies in MC tests is apparently rather rare (Mueller, 1975; Rich & Johanson, 1990; Hansen & Lee, 1997). For instance, in a review of 100 multiple-choice tests from a variety of sources, Poundstone (2014) found that AOTA (or NOTA) was the key in 52% of the items in which it appeared as a response option. It is no wonder that test-takers are often advised that, when in doubt, choose AOTA (Rush et al., 2016).

For a test composed of five-response MC items, the effect of guessing will be minimized if AOTA is the key 20% of the time and a distractor 80% of the time in which it is used as a response option. Of the 50 items that included AOTA in the current study, AOTA was the key 10 times. This perfect match with the optimal frequency was a happy coincidence, as these 50 items were spread across 14 different tests over a span of eight years. It is unlikely that any students would have been exposed to this exact percentage of AOTA keys in the particular subset of tests that they took. Nevertheless, the attention to not overusing or underusing AOTA as a key likely contributed to the AOTA items being less susceptible to guessing in the current study.

5. Conclusions and Limitations

One factor that likely contributed to the high discriminabilities of the MC items with AOTA (as well as the items with negative stems) is that all of the items in this study were written with the goal of creating the most effective way to test mastery of a construct covered in a course. The type of construct and the material studied by the students dictated the format of the stem and responses. In contrast, in controlled experimental studies of the use AOTA response options, pairs of items are written to test the ostensibly identical construct with and without the use of AOTA (Dudycha & Carpenter, 1973). Likewise, experimental studies on stem orientation use pairs of items to test an ostensibly identical construct using a positively worded and a negatively worded stem. Although the goals of objectivity and statistical rigor in such studies are laudable, it is impossible in practice to write equivalent items both negatively and positively, or both with and without AOTA (particularly if AOTA is the key). The response options will necessarily differ between the items of a pair. In addition, the act of rewriting items for an experimental test may itself lead to writing flaws, such as awkward constructions and obviously incorrect distractors, that decrease the discriminability of the items (Rich & Johanson, 1990; Frary, 1991; DiBattista et al., 2014). Therefore, the AOTA and negative-stemmed items analyzed in controlled experiments are often not representative of the items that would be written by a teacher on a test of classroom mastery. In that sense, the items used in the current study, though it is observational and unbalanced, result in a more authentic comparison of different forms of MC items (Owen & Froman, 1987; Tarrant, Ware, & Mohammed, 2009; Funk & Dickson, 2011).

The MC items in the current study originated mainly in biology courses. Many biological constructs are conveniently organized in hierarchies (e.g., taxonomic classes nested within phyla). This hierarchical nature may make it easier to come up with more correct responses than plausible incorrect responses in MC items—a fact that would naturally lead to writing a negatively oriented stem or including AOTA as a response option. Hierarchical constructs are certainly not restricted to biology, however, or even to scientific fields in general. I would expect that negative stems and AOTA can be used just as effectively in organically derived MC items in other science or humanities courses as they were in the biology courses included in this study.

Finally, I echo the cautions of other authors who suggest that any violation of common MC-item-writing guidelines should be done only with caution. In particular, constructing an MC item with a negative stem requires more attention to detail to avoid confusing or ambiguous wording. Response options that create double negatives should be strictly avoided. When writing an item with AOTA, extra care must be taken to avoid cueing that makes the correct answer too obvious. In addition, one should take care not to overuse AOTA as either the key or as a distractor: Its frequency as a key should be inversely proportional to the number of response options contained in the MC items on a test. In short, MC items that use negative stems or AOTA as a response option may be more susceptible to writing problems that lower the discriminability of items and thus the reliability of tests. However, the mere fact that the stem is oriented negatively or that the response options include AOTA does not make an MC item flawed. The results of this study show that, with careful attention to detail, negative stems and AOTA can be employed to write highly effective MC items for use in college classes.

Acknowledgments

The Departments of Biology and Environmental Studies at Roanoke College provided financial and logistical support for this study. I thank S.E. Wise for constructive comments on the manuscript text.

References

Abdulghani, H. M., Ahmad, F., Irshad, M., Khalil, M. S., Al-Shaikh, G. K., Syed, S., … Haque, S. (2015). Faculty development programs improve the quality of multiple choice questions item’s writing. *Scientific Reports, 5*, 9556. https://doi.org/10.1038/srep09556

Aiken, L. R. (1987). Testing with multiple-choice items. *Journal of Research and Development in Education*, *20*(4), 44-58.

Cassels, J. R. T., & Johnstone, A. H. (1984). The effect of language on student performance on multiple choice tests in chemistry. *Journal of Chemical Education*, *61*(7), 613-615. https://doi.org/10.1021/ed061p613

Chapman, J. C., & Toops, H. A. (1919). A written trade test: Multiple choice method. *Journal of Applied Psychology*, *3*(4), 358-365. https://doi.org/10.1037/h0073002

Chéron, M., Ademi, M., Kraft, F., & Löffler-Stastka, H. (2016). Case-based learning and multiple choice questioning methods favored by students. *BMC Medical Education*, *16*, 41. https://doi.org/10.1186/s12909-016-0564-x

Chiavaroli, N. (2017). Negatively-worded multiple choice questions: An avoidable threat to validity. *Practical Assessment, Research & Evaluation*, *22*(3), 1-14.

DiBattista, D., Sinnige-Egger, J.-A., & Fortuna, G. (2014). The “none of the above” option in multiple-choice testing: An experimental study. *The Journal of Experimental Education*, *82*(2), 168-183. https://doi.org/10.1080/00220973.2013.795127

Dudycha, A. L., & Carpenter, J. B. (1973). Effects of item format on item discrimination and difficulty. *Journal of Applied Psychology*, *58*(1), 116-124. https://doi.org/10.1037/h0035197

Frary, R. B. (1991). The none-of-the-above option: An empirical study. *Applied Measurement in Education*, *4*(2), 115-124. https://doi.org/10.1207/s15324818ame0402_2

Frey, B. B., Petersen, S., Edwards, L. M., Pedrotti, J. T., & 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

Funk, S. C., & Dickson, K. L. (2011). Multiple-choice and short-answer exam performance in a college classroom. *Teaching of Psychology, 38*(4)*,* 273-277. https://doi.org/10.1177/0098628311421329

Haladyna, T. M. (1997). *Writing Test Items to Evaluate Higher Order Thinking*. Needham Heights, MA: Allyn & Bacon.

Haladyna, T. M. (2004). *Developing and validating multiple-choice test items* (3rd ed.). New York, NY: Routledge.

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

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

Haladyna, T. M., & Rodriguez, M. C. (2013). *Developing and Validating Test Items*. New York: Routledge.

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*(3), 309-334. https://doi.org/10.1207/S15324818AME1503_5

Hansen, J. D., & Lee, D. (1997). Quality multiple-choice test questions: Item-writing guidelines and an analysis of auditing testbanks. *Journal of Education for Business*, *73*(2), 94. https://doi.org/10.1080/08832329709601623

Harasym, P. H., Price, P. G., Brant, R., Violato, C., & Lorscheider, F. L. (1992). Evaluation of negation in stems of multiple-choice items. *Evaluation & the Health Professions*, *15*(2), 198-220. https://doi.org/10.1177/016327879201500205

Ibbett, N. L., & Wheldon, B. J. (2016). The incidence of clueing in multiple choice testbank questions in accounting: Some evidence from Australia. *e-Journal of Business Education & Scholarship of Teaching*, *10*(1), 20-35.

Kelly, F. J. (1916). The Kansas Silent Reading Tests. *Journal of Educational Psychology*, *7*, 63-80. https://doi.org/10.1037/h0073542

Moncada, S. M., & Moncada, T. P. (2010). Assessing student learning with conventional multiple-choice exams: Design and implementation considerations for business faculty. *International Journal of Education Research*, *5*(2), 15-29.

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

Mueller, D. J. (1975). An assessment of the effectiveness of complex alternatives in multiple choice achievement test items. *Educational and Psychological Measurement*, *35*, 135-141. https://doi.org/10.1177/001316447503500115

Owen, S. V., & Froman, R. D. (1987). What’s wrong with three-option multiple choice items? *Educational and Psychological Measurement*, *47*(2), 513-522. https://doi.org/10.1177/0013164487472027

Poundstone, W. (2014). *How to Predict the Unpredictable: The Art of Outsmarting almost Everyone*. London: Oneworld Publications.

Rich, C. E., & Johanson, G. A. (1990). *An item-level analysis of "none of the above"*. Paper presented at the Annual Meeting of the American Educational Research Association, Boston.

Rodriguez, M. C. (1997). *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.

Rodriguez, M. C., & Albano, A. D. (2017). *The College Instructor's Guide to Writing Test Items: Measuring Student Learning*. New York, NY: Taylor & Francis Group.

Rush, B. R., Rankin, D. C., & White, B. J. (2016). The impact of item-writing flaws and item complexity on examination item difficulty and discrimination value. *BMC Medical Education*, *16*, 250. https://doi.org/10.1186/s12909-016-0773-3

Tamir, P. (1993). Positive and negative multiple choice items: How different are they? *Studies in Educational Evaluation*, *19*, 311-325. https://doi.org/10.1016/S0191-491X(05)80013-6

Tarrant, M., Knierim, A., Hayes, S. K., & Ware, J. (2006). The frequency of item writing flaws in multiple-choice questions used in high stakes nursing assessments. *Nurse Education Today*, *26*, 662-671. https://doi.org/10.1016/j.nedt.2006.07.006

Tarrant, M., Ware, J., & Mohammed, A. M. (2009). An assessment of functioning and non-functioning distractors in multiple-choice questions: A descriptive analysis. *BMC Medical Education*, *9*(40). https://doi.org/10.1186/1472-6920-9-40

Tollefson, N. (1987). A comparison of the item difficulty and item discrimination of multiple-choice items using the "none of the above" and one correct response options. *Educational and Psychological Measurement*, *47*, 377-383. https://doi.org/10.1177/0013164487472010

Towns, M. H. (2014). Guide to developing high-quality, reliable, and valid multiple-choice assessments. *Journal of Chemical Education*, *91*, 1426-1431. https://doi.org/10.1021/ed500076x

Violato, C., & Harasym, P. H. (1987). Effects of structural characteristics of stem format of multiple-choice items on item difficulty and discrimination. *Psychological Reports*, *60*, 1259-1262. https://doi.org/10.2466/pr0.1987.60.3c.1259