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 Advantages of Five-Option Multiple-Choice Items in Classroom Tests of Student Mastery

Michael Joseph Wise[1*]

[1] Biology and Environmental Studies Department, Roanoke College, 221 College Lane, Salem, Virginia, USA

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

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

Received: September 24, 2020 Accepted: October 9, 2020 Online Published: October 11, 2020

doi:10.22158/jetss.v2n4p59 URL: http://dx.doi.org/10.22158/jetss.v2n4p59

Abstract

The effectiveness of multiple-choice (MC) items depends on the quality of the response options—particularly how well the incorrect options ("distractors") attract students who have incomplete knowledge. It is often contended that test-writers are unable to devise more than two plausible distractors for most MC items, and that the effort needed to do so is not worthwhile in terms of the items' psychometric qualities. To test these contentions, I analyzed students' performance on 545 MC items across six science courses that I have taught over the past decade. Each MC item contained four distractors, and the dataset included more than 19,000 individual responses. All four distractors were deemed plausible in one-third of the items, and three distractors were plausible in another third. Each increase in plausible distractor led to an average of a 13% increase in item difficulty. Moreover, an increase in plausible distractors led to a significant increase in the discriminability of the items, with a leveling off by the fourth distractor. These results suggest that—at least for teachers writing tests to assess mastery of course content—it may be worthwhile to eschew recent skepticism and continue to attempt to write MC items with three or four distractors.

Keywords

classroom testing, difficulty, discriminability, distractor, multiple-choice item, number of response options

1. Introduction

Multiple-choice items are popular not only for standardized tests, but also for classroom assessment of mastery of course content. The conventional format for a multiple-choice (MC) item is a question (or a

statement) called the “stem”, followed by a list of potential responses. The correct response option is called the “key”, and the incorrect options are called “distractors”. Over the past few decades, testing experts have developed best-practice guidelines for the writing of MC items (Haladyna & Downing, 1989a; Haladyna & Rodriguez, 2013; Moreno, Martinez, & Muniz, 2015; Rodriguez & Albano, 2017). Nevertheless, there is still no consensus regarding one very basic issue—the number of distractors that should be included per item (Haladyna, Downing, & Rodriguez, 2002; Rodriguez, 2005; Kilgour & Tayyaba, 2016; Gierl, Bulut, Guo, & Zhang, 2017; Royal & Stockdale, 2017).

The traditional advice is to include as many distractors as possible to minimize the probability that naive test-takers will choose the correct response due to chance alone (Haladyna & Downing, 1989b; Haladyna et al., 2002; Zoanetti, Beaves, Griffin, & Wallace, 2013). For instance, the probability of choosing the correct response by random guessing decreases from 0.5 for one-distractor items (such as true or false), to 0.33 for two-distractor MC items, to 0.25, 0.20, and 0.17 for three- four- and five-distractor items, respectively. These numbers show an obviously decreasing return on investment for additional distractors. Mainly due to concerns about guessing, most test-writers tend to avoid using MC items that do not contain at least four or five options (i.e., three or four distractors) (Haladyna et al., 2002; Frey, Petersen, Edwards, Pedrotti, & Peyton, 2005; Thorndike & Thorndike-Christ, 2010).

In contrast to this tradition of maximizing the number of distractors per item, recent research appears to be converging on a recommendation of three-option MC items (with one key and two distractors) (Bruno & Dirkzwager, 1995; Haladyna et al., 2002; Rodriguez, 2005; Tarrant, Ware, & Mohammed, 2009; Kilgour & Tayyaba, 2016; Vegada, Shukla, Khilnani, Charan, & Desai, 2016; Gierl et al., 2017). The main benefit of limiting the items to two distractors is increased efficiency. Specifically, it takes less time for a writer to come up with two reasonable distractors than three or four, and it takes test-takers less time to answer each item if there are fewer distractors to read. With fewer distractors, more items can be included on a single test, which will tend to make the tests more reliable assessments of mastery.

Perhaps the most compelling argument for a test-writer to use three-option MC items is the perceived difficulty of coming up with more than two distractors that sound reasonable enough for test-takers to choose (i.e., functional distractors) (Delgado & Prieto, 1998; Haladyna et al., 2002; Tarrant et al., 2009; Dehnad, Nasser, & Hosseini, 2014; Kilgour & Tayyaba, 2016). After all, a five-option MC item is no better than a three-option item if two of the distractors are too implausible to tempt even the most naive of test-takers. Such a five-option item would be less efficient than the three-response item while presenting no advantage in terms of minimizing the effect of guessing, and potentially decreasing the validity of the item and the overall reliability of the test (Rodriguez, 2005; Tarrant et al., 2009; Towns, 2014; Rush, Rankin, & White, 2016).

Accepting for the moment that a test-writer can devise multiple plausible distractors for at least some MC items on a test, then the most pertinent question is how arbitrarily minimizing the number of response options would affect the performance of an MC item in terms of psychometric variables.

Specifically, would decreasing the number of response options change the difficulty or discriminability of MC items? Although previous research on these questions provides limited and sometimes contradictory advice, it seems likely that as the number of response options is decreased, the difficulty and discriminability of MC items will also tend to decrease (Haladyna & Downing, 1989b; Haladyna et al., 2002; Rodriguez, 2005). Clearly, more empirical evidence from organically derived MC items (i.e., from real classroom tests, rather than experimentally altered items) is required in order to give the best advice to teachers regarding the number of distractors they should include in their MC items.

The considerations that go into choosing the number of distractors and the constraints on test-writers depend on the type of test and the goals of the test-writer. For instance, the number of response options that is ideal for a large standardized test may differ from the optimal number for smaller classroom exams. The current study was undertaken from the perspective of a teacher writing MC items to assess students' mastery of content covered in class. I analyzed data from more than 19,000 responses across 545 five-option MC items to address the following specific questions regarding the number of distractors: 1) How many different distractors were chosen per MC item? 2) When more than one distractor was selected for an MC item, with what frequency were the different distractors chosen?

3) How did the number of distractors chosen relate to the difficulty of the items? and 4) How did the number of distractors chosen relate to the effectiveness of the items, as measured by a discrimination index?

2. Methods

2.1 The Data Set

The analyses in this paper include testing data from six courses that I taught at Roanoke College from 2012 to 2019 (Table 1). These courses include three introductory courses for biology majors, two versions of a general ecology course, an introductory environmental science course, and a general-education course for non-science majors. In total, 545 multiple-choice items were analyzed from hourly, midterm, and final exams given in these courses (Table 1). Each of these MC items included one key and four distractors. For each MC item, I recorded the responses (viz., A, B, C, D, or E) from every student (499 total) who took the exams—for a total of 19,150 responses (minus a small handful of inadvertent non-responses).

2.2 Number of Plausible Distractors

The first question I addressed was the relative frequency of MC items having one, two, three, or four plausible distractors. A common cutoff for a distractor to be considered "plausible" is that at least 5% of test-takers must choose that response. Because my classes were small (ranging from 23 to 51 students), I considered a distractor to be plausible if at least one student chose it. Because many MC items on the tests were intended to be relatively easy, I also looked at the number of distractors chosen per item for the subset of items for which fewer than 80% of the students chose the correct response. In other words, I addressed the question of how many different distractors were chosen per item after

excluding the 182 easiest items, because few students choose distractors for the easy items—often by design.

Table 1. Courses that Served as the Source of Data for the Analyses. Within a Course, Data for any MC Item that Appeared on Two Separate Tests (e.g., on both an Hourly Exam and the Final Exam) were Included for only the First Appearance of the Item. BIOL 205 was Substantially Revised between 2012 and 2015, and the Data Set Excludes any MC Items from the 2015 Exams that also Appeared on the Final Exam of 2012. The Data Set Thus Includes a Single Appearance of a Total of 545 MC Items across six Courses (Including two Versions of BIOL 205)

Course Number and Name	Year	Exam	Students	MC Items
BIOL 205. General Ecology	2012	Final	34	57
BIOL 120. Principles of Biology	2013	Final	48	42
BIOL 205. General Ecology	2015	Midterm	51	31
		Final	44	41
INQ 251. Bugs in the System	2016	Final	23	50
BIOL 180. Exploring Biological Diversity	2018-9	Hourly 1	35	21
		Hourly 2	34	17
		Hourly 3	35	18
		Hourly 4	33	18
		Hourly 5	31	28
		Final	31	49
ENSC 101. Introduction to Environmental Science	2019	Hourly 1	34	28
		Hourly 2	33	25
		Final	33	120

2.3 Difficulty and Discrimination Index

The "difficulty" of an item is traditionally defined (somewhat counter-intuitively) as the proportion of test-takers who chose the correct response, and it is often abbreviated as "*p*". Thus, for an item that all students answered correctly, $p = 1$, while for an item that no students answered correctly, $p = 0$.

The "discriminability" of an item can be quantified by how well students perform on the item, compared to their overall performance on a test. A highly effective item in this sense will be answered correctly by students who do relatively well on the test as a whole, and incorrectly by students who do relatively poorly overall. In this study, I quantified the discriminability of each item by a standard discrimination index (*DI*) that was calculated with a point-biserial regression (Towns, 2014). Specifically, for each item, *DI* is equivalent to a Pearson product-moment correlation between whether each student answered the item correctly (1) or incorrectly (0) and the student's total score on all the

items of an exam. The higher the correlation, the more discriminating the item, and (loosely speaking) the more likely it was that the item did a good job at assessing mastery. While a *DI* can theoretically range from -1 to 1, any $DI > 0.2$ generally indicates that the item was sufficiently discriminating, and an item with $DI > 0.4$ is considered excellent (Towns, 2014).

To explore the relationship between the number of plausible distractors and the difficulty and discriminability of MC items, I ran two linear regressions, one with difficulty (p) and one with *DI* as the response variable. The predictor variable for both regressions was the number of plausible distractors. I excluded from these regressions the 11 MC items for which no distractors were chosen by any student because such items are meaningless in terms of discriminability. All of the statistical analyses reported in this paper were performed using JMP_{IN} v. 4.0.4 (SAS Institute, Cary, North Carolina, USA).

3. Results

3.1 Number of Plausible Distractors

The number of different distractors chosen per item covered the entire gamut (Figure 1). Two or fewer distractors were chosen for ~35% of the MC items, including ~2% for which every student chose the correct response. Nearly two-thirds of the items contained at least three plausible distractors, with nearly one-third containing four plausible distractors.

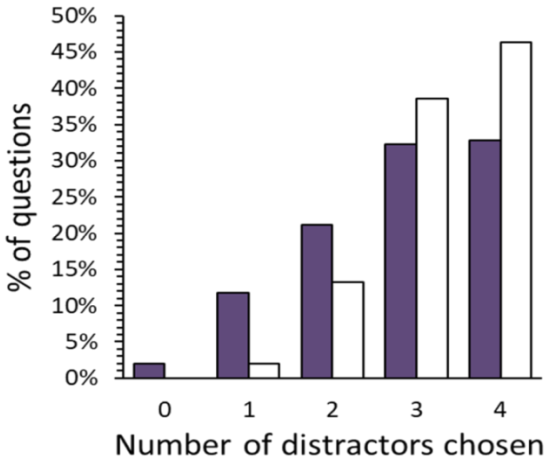

Figure 1. Frequency Distribution of Number of Distractors Chosen by at Least One Student per MC Item (i.e., Plausible Distractors). Dark (Purple) Columns Represent the Complete Data Set of 545 MC Items. Empty (White) Columns Represent the 363 Items for Which < 80% of Students Chose the Correct Option

When the 182 "easy" MC items (with $p \geq 0.8$) were excluded from consideration, the proportions of items with higher numbers of distractors chosen increased substantially (unfilled bars, Figure 1). For the 363 items with $p < 0.8$. only 15% had two or fewer plausible distractors, while nearly half of the

items included four plausible distractors.

Although more than one distractor was chosen for most items, the frequency with which different distractors were chosen varied substantially (Figure 2). That is, only rarely were there multiple equally plausible distractors for an item. For instance, when there were two plausible distractors, the second-most plausible distractor was chosen only half as frequently as the most plausible distractor (Figure 2). When three or more distractors were chosen per item, the third-most plausible distractor was chosen by 16% of the students who chose an incorrect response. When all four distractors were chosen for an item, a single distractor accounted for nearly half of the incorrect responses. Nevertheless, even the least plausible distractor made up a substantial minority (10% on average) of the incorrect responses when all four distractors were chosen for an item (Figure 2).

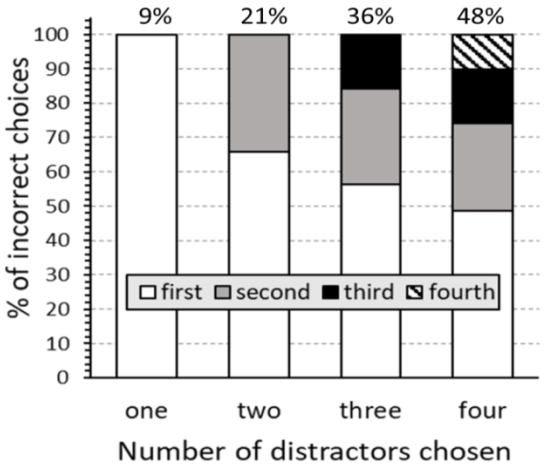

Figure 2. Stacked-column Graph of the Relative Plausibilities of Different Distractors of a Multiple-choice Item as a Function of How Many Distractors were Chosen for that Item. For Instance, for the MC Items in which Four Different Distractors were Chosen by Students in a Class, the most Plausible Distractor was Chosen 49% of the Time, and the Fourth most Plausible Distractor was Chosen 10% of the Time when Students Chose an Incorrect Option. The Percentages above the Columns Indicate the Mean Percentage of Students who Answered the Items Incorrectly as a Function of How Many Different Distractors were Chosen across the Entire Class

3.2 Difficulty and Discrimination Index

The difficulties (*p*) of the MC items were significantly correlated with the number of plausible distractors (Figure 3A; $r = -0.61$, $P < 0.0001$). Because *p* measures the proportion of correct responses, this result shows that the more plausible distractors an item contained, the larger the mean proportion of incorrect responses. Specifically, the regression equation estimates that for every plausible distractor, the proportion of incorrect responses increased by a mean of 0.13 (prediction equation: Difficulty = 1.04 – 0.13 x Distractors).

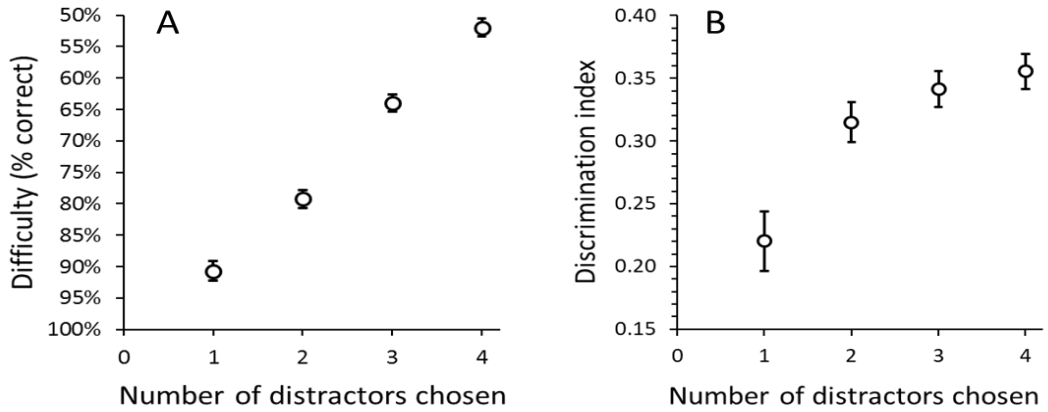

Figure 3. Relationship between Number of Plausible Distractors and A. Difficulty (*p*) and B. Discriminability (*DI*) of the 545 Multiple-choice Items. Points and Bars Represent Means ± 1 SEM

The number of plausible distractors per item was significantly positively correlated with the discriminabilities of the items (Figure 3B; $r = 0.20$, $P < 0.0001$). The linear regression equation estimates that for every plausible distractor, the *DI* increased by a mean of 0.04. However, the relationship was not strictly linear, as the effect of increasing the number of distractors lessened as more distractors were included (Figure 3B). In particular, there was a large jump (43%) in *DI* between having one and two plausible distractors, but much smaller jumps in *DI* from two-to-three (8%) or from three-to-four (4%) plausible distractors.

4. Discussion

4.1 Number of Distractors per MC Item

This study demonstrated that it is not unrealistic to expect to be able to write multiple-choice items that have more than two plausible distractors for classroom exams**.** This finding runs counter to a growing contention that the challenge in devising more than two distractors is severe enough to preclude the use of MC items with more than three response options (Delgado & Prieto, 1998; Haladyna et al., 2002; Tarrant et al., 2009; Dehnad et al., 2014; Kilgour & Tayyaba, 2016). Each of the 545 MC items in my dataset contained four distractors (in addition to the key), and each distractor was intended to be a reasonable-sounding response that would tempt students who had incomplete knowledge of the construct tested by the item. One-third of the MC items ended up containing two or fewer plausible distractors, defined as a distractor that was chosen by at least one student in a class. However, nearly one-third of the MC items contained three plausible distractors, and another third contained four plausible distractors.

In any class, there are concepts that are so fundamental that we expect all students should master them. Furthermore, many classes target a median grade of a B+/C- (i.e., ~80%). Such a target requires that there will have to be quite a few “easy” items. By definition, very few distractors will be chosen for

easy items. If we take into account these considerations and eliminate the 182 easiest items in this study (i.e., those for which the difficulty, *p*, was greater than or equal to 0.8), the conclusion that it is not prohibitively difficult to write multiple plausible distractors is even more strongly supported. Specifically, 85% of the 353 "not easy" items had at least three plausible distractors, with four different distractors chosen for more than half of those 85%.

4.2 The Role of Plausible Distractors in the Difficulty of MC Items

The difficulty of an MC item was very strongly associated with the number of plausible distractors that the item contained. However, it is not possible to discern a clear cause-and-effect relationship between the variables. Consider an item that addresses a relatively simple concept. Few students are likely to choose a distractor for a simple item, which means that fewer distractors are likely to have been chosen even once. At the same time, if the test-writer includes additional tempting distractors for a given item, then it is more likely that a student will stray from choosing the key. In the first scenario, the difficulty of the item can be seen as a cause of the number of distractors that are chosen. In the second, the difficulty can be seen as an effect of the number of plausible distractors. Neither conclusion is incorrect, as the difficulty and the number of plausible distractors are intimately connected. Regardless, it is hard to argue against the proposition that if the number of distractors were decreased, the key would be chosen more frequently (even by random chance alone), thereby making an item less difficult.

Despite this apparent truism, empirical studies have shown inconsistent results concerning the effect that the number of distractors has on the difficulty of an MC item. For instance, Haladyna and Downing (1993) reviewed four standardized MC tests and found no relationship between the number of plausible distractors and item difficulty. In an updated review, Haladyna et al. (2002) reported that five studies concluded that fewer distractors made for easier items, while two other studies found the opposite pattern. In a comprehensive meta-analysis of 56 independent trials from 27 studies, Rodriguez (2005) found that decreases in the number of response options in MC items led to significant decreases in difficulty.

4.3 The Role of Plausible Distractors in the Discriminability of MC Items

In the current study, the number of plausible distractors per item was also strongly correlated with the discriminability of the items. Unlike the relationship between the number of plausible distractors and item difficulty, the relationship with discriminability was not linear. Instead, the positive effect of the number of distractors on the discrimination index (*DI*) quickly leveled off. Specifically, there was a 43% increase in *DI* from one-to-two plausible distractors, an 8% increase from two-to-three plausible distractors, and only a 4% increase from three-to-four plausible distractors. Because the discriminability of items is of paramount importance in designing an effective and reliable test of mastery, a 4% increase may be well worth the effort needed to come up with a fourth plausible distractor. However, the pattern shown in this study suggests that going beyond four plausible distractors is likely to produce ever-diminishing gains in discriminability.

The pattern of increasing discriminability with a greater number of distractors is largely consistent with other studies reported in the literature, though exceptions exist. For instance, in Haladyna and Downing's (1993) review, all four MC tests that examined this relationship found higher *DI* for items with greater numbers of plausible distractors. However, the updated review of Haladyna et al. (2002) reported one study with no effect of the number of distractors on *DI*, and another study that found that *DI* increased with a decrease in the number of response options. In his comprehensive meta-analysis, Rodriguez (2005) found that decreases in the number of response options almost always resulted in a significant decrease in the discriminability of MC items.

Inferences regarding the effect of the number of distractors on the discriminability of MC items depend heavily on the design of the study. Very often, studies that have found three-option items to be as discriminating as five-option items have employed the strategy of giving the five-option test first, then selectively removing the poorest-functioning distractors, then retesting students (sometimes the same ones) with the resultant three-option tests (Owen & Froman, 1987; Delgado & Prieto, 1998; Nwadinigwe & Naibi, 2013; Dehnad et al., 2014; Schneid, Armour, Park, Yudkowsky, & Bordage, 2014; Kilgour & Tayyaba, 2016). Such a procedure may be relevant for teachers who teach the same course and use the same test items year after year. Even so, the discriminabilities of the MC items are likely to be even better if—instead of reducing the number of distractors—the least-functioning distractors were replaced with novel distractors, which at least have a chance of being plausible (Zoanetti et al., 2013).

4.4 Different Considerations for Standardized Tests versus Classroom Tests

The MC items analyzed in this study were written to assess students' mastery of content in my courses. I did not design the items to test hypotheses regarding the optimal number of distractors (or any other hypotheses regarding the construction of MC items). Therefore, this study was not an experiment in which the numbers of distractors for given stems were deliberately altered in a replicated fashion to discern their effects on psychometric qualities. As such, the analyses are necessarily post-hoc. Nevertheless, the number of MC items (545) and responses (>19,000) should make my inferences highly robust. Moreover, the organic origin of the MC items in the dataset lends authenticity to the inferences. In fact, several researchers have lamented the shortage of empirical studies on this type of authentic classroom data (Owen & Froman, 1987; Tarrant et al., 2009; Funk & Dickson, 2011). It should be stressed, however, that these inferences only strictly apply to similar testing situations—that is, testing of mastery of content delivered in a college course.

Classroom tests of course mastery differ in scope, character, and purpose from standardized tests in several ways that may influence guidance regarding the number of distractors that should be included in MC items. In short, there are reasons to consider including more distractors in classroom tests than may be optimal for standardized tests. Below, I summarize three of these reasons.

(1) Standardized tests are likely to cover a much wider range of content, and thus contain many more items, than a single classroom test. The most direct way to maximize the number of items that can be

answered during a given amount of time is to minimize the number of response options per test item. In fact, a premium on efficiency is a main reason for the recent call for three-option items over the traditional four- or five-option items (Owen & Froman, 1987; Rodriguez, 2005; Nwadinigwe & Naibi, 2013; Dehnad et al., 2014; Schneid et al., 2014; Vegada et al., 2016). Because having fewer response options raises the risk of guessing, more items are required for a test with fewer response options to maintain the reliability of the test (Rodriguez, 2005; Royal & Stockdale, 2017). For instance, Zimmerman and Williams (2003) calculated that if a test included only items with three response options, it would need to include a minimum of 80 items to be sufficiently reliable. Except for comprehensive final exams, classroom tests are unlikely to include that many MC items. Similarly, except for very large lecture classes, tests are likely to contain a variety of constructed-response items in addition to a limited number of MC items. In my classes, it is quite uncommon for students to be stressed for time during tests due to having too many MC items to answer. Therefore, I have experienced no need to minimize the number of distractors per MC item for the sake of test-taking efficiency.

(2) Standardized tests are given to vast populations of test-takers who represent a wide range of knowledge and experiences. In contrast, a classroom is composed of a relatively small number of students, all who have ostensibly had equal exposure to the same material covered on a test in their class. Teachers know what content has been covered in the course, and they have a good sense of their students' common misunderstandings regarding various constructs. Teachers can take advantage of these misconceptions to construct multiple distractors that are likely to lure students who have incomplete mastery, thus leading to highly discriminating MC items. It is less practical to devise a set of multiple distractors that will tempt the multitude of diverse, anonymous takers of standardized tests. In fact, there is no shortage of evidence of the use of flawed items and implausible distractors in a diversity of MC tests (Haladyna & Downing, 1993; Tarrant, Knierim, Hayes, & Ware, 2006; Tarrant et al., 2009; Kilgour & Tayyaba, 2016; Rush et al., 2016). Therefore, the rationale to minimize response options due to the difficulty of devising distractors may apply much more to standardized tests than to classroom tests.

(3) Not only are standardized tests given to multitudes of people, but they are used multiple times, year after year. Therefore, there are plenty of chances to identify and eliminate poorly functioning distractors so that only the most plausible distractors appear in revised versions of the items. Classroom teachers may reuse MC items from one class to another, but the data available to decide which distractors are implausible are much more restricted than for standardized tests. When writing new MC items, there is no way for a teacher to know for sure which reasonable-seeming distractors will end up luring no one. In addition, an unchosen distractor in one year may be a more attractive response in another year by chance alone. It would be self-defeating to eliminate potentially plausible distractors for the sole purpose of limiting the item to an arbitrary number of response options.

4.5 Recommendations for Classroom Teachers

The target audience for this paper is teachers who use, or are considering using, multiple-choice items in their classroom testing. Fortunately, there are many highly accessible sources for guidance on the writing of high-quality MC items (e.g., Haladyna & Downing, 1989a; Haladyna et al., 2002; Frey et al., 2005; Moreno, Martinez, & Muniz, 2006; Towns, 2014; Gierl et al., 2017; Rodriguez & Albano, 2017). I conclude this paper with several pieces of advice to teachers based on previously published guidelines, the quantitative results of this study on distractors, and my experience writing MC items for tests across a variety of courses:

(1) Aim to include three or four distractors in your MC items. When teaching about a topic, keep track of misconceptions expressed by students, as these misconceptions can be the best fodder for MC-item distractors. The results of the current study show that it is not unreasonable to expect to come up with multiple plausible distractors for most MC items on content taught in a class. This study also shows that the more plausible distractors an item contains, the more discriminating the item is likely to be. The improvement from three-to-four plausible distractors was not as large as from one-to-two, or from two-to-three, but it is worth the effort. The only distractor that is assured to be non-functional is the one that is never written.

(2) Most teachers include at least some easy items on each test. For instance, there are likely to be some concepts that are so fundamental that you expect all of your students to know the answer. For such items, most of your distractors will go unchosen by all the students. Unless you are content with half of your students failing a test, then you will need to accept that some of your items will contain nominally implausible distractors. This does not necessarily mean that the item is flawed. The goals of boosting student performance and confidence can sometimes be more important than making sure that every item on a test is maximally discriminating and that all the distractors are chosen by a minimum percentage of your students for every MC item.

(3) Avoid unintentional cues in the text of your distractors that may help students guess the correct response option. Foremost, all of your response options should be homogenous in content and grammatically consistent with the text of the stem. Failure on either of these fronts will ruin the appeal of a distractor. Also, avoid repeating a phrase from the stem in the correct response option (the key), and avoid making the key longer than the distractors. Failure on either of these fronts will make the key more obvious to guessers.

(4) One tempting way to increase the number of distractors in an MC item is to include “all of the above” (AOTA) or “none of the above” (NOTA) as a response option. However, most experts agree that these “inclusive” response options should only be used with caution—or never at all (Harasym, Leong, Violato, Brant, & Lorscheider, 1998; Pachai, DiBattista, & Kim, 2015; Ibbett & Wheldon, 2016; Rush et al., 2016). On the one hand, MC Items with inclusive response options are prone to misreading by students. On the other hand, these items are often susceptible to guessing by test-wise students. While the first issue increases item difficulty and the second issue decreases difficulty, both issues have

the undesirable tendency to decrease the discriminability of an item. Therefore, if AOTA or NOTA is added to an item only to increase the number of distractors, then they are more likely to make the item worse rather than better. Nevertheless, if the subject matter tested by an item suggests an organic use of an inclusive option, then its use can be beneficial as long as the item is written very carefully to maximize clarity and minimize unintentional cuing (Frary, 1991; Rodriguez, 1997; Wise, 2020).

A final consideration regarding the use of an inclusive response option is that it should serve as the key with a frequency that is inversely proportion to the number of response options that the items on the test contain (Hansen & Lee, 1997; DiBattista, Sinnige-Egger, & Fortuna, 2014). For instance, if there are five response options in the MC items on a test, then AOTA (or NOTA) should be used as the key 20% of the time and as a distractor 80% of the time that it is used as a response option.

(5) Even more often than AOTA or NOTA, test experts caution against using complex response options, such as “Both A and B”. While including options that are combinations of other options is an easy way to increase the number of distractors in an item, such an item is fraught with possibilities for ambiguity, misinterpretation, and frustration on the part of the test-takers. Thus, the increase in the number of distractors that results from combining response options would likely lead to an item being less, rather than more, discriminating.

(6) Humor is an important part of many teachers’ style. Even on tests, teachers may be tempted to include an obviously incorrect but hopefully funny response option for a MC item. The jury is still out on whether this is an advisable practice (McMorris, Boothroyd, & Pietrangelo, 1997; Haladyna et al., 2002). Humor on an exam may ease the tension for some students, but other students may not appreciate the teacher making light of what to them is a high-stakes judgment by someone who holds all the power. With welcome increases in student diversity (e.g., students from a variety of cultures, or students on different positions on the autism spectrum), it is becoming more likely that many students may not understand the intended humor. Students who do not get the joke may feel that the joke is ultimately on them. From a practical standpoint, an intentionally ludicrous distractor will do nothing to increase the discriminability of a test item. Combined with the risk of upsetting some students, the lack of concrete benefits suggests that teachers forego the humorous distractor and replace it with a distractor that has a better chance at being plausible.

(7) Once a solid list of distractors has been written for a MC item, you need to decide where in the order of response options the key should appear. A common piece of advice is to “balance the key”, which means to avoid putting the correct response option in the same location too often (e.g., by making “C” the key for half of your items). My advice is to eliminate any subjectivity in deciding where to place the key by using a computerized randomization procedure. (A procedure for randomizing key locations using Excel is detailed in the Appendix.) After writing the items and deciding the order in which the items will appear on the test, adjust the key for each item to match the position suggested by your randomization procedure. Of course, there will be cases in which you will need to override the randomized position. Specifically, if there is a logical or numerical order to the

response options, then this order should take precedence over randomness. Also, if the key is "all of the above" or "none of the above", then you may need to override the randomization to make the key the last option.

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

Bruno, J. E., & Dirkzwager, A. (1995). Determining the optimal number of alternatives to a multiple-choice test item: An information theoretic perspective. *Educational and Psychological Measurement, 55*(6), 959-966. https://doi.org/10.1177/0013164495055006004

Dehnad, A., Nasser, H., & Hosseini, A. F. (2014). A comparison between three- and four-option multiple choice questions. *Procedia - Social and Behavioral Sciences*, *98*, 398-403. https://doi.org/10.1016/j.sbspro.2014.03.432

Delgado, A. R., & Prieto, G. (1998). Further evidence favoring three-option items in multiple-choice tests. *European Journal of Psychological Assessment*, *14*(3), 197-201. https://doi.org/10.1027/1015-5759.14.3.197

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

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

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*(6), 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*, *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., & Downing, S. M. (1993). How many options is enough for a multiple-choice test item? *Educational and Psychological Measurement*, *53*, 999-1010. https://doi.org/10.1177/0013164493053004013

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., Leong, E. J., Violato, C., Brant, R., & Lorscheider, F. L. (1998). Cuing effect of "all of the above" on the reliability and validity of multiple-choice questions. *Evaluation & the Health Professions*, *21*(1), 120-133. https://doi.org/10.1177/016327879802100106

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.

Kilgour, J. M., & Tayyaba, S. (2016). An investigation into the optimal number of distractors in single-best answer exams. *Advances in Health Science Education*, *21*, 571-585. https://doi.org/10.1007/s10459-015-9652-7

McMorris, R. F., Boothroyd, R. A., & Pietrangelo, D. J. (1997). Humor in educational testing: A review and discussion. *Applied Measurement in Education*, *10*(3), 269-297. https://doi.org/10.1207/s15324818ame1003_5

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

Moreno, R., Martinez, R. J., & Muniz, J. (2015). Guidelines based on validity criteria for the development of multiple choice items. *Psicothema*, *27*(4), 388-394.

Nwadinigwe, P. I., & Naibi, L. (2013). The number of options in a multiple-choice test item and the psychometric characteristics. *Journal of Education and Practice*, *4*(28), 189-196.

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

Pachai, M. V., DiBattista, D., & Kim, J. A. (2015). A systematic assessment of "none of the above" on multiple choice tests in a first year psychology classroom. *The Canadian Journal for the Scholarship of Teaching and Learning*, *6*(3), Article 2. https://doi.org/10.5206/cjsotl-rcacea.2015.3.2

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. (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. https://doi.org/10.1111/j.1745-3992.2005.00006.x

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.

Royal, K. D., & Stockdale, M. R. (2017). The impact of 3-option responses to multiple-choice questions on guessing strategies and cut score determination. *Journal of Advances in Medical Education & Professionalism*, *5*(2), 84-89.

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

Schneid, S. D., Armour, C., Park, Y. S., Yudkowsky, R., & Bordage, G. (2014). Reducing the number of options on multiple-choice questions: Response time, psychometrics and standard setting. *Medical Education, 48,* 1020-1027. https://doi.org/10.1111/medu.12525

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

Thorndike, R. M., & Thorndike-Christ, T. M. (2010). *Measurement and Evaluation in Psychology and Education* (8th ed.). Boston, MA: Pearson Education, Inc.

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

Vegada, B., Shukla, A., Khilnani, A., Charan, J., & Desai, C. (2016). Comparison between three option, four option and five option multiple choice question tests for quality parameters: A randomized study. *Indian Journal of Pharmacology*, *48*, 571-575. https://doi.org/10.4103/0253-7613.190757

Wise, M. J. (2020). The effective use of negative stems and "all of the above" in multiple-choice tests of in college courses. *Journal of Education, Teaching and Social Studies*, *2*(4), 47-58. https://doi.org/10.22158/jetss.v2n4p47

Zimmerman, D. W., & Williams, R. H. (2003). A new look at the influence of guessing on the reliability of multiple-choice tests. *Applied Psychological Measurement*, *27*(5), 357-371. https://doi.org/10.1177/0146621603254799

Zoanetti, N., Beaves, M., Griffin, P., & Wallace, E. M. (2013). Fixed or mixed: A comparison of three, four and mixed-option multiple-choice tests in a fetal surveillance education program. *BMC Medical Education, 13,* 35. https://doi.org/10.1186/1472-6920-13-35

Appendix. Randomizing the position of the correct response option (the "key") for a list of multiple choice items using Excel. For this example, consider 20 five-option MC items. In column A of the spreadsheet, fill the first 20 rows with the numbers 1-20. In Cell B1, type "=Rand()" (omitting the quotation marks; capitalization is irrelevant). Then copy and paste that formula to fill the first 20 rows of Column B. In Cell C1, type "=If(B1>0.8,"A",If(B1>0.6,"B",If(B1>0.4,"C",If(B1>0.2, "D","E"))))". Then copy and paste that formula to fill the first 20 rows of Column C. Column C will then contain the potential randomized location for the keys for your 20 MC items. If the order is in some way unappealing (e.g., the first three keys are all "A"), then you can re-randomize the list by typing anything in any cell. For instance, typing "1" in Cell A1 and pressing Enter will cause the list to re-randomize. Once you are content with the randomized list, highlight and copy Cells C1 through C20. Then perform a "paste special" operation to paste the letters (rather than the formulas) into Cells D1 through D20. The letters pasted into Column D will not change automatically when any other operation is performed on the spreadsheet.