Routledge
Taylor & Francis Group

APPLIED MEASUREMENT IN EDUCATION

Applied Measurement in Education

ISSN: 0895-7347 (Print) 1532-4818 (Online) Journal homepage: https://www.tandfonline.com/loi/hame20

Are Multiple-choice Items Too Fat?

Thomas M. Haladyna, Michael C. Rodriguez & Craig Stevens

To cite this article: Thomas M. Haladyna, Michael C. Rodriguez & Craig Stevens (2019) Are Multiple-choice Items Too Fat?, Applied Measurement in Education, 32:4, 350-364, DOI: 10.1080/08957347.2019.1660348

To link to this article: https://doi.org/10.1080/08957347.2019.1660348

APPLIED MEASUREMENT IN EDUCATION
2019, VOL. 32, NO. 4, 350–364
https://doi.org/10.1080/08957347.2019.1660348

Routledge
Taylor & Francis Group

Are Multiple-choice Items Too Fat?

Thomas M. Haladyna[a], Michael C. Rodriguez[b], and Craig Stevens[c]

[a]Arizona State University; [b]University of Minnesota; [c]National Board of Medical Examiners

ABSTRACT

The evidence is mounting regarding the guidance to employ more three-option multiple-choice items. From theoretical analyses, empirical results, and practical considerations, such items are of equal or higher quality than four- or five-option items, and more items can be administered to improve content coverage. This study looks at 58 tests, including state achievement, college readiness, and credentialing tests. The evidence here supports previous assertions. The article also clarifies distractor functioning criteria and offers a typology of items via distractor functioning.

Among the many testing programs in the world, the multiple-choice (MC) item remains the most used for many good reasons. For measuring knowledge and cognitive skills, MC is very efficient and effective. Also, for measuring certain types of complex thinking, the MC item format is used very successfully (see Haladyna & Rodriguez, 2014; Rodriguez, 2002). A lot of content can be covered quickly resulting in highly reliable scores when MC items are well written.

The quintessential MC item has a stem consisting of a complete question or a partial sentence completed by the options, followed by a correct option and three to four incorrect options, referred to as distractors, distracters, misleads, wrong answers, and foils. *Distractor* will be used in this paper.

The motivation for this study is the idea that traditional four- and five-option MC items are simply too fat. We need slimmer items with fewer distractors that discriminate highly in the range of ability being tested. Slimmer items have many advantages. First, it takes less time to construct a slimmer item. Second, it takes less time to respond to slimmer items, and under the condition of maintaining total testing time, more slim items can be administered. As a result, content-related validity evidence (content coverage) improves and adding more items improves reliability. Third, and most importantly, many MC distractors do not distract, affecting the item's ability to discriminate among examinees with varying degrees of ability. Item discrimination is directly related to reliability (Ebel, 1967). Thus, if distractors do not distract adequately, item discrimination and reliability suffer (Author, Submitted for publication).

First, we address the rationale for distractor analysis. Then we present a trio of perspectives for reducing the number of distractors used in a typical MC item. Finally, we present an empirical analysis of 58 high quality large-scale tests spanning a large range in ability, ages, subject-matters, and test score uses. We end with some observations and recommendations.

1. Why Study Distractor Discrimination?

MC items are designed to measure important facts, concepts, principles, or procedures stated or implied in test specifications. Once a subject-matter-expert (SME) committee has agreed that the content of each item is representative of what it is supposed to measure as reflected in the test specifications, the item is field-tested. Its difficulty and discrimination are estimated. We want items that have high discrimination

CONTACT Thomas M. Haladyna tomhaladyna@gmail.com Arizona State University, 10828 N. 9th Place, Phoenix 85020
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/hame.

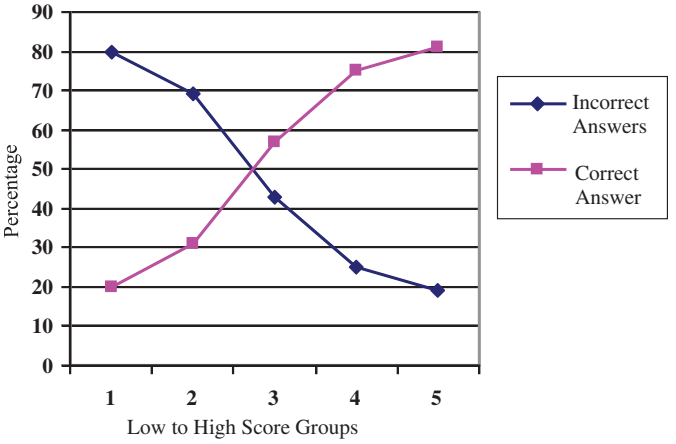


Figure 1. Trace lines for right answer and collective wrong answers.

with a desired range of difficulty that maximize reliability. This is true of all forms of tests for all purposes – we need items that contribute to and are correlated with total scores to support the meaningfulness of the composite score (resulting from the aggregation of item scores).

There is a reciprocity between the right answer and its team of distractors. A traditional item discrimination index is the point-biserial (PB) correlation between the right answer and the total score (minus the contribution of that item to total score). The magnitude of the discrimination index for any item will be equal to the magnitude of the discrimination of the set of distractors. For example, if a discrimination index is .45 for the right answer, the collective discrimination index for the distractors will be −.45 (that is, the discrimination of selecting a distractor or the distractors as a set, as the sum of probabilities of selecting each option must sum to 1 at each point on the total score scale). Figure 1 illustrates the reciprocity between the discrimination of the correct answer and all distractors combined; the trace line for the correct answer is a mirror image of the trace line for the collective item responses for distractors.

As there is a functional association between the discrimination of the correct answer and the collective discrimination of the distractors, a weak distractor weakens this association. A strong team of distractors will have a high discrimination when compared to a weaker team. One study demonstrated that when a weak distractor is removed from the team of distractors, item discrimination improves (author). This outcome is predictable. If a distractor does not discriminate, its removal should improve the overall distractability – discrimination.

2. Theoretical, Empirical, and Practical Considerations in MC Item Design

2.1. *Theoretical Analysis*

Since the 1920s, researchers have been investigating the optimal number of options on a MC test, agreeing that three options is optimal (see Rodriguez, 2005, for a meta-analysis of this body of research). Although most of those studies were empirical experimental studies varying the number of options per item across randomly administered forms, some studies employed theoretical approaches searching for analytical solutions to the question. These analytical solutions are described here.

The first researcher employing an analytical approach, then doctoral student, Amos Tversky (1964), provided a mathematical proof for his theorem focused on discrimination, power, and uncertainty, three functionally related concepts. Given a fixed number of alternatives on the entire test, the discrimination function maximizes at 2.718 (the value of *e*) alternatives per item, a point residing near three options per item. He argued that "whenever the amount of time spent on the test

is proportional to its total number of alternatives, the use of three alternatives at each choice point will maximize the amount of information obtained per time unit" (p. 390).

Grier (1975, 1976) agreed that three options is optimal, with important observations to note. First, item analysts should consider the properties of the test and its administration. For instance, word length, time-to-respond, and discrimination are factors to consider. But all things being equal, the two- or three-option test will achieve higher reliability than a four- or five-option test for a fixed time period (for longer tests) with the important caveat that we can use more two- and three-option items on this test. More items yield more information.

The time required to read an item may neutralize any advantage achieved with a three- or two-option item. Tversky's model assumed that all items are equal with respect to response time. Thus, the finding that three options is optimal should be tempered by the fact that some items are designed to take more time to administer (e.g., lengthy options). For Grier, time was an important variable in weighing the value to two- or three-option versus four- or five-option items.

Lord (1977) compared four theoretical models that informed the argument about the optimal number of items. His observations provide a capstone on the theoretical work on this issue. He compared the two previous models with two new models. Lord's first new model examined the influence of random guessing. In his second model, he used the three-parameter item characteristic function and discovered that maximum information is provided in the lower third of the distribution when a test has more options. This observation is illustrated with the trace line in Figure 2. High-scoring examinees usually narrow the choice to one or two options and choose the correct one most often. Options C and D are the least plausible distractors for the highest-scoring group (5). In the lowest scoring group (1), all options are used. Low-scoring examinees usually use the entire range of options because they cannot eliminate or distinguish options or are guessing randomly. Thus, the use of three options is optimal for most examinees who score in the midrange of the test (Lord, 1977); two options is sufficient for high-scoring examinees; if precision is required in the lower-third of the distribution, four or five options is better.

Finally, Bruno and Dirkzwager (1995) used an information-theoretic perspective to identify the optimal number of options. They found maximum information was obtained with three-option items, where each option had equal probability of selection. This resulted in the context where examinees assigned personal probabilities for each alternative as with partial knowledge.

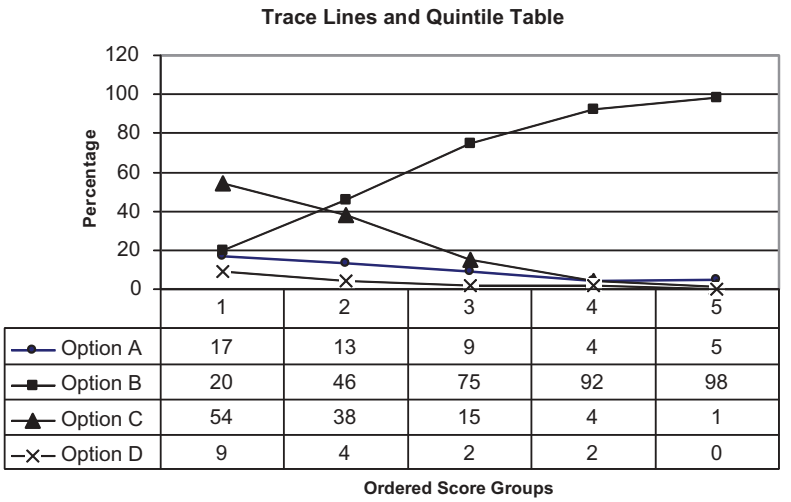


Figure 2. Trace lines and quintile table for well-behaved distractors, with correct option B.

2.2. *Empirical Evidence*

One of the most empirically studied item-writing guidelines concerns the number of MC options. The typical guideline is: Write as many options as feasible (Haladyna & Downing, 1989a). After their review of the empirical literature, Haladyna and Downing (1989b) revised the guideline: Develop as many functional distractors as feasible. In the most comprehensive review of the technology of distractor development, Gierl, Bulut, Guo, and Zhang (2017) describe the complex association between the correct answer and distractors. They stated that the primary purpose of distractor analysis is to eliminate non-functioning distractors to improve discrimination. In their review, they reported that simply having more distractors serves no purpose if the distractors are not attracting test takers with misinformation or less ability. In other words, distractors need to discriminate or they are useless.

Rodriguez (2005) completed a meta-analysis of the empirical research on this question, including 27 published articles (56 independent studies) from 1925 to 1999. The meta-analysis included experimental studies with carefully designed test forms that reduced the number of options systematically, randomly assigning them to test takers, and evaluating the impact on item and test score statistics. Most studies evaluated the effect of number of options on the item *p*-value and item discrimination (item-total correlations). Many studies also evaluated the impact of changing number of options on test score reliability; few evaluated the impact on test score validity coefficients (correlations with other measures). In comparing the use of three-option items to two-, four-, and/or five-option items, original study authors were nearly unanimous in support of the three-option item (with the exception of one author). The empirical evidence showed no loss to item quality by reducing the number of options from four or five to three, and a slight improvement in test score reliability when reducing the number of options from four to three. Haladyna and Rodriguez (2013) reviewed an additional decade of research on the topic and offered a revised guideline: Use only options that are plausible and discriminating; three options are usually sufficient.

Attention to this item-writing guideline has increased in the past two decades, including a broader arena of tests than those included in the 2005 (Rodriguez) meta-analysis. The recently examined tests include a primary school mathematics achievement test (Nwadinigwe & Naibi, 2013), modified state K-12 tests for students with disabilities (Rodriguez, Kettler, & Elliott, 2014), the ACT test for college admissions (Edwards, Arthur, & Bruce, 2012), a vocabulary test for English as a foreign language among college students (Baghaei & Amrahi, 2011), English for Science and Technology exam with college students (Berrios, Rojas, Cartaya, & Casart, 2005), a College Scholastic Ability Test (Lee & Winke, 2012), psychology course tests (Taylor, 2005), health sciences and medical school course tests (Caldwell & Pate, 2013; Dehnad, Nasser, & Hosseini, 2014; Kilgour & Tayyaba, 2016; Salazar Blanco, Vélez, & Zuleta Tobón, 2015; Schneid, Armour, Park, Yudkowsky, & Bordage, 2014; Vegada, Shukla, Khilnani, Charan, & Desai, 2016; Zoanetti, Beaves, Griffin, & Wallace, 2013), and nursing course tests (Redmond, Hartigan-Rogers, & Cobbett, 2012; Tarrant, Ware, & Mohammed, 2009). One clear shift is the great interest in the medical and health sciences fields for the potential of reducing the number of options. These authors consistently refer to the cost of creating additional distractors and the potential for including more items with fewer distractors, improving content coverage. In professional preparation program courses and licensure and certification testing programs, it is imperative that tests cover the relevant content as specified in the professional content standards of what professionals should know and be able to do.

These authors uniformly supported the development and use of three-option items. They found similar item discrimination among items with differing numbers of options and sometimes, improved item discrimination and/or score reliability with fewer options (Berrios et al., 2005; Dehnad et al., 2014; Nwadinigwe & Naibi, 2013; Rodriguez et al., 2014; Tarrant et al., 2009; Zoanetti et al., 2013). In some cases, the three-option forms resulted in higher mean scores (Edwards et al., 2012; Kilgour & Tayyaba, 2016; Lee & Winke, 2012; Nwadinigwe & Naibi, 2013; Rodriguez et al., 2014; Tarrant et al., 2009), but in many there was no change in mean scores (Baghaei & Amrahi, 2011; Dehnad et al., 2014; Redmond et al., 2012; Salazar Blanco et al., 2015; Schneid et al., 2014; Taylor, 2005; Vegada et al., 2016). This occasional

increase in mean score reflects a slight increase in item p-values, where some studies found items with fewer options to be slightly easier. In his meta-analysis, Rodriguez (2005) found an average .04 increase in item p-values when decreasing the number of options from 4 to 3. These more recent researchers also uniformly found that three-option items take less time to answer, providing support for the argument that more items could be included within the same time limits and improve content coverage.

Another interesting development is the use of differential distractor functioning (DDF) analysis, a new take on differential item functioning (DIF) as a way to more deeply explore sources of measurement invariance (Green, Crone, & Folk, 1989; Penfield, 2008). A direct evaluation of DDF provides important information regarding the potential causes of DIF: DIF may be due to a problem with the correct option or the distractors (Penfield, 2010; Suh & Bolt, 2011), emphasizing the importance of distractor quality. Writing more options for the sake of writing four- or five-option items (perhaps as required in item specifications or directions to item developers) potentially might result in DDF, as such options are less plausible and potentially differentially construct-irrelevant to diverse test takers.

We find no new compelling empirical evidence to support the operational default standard of writing four- or five-option items.

2.3. *Logistical/Practical Considerations*

Whether for standardized or classroom tests, item writers agree that writing four- or five-option MC items is harder than writing two- or three-option MC items. One hardly needs to provide empirical evidence for that assertion, especially when each distractor is intended to be plausible to low-performing examinees and implausible to higher performing examinees. Distractors should represent common errors. If SMEs judge each option to be plausible common errors, empirical evidence should support that judgment. As pointed out in past studies and in this current study, many distractors are simply not working as intended. It is surprising and disappointing that past and current research has revealed that many distractors fail to perform as expected, but test developers continue to use weak distractors.

3. Guessing as an Argument for More Distractors

Perhaps the main reason to advocate for four- or five-option items is that such items reduce the chance of lucky, random guessing. One might achieve a high score by guessing on a three-option MC test, whereas with a five-option MC test, this possibility is reduced.

Guessing comes in two forms, random and strategic guessin. *Random guessing* occurs when all options are equally plausible (to the examinee) and the true score of the examinee for that item is zero. In other words, the examinee has no idea what the right answer is and simply marks one of the options. How serious is the threat of lucky guessing? Random error is the result of random guessing. If an examinee's true score is zero (the floor of the scale), the probability of getting an item correct or a certain percentage of items correct or higher solely by random guessing varies as a function of the number of options.

Table 1 was developed using a binomial probability calculator (statrek.com). The table contains two results. The first lists the probabilities of getting a score of 40% or higher solely by chance for three test lengths (50, 100, and 200). The second lists the probabilities of getting a score of 50% or higher for the same three test lengths (50, 100, and 200). What Table 1 reveals is that in most instances, it is very improbable for a lucky guesser to exceed 40% in most instances. It is even more improbable for a lucky guesser to exceed 50% or higher in all instances. If a cut score is set at 40%, than a three-option item might reward a lucky guesser about 8% of the time on a 100 item test. If the cut score is set at 50%, the three-option item format works very well.

Strategic guessing is another matter. Whereas random guessing introduces random error into test scores as reflected by the reliability estimate, strategic guessing involves partial knowledge. When an

Table 1. Influence of random guessing for MC items with three to five options.

Probability of Getting a Score at or Above 40% when True Score is Zero			
	Test Length		
Options	50	100	200
Three	.18260	.08488	.02236
Four	.01392	.00068	.00000
Five	.00093	.00000	.00000
Probability of Getting a Score of 50% or Higher when True Score is Zero			
	Test Length		
Options	50	100	200
Three	.00422	.00032	.00000
Four	.00004	.00001	.00000
Five	.00000	.00000	.00000

examinee has partial knowledge, one or more distractors may be eliminated due to this fact, so that the MC item essentially has fewer options. Then the examinee has a higher probability of guessing the right answer. Strategic guessing is at least partially construct-relevant because it is informed by construct knowledge as opposed to luck. We argue that random guessing has very little influence on test scores, especially when the test has more than 100 items. But note that if a three-option item is written with two effective distractors (plausible alternatives), then the effect of strategic guessing would be no different than if the item was a five-option item with two nonfunctioning distractors (essentially a three-option item).

4. How Should We Evaluate Distractor Performance

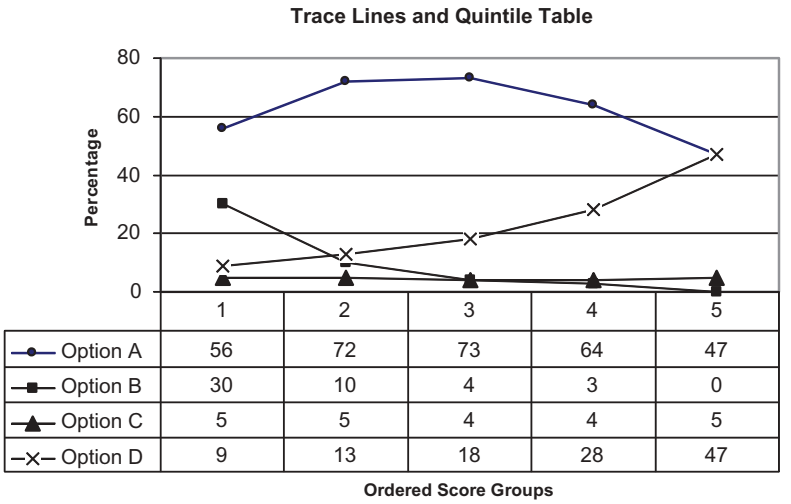
A distractor should be plausible enough to be chosen by examinees with low ability and implausible to those examinees with high ability. Otherwise, the distractor would not discriminate as it should. In the development of any test item, test developers benefit by studying the plausibility of each distractor both via SME consensus and statistical properties (both of which constitute validity evidence). SMEs judge whether a distractor is a misconception or common error or is plausible in other ways. Their judgment can be supplemented by item analysis, graphical illustration (trace lines), descriptive statistics, multivariate methods, and item response theory.

4.1. *Point Biserial/Biserial Correlations*

The correlation of the distractor response to total score is the product-moment or PB correlation. The biserial (BIS) is an alternative procedure, which assumes an underlying continuum in the dichotomous item score. A scatter plot of the PB and BIS for any set of item responses for a test will show a .99 correlation and the fact that the two methods are on different scales. Regarding information found in the PB and BIS, there may be little difference except at the tails of the distributions where a small separation is shown. Attali and Fraenkel (2000) made an important observation that the discrimination index for a distractor based on the PB should use the mean of the correct option and the mean of the distractor being evaluated. Otherwise, a distractor discrimination will be influenced by the means of other distractors.

4.2. *Trace Lines*

One of the best ways to understand the discriminating ability of distractors is with trace lines (Thissen, Steinberg, & Fitzpatrick, 1989; Wainer, 1989). Trace lines simply illustrate what a table of option response frequencies organized by ordered score groups informs. Trace lines can be derived



	1	2	3	4	5
Option A	56	72	73	64	47
Option B	30	10	4	3	0
Option C	5	5	4	4	5
Option D	9	13	18	28	47

Figure 3. Trace lines and quintile table with a poorly behaving distractor (A) and correct option D.

from these tables of option selection frequencies as a function of five or more ordered score groups. Figures 2 and 3 contain tables for five ordered score groups (1 to 5) with a four-option item from an operational test. In Figure 2, Options A, C, and D show higher selection frequencies for the lowest score group and lower frequencies for the highest score group. The correct option has the opposite characteristic.

In Figure 1, the collective trace line for all distractors (an incorrect response) is monotonically decreasing. Poorly discriminating distractors detract from the collective discriminating power of all distractors. In Figure 2, the correct answer (B) is monotonically increasing in score groups ordered low to high, and the three distractors are monotonically decreasing. For advocates for four- and five-option MC items, Figure 2 is desired. Figure 3 shows a misbehaving distractor (A); two distractors have very low selection frequencies (B and C). Note that the trace line is built based on performance in ordered score groups. For short to moderate length tests, five score groups should be used. For very long tests, ten score groups work well. Figures 2 and 3 illustrate what statistical analysis tells us: Some distractors work quite well and others fail to perform as expected.

4.3. *Frequency of Response*

Another criterion often used in distractor evaluation is the frequency of response. If one or two distractors are mostly ignored as a choice, perhaps the distractor is too implausible. Most researchers of distractor performance have chosen the arbitrary, subjective cut score of 5% to suggest that a distractor is so implausible that it is ignored by most examinees. When the 5% criterion is used for items with p-values above .90, virtually all distractors are flagged for low response rates. In the study reported in this paper, we use a sliding scale for low-frequency distractors relative to the item p-value (described below).

4.4. *Choice Mean*

The choice mean is the average score for all examinees who choose each option. The choice mean for the correct answer should be higher than the means for the distractors. When distractors have very high choice means, they are not discriminating. Fortunately, the association between choice mean and the traditional PB discrimination is very high. An analysis of variance is one way to capture the effectiveness of all options (correct answer and distractors). The sums-of-squares divided by total

variance is a discrimination index. This index is a multiple correlation with the MC options as the independent variable and test score as the dependent variable. An examination of the PB formula shows that it is a standardized version of the choice means of the correct and incorrect answer.

4.5. *Multivariate Methods and Item Response Theory*

Multivariate methods have been considered as a means for distractor evaluation (Haladyna & Rodriguez, 2013, pp. 352–355). Comparative studies of the efficacy of these methods with more conventional methods including descriptive statistics have not been reported. Thus, it would be difficult to support the use of any of these methods without information bearing on how well or better they identify poorly performing distractors. Similarly, item response theory provides information relative to distractor performance given estimation of the underlying trait level (typically referred to as theta or ability estimate). In some models, such as the nominal response model, item response characteristic curves can be estimated illustrating the probability of selecting each option conditional on the level of underlying trait (more detailed and continuous than typical trace lines based on score groups).

5. Criteria for Evaluating Distractors

The criteria used to evaluate distractors in this study are decidedly subjective but augmented with a rationale, experience, and previous research. The two criteria for detecting distractors that do not perform adequately include low selection frequency and discrimination.

5.1. *Low-Selection Frequency*

If a distractor is seldom chosen by examinees, it appears to be implausible. Haladyna and Downing (1993) used the criterion of selection frequency of 5%. However, as noted previously, where item *p*-values exceed .90, 5% might eliminate all distractors. Therefore, a sliding scale is used to identify low frequency distractors. The scale is based on the work of Raymond, Stevens, and Bucak (2019). They used a formula, and in this study, a table was developed from their formula. Table 2 contains the sliding scale for the low-selection frequency criteria.

6. Option Discrimination

In this study, we use the PB coefficient to evaluate the usefulness of a distractor. We hypothesize that a useful distractor should have a statistically significant negative correlation. However, with large samples, even the smallest correlation coefficient can be statistically significant. We also evaluate the magnitude of the correlation as an effect size. The boundary value for effect size was set at .15. If a distractor has a PB correlation between −.15 and .15, it is ruled non-functioning. Such a distractor accounts for as much as 2.25% of the variance of the total scores. Distractor discrimination of this small magnitude does not contribute to an overall discrimination for a team of distractors. Positive correlations of distractor selection and total score are indicators of a poorly performing distractor, possibly indicating a miskey, and are very rare in professionally, well-developed tests. The rules are as follows:

Table 2. Criteria for identifying low-frequency distractors given item *p*-value.

	Item *p*-value			
	> .90	> .80 and ≤ .90	> .70 and ≤ .80	≤ .70
Distractor frequency	2%	3%	4%	5%

(1) If the statistically significant PB is ≥ −.15 and ≤.15, distractor is non-functional.
(2) If the statistically significant PB is < −.15, distractor is functional.
(3) If the PB is not statistically significant, the distractor is non-functional.

As noted previously, judging the functionality of a distractor is subjective. The application of .15 seems reasonable, if and only if, that coefficient is decidedly lower than other distractors for that item. The main point here is that because distractors work as a team, having a weak teammate limits the success potential of the entire team. That poor teammate needs to be replaced and sent back to the minor leagues.

7. Empirical Evaluation of Several Well-Developed Tests

We evaluated the performance of distractors of tests from four large-scale testing programs. For each test, we asked a consistent set of questions, based on the premise that MC items are customarily too fat:

(1) What percentage of distractors have low frequency?
(2) What percentage of distractors have low discrimination?
(3) What percentage of distractors are judged to be non-functioning?
(4) Is there a pattern of distractor performance as a function of test or subject matter?
(5) Do the results confirm the typology for classifying MC items as a function of how distractors perform?

The results are presented by data set.

7.1. *Sample*

To conduct this study, we made a number of requests of large-scale testing programs to provide item-level response data, in order to cover the range of large-scale testing programs, including K-12, college admissions, and professional licensure and certification. Nearly all of our requests were granted (none were outright denied, a couple were never fulfilled). A total of 58 standardized large-scale achievement tests were used in this study. The first set includes seven reading and seven mathematics tests for grades three through eight and grade 10 from a statewide achievement testing program. The second set includes five forms of a large-scale college readiness test consisting of four subject matters: English, Mathematics, Reading, and Science. The third set includes five forms of a large-scale preliminary college admissions test consisting of four subject matters: English, Mathematics, Reading, and Science. The fourth set includes four forms of a high-stakes credentialing test.

The results section provides descriptive statistics for tests used for the study. The tables do not divulge specific names of the test, sponsors, or other private information to protect their anonymity. All tests were professionally developed, meet high standards for item development, and have been field tested and evaluated. The first two data sets were operational test results. For the second two data sets, student responses had significant omissions and not-reached rates, so only valid, complete student responses were used in this analysis. Despite being operational tests, some items were eliminated for further evaluation due to poor discrimination. These items would not ordinarily be part of an operational test. The fourth data set consists of four mutually exclusive forms, from a high stake credentialing exam, evaluating only the five-option items from each form.

8. Results and Discussion

8.1. *Data Set 1, Reading Achievement Tests*

Table 3 includes results for a state reading test for grades three to eight and ten. The means, standard deviations, and alpha reliability estimates for grades three to eight were stable and consistent. For grade ten, performance was lower. Of the 1,986 distractors evaluated from these four-option items, 34% were judged to be non-functioning. Low-frequency distractors were less numerous (12%) compared to low discriminating distractors (22%). The percentage of low-frequency and low-discriminating distractors varied considerably and unexplainably from grade to grade. Item developers for each grade used the same item-writing protocols, and items were field-tested the same way for each grade.

8.2. *Data Set 1, Mathematics Achievement Tests*

Table 4 includes results for the companion subject matter–mathematics. All items had four options. Except third grade, the means were very similar across the grades. Standard deviations were similar and alpha reliability estimates were consistently high. The number of items and distractors analyzed varied from the total number possible because some items did not perform well (overall as field-test items) and were eliminated from the item analysis. Of the 1,674 distractors evaluated, 40% were judged as non-functioning; 4% were low frequency and 36% were low discriminating. Grades three, four, and seven showed the greatest number of non-functioning distractors; grades eight and ten had the fewest. Overall, more than one-third of distractors in this professionally developed testing program were non-functioning. Compared to the companion reading test, the mathematics test had more non-functioning distractors. The reading test had more low-frequency distractors. With reading, the upper grades tests had the most non-functioning distractors, whereas with the mathematics test the frequency of non-functioning distractors was more evenly distributed across grades.

Table 3. Statewide achievement reading tests score and item statistics (n = 1,000).

Grade	*M*	*SD*	Reliability estimate	Distractors evaluated	Low frequency	Low discrimination	Non-functioning distractors
3	.65	.18	.94	282	12%	12%	23%[a]
4	.62	.19	.95	282	10%	20%	30%
5	.63	.18	.95	330	12%	19%	31%
6	.62	.18	.95	327	11%	18%	29%
7	.62	.17	.94	306	15%	26%	41%
8	.62	.16	.93	282	16%	25%	41%
10	.48	.18	.89	177	2%	42%	45%[a]
Total				1,986	12%	22%	34%

[a] Total percentage different than components due to rounding. Each test item had four options.

Table 4. Statewide achievement mathematics tests score and item statistics (n = 1,000).

Grade	*M*	*SD*	Reliability estimate	Distractors evaluated	Low frequency	Low discrimination	Non-functioning distractors
3	.59	.17	.92	252	10%	36%	46%
4	.50	.17	.91	234	5%	44%	48%[a]
5	.49	.18	.92	231	4%	36%	41%[a]
6	.50	.18	.92	240	5%	33%	38%
7	.48	.18	.93	240	2%	42%	45%[a]
8	.46	.18	.91	228	2%	31%	33%
10	.46	.17	.92	249	2%	28%	30%
Total				1674	4%	36%	40%

[a] Total percentage different than components due to rounding. Each test item had four options.

Table 5. Pre-college readiness tests score and item statistics (n = 1,000).

Subject	M	SD	Reliability estimate	Distractors evaluated	Low frequency	Low discrimination	Non-functioning distractors
English 1	.65	.19	.88	120	8%	17%	24%[a]
English 2	.65	.19	.88	120	13%	17%	31%[a]
English 3	.64	.19	.87	120	17%	19%	36%
English 4	.65	.19	.87	120	13%	17%	31%
English 5	.62	.19	.88	120	17%	15%	32%
Mathematics 1	.56	.19	.84	120	15%	28%	42%[a]
Mathematics 2	.51	.20	.85	120	14%	27%	41%
Mathematics 3	.54	.19	.82	116	20%	34%	53%[a]
Mathematics 4	.55	.20	.81	108	19%	27%	47%[a]
Mathematics 5	.52	.20	.85	90	18%	37%	54%[a]
Reading 1	.54	.21	.85	90	1%	24%	26%[a]
Reading 2	.60	.24	.90	90	3%	12%	16%[a]
Reading 3	.70	.22	.89	90	17%	17%	33%[a]
Reading 4	.67	.24	.91	87	9%	9%	18%
Reading 5	.63	.21	.87	90	8%	14%	22%
Science 1	.58	.20	.83	81	1%	17%	19%[a]
Science 2	.53	.22	.86	84	2%	14%	17%[a]
Science 3	.57	.19	.82	81	2%	26%	28%
Science 4	.56	.20	.83	84	6%	26%	32%
Science 5	.57	.21	.85	84	5%	23%	27%[a]
Total				2,015	11%	21%	33%[a]

[a] Total percentage different than components due to rounding. English, Reading and Science had four options per item. Mathematics had five options per item.

8.3. *Data Set 2, College Readiness Test*

Table 5 includes means varying from .51 to .70 with mathematics and science being the most difficult and English and reading being the least difficulty. Alpha reliability estimates ranged from .81 to .91. The mathematics test had five-option items and the other subjects had four-option items. Of the 2,015 distractors evaluated, 33% were judged to be non-functioning; 11% had low frequency, 21% had low-discrimination. Guessing was more randomly distributed and low discrimination was the greater problem with this battery of tests.

English and mathematics had the highest number of low-frequency distractors (Table 6). Mathematics had the highest frequency of low-discrimination distractors. Thus, the fact that the mathematics test had five options appears to have a negative effect on distractors. There is no distinct pattern regarding the number of low-frequency, low-discriminating, and total non-functioning distractors for these data.

8.4. *Data Set 3, Preliminary College Admissions Test*

Table 7 includes the results for the second test battery. Results are almost identical to the previous data set. Mean scores varied between .49 and .61. Reliability estimates also varied widely from .76 to .91. Of the 2,240 distractors evaluated, 32% were judged to be non-functioning. Only 11% of these distractors had low frequency, and 22% had low discrimination. When disaggregating these results by subject matter, we see a similar result as with the previous data set. The five-option mathematics items had the

Table 6. Pre-college readiness tests non-functioning distractors by subject matter.

Subject	Distractors evaluated	Low frequency	Low discrimination	Non-functioning distractors
English	717	13%	17%	31%
Mathematics	576	17%	30%	48%
Reading	387	8%	15%	23%
Science	429	3%	21%	25%

Table 7. College readiness tests score and item statistics (n = 1,000).

Test subject	*M*	*SD*	Reliability estimate	Distractors evaluated	Low frequency	Low discrimination	Non-functioning distractors
English 1	.58	.20	.91	147	12%	15%	27%
English 2	.56	.19	.88	141	13%	25%	38%
English 3	.57	.18	.89	144	9%	11%	20%
English 4	.59	.19	.89	147	12%	20%	33%[a]
English 5	.59	.18	.89	144	15%	19%	34%
Mathematics 1	.53	.20	.87	152	16%	20%	36%
Mathematics 2	.49	.19	.85	144	18%	40%	58%
Mathematics 3	.59	.18	.84	140	23%	17%	40%
Mathematics 4	.59	.20	.86	136	19%	28%	47%
Mathematics 5	.54	.19	.88	156	17%	33%	50%
Reading 1	.57	.21	.82	75	9%	1%	11%[a]
Reading 2	.50	.21	.81	75	4%	24%	28%
Reading 3	.61	.21	.83	72	1%	4%	5%
Reading 4	.56	.22	.84	75	3%	19%	21%[a]
Reading 5	.57	.22	.84	72	1%	15%	16%
Science 1	.51	.21	.84	87	2%	8%	11%[a]
Science 2	.50	.18	.76	78	1%	37%	38%
Science 3	.54	.20	.84	87	6%	9%	15%
Science 4	.51	.20	.84	87	3%	24%	28%[a]
Science 5	.54	.23	.87	81	1%	14%	15%
Total (median)				2,240	11%	20%	32%[a]

[a] Total percentage different then components due to rounding. English, Reading and Science had four options per item. Mathematics had five options per item.

highest percentage of low-frequency and non-discriminating distractors (see Table 8); Reading and science had the most robust number of four-option items that worked well.

8.5. *Data Set 4, Credentialing Test*

Table 9 includes information about the four forms of a high-stakes credentialing test. The means and standard deviations were consistent. Test difficulty was moderate. Reliability estimates were consistently high. Of the 3,360 distractors evaluated more than 52% were found to be non-functioning. Low discrimination was infrequent (4% to 6%) but with these five-option items, low frequency distractors were numerous (49% overall).

9. Conclusions

This critical evaluation and empirical study of a large number of diverse tests support the premise that tests have many poorly performing distractors. With a set of state student achievement tests in reading and mathematics, between 33% and 38% of all distractors were poorly performing. For the next two sets of college-readiness tests, results were more variable. Five-option mathematics items performed poorly with respect to distractors. The other subject matters had many poorly performing distractors, but the mean frequency of these non-functioning distractors ranged between 20% and 24%. The high-stakes credentialing test had many distractors. These low-frequency distractors were ignored by the majority of examinees who are likely well-trained and knowledgeable.

Table 8. College readiness tests non-functioning distractors by subject matter.

Subject	Distractors evaluated	Low frequency	Low discrimination	Non-functioning distractors
English	723	11%	20%	30%
Mathematics	728	19%	28%	46%
Reading	369	4%	13%	17%
Science	420	3%	19%	22%

Table 9. Professional credentialing tests score and item statistics (n = 1,204 to 1,215).

Form	*M*	*SD*	Reliability estimate	Distractors evaluated	Low frequency	Low discrimination	Low frequency and discrimination	Non-functioning distractors
1	.74	8.5	.88	828	48%	6%	3%	51%
2	.73	8.9	.88	824	49%	6%	3%	52%
3	.75	8.7	.89	828	49%	4%	2%	51%
4	.74	8.3	.88	880	51%	6%	2%	55%
Total				3,360	49%	6%	3%	52%

The percent of items that are low frequency or low discrimination are included in each respective category (they overlap). Values may not sum to 100 due to rounding.

Although not a direct or generalizable result from this evaluation of distractor functioning, we noted a difference in average scores (essentially average item p-values) across forms and testing purposes. Item difficulty does restrict the potential functioning of distractors, as much harder items invite guessing and much easier items limit the frequency with which a distractor may be selected. Although there was substantial variation in item p-values within forms, we noticed some slight variation in overall form means. For the K-12 testing program forms, average scores ranged from .48 to .65 (mostly .62 to .65) for reading and .46 to .59 for mathematics. For the college readiness testing program, average scores ranged from .51 to .70 for one set of forms and from .49 to .61 for the other. These largely overlap with the K-12 test forms. Finally, the average scores for the credentialing test ranged from .73 to .75, higher than the other forms, but not substantially so.

Based on this study and experience, Table 10 provides a typology we used for evaluating distractors and the correct answer. In item development and with item analysis, this typology provides a useful way to evaluate distractors. The typology is new. A detailed discussion of how to evaluate distractor functioning appears in Haladyna and Rodriguez (2013, pp. 352–355), but no firm guidelines were presented. For low frequency distractors, the customary 5% criterion was abandoned, and a sliding scale was used as suggested by Raymond et al. (2019). For discrimination, some boundary values were chosen subjectively based on the idea of effect size. Correlations of .15 or less have a small effect and thus form the basis for rejecting a distractor. The sign of the correlation is irrelevant as any value between −.15 and .15 is a very low correlation.

Once a distractor has been identified as non-discriminating or of low frequency, it should either be removed or replaced and re-evaluated. As an item-writing policy, creating more than two distractors may be a good strategy, if and only if, each distractor is a plausible incorrect answer that is selected often by low-performing examinees and ignored by high-performing examinees. Gierl et al. (2017) and Haladyna and Rodriguez (2013) supported this approach to item development, arguing for careful design and selection of distractors to support the goals and purpose of the test.

Table 10. A typology for evaluating distractor performance.

Type	Description
Correct answer	The correct answer should have a monotonically increasing trace line. The discrimination index is positive and greater than .15.
Functional distractor	A functional distractor has a monotonically decreasing trace line. The discrimination index should be negative and greater in magnitude than .15.
Non-functioning distractor	A non-functioning distractor should have a non-monotonic trace line. Evidence of this is found in the PB discrimination index or the trace line. This index is between −.15 and .15. If this index is greater than .15, then the distractor is mimicking the correct answer–a double key possibility.
Low-frequency distractor	Any distractor that has a low frequency is likely to be implausible and simply exists as a place to mark a random guess. We used a sliding scale instead of the more traditional 5% criterion to determine low frequency.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

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/jedm.2000.37.issue-1

Author. *Using full-information item analysis to evaluate distractors.*

Baghaei, P., & Amrahi, N. (2011). The effects of the number of options on the psychometric characteristics of multiple choice items. *Psychological Test and Assessment Modeling*, *53*, 192–211.

Berrios, G., Rojas, C., Cartaya, N., & Casart, Y. (2005). Effect of the number of options on the quality of EST reading comprehension multiple-choice exams. *Paradigma*, *26*(1), 1–18.

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

Caldwell, D. J, & Pate, A. N. (2013). *Effects Of Question Formats on Student and Item Performance. American Journal Of Pharmaceutical Education*, *77*(4), . doi: 71 doi:10.5688/ajpe77471

Dehnad, A., Nasser, H., & Hosseini, A. F. (2014). A comparison between three-and four-option multiple choice questions. *Social and Behavioral Sciences.*, *98*, 398–403.

Ebel, R. L. (1967). The relationship of item discrimination to test reliability. *Journal of Educational Measurement*, *4*(3), 125–128. doi:10.1111/j.1745-3984.1967.tb00579.x

Edwards, B. D., Arthur, W., & Bruce, L. L. (2012). The three-option format for knowledge and ability multiple-choice tests: A case for why it should be more commonly used in personnel testing. *International Journal of Selection & Assessment*, *20*(1), 65–81. doi:10.1111/ijsa.2012.20.issue-1

Gierl, M. J., Bulut, O., Guo, Q., & Zhang, X. (2017). Developing, analyzing, and using distractors for multiple-choice tests in education: A comprehensive review. *Review of Educational Research*, *87*(6), 1082–1116. doi:10.3102/0034654317726529

Green, B. F., Crone, C. R., & Folk, V. G. (1989). A method for studying differential distractor functioning. *Journal of Educational Measurement*, *26*, 147–160. doi:10.1111/jedm.1989.26.issue-2

Grier, B. (1976). The optimal number of alternatives at a choice point with travel time considered. *Journal of Mathematical Psychology*, *14*, 91–97. doi:10.1016/0022-2496(76)90016-X

Grier, J. B. (1975). The number of alternatives for optimum test reliability. *Journal of Educational Measurement*, *12*, 109–112. doi:10.1111/jedm.1975.12.issue-2

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

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

Haladyna, T. M., & Downing, S. M. (1993). How many options is enough for a multiple-choice test item. *Educational and Psychological Measurement*, *53*(4), 999–1010. doi:10.1177/0013164493053004013

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

Haladyna, T. M, & Rodriguez, M. C. (2014). *Developing and validating test items*. New York, NY: Routledge. doi:10.4324/9780203850381

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

Lee, H., & Winke, P. (2012). The differences among three-, four-, and five-option-item formats in the context of a high-stakes English-language listening test. *Language Testing*, *30*(1), 99–123. doi:10.1177/0265532212451235

Lord, F. M. (1977). Optimal numbers of choice per item—a comparison of four approaches. *Journal of Educational Measurement*, *14*, 33–38. doi:10.1111/j.1745-3984.1977.tb00026.x

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

Penfield, R. D. (2008). An odds ratio approach for assessing differential distractor functioning effects under the nominal response model. *Journal of Educational Measurement*, *45*(3), 247–269. doi:10.1111/jedm.2008.45.issue-3

Penfield, R. D. (2010). Modeling DIF effects using distractor-level invariance effects: Implications for understanding the causes of DIF. *Applied Psychological Measurement*, *34*(3), 151–165. doi:10.1177/0146621609359284

Raymond, M. R., Stevens, C., & Bucak, S. D. (2019). The optimal number of options for multiple-choice questions on high-stakes tests: Application of a revised index for detecting nonfunctional distractors. *Advances in Health Sciences Education*, *24*, 141–150. doi:10.1007/s10459-018-9855-9

Redmond, S. P., Hartigan-Rogers, J. A., & Cobbett, S. (2012). High time for a change: Psychometric analysis of multiple-choice questions in nursing. *International Journal of Nursing Education Scholarship*, *9*(1), 1–16. doi:10.1515/1548-923X.2487

Rodriguez, M. C. (2002). Choosing an item format. In G. T. Tindal & T. M. Haladyna (Eds.), *Large-scale assessment programs for all students: Validity, technical adequacy, and implementation* (pp. 213–231). Mahwah, NJ: Lawrence Erlbaum Associates.

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

Rodriguez, M. C., Kettler, R. J., & Elliott, S. N. (2014). Distractor functioning in modified items for test accessibility. *Sage Open*, *4*(4), 1–10. doi:10.1177/2158244014553586

Salazar Blanco, O. F., Vélez, C. M., & Zuleta Tobón, J. J. (2015). Evaluación de conocimientos con exámenes de selección multiple: tres or quarto opciones de respuesta? [Knowledge assessment with multiple-choice tests: Three or four response options?]. *IATREIA*, *28*(3), 300–311.

Schneid, S. D., Armour, C., Park, Y. S., Yudkowsky, R., & Bordage, G. (2014). Reducing the number of options on multiple-choice questions: Response time, psychometrics and standard setting. *Medical Evaluation*, *48*, 1020–1027.

Suh, Y., & Bolt, D. M. (2011). A nested logit approach for investigating distractors as causes of differential item functioning. *Journal of Educational Measurement*, *48*(2), 188–205. doi:10.1111/jedm.2011.48.issue-2

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*, 40–47. doi:10.1186/1472-6920-9-40

Taylor, A. K. (2005). Violating conventional wisdom in multiple choice test construction. *College Student Journal*, *39* (1), 141–148.

Thissen, D, Steinberg, L, & Fitzpatrick, A. R. (1989). Multiple-choice models: the distractors are also part of the item. *Journal Of Educational Measurement*, *26*(2), 61-76. doi:10.1111/jedm.1989.26.issue-2

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

Vegada, B., Shukla, A., Khilnani, A., Charan, J., & Desai, C. (2016). Comparison between three option, four option and five option multiple choice question tests for quality parameters: A randomized study. *Indian Journal of Pharmacology*, *48*(5), 571–575. doi:10.4103/0253-7613.190757

Wainer, H. (1989). The future of item analysis. *Journal of Educational Measurement*, *26*, 191–208. doi:10.1111/jedm.1989.26.issue-2

Zoanetti, N., Beaves, M., Griffin, P., & Wallace, E. M. (2013). Fixed or mixed: A comparison of three, four and mixed-option multiple-choice tests in a fetal surveillance education program. *BMC Medical Education*, *13*, 35–45. doi:10.1186/1472-6920-13-35