Advances in Health Sciences Education (2019) 24:141–150
https://doi.org/10.1007/s10459-018-9855-9

The optimal number of options for multiple-choice questions on high-stakes tests: application of a revised index for detecting nonfunctional distractors

Mark R. Raymond[1] · **Craig Stevens**[1] · **S. Deniz Bucak**[1]

Received: 24 April 2018 / Accepted: 19 September 2018 / Published online: 25 October 2018

Abstract
Research suggests that the three-option format is optimal for multiple choice questions (MCQs). This conclusion is supported by numerous studies showing that most distractors (i.e., incorrect answers) are selected by so few examinees that they are essentially nonfunctional. However, nearly all studies have defined a distractor as nonfunctional if it is selected by fewer than 5% of examinees. A limitation of this definition is that the proportion of examinees available to choose a distractor depends on overall item difficulty. This is especially problematic for mastery tests, which consist of items that most examinees are expected to answer correctly. Based on the traditional definition of nonfunctional, a five-option MCQ answered correctly by greater than 90% of examinees will be constrained to have only one functional distractor. The primary purpose of the present study was to evaluate an index of nonfunctional that is sensitive to item difficulty. A secondary purpose was to extend previous research by studying distractor functionality within the context of professionally-developed credentialing tests. Data were analyzed for 840 MCQs consisting of five options per item. Results based on the traditional definition of nonfunctional were consistent with previous research indicating that most MCQs had one or two functional distractors. In contrast, the newly proposed index indicated that nearly half (47.3%) of all items had three or four functional distractors. Implications for item and test development are discussed.

Keywords Assessment · Multiple-choice questions · Test development · Item-writing guidelines · High-stakes testing

Introduction

Theory and analytical work on multiple-choice questions (MCQs) suggest that three options—one correct answer and two distractors—are optimal for the single-best answer format (Grier 1975; Lord 1944; Tversky 1964). Empirical support for this recommendation

✉ Mark R. Raymond
mraymond@nbme.org

[1] National Board of Medical Examiners, 3750 Market Street, Philadelphia, PA 19104, USA

is provided by a thorough meta-analysis completed by Rodriguez (2005) who found that most distractors are selected by so few examinees that they are essentially nonfunctional. In addition, several studies have shown that eliminating one option from a four-option MCQ, or two options from a five-option MCQ, has a negligible impact on measurement precision (Delgado and Prieto 1998; Edwards et al. 2012; Rodriguez 2005; Tarrant et al. 2009; Kilgour and Tayyaba 2016). Test items with fewer options should, in theory, reduce the amount of reading time per item, thereby decreasing exam speededness or allowing more test items to be covered on an exam (Tversky 1964). In addition, reducing the number of options should decrease the time and effort required to write, review, and edit test items.

Despite these findings, the five-option format is still common in health professions education (Kilgour and Tayyaba 2016), implying that practice lags behind research. One reason for the popularity of the five-option format is that, until recently, most studies have been conducted in contexts other than health professions education (see Abdulghani et al. 2017; Kilgour and Tayyaba 2016; Rogausch et al. 2010; Schneid et al. 2014 for exceptions). For example, the vast majority of studies included in the Rodriguez (2005) meta-analysis were conducted in educational settings with school-aged children. Another possible explanation for the persistence of the five-option MCQ is its continued appearance on high-stakes examinations in medical education. By replicating the formats used on credentialing tests, educators can ensure that students will be exposed to the same formats that later will be used for making high-stakes decisions. Meanwhile, credentialing agencies may be reluctant to adopt MCQs with three or four options because most research has been conducted on relatively short, locally-developed exams. It is conceivable that high-stakes tests consist of higher quality items than local tests due to the extensive item review procedures employed by credentialing agencies (Abozaid et al. 2017; Jozefowicz et al. 2002; Wallach et al. 2006). Of the 43 studies included in the Rodriguez (2005) meta-analysis, only seven involved credentialing exams. Since that publication, one study looked specifically at credentialing tests (Rogausch et al. 2010). Therefore, one motivation for the present study was to determine the extent to which previous research supporting the three-option format generalizes to high-stakes in the health professions.

Another limitation of prior research arises from the method for identifying nonfunctional distractors. A nonfunctional distractor (NFD) is typically defined using two criteria (Kilgour and Tayyaba 2016; Rodriguez 2005; Tarrant et al. 2009; Wakefield 1958). The first is based on the proportion of examinees who choose each distractor. This is known as the p value for distractor j and is designated here as p_{dj}. It is commonly accepted that a distractor is nonfunctional if $p_{dj} < .05$. The second criterion for nonfunctional is based on the correlation of distractor j with the total score, designated as r_{djt}. Nonfunctional distractors are those for which $r_{djt} \geq 0$. In words, NFDs are those distractors chosen by fewer than 5% of examinees, or that exhibit a positive correlation with total score.

The problem with this definition is that the number of examinees choosing a distractor depends on overall item difficulty; therefore, easy items will have fewer examinees available to select distractors. This fact was noted in an extensive literature review by Gierl et al. (2017), and was empirically demonstrated on medical school tests by Abdulghani et al. (2014). The threshold of $p_{dj} < .05$ is not sensitive to the fact that easy items, which may otherwise be effective, could have no functional distractors. Consider an item with five options (i.e., four distractors) where the 82% of examinees choose the correct response ($p_c = .82$). If each distractor performs equally well and draws one-fourth of the incorrect responses, then $p_{dj} = .045$; in this instance, none of the distractors would be declared functional. Many educational tests and most professional credentialing tests are designed as mastery tests; as such they consist of items that most examinees are expected to answer correctly. Although tests used for other

purposes (e.g., college admissions) should and do consist of difficult items, it is common for mastery tests to have p values in the .80 s and .90 s. A consequence of the traditional definition of nonfunctional is that any item for which $p_c > .90$ will have, at most, just one functional distractor. Thus, the traditional definition of nonfunctional does not seem appropriate for mastery tests, prompting some scholars to suggest that the rule be ignored for easy items (Gierl et al. 2017).

Prior to drawing any firm conclusions about the optimal number of options optimal for mastery tests, it is important to conduct research that utilizes an index consistent with the purpose of such tests. The dependency between item difficulty and distractor functionality was recognized by Rogausch et al. (2010) who evaluated distractors by applying two criteria, $p_{dj} < .05$ and $p_{dj} < .01$. They demonstrated that the two criteria provided very different outcomes, and ultimately argued that a threshold of .01 be used to define nonfunctional distractors on high-stakes exams in medical education. We suggest a different approach. Rather than requiring test developers to choose between the two thresholds (.05 or .01), or disregarding the rule for easy items, an alternative strategy is proposed below—one that uses a single threshold but allows it to vary with overall item difficulty.

The present study has two objectives. The first is to extend previous research by studying distractor functionality within the context of a professionally-developed credentialing test that employs item writing procedures recognized for their rigor (Abozaid et al. 2017). The second purpose is to propose and evaluate an index of *nonfunctional* that is sensitive to item difficulty, and to compare that index to the traditional definition of $p_{dj} < .05$. The study is replicated across four test forms of a large-scale test, allowing the inclusion of several hundred MCQs.

Method

Participants and test items

The study included live (scored) items from four test forms of an examination for physician licensure. The number of examinees completing each test form ranged from 1204 to 1237. Each form consisted of approximately 320 scored and unscored (experimental) test items, with most items written as single best answer clinical vignettes followed by three to five options. This study included only scored items with five-options (four distractors and one correct answer) as there were too few items with three and four options for systematic study. The final sample included 840 five-option items spread across the four test forms, with each reduced form containing from 206 to 220 items. Mean scores on the four reduced forms ranged from .734 to .748 on a proportion correct scale. SDs ranged from .083 to .089, and coefficient alphas were between .880 and .890.

Analyses

An index was developed to address the limitations of the traditional definition of NFD. The proposed index: (a) builds on the previous definition of NFD; (b) is conditional on an item's p value such that the easier the item, the lower the threshold; and (c) applies to any dichotomously-scored MCQ and is not specific to any class of tests. Specifically, the threshold p value for designating a distractor as NFD was determined by:

$$p_{nfd} = 0.1 - \left(p_c * 0.1\right) \quad (1)$$

where p_c is the proportion choosing the correct response. A distractor was designated as nonfunctional if $p_{dj} < p_{nfd}$. It is apparent that p_{nfd} varies across items, and that as $p_c \rightarrow 1.0$, $p_{nfd} \rightarrow 0.0$. This definition has the desirable property that across the range of p values, more difficult items have larger values of p_{nfd} than easier items, with the central value of p_{nfd} equal to the traditional criterion of .05. For example, three items with p values of .10, .50, and .90 would have corresponding values of p_{nfd} equal to .09, .05, and .01. Equation (1) also preserves the traditional definition of NFD for optimally difficult items. That is, the point at which item variance is maximized (i.e., $p_c = .50$) is also that point at which $p_{nfd} = .05$.

For this study, the number of nonfunctional distractors based on Eq. (1) was compared to the number of distractors based on the traditional constant of .05. For completeness we also tabulated the number of distractors where or $r_{djt} \geq 0$, although that index is of secondary interest.

Finally, we studied the effect of eliminating distractors on test difficulty and score reliability. It is known that reducing the number of options can have a negative impact on test score reliability assuming that each distractor is equally functional (Lord 1944). This part of the analysis sought to determine the effect of distractor elimination using two methods: deletion of the least popular distractor and random deletion. The former method assumes that item writers are sensitive to distractor quality when producing items and, if asked to write items with fewer distractors, will avoid writing the worst distractor. The latter method (random deletion) does not make this assumption—it represents a worst-case mechanism for reducing distractors because they are eliminated blindly. Under both scenarios, the responses for examinees who chose an option that had been eliminated were assigned at random the one of the remaining options, including the correct answer. The distractor reduction process was first done by reducing the number of distractors from four to three, and then again from four to two. Similar methods have been employed in other studies (e.g., Kilgour and Tayyaba 2016; Tarrant et al. 2009).

Results

We first summarize the total number of nonfunctional distractors based on the traditional and newly proposed definitions, and then more closely examine the number of functioning distractors at the item level. We next evaluate the consequences of dropping distractors. For each analysis we present the results and interpret them in the context of previous studies. Although this organization breaks with convention, it allows the Discussion section to address more general themes.

A total of 3360 distractors were studied across 840 items. The mean p_c across all items was 0.743, indicating that the present test forms were easier than tests reported in most previous studies. Table 1 summarizes results averaged over the four test forms. The first column of Table 1 presents the results based on the traditional threshold of 0.05 for nonfunctional, while the second column shows outcomes based on the newly proposed definition presented in Eq. (1).

The percentage of nonfunctional distractors based on the traditional index of $p_{dj} < .05$ was 58.9%. This is comparable to other studies of the five-option MCQ in medical education. Kilgour and Tayyaba (2016) and Rogausch et al. (2010) reported nonfunctional distractor rates of 55.3% and 68.3%, respectively. The present value cannot be directly compared to the findings of Tarrant et al. (2009) because that study examined MCQs with four options. The important result in Table 1 is that the newly proposed definition based on

Table 1 Number and percent of nonfunctioning distractors for the traditional and new method based on distractor frequency and discrimination

Criteria	Traditional method $p_{dj} < .05$	New method $p_{dj} < p_{nfd}$
Distractor frequency (p_{dj})	1978 (58.9%)	1167 (34.7%)
Distractor discrimination ($r_{djt} \geq 0$)[a]	197 (5.9%)	197 (5.9%)
Either frequency or discrimination	2083 (62.0%)	1295 (38.5%)
Total N of distractors	3360 (100%)	3360 (100%)

[a]There is no new method for defining distractor discrimination so the two columns are the same

p_{nfd} identifies 34.7% of items as being nonfunctional. Even though this outcome is more encouraging than the 58.9% based on the traditional definition, by either definition a significant number of distractors attract very few examinees. The other notable finding in Table 1 is that only 5.9% of distractors were designated as nonfunctional based solely on the $r_{djt} \geq 0$.

Table 2 provides a more detailed look at the number and percent of functioning distractors per item based on distractor p values. The results using the traditional method are comparable to those reported in previous studies. For example, the present study found that all four distractors were functional for only 2.5% of items, which is similar to studies where all four distractors were functional only for 2.7% and 2.8% of items (Kilgour and Tayyaba 2016; Rogausch et al. 2010). At the other end of the spectrum, the present study identified 11.5% of the items as having no functional distractors by the traditional criterion. This outcome also compares to values of 14.2% and 12.5% reported in previous studies (Kilgour and Tayyaba 2016; Tarrant et al. 2009). The cumulative percent column is also revealing. It indicates that by the traditional method only 13.1% of all items had three or four functioning distractors, suggesting that the third and fourth distractors could be dropped from all items with only a minor loss of information. This outcome would support a three-option MCQ.

The right-most columns in Table 2 based on the new method paint a more optimistic picture of the distractors. First, there are, quite literally, only a handful of items with no functioning distractors. Second, the finding that more than one-third of items (34.4%) have three functional distractors suggests that there is merit in retaining the third distractor. The cumulative percent column further indicates that 47.3% of items had three or four functional distractors, a finding that supports the four-option MCQ. Third, the finding that all four distractors function for only 13% of items suggests that a distractor could be dropped

Table 2 Functioning distractors per item across all items for the traditional method and new method (N = number; cum% = cumulative percent)

Functioning distractors per item	Traditional method $p_{dj} < .05$			New method $p_{dj} < p_{nfd}$		
	N	%	Cum%	N	%	Cum%
Four	21	2.5	2.5	109	13.0	13.0
Three	89	10.6	13.1	289	34.4	47.3
Two	293	34.9	48.0	326	38.8	86.2
One	340	40.5	88.5	110	13.1	99.3
None	97	11.5	100.0	6	0.7	100.0
Total N of items	840			840		

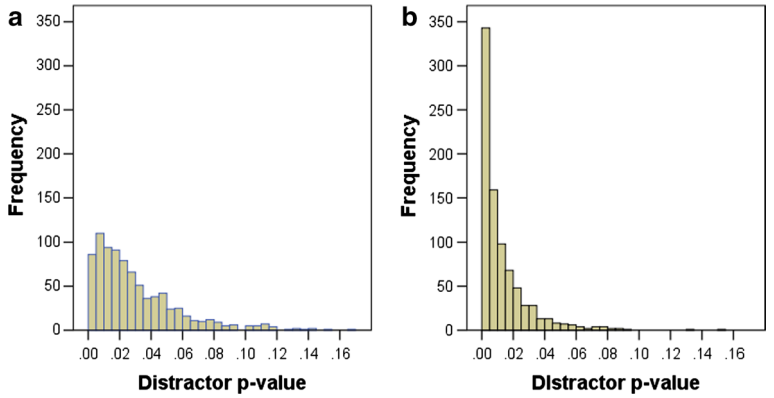

Fig. 1 Frequency distribution of distractor p values for the third most popular (panel a) and fourth most popular (panel b) distractors for 840 items. A three-option format would exclude the distractors represented in both panels, while a four-option format would exclude distractors represented only by panel b

Table 3 Test score mean, average inter-item correlation, and reliability (coefficient alpha) after eliminating distractors

Distractors eliminated	Method of elimination	Test mean	Inter-item correlation	Test score reliability
None	–	.743	.0368	.884
One	Least popular	.746	.0355	.880
	Random	.759	.0338	.874
Two	Least popular	.757	.0320	.867
	Random	.783	.0291	.855

with minimal loss. The findings in Table 2 based on the new method challenge previous findings in support of the 3-option format (Rodriguez 2005). More generally, the cumulative percent columns indicate that the new method is roughly one distractor more liberal than the traditional method.

The frequency distributions in Fig. 1 provide a closer look at the plausibility of the third and fourth distractors. Each graph represents 840 items. It is evident that the third most popular distractor (Fig. 1a) still attracts a fair proportion of examinees for a sizable number of items. The mean and median of the distribution are .030 and .022, respectively. That is, the third "best" distractor for half of the 840 items attracted 2.2% or more of examinees. However, the fourth (or least popular) distractor (Fig. 1b) drew very few examinees for most items. The mean and median values of p_{dj} for the fourth distractor are .013 and .007. It is admittedly a judgment call, but the spike in Fig. 1b offers little support for the fourth distractor, while the flatter distribution in Fig. 1a implies that there may be merit in retaining the third distractor.

Table 3 displays the new test mean, average inter-item correlation, and reliability (coefficient alpha) after dropping one or two distractors using either systematic elimination or random elimination of distractors. As expected, dropping distractors made the test easier and slightly less reliable. The most noticeable impact on reliability occurs when using random elimination to drop two distractors; otherwise the changes are negligible. These

findings are consistent with previous research (Kilgour and Tayyaba 2016; Pappenberg and Musch 2017; Rodriguez 2005; Rogausch et al. 2010; Tarrant et al. 2009), and suggest that even if item writers or reviewers are not completely effective at deciding what options not to write (systematic elimination), and the reduction in distractors occurs by a more or less random mechanism, test reliability still will not suffer much.

Discussion

This study produced four notable outcomes, three of which are consistent with previous research. First, using the traditional definition of nonfunctional ($p_{dj} < .05$) the present study found that the vast majority of MCQs have one or more nonfunctional distractors. In particular, there was evidence that according to the traditional definition MCQs typically consist of only two effective distractors. Second, compared to previous studies, the distractors studied here were less likely to be flagged because of positive correlations with the total score ($r_{djt} \geq 0$). We attribute this outcome to the extensive review and pretesting of items prior to live administration, during which items with $r_{djt} \geq 0$ are revised or deleted. Third, eliminating the least functional distractor had almost no impact on test score reliability. The fourth finding was that using the new threshold of distractor functionality, p_{nfd}, instead of the constant value of .05 suggested that many MCQs consist of three effective distractors. Neither threshold ($p_{dj} < .05$ or p_{nfd}) supported the four distractor (five-option) format. However, there is some question as to whether the evidence favors three-options or four-options. The remainder of the paper addresses this issue and discusses its implications for testing practice.

The traditional method ($p_{dj} < .05$) identified 58.9% of all distractors as being nonfunctional, and also found that only 13.1% of items had three or four functional distractors. These outcomes support a three-option MCQ—that is, an item with two distractors and one correct answer. In contrast, setting the threshold at p_{nfd}, indicated that 34.7% of all distractors would be declared nonfunctional, and that nearly half (47.3%) of the items had three or four functional distractors. These outcomes suggest that there may be merit to the four-option MCQ. Figure 1 illustrated that the third distractor was quite functional for many items, and offered additional support for the four-option format. Whether the data support three or four options is a matter of judgment. However, our opinions align with those of Rogausch et al. (2010) who argued that options which attract even a few low-performing examinees fulfill an important role.

Regardless of the number of options that one deems optimal, we advocate use of the variable threshold proposed in Eq. (1) over the traditional fixed index of 0.05 because the former acknowledges that item difficulty naturally constrains distractor performance. It is also important to acknowledge that any cut-off value, whether variable or fixed, is arbitrary. In our experience with many item writers and reviewers for numerous mastery tests, distractors with p_{dj} in the range of .02 to .05 typically have been viewed as making a positive contribution to an item's quality assuming that $r_{djt} < 0$ (Rogausch et al. 2010). This may explain the reluctance of educators and testing agencies to adopt a three-option MCQ. As a practical matter, convincing educators and testing agencies to convert from five options to three options will be challenging because it seems such a radical change. However, moving from five options to four may be perceived as less jarring to stakeholders (Cizek and O'Day 1994 document the transition of one medical specialty board from five to four options).

There are structural reasons why four options can be more effective than five. *In many instances* an even number of options will be more resistant to the types of item writing flaws that promote test-taking strategies irrelevant to what the test is intended to measure. One of those flaws is the convergence strategy (Smith 1982), which occurs when the correct answer is also the option that has the most in common with all other options. It is often easier to balance the option set with an even number of distractors. In addition, an even number of distractors is the natural result of simultaneously varying two factors, concepts, or physical properties. In the following example, the two factors that vary are imaging parameters.

> *Clinical scenario describing an elderly patient who is scheduled for a mammogram….* Given the client's age and body habitus, what change in exposure factors will most likely produce an image with acceptable quality?
>
> A. increase kVp
>
> B. decrease kVp
>
> C. increase focal spot
>
> D. decrease focal spot

Numerous concepts and principles lend themselves to option sets that consist of one or more parameters (e.g., blood pressure, wave frequency) that vary in direction—they can increase or decrease. For many such items, effective options are pairs of opposites, and the logical number of options will be an even number. Of course, opposites can be avoided by including additional parameters (e.g., mAs, distance) in the option set, assuming the additional distractors are plausible. To be sure, there also are occasions when items are best served by a three-option format, as when the options consist of the direction of a single parameter of interest (e.g., blood pressure), and a fully plausible option set might consist of "increases," "decreases," and "no change." Our point is that effective MCQs can be written with four options; however, some content will naturally have three options as the upper limit.

There is an additional factor to consider when weighing the pros and cons three and four options. Most published studies involved test items developed by teachers and researchers (Rodriguez 2005), and one might reasonably question the type of item-writing training they received. Writing effective distractors is challenging. While most texts offer general guidelines for developing distractors, they devote very little attention to *strategies* for creating those distractors, although there are exceptions (Gierl et al. 2017; Roid and Haladyna 1982). Indeed, there is evidence to suggest that rigorous item writing procedures can yield effective distractors. Haladyna and Rodriguez (2013) cite items from the ACT assessment where rationales are given for each option; the process of producing rationales compels item writers to fully attend to each distractor. Haladyna and Rodriguez (2013) go on to note that "If justifications were given for all distractors written, four- and five-option items might be more effective than they are currently" (p. 106). The benefits of thorough training are also documented by Abdulghani et al. (2017) who demonstrated the long-term effectiveness of faculty development workshops aimed at reducing the number of NFDs. Furthermore, early research suggests that automated item generation (AIG)—a process that steps item writers through a cognitive task analysis—can produce multiple effective distractors. The Medical Council of Canada pilot tested 22 five-option MCQs produced by AIG and found that all but three of the 110 options were selected by some examinees (Gierl et al. 2016). We suggest that in those instances where the data indicate that distractors are

ineffective, the solution is not to write fewer distractors, but rather to adopt item development procedures that result in more effective distractors. Training item writers in methods such as cognitive task analysis (Gierl et al. 2017), writing explanations (Haladyna and Rodriguez 2013), and concept mapping (Fisher et al. 2000) may prove fruitful.

Although there is an abundance of research on the optimal number of options for MCQs, the logical argument and empirical results presented here suggest that there could be merit in re-analyzing previous studies using the index of NFD proposed in this paper. We suspect that the conclusions of some of those studies might change. Other useful research might include experiments that randomly assign students to items with different numbers of options (Pappenberg and Musch 2017); studies to evaluate the ability of item authors or reviewers to eliminate the least effective distractor from existing items; and studies that compare strategies (e.g., concept mapping; cognitive modeling) for teaching content experts how to produce plausible, instructionally-meaningful distractors.

In conclusion, the present study found little support for the conventional practice of five options, challenged the recommendation from previous research that three options are optimal, and proposed the continued use of the four-option format. The large number of items studied here, and the adequate sample sizes, provide stable results that should generalize to other large-scale testing programs. However, given that the items were drawn from a high-quality item pool for one testing program, the findings may not generalize to teacher-made tests or other contexts lacking in item writer training and item review processes.

Acknowledgements The authors express their gratitude to NBME for supporting this research. However, the opinions expressed here are those of the authors and do not necessarily reflect the position of NBME or the United States Medical Licensing Examination.

Compliance with ethical standards

Conflict of interest The authors have no conflicts of interest to report. After IRB review by the American Institutes of Research, it was determined that this research is exempt from IRB review and oversight.

References

Abdulghani, H. M., Ahmad, F., Ponnamperuma, G. G., Khalil, M. S., & Aldrees, A. (2014). The relationship between non-functioning distractors and item difficulty of multiple choice questions: A descriptive analysis. *Journal of Health Specialties, 2,* 148–151.

Abdulghani, H. M., Irshad, M., Haque, S., Ahmad, T., Sattar, K., & Khalil, M. S. (2017). Effectiveness of longitudinal faculty development programs on MCQs items writing skills: A follow-up study. *PLoS ONE, 12*(10), e0185895. https://doi.org/10.1371/journal.pone.0185895.

Abozaid, H., Park, Y. S., & Tekian, A. (2017). Peer review improves psychometric characteristics of multiple choice questions. *Medical Teacher, 39,* s50–s54.

Cizek, G. J., & O'Day, D. (1994). Further investigation of nonfunctioning options in multiple-choice test items. *Educational and Psychological Measurement, 54*(4), 861–872.

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.

Edwards, B. D., Arthur, W., & Bruce, L. L. (2012). The 3-option format for knowledge and ability multiple-choice tests: A case for why it should be more commonly used in personnel testing. *International Journal of Selection and Assessment, 20*(1), 65–81.

Fisher, K. M., Wandersee, J. H., & Moody, D. E. (2000). *Mapping biology knowledge*. Boston, MA: Kluwer Academic.

Gierl, M. J., Balut, 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.

Gierl, M. J., Lai, H., Pugh, D., Touchie, C., Boulais, A. P., & De Champlain, A. (2016). Evaluating the characteristics of generated multiple-choice test items. *Applied Measurement in Education, 29*(3), 196–210.

Grier, J. B. (1975). The number of alternatives for optimum test reliability. *Journal of Educational Measurement, 12,* 109–112.

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

Jozefowicz, R. F., Koeppen, B. M., Case, S., Galbraith, R., Swanson, D., & Glew, R. H. (2002). The quality of in-house medical school examinations. *Academic Medicine, 77,* 156–161.

Kilgour, J. M., & Tayyaba, S. (2016). An investigation into the optimal number of distractors in single-best answer exams. *Advances in Health Sciences Education, 21,* 571–585.

Lord, F. M. (1944). Reliability of multiple choice tests as a function of number of choices per item. *Journal of Educational Psychology, 35,* 175–180.

Pappenberg, M., & Musch, J. (2017). Of small beauties and large beasts: The quality of distractors on multiple-choice tests is more important than their quantity. *Applied Measurement in Education, 30*(4), 273–286.

Rodriguez, M. C. (2005). Three options are optimal for multiple-choice test items: A meta-analysis of 80 years of research. *Educational Measurement: Issues and Practice, 24*(2), 3–13.

Rogausch, A., Hofer, R., & Krebs, R. (2010). Rarely selected distractors in high stakes medical multiple choice examinations and their recognition by item authors: A simulation and survey. *BMC Medical Education, 10,* 85.

Roid, G. H., & Haladyna, T. M. (1982). *A technology for test-item writing*. New York: Academic Press.

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*(10), 1020–1027.

Smith, J. K. (1982). Converging on correct answers: A peculiarity of multiple-choice items. *Journal of Educational Measurement, 19*(3), 211–220.

Tarrant, M., Ware, J., & Mohammed, A. M. (2009). An assessment of functioning and nonfunctioning distractors in multiple-choice questions: A descriptive analysis. *BMC Medical Education, 9,* 40–47.

Tversky, A. (1964). On the optimal number of alternatives at a choice point. *Journal of Mathematical Psychology, 1,* 386–391.

Wakefield, J. A. (1958). Does the fifth choice strengthen a test item? *Public Personnel Review, 19,* 44–48.

Wallach, P. M., Crespo, L. M., Holtzman, K. Z., Galbraith, R. M., & Swanson, D. B. (2006). Use of a committee review process to improve the quality of course examination. *Academic Medicine, 77*(2), 156–161.