Review of Educational Research
December 2017, Vol. 87, No. 6, pp. 1082–1116
DOI: 10.3102/0034654317726529

Developing, Analyzing, and Using Distractors for Multiple-Choice Tests in Education: A Comprehensive Review

Mark J. Gierl, Okan Bulut, Qi Guo, and Xinxin Zhang
University of Alberta

Multiple-choice testing is considered one of the most effective and enduring forms of educational assessment that remains in practice today. This study presents a comprehensive review of the literature on multiple-choice testing in education focused, specifically, on the development, analysis, and use of the incorrect options, which are also called the distractors. Despite a vast body of literature on multiple-choice testing, the task of creating distractors has received much less attention. In this study, we provide an overview of what is known about developing distractors for multiple-choice items and evaluating their quality. Next, we synthesize the existing guidelines on how to use distractors and summarize earlier research on the optimal number of distractors and the optimal ordering of distractors. Finally, we use this comprehensive review to provide the most up-to-date recommendations regarding distractor development, analysis, and use, and in the process, we highlight important areas where further research is needed.

Keywords: distractor, multiple choice, test, item development

Multiple-choice testing could easily be considered one of the most enduring and successful forms of educational technology that remains in practice today. Fredrick J. Kelly is often cited as the developer of the multiple-choice item format (Rogers, 1995). In 1916, he published the *Kansas Silent Reading Test* in the *Journal of Educational Psychology* where students who wrote the test were required to circle the correct answer rather than writing their answer for each item. The multiple-choice item format was an important breakthrough in educational testing because it served as an objectively scored task that used a structured-response format where the student was presented with one correct option and two or more incorrect options or distractors. The task was to select the correct option. As shown in Figure 1, "yellow is the correct option for Item 1, while "red" and "green" are the incorrect options or distractors. Similarly, "orange" is the correct option for Item 2, and "apples" is the distractor. Kelly (1916) went further claiming that a multiple-choice

1. I have red, green and yellow papers in my hand. If I place red and green papers on the chair, which color do I still have in my hand?

Red Green Yellow

2. Think of the thickness of the peelings off apples and oranges. Put a line around the name of the fruit having the thinner peeling.

Apples Oranges

FIGURE 1. *Two multiple-choice items from the Kansas Silent Reading Test.*

item, as he conceived it, must also satisfy three criteria: (a) the item should be interpreted by all students in the same way; (b) the item should target a single problem so that its answer would be completely right or completely wrong, and not partly right and partly wrong; and (c) the difficulty level of the item should not depend on either obscure words or unintentional cues in the stem. These criteria serve as the first guidelines for developing multiple-choice items.

In the 1920s, the College Board's Scholastic Aptitude Test and Lewis Terman's Stanford–Binet Intelligence Test adopted Kelly's multiple-choice item format largely because of the benefits produced from using a structured-response format. Multiple-choice items and other selected-response formats (e.g., true–false questions) could be scored easily and efficiently using a stencil.

In 1934, another significant breakthrough occurred. IBM introduced a "test-scoring machine" that electronically sensed the location of lead pencil marks on a scanning sheet, which further increased the efficiency of scoring multiple-choice items and permitting, for the first time, large-scale educational testing as we know it today. The IBM machine was used in 1936 to score tests for the New York State Regents and the Providence Rhode Island public schools (Lemann, 1999). This combination of using a structured-response format to administer an objective task that, in turn, was scored with a machine has been used billions of times at every level in our educational system for virtually all content areas to implement multiple-choice testing.

Today, a typical North American student takes hundreds of multiple-choice tests and answers thousands of multiple-choice items as part of her K–12 educational experience. Chingos (2012) reported that one third of the U.S. states use multiple-choice items exclusively for assessing fourth-grade and eighth-grade students' math and reading skills. Similarly, in higher education, a multiple-choice test remains the most widely used assessment format for measuring students' knowledge, especially in introductory courses with a large group of students. Multiple-choice testing is also used extensively for international assessments. For instance, in the 2015 administration of the Trends in International Mathematics and Science Study, half of the mathematics and science items used the multiple-choice format (Mullis, Cotter, Fishbein, & Centurino, 2016). In the 2015 administration of the Programme for International Student Assessment, two thirds of the items in reading, mathematics, and science assessments were in the multiple-choice format (OECD, 2016).

Multiple-choice items are widely used in educational testing because they permit the direct measurement of many knowledge, skills, and competencies across a broad range of disciplines and content areas including the ability to understand concepts and principles, make judgments, draw inferences, reason, complete statements, interpret data, and apply information. Multiple-choice items are efficient to administer, they are easy to score objectively, and they can be used to sample a wide range of content domains in a relatively short time using a single test administration (Haladyna & Rodriguez, 2013; Rodriguez, 2016). Compared with essays and other constructed-response tasks, which are prone to subjective scoring and require more time for recording answers, multiple-choice items can be scored more accurately, and they require students to spend less time on recording answers (Haladyna, 2004). Because of these noteworthy benefits, multiple-choice testing is considered to be an economical form of educational assessment. Olson (2005) claimed, for example, that it would cost the United States $1.9 billion to meet testing requirement for 6 years using machine-scored multiple-choice testing. The costs increase dramatically when other item formats are used. Lau, Lau, Hong, and Usop (2011) reported that it would cost $3.9 billion if both multiple-choice and open-ended items were used, and up to $5.3 billion if tests with human-scored, written-responses items were administered.

Multiple-choice testing is considered by many to be an effective form of educational assessment. Downing (2006a), in his seminal chapter in the *Handbook of Test Development*, went further to claim that selected-response items, like multiple-choice items, are the most appropriate item format for measuring cognitive achievement or ability, especially higher order cognitive skills, such as problem solving, synthesis, and evaluation. He also claimed that this item format is both useful and appropriate for creating exams intended to measure a broad range of knowledge, ability, or cognitive skills across many domains (see also Downing, 2006b; Haladyna, 2004).

A multiple-choice item consists of the stem, the options, and any auxiliary information. The stem contains context, content, and/or the question the student is required to answer. The options include a set of alternative answers with one correct option and one or more incorrect options or distractors. Auxiliary information includes any additional content, in either the stem or option, required to generate an item, including text, images, tables, graphs, diagrams, audio, and/or video. To answer a multiple-choice item, the student is presented with a stem and two or more options that differ in their relative correctness. Students are required to make a distinction among response options, several of which may be partially correct, in order to select the best or most correct option. Hence, students must use their knowledge and problem-solving skills to identify the relationship between the content in the stem and the correct option. The incorrect options are called distractors because they are considered to be "distracting" to students with partial knowledge due to their plausibility to yield the correct option.

In 1989, Thissen, Steinberg, and Fitzpatrick published an article describing a method for modeling multiple-choice item performance (this method is described later in our review) with the title "Multiple-Choice Models: The Distractors Are

Also Part of the Item." Thissen et al. (1989) claimed that most test developers and users believed that the stem and the correct option serves as the most important part of the multiple-choice item. The intention for using a provocative title in their 1989 article was to draw attention to the importance of distractors in the multiple-choice format. Distractors are an important part of a multiple-choice item for at least three reasons. First, distractors require a significant amount of time and resources during the item development process because they must be written by content specialists. For each multiple-choice item, one correct option is required. But two (i.e., three-option item), three (i.e., four-option item), or four (i.e., five-option item) incorrect options—all of which must be plausible but incorrect—are also produced for each item. When many items are needed, distractor development becomes a formidable task for the content specialist. For example, when 100, five-option, multiple-choice items are created, the content specialist is required to create 100 stems, 100 correct options, and 400 distractors.

Second, distractors create an important part of the context required to solve a multiple-choice item that can affect item quality and learning outcomes. Within this context, a complex relationship exists between the correct and incorrect options due to the fact that students are required to make a distinction among response options in order to select the correct response (cf. Hambleton & Jirka, 2006). This complex relationship is fostered by the effects of partial knowledge in response performance that, in turn, interacts with the plausibility of each distractor can affect the psychometric properties of the correct and incorrect options (e.g., Bock, 1972; Dorans, Schmitt, & Bleistein, 1992; Haladyna, 2016; Haladyna & Rodriguez, 2013; Penfield, 2008; Thissen et al., 1989; Wainer, 1989). A substantial body of empirical research also indicates that this complex relationship between the correct and incorrect options can affect learning. The "testing effect" occurs when an assessment is used to enhance memory retention (Roediger & Karpicke, 2006). Empirical studies have demonstrated that multiple-choice items can produce the testing effect by eliciting beneficial retrieval processes that, in turn, result in improved performance on subsequent examinations (Butler & Roediger, 2008; Fazio, Agarwal, Marsh, & Roediger, 2010; Marsh, Roediger, Bjork, & Bjork, 2007; Roediger & Marsh, 2005). However, the benefits of the testing effect are contingent on the quality of the distractors. Little and Bjork (2015; see also Little, Bjork, Bjork, & Angello, 2012) reported that competitive multiple-choice items (i.e., items where the distractors are plausible and share important information with the correct option) elicit beneficial retrieval processes when the information related to both the correct option and the distractors is evaluated on future exams. But, when the distractors are not related to the correct option they can introduce misinformation into the assessment process, thereby decreasing memory retention (Bishara & Lanzo, 2015; Butler, Marsh, Goode, & Roediger, 2006; Butler & Roediger, 2008; Odegard & Koen, 2007; Roediger & Marsh, 2005). The number of unrelated distractors used for an item can also adversely affect memory retention (Brown, Schilling, & Hockensmith, 1999; Butler et al., 2006).

Third, distractor analysis can help test developers and instructors understand why students produce errors and thereby guide our diagnostic inferences about test performance. For example, distractors can provide information about student misconceptions that, in turn, can specify the type of instruction that is needed to

overcome these errors in thinking, reasoning, and problem solving (e.g., Briggs, Alonzo, Schwab, & Wilson, 2006). Based on the results from distractor analysis, instructors can identify the content areas that need instructional improvement and provide students with remedial instruction in those content areas.

Since the publication of Frederick Kelly's (1916) Kansas Silent Reading Test in the *Journal of Educational Psychology*, there has been continuous growth in what could now be considered a vast literature on multiple-choice testing. But a review of this literature reveals one noteworthy finding: While the development, analysis, and use of the stem and the correct response option are well documented (cf. Thissen et al., 1989), there is comparatively little research on the incorrect options or distractors. This imbalance in the literature is also apparent in the practice of item development where the task of creating the stem and correct option is largely well described, in our experience. But by comparison, the task of creating the distractors is poorly described. Distractor development, in fact, is often considered by content specialists to be the most daunting and challenging component of writing a multiple-choice item.

Purpose of the Review

In this review, we present a comprehensive survey of the literature on multiple-choice testing focused, specifically, on the development, analysis, and use of distractors in the context of educational assessment. The purpose of our review is to provide the most up-to-date recommendations on distractor development, analysis, and use. In the process, we also identify areas where research on distractors is lacking. To achieve this goal, we present our review in two sections. The first section titled "Creating and Analyzing Distractors for Multiple-Choice Items" is focused on the initial process of writing distractors. We also review the literature on how to evaluate the quality of these newly created distractors. The second section is titled "Considerations When Using Distractors." In this section, we review three different topics. We synthesize the existing guidelines on how to effectively use distractors in multiple-choice items. We summarize research on the optimal number of distractors. We also describe the research on the optimal position for distractors. Finally, we conclude our review with recommendations from the literature for distractor development, analysis, and use, and in the process, we highlight one important area where further research is needed.

Method

The focus of the literature review was to access full-text documents using various search terms or keywords. In our review, the phrase *multiple choice* was used with the terms and phrases *distractors, alternative, option, cognition, learning, psychometric, item development*, and *test development* to identify the initial list of citations. The first round of review on the distractor literature was conducted by searching the following databases:

- *Education Index Retrospective 1929–1983*, which provides a vast record of important education literature. It also contains both historical and updated subject headings.
- *Education Research Complete*, a bibliographic and full-text database covering scholarly research and information relating to all areas of education.

- *ERIC* (Education Resources Information Center), a database on indexed and full-text educational reports, evaluations, and research, which features journals included in the *Current Index of Journals in Education* and *Resources in Education Index.*
- *ProQuest Education Journals*, a database that includes more than 1,000 full-text journals and 18,000 dissertations, supporting research on the theory and practice of education.
- *JSTOR*, which contains full-text academic journals in the humanities, social sciences, and sciences.

A second round of review was also conducted using the compiled bibliography derived from the first round, garnering even more articles, and continuing iteratively throughout the review process. The articles included in our review are readily available in the academic literature and have all undergone some form of scholarly peer review prior to their publication or presentation. Because of this requirement, some documents were excluded. Test development is a standardized process that requires iterative refinements (Schmeiser & Welch, 2006). Because this process must yield fair and equitable assessment tasks, it is often standardized through the use of test development guides (Haladyna & Rodriguez, 2013). These guides, which are often paired with item-writing training programs, provide content specialists with information to structure their task. Test development guides provide a summary of best practices, common mistakes, and general expectations, often written specifically for a testing company or organization, that help ensure that the content specialists have a shared understanding of their tasks and their responsibilities.

During our search, we identified many different test development guides that could potentially include information on distractors. However, these guides have been excluded from this review for two reasons. First, there is no way to properly sample the test development guides from testing companies and organizations because some groups publicly share their documents while others do not. Second, the guides, while providing potentially important suggestions on the practice of item writing, have not undergone a peer review. Because of these two limitations, we cannot evaluate the representativeness or quality of the content. As a result, these types of test development guides have been excluded from our study.

In total, 834 articles, books, and conference proceedings were collected. From this larger set, a total of more than 375 documents met the topical relevance criteria for inclusion in the literature review. The majority of the documents were journal articles (239), followed by books and book chapters (40) and conference proceedings and research reports (85). The final set of 105 documents that we cited in our review appear in the reference section along with a digital object identifier or a direct link so that each article can be accessed directly from their primary source in the literature.

Results

Creating and Analyzing Distractors for Multiple-Choice Items

To begin, we present our summary of the literature on distractor development and analysis. That is, we review the literature on how to create distractors for multiple-choice items. Once the distractors are initially created, we describe the

most common psychometric and statistical methods that are used to evaluate the quality of the distractors.

Developing Distractors for Multiple-Choice Items in Education

Two general strategies have been consistently described and advocated for distractor development. The first strategy, which is the most common one, focuses on a list of plausible but incorrect alternatives linked to common misconceptions or errors in thinking, reasoning, and problem solving (Case & Swanson, 2001; Collins, 2006; de la Torre, 2009; Haladyna & Downing, 1989; Moreno, Martínez, & Muñiz, 2006, 2015; Rodriguez, 2011, 2016; Tarrant, Ware, & Mohammed, 2009; Vacc, Loesch, & Lubik, 2001). Misconceptions can be identified by looking at students' answers from constructed-response or open-ended items (e.g., Briggs et al., 2006) or from studies of student response processes using verbal reports (e.g., Haladyna & Rodriguez, 2013). If outcomes from these two methods are not available, then lists of alternatives can be developed by experienced content specialists using the responses obtained from questions such as "What do students usually confuse this concept or idea with?" "What is a common error for solving this problem?" or "What are the common misconceptions in this field?" (Collins, 2006). Haladyna and Rodriguez (2013) provided the following concise description that serves as the common and most up-to-date recommendation on how to create distractors for multiple-choice test items:

> The most effective way to develop plausible distractors is to either obtain or know what typical learners will be thinking when the stem of the item is presented to them. We refer to this concept as a common error. Knowing common errors can come from a good understanding of teaching and learning for a specific grade level; it can come from think aloud studies with students; or it can come from student responses to a constructed-response format version of the item without options. (p. 106)

The second strategy focuses on similarity. Content specialists are instructed to create distractors that are similar in content and structure relative to the correct option (Ascalon, Meyers, Davis, & Smits, 2007; Case & Swanson, 2001; Guttman, Schlesinger, & Schlesinger, 1967; Hoshino, 2013; Lai et al., 2016; Mitkov & Ha, 2003; Owens, Hanna, & Coppedge, 1970; Towns, 2014). Content similarity includes incorrect options that are comparable with but different from the correct option. For numeric options, once the correct answer is calculated, factors can be removed or inverted and the answer recalculated to produce a distractor. For key feature options, similarity includes distractors that fall into the same category as the correct option, such as the same concept, topic, or idea. Similarity can also be specified using computational tools like semantic relatedness where hypernyms or hyponyms are identified that can serve as distractors for the correct option (e.g., Mitkov & Ha, 2003; Mitkov, Ha, Varga, & Rello, 2009). Structural similarities include distractors that share characteristics with the correct option such as length, complexity, formatting, and grammar.

Distractor Analysis

Once the distractors are written, their quality must be evaluated. Distractor quality is typically evaluated using the results from an item analysis of the

distractors, which is also known as distractor analysis. Two types of distractor analyses are often conducted with multiple-choice items. The first, which we call traditional distractor analysis, is based on either classical test theory (CTT) or item response theory (IRT). The second, which we characterize as contemporary distractor analysis, is based on more recent developments in psychometric theory, including cognitive diagnostic analyses.

Traditional distractor analysis. Traditional distractor analysis can be conducted with either CTT or IRT. In CTT distractor analysis, the major purpose of the analysis is to guide item revision (Haladyna, 2016). Although CTT distractor analysis has also been used for determining the optimal number of distractors and weighting distractors (Haladyna & Rodriguez, 2013), the primary purpose of CTT distractor analysis is to eliminate nonfunctioning distractors and to improve the discrimination power of multiple-choice items in distinguishing low- from high-ability students.

The most basic CTT distractor analysis is to examine the percentage of students that choose each distractor. The aim of this analysis is to detect low-frequency distractors called "nonfunctioning distractors." According to Haladyna and Downing (1993), if less than 5% of the students choose a distractor, it is considered a low-frequency distractor. Content specialists should consider either removing such a distractor from the item or revising it to improve item discrimination. However, under certain circumstances, this suggestion can be ignored. For example, if a multiple-choice item is very easy (e.g., more than 90% of the students answered the item correctly), then most of the distractors are expected to have a low frequency. However, such easy items can still be retained in the test to maintain the content coverage or to meet content requirements in the test blueprint.

Wainer (1989) discussed using trace line plots to visualize the relationship between students' abilities and distractor selection percentages. A trace plot can easily be used to identify nondiscriminating and nonfunctioning distractors (Haladyna, 2016). An example of a trace line plot is shown in Figure 2. The horizontal axis represents students' total score, which is divided into five ordinal categories from the lowest group to the highest group. The vertical axis represents the percentage of students who choose a particular option. In this example, Option A is the correct answer. As expected, the percentage of students who select Option A increases as student ability increases. Option B is an example of a well-functioning distractor. The percentage of students who select Option B decreases as student ability increases. Option C is an example of nondiscriminating distractor, which has a relatively constant selection rate across different ability levels. Option D is an example of nonfunctioning distractor, which has a selection percentage smaller than 5% across all ability groups. The omit trace line represents students who do not make any selection in the item. As demonstrated in the example, the trace line plot can be useful when identifying nondiscriminating and nonfunctioning distractors. Based on this visual item analysis, content specialists can readily identify nonfunctioning distractors and replace them with better distractors.

While the trace line plot provides a way to visualize distractor distributions across different ability groups, it does not provide an objective way to determine whether a trace line of a distractor is flat (i.e., the percentage of selecting

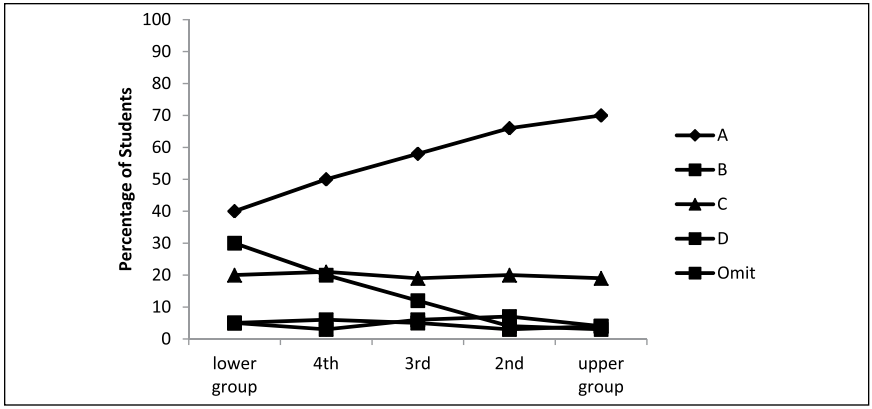

FIGURE 2. *A trace plot for a hypothetical multiple-choice item.*
Note. The correct response option is A.

the distractor does not change depending on the ability). In order to address this problem, Haladyna and Downing (1993) used the chi-square goodness-of-fit test to test whether the slope of a trace line is significantly different from 0 (i.e., whether it is flat). In the chi-square goodness-of-fit test, the observed frequencies of distractors are compared with the expected frequency of distractors to compute the chi-square statistic. The formula is given as

$$\chi_k^2 = \sum_{c=1}^{C} \frac{\left(O_{ck} - E_{ck}\right)^2}{E_{ck}} \quad \text{and} \quad df = C - 1,$$

where O_{ck} is the observed frequency of students with ability level c who choose option k in the item, E_{ck} is the expected frequency of students with ability level c who choose option k, df is the degrees of freedom, and C is the total number of student ability levels. If the chi-square value of a particular distractor (χ_k^2) is significant, then it is considered to be a well-discriminating distractor.

Another useful tool for examining distractors is the choice mean of a distractor, which is the mean of total scores of all the students who choose the distractor (Haladyna & Rodriguez, 2013). For a well-discriminating item, the choice mean of the correct option is expected to be higher than the choice mean of any distractor. If the choice mean of a distractor is higher than the choice mean of the correct option, then the distractor needs to be evaluated for content accuracy. If the choice means of all response options are similar, then the item does not appear to be discriminating adequately.

A more objective way to evaluate the choice mean of a distractor is the point–biserial correlation (i.e., the correlation between a dichotomous item and continuous total test score) or biserial correlation (i.e., the correlation between a latent item score and continuous total test score; Attali & Fraenkel, 2000). Attali and Frankel (2000) pointed out that when computing the point–biserial correlation for a distractor, researchers should contrast students who choose the distractor with

the students who choose the correct option rather than the students who do not choose the distractor. The point–biserial formula is

$$PB_{DC} = \frac{M_D - M_{DC}}{S_{DC}} \sqrt{\frac{P_D}{P_C}},$$

where M_D is the choice mean of distractor *D*, M_{DC} is the mean of the total scores of students who chose the distractor or the correct option, S_{DC} is the standard deviation on the criterion of students who chose the distractor or the correct option, and P_D and P_C are the proportion of students who selected the distractor and the correct option, respectively. According to Attali and Fraenkel (2000), an item with a PB_{DC} value greater than −0.05 should be considered as not discriminating adequately, while any value lower than −0.05 can be accepted.

Compared with CTT distractor analysis, IRT distractor analysis not only assesses whether a distractor is functioning properly but also allows the analyst to use distractors for estimating students' abilities. There are two widely used IRT models for distractor analysis: the nominal-response model (Bock, 1972) and the graded-response model (Samejima, 1979). Bock (1972) proposed the nominal-response model to analyze distractors in multiple-choice items. Instead of traditional IRT models that estimate the probability of responding to an item correctly, the nominal-response model estimates the probability of choosing each multiple-choice option without assuming any ordering among the options. The nominal-response model can be written as

$$P\left(x_j = k \mid \theta\right) = \frac{exp\left(a_k\left(\theta - b_k\right)\right)}{\sum_{h=1}^{m_j} exp\left(a_h\left(\theta - b_h\right)\right)},$$

where $P(x_j = k \mid \theta)$ is the probability of choosing option *k* in item *j* given the student's ability θ (typically ranging from −4 to 4), a_k is the item discrimination for distractor *k*, b_k is the difficulty of distractor *k*, and m_j is the total number of options for item *j*.

The major disadvantage of Bock's (1972) model is that as student ability decreases, the probability of choosing one particular distractor is expected to increase and eventually approach 1. However, this may be unlikely in practice since students with very low ability are likely to guess the correct option randomly. To overcome this limitation, Samejima (1979) proposed a variant of the nominal-response model that takes into account the proportion of students who randomly guess the correct option. The model assumes that there exists a latent category of *don't know* (DK). Students who belong to the DK category will randomly guess, and the probability of guessing is considered when modeling the probability of selecting each option. Samejima's (1979) graded response model is

$$P\left(x_j = k \mid \theta\right) = \frac{exp\left(a_k\left(\theta - b_k\right)\right)}{\sum_{h=1}^{m_j} exp\left(a_h\left(\theta - b_h\right)\right)} + d_k \frac{exp\left(a_0\left(\theta - b_0\right)\right)}{\sum_{h=1}^{m_j} exp\left(a_h\left(\theta - b_h\right)\right)},$$

where d_k is fixed to $1/m_j$ to represent the assumption that students will randomly guess if they belong to latent category *DK*, and

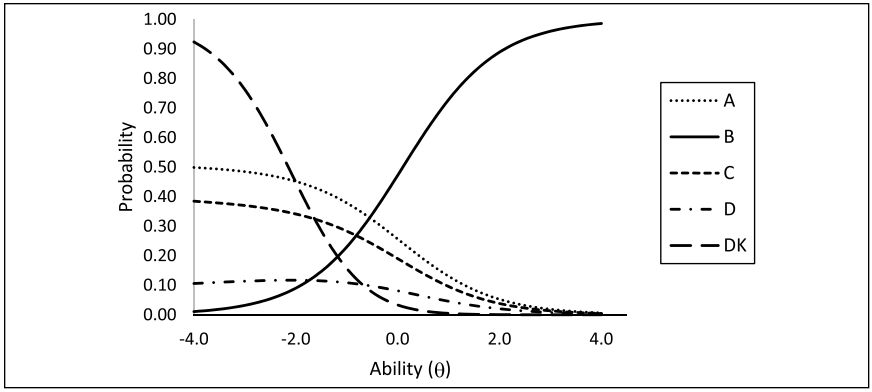

FIGURE 3. *Item characteristic curves for each response option in a multiple-choice item.* *Note.* The correct response option is B; DK: Don't know.

$$\frac{exp\left(a_0\left(\theta - b_0\right)\right)}{\sum_{h=1}^{m_j} exp\left(a_h\left(\theta - b_h\right)\right)}$$

represents the probability a student belongs to latent category DK.

The two IRT models presented above can be visualized using an item characteristic curve (ICC) plot. A sample ICC plot is shown in Figure 3. In this example, Option B is the correct answer. Figure 3 shows that as student ability increases, the probability of choosing Option B increases. Option D is an example of a distractor that attracts low-ability students. As ability increases, the probability of choosing Option D decreases. Option C is an example of a distractor that attracts students with partial knowledge. The probability of choosing Option C peaks at $\theta = -1$, but then decreases for lower or higher ability levels. Option A is an example of a non-discriminating distractor. The probability of choosing Option A is relatively constant across ability levels. For latent category DK, as ability decreases, the probability of belonging to the DK category approaches to 1. Samejima's (1979) model was later extended by Thissen et al. (1989) to allow d_k to be estimated rather than fixed within the model. However, in Thissen et al.'s (1989) model, additional constraints are necessary to estimate d_k. The details can be found in the original article.

In addition to evaluating distractor properties and estimating student ability, the nominal-response model can also be used to identify differential distractor functioning (DDF), which occurs when there are conditional between-group differences in the probabilities associated with each of the distractors (Dorans et al., 1992; Penfield, 2008). For example, using gender as a grouping variable, male and female students with the same ability may have different probabilities of choosing a specific distractor. The mathematical representation of DDF using the nominal-response model is shown as

$$P\left(x_j = k \mid \theta\right) = \frac{exp\left(-\beta_k - \alpha_k\theta - G\omega_k\right)}{1 + \sum_{h=1}^{n_j} exp\left(-\beta_h - \alpha_h\theta - G\omega_h\right)},$$

where $P\left(x_j = k \mid \theta\right)$ is the probability of a student with ability θ choosing distractor k, G is a dichotomous variable, where "0" represents the focal group and "1" represents the reference group, ω_k is the DDF effect, and n_j is the number of distractors (i.e., it does not include the correct option). There are several methods to estimate ω_k and its statistical significance (e.g., Dorans et al., 1992; Penfield, 2008; Thissen, Steinberg, & Wainer, 1993). Of these, only Penfield's (2008) odds ratio approach is capable of estimating DDF effects that are consistent with Bock's (1972) nominal-response model. Penfield (2010a) provided a user-friendly computer program called *DDFS: Differential Distractor Functioning Software* that is capable of estimating DDF effects in multiple-choice items. In a follow-up study, Penfield (2010b) examined the relationship between differential item functioning (DIF) and DDF. In the same vein as DDF, DIF also identifies group differences on multiple-choice items. As opposed to DDF that focuses on distractors, DIF examines whether students from two groups that have similar ability have the same probability of answering an item correctly. Penfield (2010b) suggested that DDF analysis is not only important for checking distractor quality but also important for understanding DIF in general.

In contrast to the nominal-response model that uses options as nominal categories, the partial credit model (Masters, 1982) considers response options as ordinal (i.e., ordered) categories. This implies that a hypothesis about which distractor requires more partial knowledge is required. Such hypotheses can be formed by studying distractor ICCs (Andrich & Styles, 2011) or by constructing the multiple-choice items using a cognitive model (Briggs et al., 2006). The partial credit model can be used to evaluate distractors as follows:

$$P\left(x_j = k \mid \theta\right) = \frac{exp\left(k * \theta - \sum_{i=0}^{k} b_i\right)}{\sum_{g=0}^{m_j} exp\left(k * \theta - \sum_{i=0}^{g} b_i\right)},$$

where the symbols can be interpreted as with the nominal-response model and, in turn, produce the ICC. But when compared with the nominal-response model, the partial credit model has the advantage of having fewer parameters, thereby making it easier to estimate in some testing situations (e.g., when only a small sample size is available).

Contemporary distractor analysis. While the purpose of distractors in traditional multiple-choice educational tests is simply to "distract" students with insufficient ability or partial knowledge, contemporary researchers have begun to see distractors as part of the assessment that may provide useful diagnostic information about students' problem-solving skills (Briggs et al., 2006). For example, in a classroom assessment, distractors selected by students due to their misconceptions can inform the instructor about which skills need to improve in order to eliminate those misconceptions. In order to extract diagnostic information from distractors, cognitive diagnostic models (CDMs) have been proposed to systematically develop and analyze distractors. This section of our review will provide an overview of some CDM approaches to distractor analysis. The focus will be on

conceptual models and applications rather than mathematical or technical details, which can be found in the original articles.

Early attempts at extracting diagnostic information from distractors include Two-Tier Items (Treagust, 1995), the Force Concept Inventory (Hestenes, Wells, & Swackhamer, 1992), and the project STAR Astronomy Concept Inventory (Sadler, 1998). In Two-Tier Items, each item has two tiers. The first tier works like a normal multiple-choice item with four- or five-response options, and the second tier of the item asks students why they selected a particular first tier option. In the Force Concept Inventory developed by Hestenes et al. (1992), students were first given open-ended physics questions, and then distractors were constructed from the most common misconceptions in students' written responses to create the open-ended tasks. The most important contribution from research on the Force Concept Inventory is that it provides an objective way to find out students' misconceptions. However, as pointed out by Briggs et al. (2006), inferences about student misconceptions in the Force Concept Inventory are often based on the response to a single item and such inferences can lack reliability. Similar to the Force Concept Inventory, the distractors in the Project STAR Astronomy Concept Inventory were developed based on open-ended tasks as well as on a literature review and an interview of teachers and students (Sadler, 1998). The most notable contribution of the research from Project STAR is that it was the first study that used IRT to model students' misconceptions. Sadler (1998) found that contrary to the conventional wisdom, the probability of choosing an incorrect response did not necessarily decrease as the ability increased and all items in the inventory had at least one distractor that violated this rule. Sadler (1998) further hypothesized that misconceptions should be perceived as developmental states rather than barriers to learning.

Based on these early CDM approaches, Briggs et al. (2006) proposed the *ordered multiple-choice items* method. Briggs et al. (2006) suggested that distractors should be developed based on a learning model that specifies how students progress through different stages of understanding a particular topic. Such a learning model was referred to as the construct map. To illustrate this approach using a hypothetical example, suppose the content area is multiple-digit integer subtraction. Level 0 of the construct map could be that students do not know how to perform multiple-digit integer subtraction at all. Level 1 could be that students understand how to subtract two multiple-digit integers when each digit of the second number is smaller than its corresponding digit of the first number (e.g., 382 − 131). Level 2 could be that students have partial understanding that borrowing is required when they need to subtract two multiple-digit numbers and some digits of the second number are larger than their corresponding digits in the first number (e.g., 236 − 179). Level 3 could be that students completely understand how to perform two multiple-digit number subtraction using borrowing when some digits of the second number are larger than their corresponding digits in the first number.

In practice, construct maps can be developed from outcomes described in the research literature and from curriculum standards. After developing the construct map, distractors are created based on each level of understanding in the construct map. Continuing from the previous example, for the problem of 236 − 179, a

Level-1 distractor could be 143 since the students with Level 1 understanding can only subtract a smaller number from a larger number. A Level 2 distractor could be 67, since students with Level 2 understanding could not perform borrowing consistently. The Level 3 distractor, in this example, is the correct option, 57.

After developing and administering the ordered multiple-choice items, instructors or researchers need to make inferences about students' misconceptions based on their responses. Briggs et al. (2006) proposed two methods: a practical method and a psychometric method. The practical method simply involves counting how many times a student chooses the distractor of each level. For example, suppose a test consists of 10 items, and each item consists of distractors that correspond to a different level of understanding in the construct map. If a student selects Level 1 distractor one time, and Level 2 distractor eight times, and Level 3 distractor one time, then the student most likely has Level 2 understanding. The psychometric method involves analyzing the data using an ordinal IRT model called the ordered partition model (Wilson, 1992), which can be used to estimate a student's ability and provide the probability of choosing each level distractor given the student's ability. The ordered multiple-choice items method can be limited in practice because it assumes that the learning of a particular content area is based on a single latent ability (i.e., unidimensional) and that it is possible to capture students' strengths and weaknesses using a single latent ability.

To address the limitations of the ordered multiple-choice items method, de la Torre (2009) proposed an attribute-based CDM to develop and analyze distractors. An attribute is a particular knowledge, skill, or cognitive process that is required to correctly solve a type of problem. While attributes may have some overlap with levels of understanding in ordered multiple-choice items, their major difference is that levels of understanding must be organized in a unidimensional ordinal fashion. Attributes, by comparison, do not have to be organized in that specific way. Attributes can be dependent or independent, convergent, or divergent. Attributes have more flexibility than levels of understanding. For example, to correctly solve the problem, 2 + 3* 2, three attributes may be required. Attribute 1 could be the ability to do addition. Attribute 2 could be the ability to do multiplication. Attribute 3 could be the understanding of the order of algebraic operation. In this attribute-based framework, distractors can be developed based on a subset of attributes that are required to solve the item. Using the same example mentioned earlier, one subset of Attributes 1, 2, and 3 could be Attributes 1 and 2. Students with only Attributes 1 and 2 do not know that multiplication needs to be performed before addition; consequently, their answer may be 10, which can be used as a distractor. Similarly, other distractors could be developed based other subsets of Attributes 1, 2, and 3.

To make inferences about students' attribute mastery, de la Torre (2009) modified the *deterministic-input noisy "and" gate* (DINA; Junker & Sijtsma, 2001) model to allow item responses to have more than two categories. This modified model was referred to as the *multiple-choice DINA model*. There are two unique features of the multiple-choice DINA model. First, instead of estimating a unidimensional ability score for each student, the multiple-choice DINA model estimates an attribute profile, which describes the attributes a student has mastered. Second, the multiple-choice DINA model permits inferences about the student's

attribute profile based not only on the items the student correctly answered but also on the distractors the student selected. The multiple-choice DINA model is implemented in the computer software program Ox (Doornik, 2002), which is available free of charge. The program code for multiple-choice DINA model (de la Torre, 2009) can be requested from the author. The multiple-choice DINA model has been used in several recent studies. For example, Ozaki (2015) improved the efficiency of the estimation process of the multiple-choice DINA model by reducing the number of parameters. Huo and de la Torre (2014) extended the multiple-choice DINA model to make inferences based on several plausible cognitive models (i.e., different students may use different attributes to solve the same problems; thus, there are several plausible cognitive models for a problem).

Considerations When Using Distractors

In the previous section, we summarized the literature on how to create distractors for multiple-choice items, and we presented both traditional and contemporary methods for evaluating the quality of these distractors. Next, we present three separate but related considerations on how to use distractors during the test development and test assembly process. We begin with a review of the existing guidelines for distractor use. We then summarize the literature on the optimal number of distractors for a multiple-choice item. We conclude with a review of the literature on the ordering of distractors within an item.

Guidelines for Using Distractors on Educational Tests

Guidelines are detailed instructions or frameworks used by content specialists for creating test items (Haladyna & Rodriguez, 2013). We summarize the outcomes from six published item-writing guidelines that include recommendations for distractors when creating educational tests (i.e., Frey, Petersen, Edwards, Pedrotti, & Peyton, 2005; Haladyna & Downing 1989; Haladyna, Downing & Rodriguez, 2002; Haladyna & Rodriguez, 2013; Moreno et al., 2006, 2015). The results are presented in Table 1.

Haladyna and Downing (1989) developed a taxonomy of 43 multiple-choice item-writing rules from 46 textbooks on classroom assessment and other sources in the educational measurement literature. They specified the rules using "the joint evidence of author's consensus" meaning that Haladyna and Downing reviewed the manuscripts and each identified the rules. The final list consisted of rules identified by both authors. Among the 43 rules, six are directly related to distractor development and use. They are listed as Rule 1 to Rule 6 in Table 1. In a follow-up study, Haladyna et al. (2002) updated their literature review and reorganized the original taxonomy into 31 rules and validated these rules based on the evidence from 27 different textbooks on classroom assessment and 19 empirical studies. Twelve rules out of these 31 rules are applicable to distractors. They are restated as Rules 1, 2, 6, 8, 9, 10, 11, 12, 13, and 14 in Table 1. Frey et al. (2005) analyzed the references from 20 educational assessment textbooks to identify a list of item-writing rules. The rules from Frey et al. (2005) that are applicable to distractors are 1, 3, 8, 9, 10, 11, 12, 13, and 14 in Table 1. One year later, Moreno et al. (2006) proposed a new set of recommendations with 15 rules based on empirical results cited in Hoepfl (1994), Osterlind (1998), Haladyna and Downing

TABLE 1

Summary of the item-writing guidelines for developing distractors in multiple-choice items

	Haladyna and Downing (1989)	Haladyna et al. (2002)	Frey et al. (2005)	Moreno et al. (2006)	Haladyna and Rodriguez (2013)	Moreno et al. (2015)
1. Use plausible distractors	√	√	√	√	√	√
2. Use common errors or misconceptions	√	√			√	√
3. Avoid technically phrased distractors	√		√			√
4. Use familiar yet incorrect phrases						
5. Use true statements that do not correctly answer the stem	√					
6. Avoid the use of humor	√				√	
7. Develop as many effective options as possible	√					
8. Place distractors in logical or numerical order	√	√	√	√	√	√
9. Keep distractors independent, distractors should not overlap	√	√	√	√	√	√
10. Keep distractors homogeneous (content and structure)	√	√	√			√
11. Keep the length of distractors about equal	√	√	√			√
12. "None of the above" and "all of the above" should be used carefully	√	√	√	√	√	√
13. Avoid giving clues to the correct answer	√	√	√	√	√	√
14. Phrase distractors positively, avoid negatives		√	√		√	√

(1989), and Haladyna et al. (2002). Rules 1, 8, 9, 12, and 13 of Table 1 are based on Moreno et al. (2006). Haladyna and Rodriguez (2013) synthesized item-writing guidelines across all of the published studies—including the five studies included in our review—by identifying the common recommendations to produce 22 multiple-choice item-writing rules. Their recommendations correspond to Rules 1, 2, 6, 8, 9, 12, 13, and 14 in Table 1. Most recently, Moreno et al. (2015) proposed a validity-based framework to efficiently organize multiple-choice writing guidelines. In this framework, a valid item should have three properties: representativeness, clarity, and differentiation. Representativeness refers to completeness of the content in the item meaning that all the information required to solve the task is included. Clarity means that the item is clearly presented and, hence, can be easily understood by the examinee. Differentiation is the term used to describe how the content is independent from one item to the next. These three properties were identified by authors based on their review of the previous item-writing guidelines available in the literature. The authors then described and illustrated how items could be created that met each of these three validity-defining properties (see Moreno et al., 2015, for a complete description of their method). The guidelines applicable to distractors are summarized as 1, 2, 3, 8, 9, 10, 11, 12, 13, and 14 in Table 1.

Across the six studies, the rules with the most consensus in Table 1 are "use plausible distractors," "place distractors in logical or numerical order," "keep distractors independent, distractors should not be overlapping," "none-of-the-above and all-of-the-above should be used carefully," and "avoid giving clues to the right answer." Other important rules that produced slightly less consensus included "incorporate common errors of students in distractors," "keep distractors homogeneous in content and grammatical structure," and "phrase distractors positively; avoid negatives such as *not.*" For the most part, however, the guidelines are consistent across studies. Hence, they provide a strong foundation for what can be considered typical guidelines for distractor development and use in educational testing. The only important disagreement among the cited guidelines is related to the recommended number of distractors. Haladyna and Downing (1989) claim that more distractors are desirable. However, other researchers state that a specific number or range of distractors are preferred. Haladyna et al. (2002), Moreno et al. (2006, 2015), and Haladyna and Rodriguez (2013) claim that three response options (i.e., a correct response option and two distractors) is sufficient, whereas Frey et al. (2005) state that the number of distractors can range from two to four. Next, we review the literature on the number of recommended distractors in order to shed some light on this controversial topic.

Optimal Number of Distractors

The number of response options is one of the important characteristics of a multiple-choice item because it may affect the item-writing process (e.g., time, cost, and testing time) as well as the psychometric properties of items (e.g., difficulty, discrimination, and reliability). In most testing situations, it is believed that increasing the number of response options will reduce guessing, thereby making the exam scores more reliable (Haladyna & Downing, 1993; Rodriguez, 2005). For example, the probability of randomly selecting the correct option in a true/

false item is 50% (i.e., one out of two response options), whereas the probability of randomly selecting the correct response option in a multiple-choice item with five options goes down to 20% (i.e., one out of five response options). Also, increasing the number of response options can be helpful in creating more difficult multiple-choice items (Landrum, Cashin, & Theis, 1993; Rogers & Harley, 1999; Sidick, Barrett, & Doverspike, 1994).

The optimal number of response options for multiple-choice items has been widely investigated in the educational testing literature. In Haladyna et al.'s (2002) guidelines on multiple-choice item writing, the authors recommend creating as many plausible distractors as possible to create high-quality, multiple-choice items. However, research has shown that increasing the number of response options may not eliminate random guessing if plausible distractors are lacking (Delgado & Prieto, 1998; Haladyna & Downing, 1993; Rodriguez, 2005). When distractors are not functioning well, students can easily rule out implausible distractors, which improves their chance of finding the correct answer. Therefore, increasing the number of response options may not guarantee reliable and well-functioning multiple-choice items.

Multiple-choice items with four response options (i.e., three distractors and one correct answer) or five response options (four distractors and one correct answer) are recommended by most authors of measurement textbooks and are widely used in educational testing (Delgado & Prieto, 1998; Epstein, 2007; Sidick et al., 1994; Vyas & Supe, 2008). However, some studies have revealed that reducing the number of response options to three can improve the psychometric quality of a multiple-choice item. Earlier studies by Tversky (1964) and Costin (1970) provided mathematical proofs and empirical evidence indicating that items with three response options can have higher discriminative power than items with four or more response options whenever the amount of time spent on the test is proportional to its total number of alternatives. Haladyna et al. (2002) suggested that the use of four or more response options in multiple-choice items might not be desirable because it is challenging for content specialists to create three or more distractors that are highly plausible but still erroneous.

In a comprehensive meta-analysis on multiple-choice testing, Rodriguez (2005) concluded that the use of three response options instead of four or five can help content specialists strengthen several aspects of their validity-related arguments for the multiple-choice item type. The use of three response options can reduce technical flaws in writing multiple-choice items for a number of reasons including the fact that writing three response options instead of four or five response options would be less challenging and less time consuming for content specialists (Vyas & Supe, 2008); it would reduce testing time for students (Cizek, Robinson, & O'Day, 1998; Owen & Froman, 1987); it would lead to more efficient use of testing time and greater score precision per unit testing time (Swanson, Holtzman, Albee, & Clauser, 2006; Swanson, Holtzman, Clauser, & Sawhill, 2005); and it would help eliminate implausible distractors in the items and thus increase the information obtained from the items (Andrés & del Castillo, 1990; Bruno & Dirkzwager, 1995; Landrum et al., 1993; Trevisan, Sax, & Michael, 1991).

The literature on the optimal number of distractors in multiple-choice items emphasizes the importance of creating plausible but incorrect distractors for all

incorrect response options. But researchers have also highlighted the challenge of writing consistently good distractors leading to the suggestion that three response options instead of either four or five may be preferable. Levine and Drasgow (1983) argued that student ability can play an important role in determining the optimal number of response options in multiple-choice items. The amount of information from multiple-choice items can be maximized by using more response options for students with lower ability and fewer response options for students with higher ability. Therefore, the optimal number of distractors could also be determined based on student characteristics (e.g., low or high ability), the item-writing process (e.g., cost, time, availability of plausible distractors), and test administration (e.g., testing time, testing mode). But when a decision is made to reduce the number of options, the first step should always be to identify and remove the implausible and ineffective distractors (Rodriguez, 2005).

Ordering of Distractors

There has also been a considerable amount of research investigating the impact of changing distractor position on the difficulty level of multiple-choice items and test scores. Some researchers focused on the optimal ordering of distractors (e.g., Mosier & Price, 1945; Tellinghuisen & Sulikowski, 2008), whereas the other have evaluated the effects of altering the position of the correct answer relative to the positions of distractors (e.g., Attali & Bar-Hillel, 2003; Bresnoc, Graves, & White, 1989; Cizek, 1994; McNamara & Weitzman, 1945). One of the earliest studies on distractor positions was conducted by McNamara and Weitzman (1945). The researchers investigated whether the position of the correct answer in four- and five-option multiple-choice items had any impact on item difficulty. Items related to Mathematics, Physics, Principles of Flying, Aerology, and Operational of Aircraft Engines were administered to large groups of students from Navy Flight Preparatory Schools. The results from this study suggested that the position of the correct response option can influence the difficulty level of items. An interesting finding from McNamara and Weitzman's (1945) study was that when the correct answer was next to the last response option (i.e., the third position in a four-option item or the fourth position in a five-option item), the items became more difficult. Bresnoc et al. (1989) reported that undergraduate students who took an economics exam performed better when the correct answer was placed in the first position, whereas the same students performed worse when the correct answer was placed in the last position. The researchers argued that students tend to fail to choose the correct answer presented in the last position because reading the preceding distractors may lead to confusion.

In 1993, Huntley and Welch evaluated 32 mathematics items administered to 300 students in the American College Testing pretest sessions. They evaluated the impact of presenting distractors in ascending, descending, and random order on average item difficulty and item discrimination. The authors reported that although ascending, descending, or random ordering of distractors did not result in any significant impact on item difficulty, random ordering of distractors may be a disadvantage for low-ability students, and thus distractors should be placed in logically descending or ascending orders on the test (Huntley & Welch, 1993). However, this recommendation is limited to subject areas such as Mathematics,

Chemistry, and Physics in which distractors are often numerical values. If the content of the items is based on other subject areas in which distractors are words, phrases, or sentences rather than numerical values (e.g., English Language Arts, Science, and History), then the recommendation for either logical or numerical ordering of distractors may not be applicable.

Unlike the studies favoring either logical or numerical ordering of distractors (e.g., Haladyna et al., 2002; Huntley & Welch, 1993), there has also been research highlighting the benefits of randomized ordering for distractors. Because content specialists tend to develop distractors in order of plausibility, the last distractor is often the least tempting one. Or said differently, the plausibility of the distractors is likely to decrease after each distractor is created. To prevent the later distractors from being easily ruled out, Mosier and Price (1945) suggested that the randomization process should include not only the position of the correct answer but also the position of distractors. McLeod, Zhang, and Yu (2003) investigated the effects of randomizing the positions of the response options independently for each student or ordering the distractors logically or numerically within the multiple-choice item. Although the authors hypothesized that logical or numerical ordering of distractors could be an advantage to the students, they did not find any evidence in favor of logical or numerical ordering of distractors. McLeod et al. (2003) concluded that randomizing the position of distractors for every student could be a better option for reducing the possibility of cheating without adversely affecting the psychometric characteristics of the items.

Tellinghuisen and Sulikowski (2008) also examined whether the position of distractors would influence the quality of multiple-choice items. The researchers administered two versions of the American Chemical Society standardized test to 676 college students. Both versions of the test consisted of 70 items with four response options (i.e., one correct answer and three distractors). Their findings indicated that differences in student performance were somewhat related to the positions of the distractors, possibly as a result of the primacy effect. Students performed better on the items in which the correct response option appeared earlier than the distractors. The researchers argued that students tend to select the first response option as correct when all options appear equally attractive.

A subsequent study by Schroeder, Murphy, and Holme (2012) focused on the position of distractors in the American Chemical Society standardized test and found that students are less likely to choose later distractors on conceptual questions. If students find an earlier response option that they believe is correct, then they do not review the other response options carefully. Schroeder et al. (2012) also argued that if the most tempting distractor is placed earlier in the set of response options, students may question their own understanding of the content and thus choose the distractor instead of the correct answer. Therefore, the authors suggested that it is important to randomize the position of the correct answer on conceptual items.

The most common guideline on the position of distractors is to randomize the position of the correct answer relative to the position of distractors because randomization can reduce the possibility of cheating and improve test security, randomization can eliminate any advantage for students who are familiar with the order of content presented in the distractors, and randomization can reduce

guessing. Despite these advantages, researchers also cautioned that randomization may not always be feasible. For instance, Mosier and Price (1945) suggested that the position of distractors should not be randomized for certain types of items, such as items where the response options are based on a meaningful, ordered series (e.g., dates or magnitudes) and items with "none of the above" as one of the options. Furthermore, randomization can result in positioning the correct answer as the first response option too often within a test (McNamara & Weitzman, 1945). When a distractor that is highly similar to the correct answer precedes the correct answer, this can be a disadvantage, especially for low-ability students who are unable to find the correct answer and thus tend to choose the most tempting distractor without checking the other response options (Huntley & Welch, 1993; Marcus, 1963; Schroeder et al., 2012). Cizek (1994) also noted that when the positions of response options are scrambled across test forms, altering the position of the correct answer might result in unpredictable changes in item difficulty. Therefore, the positions of the correct answer and distractors should still be carefully reviewed even when randomization is used.

Discussion

Multiple-choice testing is undeniably a core part of educational testing that ranges from student assessment in today's classrooms and schools to large-scale, high-stakes certification and licensure testing within the professions. Despite using multiple-choice tests at every level of education, developing high-quality, multiple-choice items continues to be a challenge for teachers, educators, and test developers. To date, the majority of research articles, books, chapters, and conference presentations available in the literature have focused on the development, analysis, and use of the stem and the correct response option. However, the distractors, which are a crucial component of the multiple-choice item type, have received much less attention. As a result, distractors development has often been considered the Achilles' heel of the multiple-choice items. The purpose of our review was to synthesize the literature on distractors and to provide a comprehensive summary of how to develop, analyze, and use distractors for multiple-choice items on educational testing.

Recommendations for Practice

Based on our review of the literature, we present six recommendations:

1. Two different recommendations were consistently identified in the literature on how to develop distractors. The first and most common recommendation focused on identifying common misconceptions related to thinking, reasoning, and solving the problem (e.g., Case & Swanson, 2001; Moreno et al., 2015; Rodriguez, 2016; Tarrant et al., 2009). These common misconceptions can be identified either by reviewing responses to constructed-response and open-ended items or by consulting with content specialists (Briggs et al., 2006; Haladyna & Rodriguez, 2013). The second recommendation for developing distractors focused on creating plausible alternatives that are similar in content and structure relative to the correct option (e.g., Ascalon et al., 2007; Guttman et al., 1967; Lai

et al., 2016; Owens et al., 1970; Towns, 2014). Content similarity can be based on many different points of reference including semantic relatedness, key feature, and structural similarities including length, complexity, formatting, and grammar (e.g., Mitkov & Ha, 2003; Mitkov et al., 2009).

2. Once the distractors are created, they must be evaluated to discern their quality and effectiveness at differentiating students who write the test. The recommendations for distractor analysis are mixed and quite diverse but the most common recommendation requires the analyst to review the percentages of students who select each of the distractors in order to identify low-frequency distractors (Haladyna & Downing, 1993). These types of distractors can be either removed from the test or revised by the content specialist. Distractors can also be evaluated visually using trace line plots to identify those that do not differentiate low- and high-achieving students (Wainer, 1989). The chi-square goodness-of-fit test can also be used to determine whether a distractor has a flat trace line (Haladyna & Downing, 1993) meaning that it has low discrimination power.
3. For a more comprehensive and technical analysis of distractors, three methods were consistently identified. First, the nominal-response IRT model and its variants (e.g., Bock, 1972; Samejima, 1979; Thissen et al., 1989) can be used for the visual and statistical analysis of the probability of selecting each distractor depending on the student's ability level. Second, DDF methods (e.g., Dorans et al., 1992; Penfield, 2008, 2010a, 2010b; Thissen et al., 1993) are particularly useful for evaluating whether the distractors in a multiple-choice item function similarly across subgroups of students (e.g., male students vs. female students). Third, the CDM approaches (e.g., Briggs et al., 2006; de la Torre, 2009) can be used to extract diagnostic information from the items (including the distractors) by mapping out the cognitive attributes or levels of understanding required to solve each test item.
4. There are many item-writing guidelines in the published literature, most of which include a small number of recommendations for developing and using distractors. The recommendations are relatively consistent with one another. They include the following: (a) use plausible distractors in multiple-choice items, (b) place distractors in logical order, (c) keep the content within the distractors independent of one another, (d) none-of-the-above and all-of-the-above should be used carefully, (e) avoid providing inadvertent clues to the correct option in the distractors, (f) incorporate common errors of students in distractors, (g) keep distractors homogeneous in content and grammatical structure, and (h) phrase distractors positively (e.g., Frey et al., 2005; Haladyna & Downing, 1989; Haladyna et al., 2002; Moreno et al., 2006, 2015).
5. Research on the optimal number of distractors has resulted in relatively clear recommendations (e.g., Delgado & Prieto, 1998; Haladyna et al., 2002; Rodriguez, 2005). While some researchers recommend three or more distractors for multiple-choice items (Delgado & Prieto, 1998; Epstein, 2007; Sidick et al., 1994; Vyas & Supe, 2008), there is evidence and consensus that using two distractors and a correct response option is

optimal (Cizek et al., 1998; Rodriguez, 2005; Swanson et al., 2005; Swanson et al., 2006; Vyas & Supe, 2008). Using two distractors rather than three or more is not only more efficient from a development standpoint but also it is preferable for students because it reduces testing time.

6. The optimal approach for ordering distractors in multiple-choice items has also been investigated (e.g., Cizek, 1994; Haladyna et al., 2002; Mosier & Price, 1945; Schroeder et al., 2012). But, unlike the literature on the optimal number of distractors, there is no clear consensus among researchers about how distractors should be ordered. One of the recommendations is to position distractors in a logical or numerically descending or ascending order on the test (Haladyna et al., 2002; Huntley & Welch, 1993). But this approach is limited to content areas such as mathematics. The other recommendation is to randomize the order of distractors, which can reduce the possibility of cheating and improve test security (McLeod et al., 2003; Mosier & Price, 1945; Schroeder et al., 2012).

Expanding Methods for Distractor Development

Increasingly, educational testing organizations are required to create large numbers of diverse, high-quality multiple-choice items. For the most part, multiple-choice item development is still conducted using a traditional approach where a content specialist writes every item. To implement this approach for distractor development, we presented two general strategies. The first strategy focuses on creating a list of plausible but incorrect alternatives linked to common misconceptions or errors in thinking, reasoning, and problem solving. The second strategy focuses on creating plausible but incorrect alternatives that are similar in content and structure to the correct option.

Of these two strategies, the first is the most common recommendation. Misconceptions can be anchored to empirical results by reviewing the solutions from similar constructed-response items or from studies of student response processes using verbal reports. Unfortunately, the feasibility of collecting empirical data in the form of construct responses or verbal reports is limited, particularly when large numbers of diverse items must be created quickly and economically. Misconceptions can also be anchored to judgmental results by asking content specialists to identify common errors in thinking, reasoning, and problem solving. This judgmental approach is typically used in practice. But to successfully implement this judgmental approach to distractor development, at least three assumptions must be satisfied. First, plausible algorithms, rules, or procedures must be specified by content specialists. Second, plausible but incorrect distractors must be produced using these rules. Third, the misconception identified by the content specialists are, in fact, the same misconceptions held by the students. Proper alignment of the assumptions is critical for creating distractors that measure plausible misconceptions. Moreover, the alignment must occur for each distractor across every multiple-choice item. For example, if a content specialist writes 100 multiple-choice items and each item contains five options (i.e., one correct option and four distractors), then the content specialist must identify 400 plausible but incorrect alternatives that satisfy the three assumptions to serve as reasonable or believable errors in students' thinking, reasoning, and problem solving. If the task

is to create an item bank, for instance, with 3,000 five-option multiple-choice items, then the content specialist needs to write 12,000 options that satisfy the three assumptions to yield plausible distractors. We contend that writing distractors that measure plausible misconceptions and replicating this outcome consistently over large number of multiple-choice items is a challenging task. Hence, an important direction for future research is to *identify and evaluate a much broader variety of methods for creating distractors*. We identified and described two primary methods for distractor development in our review. However, more methods are needed for developing high-quality distractors. These methods could be based on different assumptions about student performance and they should yield a large number of distractors.

One promising area of research that could be used to identify and evaluate new methods for distractor development is with *automatic item generation* (AIG). AIG is a relatively new but rapidly evolving research area where cognitive theory and psychometric practice guide the production of items that are generated with the aid of computer technology (Gierl & Haladyna, 2013; Irvine & Kyllonen, 2002). Gierl and Lai (2013) described a three-step process for generating multiple-choice test items. In Step 1, content specialists identify the tasks for item generation. In Step 2, an item model is developed to specify where the content from Step 1 must be placed to generated new items. An item model is like a template that highlights the features in an item that must be manipulated to generate items. In Step 3, computer-based algorithms place the content specified in Step 1 into the item model developed in Step 2. AIG researchers are also faced with the challenging task of creating distractors for multiple-choice items because each generated item must include a stem with both a corresponding correct option and a set of incorrect options. These incorrect options could be designed from a list of plausible but incorrect alternatives linked to common misconceptions or created from plausible alternatives that are similar in content to the correct option, as recommendation in the literature. However, these two approaches for creating distractors have proven to be infeasible because AIG is a much more complex assembly task compared with traditional item development.

The content in a multiple-choice item is constantly changing in a generative item development system. While a small number of constraints are needed to ensure that information presented in the stem yields a correct response, this requirement must be counterbalanced with a much larger number of constraints that are needed to ensure the information presented in the distractors is plausible yet erroneous. Also, one correct option is required for a multiple-choice item. But three or four incorrect options must be produced for each item and then scaled across many items because one AIG item model typically yields hundreds or thousands of generated items. For instance, Gierl, Lai, and Turner (2012) described the development of one medical item model in the area of general surgery that generated 1,248 items. More recently, Gierl, Lai, Hogan, and Matovinovic (2015) described the development of 18 different mathematics item models that generated 109,300 items, meaning that each model produced, on average, 6,072 items. Because of the complexity and magnitude of the generative task, the recommendations we presented for distractor development using a traditional approach typically cannot be used for generating distractors. Instead, researchers have developed

new approaches for distractor development. Next, we provide two examples of contemporary distractor development that could serve as feasible alternative methods for creating multiple-choice distractors.

Systematic Distractor Development Using Key Features

Popham (2008) noted that solving multiple-choice items requires students to distinguish among options that differ in their *relative correctness*. That is, multiple-choice items require students to make subtle but meaningful distinctions among options, several of which may be partially correct. To build on the concept that multiple-choice items contain options differing in relative correctness, the distractors can be created using *key features*. The basic logic of this approach requires that the key features required to produce the correct option are first identified and then these same features are used again to construct the distractors. The strategy of creating distractors using key features was first proposed by Guttman et al. (1967). The authors, reacting to what they believed was an arbitrary approach to distractor development reminiscent of strategies still used today, stated, "Questions are typically constructed by a kind of trial-and-error procedure where the intuition of the investigator is subsequently checked by some form of item analysis" (p. 570). As an alternative, Guttman et al. (1967) described a more systematic approach based on the following logic:

> Distractors usually vary in the degree of their attraction for the respondent, namely the proportion of respondents who choose it when they don't choose the correct answer. It may be hypothesized that the degree of attraction of a distractor increases monotonely with its "degree of similarity" to the correct answer. A viable a priori definition of "degree of similarity" is one based only on content considerations, yet successfully predicts empirical attractiveness. (p. 571)

By "degree of similarity," Guttman et al. (1967) meant the number of key features shared between the correct response and the distractor (see p. 572 for illustration). In other words, distractors could be created by first defining the key feature in the correct option and then systematically removing one of more of these features.

Lai et al. (2016), building on the logic presented in Guttman et al. (1967), recently proposed an AIG method for systematic distractor development using key features. We illustrate Lai et al.'s (2016) approach with an example drawn from the content area of general surgery where the student is required to diagnose problems that could arise from a serious abdominal injury. To begin, the key features for the correct option—splenic rupture in this example—are identified. Five key features could be identified based on the structure of the test item where specific variables in this task—type of accident, hemodynamics, side, air entry, and Foley output—are manipulated to produce the correct option (see Figure 4). Splenic rupture is the correct option when the content for each key feature includes "highway speed roll over" (Type of Accident), "blood pressure is 75/35 and heart rate is 140" (Hemodynamics), "left side" (Side), "good air entry and a large distended abdomen with guarding" (Air Entry), and "100 cc of bloody urine" (Foley Output).

A 25-year-old male is involved in a highway speed roll over (TYPE OF ACCIDENT). Emergency Medical Services (EMS) resuscitates him with 2L crystalloid and transports him to your tertiary centre. When he arrives his blood pressure is 75/35 and his heart rate is 140 (HEMODYNAMICS). He has a Glasgow Coma Scale score of 14. He is complaining of lower-rib pain on his left side (SIDE). On examination, he has good air entry and a large distended abdomen with guarding (AIR ENTRY). A foley catheter emits 100cc of bloody urine (FOLEY OUTPUT).

What is the most likely diagnosis?

A. Aortic rupture
B. Pneumothorax
C. Splenic rupture *
D. Cardiac tamponade
E. Diaphragmatic rupture

FIGURE 4. *An example of systematic distractor development using key features.*

The same logic is also used to produce the incorrect options. Diaphragmatic rupture, as an example, shares some, but not all, of the key features with splenic rupture. For instance, a diaphragmatic rupture can be associated with the key features "highway speed roll over" (Type of Accident), "blood pressure is 75/35 and heart rate is 140" (Hemodynamics), and "100 cc of bloody urine" (Foley Output). But a diaphragmatic rupture is not associated with pain on either the left or the ride side (Side) and does not necessarily display physical examination results related to air entry and abdominal pain (Air Entry). As a result, diaphragmatic rupture is a plausible but erroneous option because it shares some but not all of the key features related to splenic rupture. This distractor will be an appealing option for a student who has partial knowledge because the distractor shares some of the key feature with the correct option. The same logic is used to identify other attractive but incorrect options that share some but not all of the key features of the correct option (e.g., aortic rupture, pneumothorax, cardiac tamponade).

The method described by Lai et al. (2016) serves as a general approach to distractor development because it can be used in diverse testing situations and across many content areas. To use this approach for distractor development, the key features leading to the correct options are first specified. Then, this list of key features is used to create plausible distractors that contain some but not all of these features. The most important characteristic of this method is that *it can be evaluated.* That is, the accuracy and plausibility of the key features can be identified and then verified by content specialists, thereby providing evidence and consensus about why a distractor could serve as a plausible incorrect option. With systematic distractor development using key features, the content specialist's task is not to identify why a student may think a distractor is correct based on a misconception but rather to identify distractors based on their relationship to key features in the correct response. The accuracy of the key features and, hence, the quality and accuracy of the distractor created using the key features can therefore be evaluated.

Gierl et al. (2016) evaluated the psychometric properties (i.e., difficulty, discrimination, and usefulness of the distractors) for automatically generated

medical items using the key features approach for distractor development. The generated items were administered to a sample of 455 medical students. The results indicate that the generated items produced a range of difficulty levels for the correct option while providing a consistently high level of discrimination. The distractors served as effective alternatives that contained information appealing to low-performing students resulting in differentiated options for each AIG item. Only 3 of the 110 generated distractors were not selected by any of the students. The other 107 generated distractors effectively differentiated low- from the high-performing medical student.

Systematic Distractor Development Using Content Similarity

Another promising strategy for distractor development, first reported in Mitkov and Ha (2003) and then expanded on in Mitkov, Ha, and Karamanis (2006), can be described as systematic distractor development using content similarity. Mitkov et al. (2006) presented a method for generating multiple-choice items using natural language processing technology focused on the premise that "distractors should be as semantically close as possible to the answer" (p. 179). To begin, key terms, concepts, and noun phrases are identified in electronic textbooks. Textbook sentences that contain these key terms are then used to create the stem and the correct option. Next, plausible distractors are identified for each stem and correct option by conducting a search using a lexical database like WordNet in order to find words that are semantically close to the correct option. If the database returns too many results, then the words appearing in the textbooks are given priority when developing distractors. If the database returns no results, then the textbook is searched for noun phrases with the same "head" and the results are used as distractors. For example, suppose the task presents the symptoms of a patient who suffers from major depression. Students are required to make a diagnosis by selecting the option that best describes the symptoms. Knowing that the correct answer is *major* depression, "depression" is the head of the key term, and a search is performed in the electronic textbook to identify noun phrases that have the head "depression." The resulting distractors might be *catatonic* depression, *chronic* depression, or *melancholic* depression because these incorrect options could serve as semantically close concepts to the correct options. With this approach, the list of distractors that is identified will be appealing options for students who have partial knowledge because the distractors serve as related but erroneous concept relative to the correct answer. Large lists of plausible distractors can be quickly identified using the database search.

The method used by Mitkov et al. (2006) serves as a general approach for distractor development because it can be used in diverse testing situations and across many content areas, as long as a database or corpora is available to guide the lexical and conceptual search. A lexical database like WordNet groups words into synonyms, provides definitions, and records quantitative relationship among the synonym to facilitate automatic text analysis. Increasingly, large databases across many disciplines and content areas can be accessed and used for this type of text and conceptual search. Systematic distractor development using content similarity is guided by key features that help identify correct options. These same key features are also used to create plausible distractors. The accuracy and plausibility

of this method can be evaluated in two ways. First, the sematic relatedness between the correct option and the distractors can be computed using sematic similarity measures such as the Lesk algorithm (Lesk, 1986), the Leacock–Chodorow measure (Leacock & Chodorow, 1998), and the Lin measure (Lin, 1997). Second, the list of distractors can be reviewed by content specialists, thereby providing consensus that a set of incorrect options is plausible and appropriate for a given multiple-choice item. Mitkov et al. (2006) compared the psychometric properties of 18 items generated with a computer to 12 items written by content specialists. The items were administered to 78 students in an undergraduate linguistics class. They reported the computer-generated items had comparable or better quality than the traditional items using measures of item difficulty, item discrimination, and distractor usefulness (see Mitkov et al., 2006).

To summarize, we identified two general strategies for distractor development in our review. But we also noted that it could be challenging to consistently apply these methods over large numbers of multiple-choice items in order to produce high-quality distractors for educational testing. To address these limitations, we described two alternative methods—systematic distractor development using key features and content similarity—that could be used to create large numbers of high-quality distractors. But more research is needed with the two specific methods we presented. Also, a broader range of distractor development procedures is needed. Hence, an important direction for future research is to identify and evaluate a more diverse range or methods for creating multiple-choice distractors.

References

References marked with an asterisk were included in this review.

*Andrés, A. M., & del Castillo, J. D. (1990). Multiple-choice tests: Power, length, and optimal number of choices per item. *British Journal of Mathematical and Statistical Psychology*, *43*, 57–71. doi:10.1111/j.2044-8317.1990.tb00926.x

Andrich, D., & Styles, I. (2011). Distractors with information in multiple choice items: A rationale based on the Rasch model. *Journal of Applied Measurement*, *12*, 67–95.

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

*Attali, Y., & Bar-Hillel, M. (2003). Guess where: The position of correct answers in multiple-choice test items as a psychometric variable. *Journal of Educational Measurement*, *40*, 109–128. doi:10.1111/j.1745-3984.2003.tb01099.x

*Attali, Y., & Fraenkel, T. (2000). The point-biserial as a discrimination index for distractors in multiple-choice items: Deficiencies in usage and an alternative. *Journal of Educational Measurement*, *37*, 77–86. doi:10.1111/j.1745-3984.2000.tb01077.x

Bishara, A. J., & Lanzo, L. A. (2015). All of the above: When multiple correct response options enhance the testing effect. *Memory*, *23*, 1013–1028. doi:10.1080/09658211.2014.946425

*Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, *37*, 29–51. doi:10.1007/BF02291411

*Bresnoc, A. E., Graves, P. E., & White, N. (1989). Multiple-choice testing: Question and response position. *Journal of Economic Education*, *20*, 239–245. doi:10.2307/1182299

*Briggs, D. C., Alonzo, A. C., Schwab, C., & Wilson, M. (2006). Diagnostic assessment with ordered multiple-choice items. *Educational Assessment*, *11*, 33–63. doi:10.1207/s15326977ea1101_2

Brown, A. S., Schilling, H. E., & Hockensmith, M. L. (1999). The negative suggestion effect: Pondering incorrect alternatives may be hazardous to your knowledge. *Journal of Educational Psychology*, *91*, 756–764. doi:10.1037/0022-0663.91.4.756

*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*, 959–966. doi:10.1177/0013164495055006004

Butler, A. C., Marsh, E. J., Goode, M. K., & Roediger, H. L., III. (2006). When additional multiple-choice lures aid versus hinder later memory. *Applied Cognitive Psychology*, *20*, 941–956. doi:10.1002/acp.1239

Butler, A. C., & Roediger, H. L., III. (2008). Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing. *Memory & Cognition*, *36*, 604–616. doi:10.3758/MC.36.3.604

*Case, S. M., & Swanson, D. B (2001). *Constructing written test questions for the basic and clinical sciences* (3rd ed.). Philadelphia, PA: National Board of Medical Examiners.

Chingos, M. M. (2012). *Strength in numbers: State spending on K–12 assessment systems*. Washington, DC: Brown Center on Education Policy, Brookings Institution.

*Cizek, G. J. (1994). The effect of altering the position of options in a multiple-choice examination. *Educational and Psychological Measurement*, *54*, 8–20. doi:10.1177/0013164494054001002

*Cizek, G. J., Robinson, K. L., & O'Day, D. (1998). Nonfunctioning options: A closer look. *Educational and Psychological Measurement*, *58*, 605–611. doi:10.1177/0013164498058004004

*Collins, J. (2006). Writing multiple-choice questions for continuing medical education activities and self-assessment modules. *Radiographics*, *26*, 543–551. doi:10.1148/rg.262055145

*Costin, F. (1970). The optimal number of alternatives in multiple-choice achievement tests: Some empirical evidence for a mathematical proof. *Educational and Psychological Measurement*, *30*, 353–358. doi:10.1177/001316447003000217

*de la Torre, J. (2009). A cognitive diagnosis model for cognitively based multiple-choice options. *Applied Psychological Measurement*, *33*, 163–183. doi:10.1177/0146621608320523

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

*Doornik, J. A. (2002). Object-oriented matrix programming using Ox (Version 3.1) [Computer software]. London, England: Timberlake Consultants Press.

*Dorans, N. J., Schmitt, A. P., & Bleistein, C. A. (1992). The standardization approach to assessing comprehensive differential item functioning. *Journal of Educational Measurement*, *29*, 309–319. doi:10.1111/j.1745-3984.1992.tb00379.x

Downing, S. M. (2006a). Selected-response item formats in test development. In S. M. Downing & T. Haladyna (Eds.), *Handbook of test development* (pp. 287–302). Mahwah, NJ: Erlbaum.

Downing, S. M. (2006b). Written tests: Constructed-response and selected-response formats. In S. M. Downing & R. Yudkowsky (Eds.), *Assessment in health professions education* (pp. 149–184). New York, NY: Routledge.

*Epstein, R. M. (2007). Assessment in medical education. *New England Journal of Medicine*, *356*, 387–396. doi:10.1056/NEJMra054784

Fazio, L. K., Agarwal, P. K., Marsh, E. J., & Roediger, H. L., III. (2010). Memorial consequences of multiple-choice testing on immediate and delayed tests. *Memory & Cognition*, *38*, 407–418. doi:10.3758/MC.38.4.407

*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. doi:10.1016/j.tate.2005.01.008

*Gierl, M. J., & Haladyna, T. (2013). *Automatic item generation: Theory and practice*. New York, NY: Routledge.

*Gierl, M. J., & Lai, H. (2013). Using automated processes to generate test items. *Educational Measurement*, *32*, 36–50. doi:10.1111/emip.12018

*Gierl, M. J., Lai, H., Hogan, J., & Matovinovic, D. (2015). A method for generating test items that are aligned to the common core state standards. *Journal of Applied Testing Technology*, *16*, 1–18.

*Gierl, M. J., Lai, H., Pugh, D., Touchie, C., Boulais, A.-P., & De Champlain, A. (2016). Evaluating the psychometric characteristics of generated multiple-choice test items. *Applied Measurement in Education*, *29*, 196–210. doi:10.1080/08957347.2016.1171768

*Gierl, M. J., Lai, H., & Turner, S. (2012). Using automatic item generation to create multiple-choice items for assessments in medical education. *Medical Education*, *46*, 757–765. doi:10.1111/j.1365-2923.2012.04289.x

*Guttman, L., Schlesinger, I. M., & Schlesinger, L. M. (1967). Systematic construction of distractors for ability and achievement test items. *Educational and Psychological Measurement*, *27*, 569–580. doi:10.1177/001316446702700301

Haladyna, T. M. (2004). *Developing and validating multiple-choice test items*. Mahwah, NJ: Lawrence Erlbaum.

*Haladyna, T. M. (2016). Item analysis for selected-response test items. In S. Lane, M. Raymond, & T. Haladyna (Eds.), *Handbook of test development* (2nd ed., pp. 392–409). New York, NY: Routledge.

*Haladyna, T. M., & Downing, S. M. (1989). Validity of a taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education*, *2*, 37–50. doi:10.1207/s15324818ame0201_4

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

*Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education*, *15*, 309–333. doi:10.1207/S15324818AME1503_5

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

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

*Hestenes, D., Wells, M., & Swackhamer, G. (1992). Force concept inventory. *The Physics Teacher*, *30*, 141–151. doi:10.1119/1.2343497

*Hoepfl, M. C. (1994). Developing and evaluating multiple choice tests. *Technology Teacher*, *53*(7), 25–26.

*Hoshino, Y. (2013). Relationship between types of distractor and difficulty of multiple-choice vocabulary tests in sentential context. *Language Testing in Asia*, *3*, 1–14. doi:10.1186/2229-0443-3-16

*Huntley, R., & Welch, C. J. (1993, April). *Numerical answer options: Logical or random order?* Paper presented at the annual of meeting of the American Educational Research Association, Atlanta, GA.

*Huo, Y., & de la Torre, J. (2014). Estimating a cognitive diagnostic model for multiple strategies via the EM algorithm. *Applied Psychological Measurement*, *38*, 464–485. doi:10.1177/0146621614533986

*Irvine, S. H., & Kyllonen, P. C. (2002). *Item generation for test development*. Hillsdale, NJ: Erlbaum.

*Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, *25*, 258–416. doi:10.1177/01466210122032064

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

*Lai, H., Gierl, M. J., Touchie, C., Pugh, D., Boulais, A., & De Champlain, A. (2016). Using automatic item generation to improve the quality of MCQ distractors. *Teaching and Learning in Medicine*, *28*, 166–173. doi:10.1080/10401334.2016.1146608

*Landrum, R. E., Cashin, J. R., & Theis, K. S. (1993). More evidence in favor of three-option multiple choice tests. *Educational and Psychological Measurement*, *53*, 771–778. doi:10.1177/0013164493053003021

*Lau, P. N. K., Lau, S. H., Hong, K. S., & Usop, H. (2011). Guessing, partial knowledge, and misconceptions in multiple-choice tests. *Educational Technology & Society*, *14*(4), 99–110. Retrieved from http://www.ifets.info/journals/14_4/10.pdf

*Leacock, C., & Chodorow, M. (1998). Combining local context and WordNet similarity for word sense identification. In C. Fellbaum (Ed.), *WordNet: An electronic lexical database* (pp. 265–283). Cambridge: MIT Press.

Lemann, N. (1999). *The big test: The secret history of the American meritocracy*. New York, NY: Farrar, Straus & Giroux.

*Lesk, M. (1986). Automatic sense disambiguation: How to tell a pine cone from an ice cream cone. In *Proceedings of the 1986 SIGDOC Conference* (pp. 24–26), Toronto, Ontario, Canada.

*Levine, M. V., & Drasgow, F. (1983). Appropriateness measurement: Validating studies and variable ability models. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 109–131). New York, NY: Academic Press.

*Lin, D. (1997). Using syntactic dependency as a local context to resolve word sense ambiguity. Paper presented at the proceedings of the 35th annual meeting of the Association for Computational Linguistics, Madrid, Spain.

Little, J. L., & Bjork, E. L. (2015). Optimizing multiple-choice tests as tools for learning. *Memory & Cognition*, *43*, 14–26. doi:10.3758/s13421-014-0452-8

Little, J. L., Bjork, E. L., Bjork, R. A., & Angello, G. (2012). Multiple-choice tests exonerated, at least of some charges fostering test-induced learning and avoiding test-induced forgetting. *Psychological Science*, *23*, 1337–1344. doi:10.1177/0956797612443370

*Marcus, A. (1963). The effect of correct response location on the difficulty level of multiple-choice questions. *Journal of Applied Psychology*, *47*, 48–51. doi:10.1037/h0042018

Marsh, E. J., Roediger, H. L., III, Bjork, R. A., & Bjork, E. L. (2007). The memorial consequences of multiple-choice testing. *Psychonomic Bulletin & Review*, *14*, 194–199. doi:10.3758/BF03194051

*Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149–174. doi:10.1007/BF02296272

*McLeod, I., Zhang, Y., & Yu, H. (2003). Multiple-choice randomization. *Journal of Statistics Education*, *11*(1). Retrieved from http://ww2.amstat.org/publications/jse/v11n1/mcleod.html

*McNamara, W. J., & Weitzman, E. (1945). The effect of choice placement on the difficulty of multiple choice questions. *Journal of Educational Psychology*, *36*, 103−113. doi:10.1037/h0060835

*Mitkov, R., & Ha, L. A. (2003). *Computer-aided generation of multiple-choice tests*. Paper presented at the proceedings of the HLT/NAACL 2003 Workshop on Building Educational Applications Using Natural Language Processing, Edmonton, Canada.

*Mitkov, R., Ha, L. A., & Karamanis, N. (2006). A computer-aided environment for generating multiple-choice test items. *Natural Language Engineering*, *12*, 177–194. doi:10.1017/S1351324906004177

*Mitkov, R., Ha, L. A., Varga, A., & Rello, L. (2009, March). Semantic similarity of distractors in multiple-choice tests: Extrinsic evaluation. In *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics* (pp. 49–56). Athens, Greece: Association for Computational Linguistics.

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

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

*Mosier, C. I., & Price, H. G. (1945). The arrangement of choices in multiple choice questions and a scheme for randomizing choices. *Educational and Psychological Measurement*, *5*, 379−382. doi:10.1177/001316444500500405

Mullis, I. V. S., Cotter, K. E., Fishbein, B. G., & Centurino, V. A. S. (2016). Developing the TIMSS advanced 2015 achievement items. In M. O. Martin, I. V. S. Mullis, & M. Hooper (Eds.), *Methods and procedures in TIMSS 2015* (pp. 1.1–1.17). Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

OECD. (2016). *PISA 2015 assessment and analytical framework: Science, reading, mathematic and financial literacy*. Paris, France: OECD Publishing. doi:10.1787/9789264255425-en

Odegard, T. N., & Koen, J. D. (2007). "None of the above" as a correct and incorrect alternative on a multiple-choice test: Implications for the testing effect. *Memory*, *15*, 873–885. doi:10.1080/09658210701746621

Olson, L. (2005). State test programs mushroom as NCLB mandates as kicks in. *Education Week*, *25*(13), 10–12. Retrieved from http://www.edweek.org/media/13testing.pdf

*Osterlind, S. J. (1998). *Constructing test items: Multiple-choice, constructed response, performance, and other formats* (2nd ed.). Boston, MA: Kluwer Academic.

*Owen, W. V., & Froman, R. D. (1987). What's wrong with three option multiple choice items? *Educational and Psychological Measurement*, *47*, 513–522. doi:10.1177/0013164487472027

*Owens, R. E., Hanna, G. S., & Coppedge, F. L. (1970). Comparison of multiple-choice tests using different types of distractor selection techniques. *Journal of Educational Measurement*, *7*, 87–90. doi:10.1111/j.1745-3984.1970.tb00700.x

*Ozaki, K. (2015). DINA models for multiple-choice items with few parameters: Considering incorrect answers. *Applied Psychological Measurement*, *39*, 431–447. doi:10.1177/0146621615574693

*Penfield, R. D. (2008). An odds ratio approach for assessing differential distractor functioning effects under the nominal response model. *Journal of Educational Measurement*, *45*, 247–269. doi:10.1111/j.1745-3984.2008.00063.x

*Penfield, R. D. (2010a). DDFS: Differential distractor functioning software. *Applied Psychological Measurement*, *34*, 646–647. doi:10.1177/0146621610375690

*Penfield, R. D. (2010b). Modeling DIF effects using distractor-level invariance effects: Implications for understanding the causes of DIF. *Applied Psychological Measurement*, *34*, 151–165. doi:10.1177/0146621609359284

*Popham, W. J. (2008). *Classroom assessment: What teachers need to know*. Boston, MA: Allyn & Bacon.

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

*Rodriguez, M. C. (2011). Item-writing practice and evidence. In S. N. Elliott, R. J. Kettler, P. A. Beddow, & A. Kurz (Eds.), *Handbook of accessible achievement tests for all student: Bridging the gaps between research, practice, and policy* (pp. 201–216). New York, NY: Springer.

*Rodriguez, M. C. (2016). Selected-response item development. In S. Lane, M. Raymond, & T. Haladyna (Eds.), *Handbook of test development* (2nd ed., pp. 259–273). New York, NY: Routledge.

Roediger, H. L., & Karpicke, J. D., III. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, *17*, 249–255. doi:10.1111/j.1467-9280.2006.01693.x

Roediger, H. L., III, & Marsh, E. J. (2005). The positive and negative consequences of multiple-choice testing. *Journal of Experimental Psychology*, *31*, 1155–1159. doi:10.1037/0278-7393.31.5.1155

Rogers, T. B. (1995). *The psychological testing enterprise: An introduction*. Belmont, CA: Wadsworth.

*Rogers, W. T., & Harley, D. (1999). An empirical comparison of three- and four-choice items and tests: Susceptibility to testwiseness and internal consistency reliability. *Educational and Psychological Measurement*, *59*, 234–247. doi:10.1177/00131649921969820

*Sadler, P. M. (1998). Psychometric models of student conceptions in science: Reconciling qualitative studies and distractor-driven assessment instruments. *Journal of Research in Science Teaching*, *35*, 265–296. doi:10.1002/(SICI)1098-2736(199803)35:3<265::AID-TEA3>3.0.CO;2-P

*Samejima, F. (1979). *A new family of models for the multiple choice item* (Research Rep. No. 79-4). Knoxville, TN: University of Tennessee.

*Schmeiser, C. B., & Welch, C. J. (2006). Test development. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 307–353). Westport, CT: National Council on Measurement in Education and American Council on Education.

*Schroeder, J., Murphy, K. L., & Holme, T. A. (2012). Investigating factors that influence item performance on ACS exams. *Journal of Chemical Education*, *89*, 346−350. doi:10.1021/ed101175f

*Sidick, J. T., Barrett, G. V., & Doverspike, D. (1994). Three-alternative multiple choice tests: An attractive option. *Personnel Psychology*, *47*, 829–835. doi:10.1111/j.1744-6570.1994.tb01579.x

*Swanson, D. B., Holtzman, K. Z., Albee, K., & Clauser, B. E. (2006). Psychometric characteristics and response times for content-parallel extended-matching and one-best-answer items in relation to number of options. *Academic Medicine*, *81*, 52–55. doi:10.1097/01.ACM.0000236518.87708.9d

*Swanson, D. B., Holtzman, K. Z., Clauser, B. E., & Sawhill, A. J. (2005). Psychometric characteristics and response times for one-best-answer questions in relation to number and sources of options. *Academic Medicine*, *80*, 93–96. doi:10.1097/00001888-200510001-00025

*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), 1–8. doi:10.1186/1472-6920-9-40

*Tellinghuisen, J., & Sulikowski, M. M. (2008). Does the answer order matter on multiple-choice exams? *Journal of Chemical Education*, *85*, 572–575. doi:10.1021/ed085p572

*Thissen, D., Steinberg, L., & Fitzpatrick, A. R. (1989). Multiple-choice models: The distractors are also part of the item. *Journal of Educational Measurement*, *26*, 161–176. doi:10.1111/j.1745-3984.1989.tb00326.x

*Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67–113). Hillsdale, NJ: Lawrence Erlbaum.

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

*Treagust, D. F. (1995). Diagnostic assessment of students' science knowledge. In S. M. Glynn & R. Duit (Eds.), *Learning science in the schools: Research reforming practice* (pp. 327–346). Mahwah, NJ: Erlbaum.

*Trevisan, M. S., Sax, G., & Michael, W. B. (1991). The effects of the number of options per item and student ability on test validity and reliability. *Educational and Psychological Measurement*, *51*, 829–837. doi:10.1177/001316449105100404

*Tversky, A. (1964). On the optimal number of alternatives at a choice point. *Journal of Mathematical Psychology*, *1*, 386–391. doi:10.1016/0022-2496(64)90010-0

*Vacc, N. A., Loesch, L. C., & Lubik, R. E. (2001). Writing multiple-choice test items. In G. R. Walz & J. C. Bleuer (Eds.), *Assessment: Issues and challenges for the millennium* (pp. 215–222). Greensboro, NC: ERIC Clearinghouse on Counseling and Student Services. Retrieved from http://files.eric.ed.gov/fulltext/ED457440.pdf

*Vyas, R., & Supe, A. (2008). Multiple choice questions: A literature review on the optimal number of options. *National Medication Journal of India*, *21*, 130–133.

*Wainer, H. (1989). The future of item analysis. *Journal of Educational Measurement*, *26*, 191–208. doi:10.1111/j.1745-3984.1989.tb00328.x

*Wilson, M. (1992). The ordered partition model: An extension of the partial credit model. *Applied Psychological Measurement*, *16*, 309–325. doi:10.1177/014662169201600401

Authors

MARK J. GIERL is a professor of educational psychology and Canada Research Chair in Educational Measurement, Department of Educational Psychology, University of Alberta, 6-110 Education North, Edmonton, Alberta, Canada T6G 2G5; email: *mark.gierl@ualberta.ca*.

OKAN BULUT is an assistant professor of educational psychology, Department of Educational Psychology, University of Alberta, 6-110 Education North, Edmonton, Alberta, Canada T6G 2G5; email: *bulut@ualberta.ca*.

QI GUO is a PhD student in the Department of Educational Psychology, University of Alberta, 6-110 Education North, Edmonton, Alberta, Canada T6G 2G5; email: *qig@ualberta.ca*.

XINXIN ZHANG is a PhD student in the Department of Educational Psychology, University of Alberta, 6-110 Education North, Edmonton, Alberta, Canada T6G 2G5; email: *xinxin4@ualberta.ca*.