Article

Distractor Functioning in Modified Items for Test Accessibility

SAGE Open
October-December 2014: 1–10
© The Author(s) 2014
DOI: 10.1177/2158244014553586
sgo.sagepub.com
SAGE

Michael C. Rodriguez[1], Ryan J. Kettler[2], and Stephen N. Elliott[3]

Abstract

Effective distractors in multiple-choice items should attract lower ability students, those with misconceptions or limited knowledge and skills, because they are based on common misconceptions or errors in logic. A large, multi-state data set collected for a quasi-experimental study on test modifications was analyzed to measure the impact on distractor functioning. The key modification of interest was the removal of the weakest of three distractors, from 39 items in reading and 39 items in mathematics. Distractor functioning was neither systematically improved nor systematically weakened through the modification process. However, more than 70% of the distractors became more discriminating. A moderate correlation between distractor selection rate and distractor discrimination, in mathematics, may have indicated that the modified items were being missed by the appropriate students. Implications of these findings for test developers are discussed.

Keywords

special education, education, social sciences, educational measurement and assessment, achievement, reliability and validity

Item writing has received a boost in attention recently, largely because of the high-stakes demands that have been placed on achievement testing through No Child Left Behind (NCLB) requirements, the increasing attention to international assessments (e.g., Programme for International Student Assessment [PISA], Trends in International Mathematics and Science Study [TIMSS]), a renewed interest in the use of assessments for formative purposes, and innovations in online item and assessment delivery. Within this attention has been a focused effort to create test items that are accessible to students with a wide range of abilities, particularly those students with disabilities. Under the NCLB Act, this attention was motivated through the allowance for an alternate assessment based on modified academic achievement standards (AA-MAS) for a small proportion of students. A large part of this renewed attention is due to the expanded use of the principles of Universal Design in the item-development process and attention to language complexity providing greater access for students learning English as a second language.

Item writing has received limited attention in the measurement research literature, where many argue that the science of item writing is underdeveloped (Haladyna & Rodriguez, 2013). Some of this work began in the 1920s, where testing specialists conducted research on variations of item formats. Only a handful of the item-writing guidelines in the current literature have been tested empirically, including guidelines regarding the number of options used, the use of none-of-the-above and all-of-the-above options, the complex type-K format, the use of negatively phrased stems, and option length (Haladyna, Downing, & Rodriguez, 2002). Issues related to the number of response options have been studied about 3 times as much as any others. In particular, Rodriguez's (2005) meta-analysis of 27 published articles addressed the question, "What is the optimal number of response options for a multiple-choice test?" Rodriguez concluded that three options are optimal for multiple-choice items in most settings. He found that moving from five or four options to three options had little to no effect on multiple-choice item difficulty and discrimination and test score reliability on average.

Attention to distractors has been the most popular area for item-writing research. The earliest leadership on item writing was provided by Ebel (1951) in his seminal chapter on the topic. Haladyna and Downing (1988) presented a framework for developing functional distractors. Haladyna and Downing (1989a, 1989b) later presented a taxonomy of item-writing guidelines based on a review of textbook author recommendations and supplemented that with validity-related evidence from the empirical research literature. This evidence was updated and comprehensively reviewed by Haladyna and Rodriguez (2013).

[1]University of Minnesota, Minneapolis, USA
[2]Rutgers, The State University of New Jersey, Piscataway, USA
[3]Arizona State University, Tempe, USA, and Australian Catholic University, Brisbane, AU

Corresponding Author:
Michael C. Rodriguez, University of Minnesota, 250 Education Sciences, 56 East River Road, Minneapolis, MN 55455, USA.
Email: mcrdz@umn.edu

From Distractors to Attractors

Several states where AA-MASs have been developed are using the reduction of an item response choice as one of their key strategies for improving accessibility and reducing the difficulty of items for students with disabilities (Lazarus, Thurlow, Christensen, & Cormier, 2007). Rodriguez's (2005) findings indicate that reducing the number of distractors does not harm the psychometric properties of the test within the general population, and may reduce the cumulative cognitive load of the test. Modifications to ensure that only plausible distractors are used are truly in the spirit of making items more accessible to students with disabilities. Through combining concepts of Universal Design (Center for Universal Design [CUD], 2008), good item-writing principles (Haladyna et al., 2002), and cognitive load theory (Clark, Nguyen, & Sweller, 2006), the use of three-option items is meaningful, with careful attention to the item distractors, or what could more meaningfully be called "attractors." The term *distractor* originated in classic item-writing literature (Ebel, 1951). The intent is to suggest that the incorrect options "distract" the student with limited knowledge and understanding. In the context of the principles described earlier, a more productive intent of an incorrect option is to attract those students with specific misconceptions or errors in knowledge, reasoning, and problem solving.

Considering the "attractor" function of the incorrect options explicitly requires greater attention to their design because they must contain information about misconceptions or errors to attract the right students. This is not different than what is found in item-writing guidelines, but in practice, most items appear to be written such that incorrect options are not functioning well, largely because they do not conform to these principles (Rodriguez, 2005). Here, we argue that the lessons learned about distractor quality from item-modification experiments across several states have important implications for item writing and test design more generally, potentially improving item quality, and thus test score quality, for all students.

Item Accessibility

Distractors constitute the most important element of a multiple-choice item. They also present the most challenging aspects of an item to the item writer. The difficulty of an item is most easily manipulated by the nature of the distractors, particularly in their proximity. Item-writing researchers have found that the plausibility and proximity (similarity) of the distractors has a much greater impact on item difficulty than do characteristics of the stem, for example, whether the stem is a complete question or open-ended statements completed by the options (Ascalon, Meyers, Davis, & Smits, 2007; Haladyna et al., 2002). Consider two versions of the same test question:

1. Who was elected President of the United States in 1932 during the Great Depression?

   A. Daniel Boone
   B. Dwight D. Eisenhower
   C. Ronald W. Reagan
   D. Franklin D. Roosevelt (correct option)

2. Who was elected President of the United States in 1932 during the Great Depression?

   A. Calvin Coolidge
   B. Herbert Hoover
   C. Franklin Roosevelt
   D. Theodore Roosevelt

First, note that the options are in alphabetical order by last name, a good item-writing technique. However, in the first version, not all options are plausible, because Daniel Boone was not a president; remaining options are quite different in plausibility, as Reagan was president far more recently than Eisenhower or Roosevelt. In the second version of the item, all options are presidents, three of which were president between 1923 and 1945. Theodore Roosevelt might be an interesting option as he had the same last name as the correct response; however, this might present a clue as there are two Roosevelt's: "It must be one of them." Nevertheless, the first version of the item is likely to be much easier than the second. Perhaps the most important consideration is that this is not a particularly interesting item as it is tapping simple recall. Consider two versions of another item (based on a similar item by Haladyna, 1999):

3. What is the most effective method to reduce the internal air temperature of a house in a humid subtropical climate?

   A. Fan
   B. Evaporative cooler
   C. Air conditioner (correct option)
   D. Dehumidifier

4. What is the most effective way to cool a home in a humid climate?

   A. Air conditioner
   B. Evaporative cooler
   C. Fan

Item 3 presents several problems. First, the stem is wordy and unnecessarily technical. One of the options, Dehumidifier, contains part of a key word in the stem, humid, which presents a clang association potentially leading to response errors that are construct-irrelevant. Depending on the region, the use of an evaporative cooler may not be familiar. This is also a curriculum issue—the simple evaporation of water tends to cool the air. So the question is a complex one, requiring students to evaluate each option, and select the most effective, as more than one will cool the air (e.g., dehumidifiers reduce the humidity, but tend to create more heat in the process). A minor point is that the options are not in any particular order, where alphabetical, numerical, or some other logical order provides for a standard method of ordering options. Item 4 corrects most of the faults of the first version and retains the complex nature of the question.

When considering modifications of items for students with disabilities facing persistent academic difficulties, the use of standard item-writing guidelines can improve the quality of resulting responses and improve measurement overall—but also for all students (Elliott et al., 2010). From these examples, it is clear that the options are an important part of the item and the functioning of the distractors is essential. The presence of obviously irrelevant options takes time (an important resource) for students to consider and reason about their relevance (which may be a more difficult task for students with learning difficulties and cognitive impairments), potentially distracts students from thinking clearly about the construct, reduces the measurement power of the item, and eliminates the opportunity to obtain additional information about students' misconceptions or reasoning errors (because they are not present in each options).

A nonfunctioning option is one that is not selected by students or does not discriminate between high and low ability students. Nonfunctioning options typically are not plausible. The elimination of nonfunctioning options promotes several goals in making test items more accessible to all students, particularly by reducing the per-item testing time, reducing the required amount of reading, and eliminating potential sources of confusion. By using the label "distractor," these issues are not central concerns to the item writer—why an option distracts a student becomes less salient. When we enter the item-writing task or item-modification task using the language of "attractors," our attention is focused on the explicit role of the attractor in presenting a plausible challenge (an attractive but incorrect alternative) to the correct response. At the same time, we need to make sure that the distractors are attracting the right students: those students with misconceptions or reasoning errors, and those students who tend to be of relatively lower ability. We want to avoid distractions and unnecessary words for students who generally are poor readers and have a history of poor test performances. By having fewer answer choices and emphasizing the attraction aspect of distractors, item reliability and test score validity can be improved for all students, not just those with disabilities (Rodriguez, 2009, 2011).

Purpose of the Current Study

The current study originated during a U.S. Department of Education (USDE) funded project, The Consortium for Alternate Assessment Validity and Experimental Studies (CAAVES), a four-state collaboration to modify items for new state alternate assessments. AA-MAS are designed to provide access to the assessment of state standards for students receiving special education services who have a history of academic difficulties and who are unlikely to attain proficiency on the general education assessment even with appropriate instruction and accommodations. Through this project, researchers and state personnel modified test items to make them more accessible, reduce cognitive load, and improve the validity of results for those students who otherwise would not be able to display their knowledge and skills.

A quasi-experimental design was employed with three groups of students: (a) students without disabilities and students with disabilities that were either (b) eligible or (c) not eligible for the AA-MAS (determined through specific participation criteria). Each student completed one set of 13 items in each of the three formats: original, modified without reading support, and modified with reading support. The original four-option multiple-choice items from an existing data bank containing item statistics were modified by removing one option that was either not functioning (based on item statistics) or was implausible (i.e., did not contain useful information about misconceptions, based on content-expert review of the options).

Additional modifications were made to reduce the complexity of language and sentence structure and provide greater access to the intent of the item, including the use of carefully selected graphics or pictures (see Figures 1 and 2, for examples, Items 5-8). The study included a carefully balanced design with rotation of modified items from the beginning to the end of the test across students, such that each student was exposed to items in original and modified format. The results of this study have been analyzed in terms of shifts in item difficulty and test score reliability (Elliott et al., 2010; Kettler et al., 2011). Although there were significant group by condition interaction effects on test score reliability, none of the differences (less than .06) were very meaningful. No matter how else the data were cut (group, condition, order of item set within the test form), all other differences were less than .02. The researchers suggested that this provided evidence to support systematic modifications without undermining score consistency. Elliott and colleagues then used the Rasch model to assess modification effects on item difficulty (Rasch was used to equate group ability distributions). They found item difficulties to be reduced through item modification on both tests and this effect was greater for eligible students than for students without disabilities. This supported the researchers' interaction hypothesis, such that students eligible to participate in AA-MAS experienced a greater benefit from the modification than students who were not eligible.

Using the Rasch model to control for individual (and group) ability, the changes in average item difficulty were evaluated for each group (see Kettler et al., 2011, for a complete review of results). Generally, the Rasch model produces a scale score with a standard deviation of 1.0 and negative change in item difficulty indicates the items become easier (requiring less ability to respond correctly). The changes for

5. Which stem-and-leaf is the best representation of the data reported here?

11, 14, 14, 15, 21, 24, 25, 25, 26, 26, 27, 29, 33, 33, 34, 34, 34, 35, 35, 37,
42, 42, 42

A.

1	1	
2	1	
3	3	
4	2	

B.

1	1, 4, 4, 5	(correct option)
2	1, 4, 5, 5, 6, 6, 7, 9	
3	3, 3, 4, 4, 4, 5, 5, 7	
4	2, 2, 2	

C.

1	1, 4, 5	
2	1, 4, 5, 6, 7, 9	
3	3, 4, 5, 7	
4	2	

D.

1	11	
2	21	
3	33	
4	42	

6. Which stem-and-leaf is the best representation of the data reported here?
11, 14, 14, 15, 21, 25, 25, 26, 26, 29, 33, 34, 34, 34, 35

A.

1	1, 4, 4, 5	(correct option)
2	1, 5, 5, 6, 6, 9	
3	3, 4, 4, 4, 5	

B.

1	1, 4, 5	
2	1, 4, 5, 6, 9	
3	3, 4, 5	

C.

1	11	
2	21	
3	33	

Figure 1. A sample mathematics test question in original format (Item 5) and modified format (Item 6).

students without disabilities (−0.03 for reading and −0.12 for mathematics) and those with disabilities but not eligible for the alternate assessment (−0.07 for reading and −0.15 for mathematics) were very small. The changes in item difficulty for students with disabilities eligible for alternate assessment were moderate and meaningful (−0.40 for reading and −0.34 for mathematics).

The central purpose of this study was to examine in detail the distractor functioning of modified test items to improve test performance of all students. It was hypothesized that employing a package of modifications to test items designed to improve accessibility would improve the psychometric quality of the item distractors, the items, and the tests for all students. The data employed to conduct the item-level analyses are pooled across student groups—as our focus is on the items and not subgroup performance.

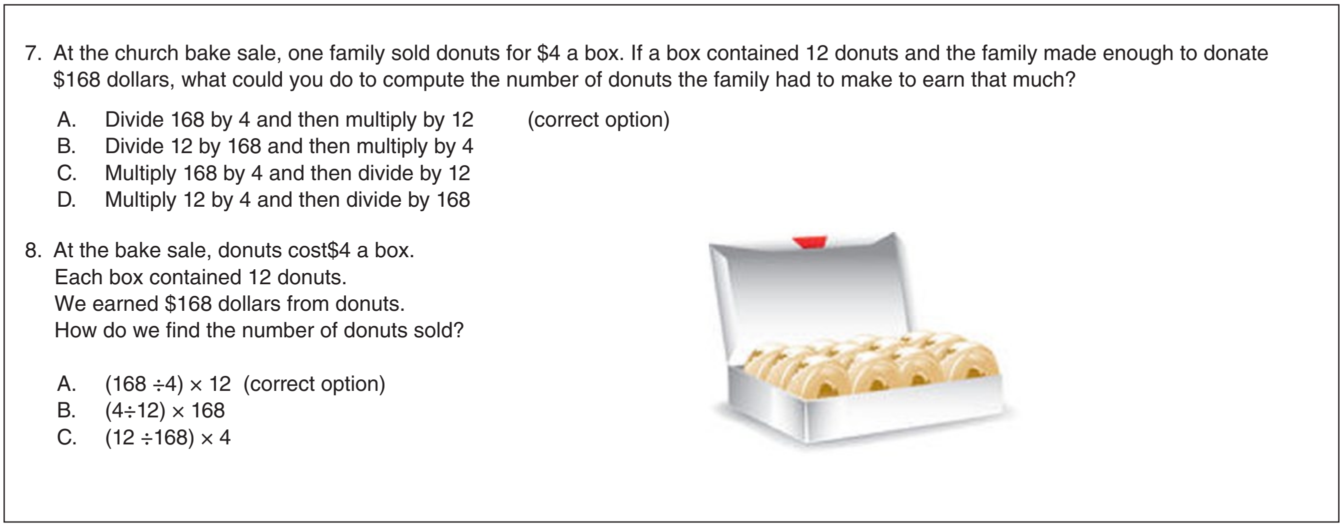
Method

Participants

The sample included 755 eighth-grade students from four states, including students with (n = 486) and without (n = 269) disabilities. The students were approximately 58% male and 69% White.

Measures

The study included mathematics and reading tests, both composed of 39 multiple-choice items. The items were provided by Discovery Education Assessment from a pool of items identified to meet common state standards in both areas. The mathematics test included 20 number items requiring decoding of mathematical symbols and basic operations and 19



7. At the church bake sale, one family sold donuts for $4 a box. If a box contained 12 donuts and the family made enough to donate $168 dollars, what could you do to compute the number of donuts the family had to make to earn that much?

   A. Divide 168 by 4 and then multiply by 12 (correct option)
   B. Divide 12 by 168 and then multiply by 4
   C. Multiply 168 by 4 and then divide by 12
   D. Multiply 12 by 4 and then divide by 168

8. At the bake sale, donuts cost$4 a box.
   Each box contained 12 donuts.
   We earned $168 dollars from donuts.
   How do we find the number of donuts sold?

   A. (168 ÷4) × 12 (correct option)
   B. (4÷12) × 168
   C. (12 ÷168) × 4

Figure 2. A sample mathematics test question in original format (Item 7) and modified format (Item 8).

data items requiring basic arithmetic operations. The reading test included 20 comprehension items and 19 vocabulary items. Each item was modified using the principles discussed above and those summarized in the *Test Accessibility and Modification Inventory* (TAMI; Beddow, Kettler, & Elliott, 2008). The TAMI includes guidance for modifying (developing) the passage and/or item stimulus, item stems, visuals, answer choices, and page format and layout.

When the tests were assembled, they contained three sets of 13 items (39 items in total), where one set was in original format, one set was modified, and the third set was modified with the addition of reading support (a recorded voice that read item directions and stems). The three forms of items were rotated across each of the three sets of 13 items and across the three positions of the test (first, second, and third set of 13 items). This balanced item order and item format. The tests were administered by computer. Coefficient alpha was .89 for reading and .85 for mathematics.

Study Design and Analyses

To extend the analyses of the item data, we examined the functioning of the options for each item. All analyses were across students (student group was not of interest for these purposes) and the item format was considered to be either original or modified. First, we reexamined the effect of modification on overall item statistics, including item difficulty and discrimination (classical test statistics). Then, the distractor discrimination values were examined for each item. The distractor discrimination index is the point-biserial correlation between the selection of a distractor (0 for not selected or 1 for selected) and the total score. Ideally, distractor discrimination indices should be negative, indicating that the selection of a distractor is associated with a lower total score overall. Item difficulty, item discrimination, and distractor discrimination were estimated using Winsteps 3.65, a Rasch analysis program, which also estimates classical test statistics.

With modification, we expected an increase in the item *p* value (more students respond correctly making the item easier) and because the measurement properties should improve, we expected an increase in the item discrimination. Finally, the two distractors remaining should result in stronger discrimination as well, where we expected the discrimination index to be negative and larger in the modified version than the original version.

Results

Item Difficulty

Classical item difficulty was examined in terms of the item *p* value (proportion correct). We found the average change in item difficulty was about 6% for mathematics and 10% for reading, making the items easier on average for both tests. By comparison, Rodriguez (2005) reported an average increase in the percent correct of about 4.4% when reducing the number of options from 4 to 3 across 36 studies in his meta-analytic review. Overall, 6 of 39 items became more difficult in mathematics; only 2 items became more difficult in reading; the remaining became easier. Table 1 contains summary statistics for mean item difficulty changes between original and modified formats.

Item Discrimination

Item discrimination was based on the corrected point-biserial correlation between the item and the total score (excluding

Table 1. Summary Statistics of Differences in Item Percent Correct From Original to Modified Format.

Subject	Minimum (%)	Maximum (%)	*M* (%)	*SD* (%)
Mathematics	−7	18	6	6
Reading	−12	35	10	10

Table 2. Summary Statistics of Differences in Item Discrimination From Original to Modified Format.

Subject	Minimum	Maximum	*M*	*SD*
Mathematics	−.30	.16	−.05	.08
Reading	−.16	.32	−.01	.11

Table 3. Summary Statistics of Differences in Distractor Discrimination From Original to Modified Format.

Subject	Minimum	Maximum	*M*	*SD*
Mathematics	−.27	.20	.04	.08
Reading	−.23	.23	.06	.09

that item). Item discrimination results were transformed to Fisher's *Z* for statistical analyses. We found the average change in item discrimination was −.05 for mathematics and −.01 for reading, both small but in the unexpected direction. By comparison, Rodriguez (2005) found an average increase in item discrimination of .03 when the number of options was reduced from 4 to 3 based on 30 studies. Overall, in mathematics, 11 of 39 items increased in discrimination; in reading, 17 items increased in discrimination. Table 2 contains summary statistics for mean item discrimination changes between original and modified formats.

Distractor Functioning

An effective distractor is one that attracts students with misconceptions or errors in thinking and reasoning, generally those with lower overall ability. There are two common indices that help us assess the effectiveness of a distractor, including the response rate (number of students selecting the distractor) and the distractor discrimination (corrected point-biserial correlation). Table 3 contains summary statistics for changes in distractor functioning from original to modified format. For both mathematics and reading, on average, distractors became more discriminating in modified format. In mathematics, 71% of all distractors across the 39 items became more discriminating in modified format; in reading, 78% of all distractors became more discriminating.

A sample item-by-item analysis of distractor functioning is provided in Tables 4 and 5. For five items, the selection frequency of every option and the option point-biserial

Table 4. Mathematics Item Response Frequencies.

			Original format			Modified format		
Item	Option	Key	Count	%	Ptbs	Count	%	Ptbs
1	A		10	4	−.17			
	B	*	161	69	.38	323	68	.40
	C		36	16	−.23	99	21	−.25
	D		25	11	−.18	53	11	−.27
2	A		27	12	−.28	63	13	−.28
	B	*	145	62	.48	315	66	.38
	C		48	21	−.26	97	20	−.21
	D		13	6	−.17			
3	A	*	139	60	.37	320	67	.35
	B		24	10	−.25			
	C		39	17	−.21	88	19	−.34
	D		31	13	−.08	67	14	−.1
4	A		35	15	−.10	95	20	−.25
	B		31	13	−.23	104	22	−.26
	C	*	126	54	.43	276	58	.41
	D		40	17	−.27			
5	A		95	41	−.26	158	33	−.27
	B		26	11	−.20	108	23	−.33
	C	*	94	40	.49	209	44	.53
	D		18	8	−.20			

Note: The asterisks in the Key column indicates the correct option (key). Ptbs = Point-biserial correlation, the item discrimination.

Table 5. Reading Item Response Frequencies.

			Original format			Modified format		
Item	Option	Key	Count	%	Ptbs	Count	%	Ptbs
1	A	*	182	76	.57	396	86	.49
	B		24	10	−.35	30	7	−.36
	C		13	5	−.27			
	D		20	8	−.27	33	7	−.31
2	A		24	10	−.28	46	10	−.32
	B	*	166	69	.56	345	75	.48
	C		40	17	−.33	68	15	−.30
	D		10	4	−.27			
3	A		58	24	−.27	46	10	−.46
	B		24	10	−.37			
	C		21	9	−.23	35	8	−.36
	D	*	136	57	.59	377	82	.61
4	A		26	11	−.17	79	17	−.36
	B		77	32	−.11			
	C	*	78	33	.30	310	68	.59
	D		59	25	−.08	68	15	−.40
5	A		8	3	−.19	22	5	−.35
	B		27	11	−.33	40	9	−.37
	C		10	4	−.29			
	D	*	194	81	.50	395	86	.53

Note: The asterisks in the Key column indicates the correct option (key). Ptbs = Point-biserial correlation, the item discrimination.

Table 6. Summary Statistics of Change in Distractor Response Rates From Original to Modified Format.

Subject	Minimum (%)	Maximum (%)	*M* (%)	*SD* (%)
Mathematics	−12	11	−0.3	4.9
Reading	−35	22	−1.2	7.3

correlations are listed, for the item in original and modified format.

In addition to examining the change in item and distractor statistics, we examined the response rate and the change in response rate for each distractor. To facilitate the comparison, we computed the percent responding to each distractor, using only the three options that remained in the modified item, but for both the original and modified items—thus, we ignored the responses to the option that was deleted and recomputed the percent responding based on the number responding to the remaining options (so proportional comparisons are equivalent). There were two reasons for considering an analysis of distractor response rates. First, if a distractor was removed, the remaining distractors should be relatively more plausible, and second, differences in response rates between distractors should be reduced—distractors should be plausible and relatively more equally attractive.

These results are less conclusive and should be interpreted with caution, considering two possible limitations. In the original version of the item, if an option was eliminated, it is not clear that the modified option response rates would be observed because modifications were also made to the language and layout of many items. The recomputation of option response rates on original items is a hypothetical comparison, ignoring the fact that those students would have selected other options (possibly the correct option) if the eliminated option was not present originally. In addition, when an option is removed, the item tends to become easier, meaning that more students are selecting the correct response and not one of the distractors. Nonetheless, the examination of distractor response rates is useful in the analysis of distractor functioning.

On the mathematics test, on average, distractor response rates decreased 0.3% (meaning, 0.3% fewer students selected the remaining distractors in the modified version than the original version), when examining only the two distractors retained through modification. On average, 10.3% of students selected one of the removed options (ranging from 1% to 22%). This suggests that although distractor response rates changed from item to item, overall there was no change in the selection of distractors.

On the reading test, the results were more varied, where on average, distractor response rates decreased by about 1.2%, but varied more significantly across items (*SD* = 7.3). Again, the distractor response rates changed very little (with fewer students selecting the retained distractors). Table 6 contains the summary statistics for this change on both tests.

Table 7. Summary Statistics of Change in Differences in Distractor Response Rates From Original to Modified Format.

Subject	Minimum (%)	Maximum (%)	*M* (%)	*SD* (%)
Mathematics	−20	11	1.0	6.7
Reading	−37	27	0.9	9.7

As a final analysis, we examined the difference in response rates between the two retained distractors (recall that each modified item has one correct option and two distractors). The question is whether the difference in distractor response rates changed given the presence (original format) or absence (modified format) of one of the options. Ideally, distractors are equally plausible. To estimate this, the difference in distractor response rates was computed (if A and B are the two distractors, then we estimated % responding to A minus % responding to B). We computed the difference between the distractor response-rate-differences for each format (difference in distractor response rates for modified version minus original version):

$$\begin{aligned}&\left|\text{Modified \% Selecting A} - \text{\% Selecting B}\right| \\ &-\left|\text{Original \% Selecting A} - \text{\% Selecting B}\right|.\end{aligned}$$

On the mathematics test, the average change in response rate differences was 1%, suggesting that overall, the difference in response rates of the two distractors did not change when a third distractor was removed. On the reading test, the average change in response rate differences was less than 1%. However, across items, there were substantial changes in distractor response rate differences, where some differences in distractor response rates decreased by up to 37% or increased as much as 27%. Table 7 contains the summary statistics for this change in response rate differences.

In mathematics, there was a small to moderate relation between improvement in distractor discrimination and an increase in the selection rate of distractors (*r* = .39). This suggests one of two things: (a) that the item distractor discrimination improved as result of attracting more of the right students (lower ability students); or (b) that because distractor response rates increased, this provided more variation in responses that resulted in a higher correlation with test performance making the distractors more discriminating.

This relation between improvement in distractor discrimination and increase in selection rate of distractors was not present in reading (*r* = .06), even though the variation in differences in distractor response rates was much greater, providing for more variance to estimate the correlation. This suggests that in mathematics, the relation was a function of the distractors attracting more of the right students. This draws attention to the importance of plausible distractors that are attractive to lower ability students—those with misconceptions and reasoning errors. This of course requires the distractors to be based on misconceptions and reasoning errors.

Discussion

Common modifications to achievement tests for students with disabilities have included the reduction of the number of multiple-choice options, simplification of language, and reduction of the number of words used. The present study focused on item response distractors and provided practical information regarding their use for all students. The experimental modifications were intended to make the items more accessible and in doing so, improve the measurement qualities of the items. This particular package of modifications was expected to make the items more discriminating, and to result in more effective distractors such that they would be selected at a higher rate and more effectively discriminate between students with higher and lower ability. A moderate decrease in difficulty was also expected.

We, in fact, found modified items were easier overall, with an average increase in correct responses of 6% for mathematics and 10% for reading. Item discrimination did not fare as well. Negligible decreases of .05 for discrimination in mathematics and .01 for discrimination in reading were observed. Overall, 11 of the 39 items increased in discrimination for mathematics and 17 of the 39 items increased in discrimination for reading.

The key analyses for this project were at the distractor level, assessing the effect of modification on the functioning of the distractors. For both mathematics and reading, the retained distractors became more discriminating. In mathematics, 79% of all distractors across the 39 items became more discriminating (an average of .04) and 71% of all distractors in reading became more discriminating (an average of .06). The response rates to distractors did not change in either mathematics (lower by 0.3% on average) or reading (lower by 1.2%). Finally, differences in response rates to the two distractors within an item did not change either, where the change was 1% or less for both mathematics and reading.

These results are similar to those found in a study of nursing assessments. When removing the distractor with the lowest response rate, three-option items contained more functioning distractors than four-option items (even with fewer distractors), and the two remaining distractors became more discriminating when the least functioning distractor was removed (Tarrant & Ware, 2010).

The analyses in this study included the examination of distractor functioning. Distractor functioning is difficult to assess as most of the statistics typically reported are affected by the difficulty of the item. For example, if the item is very easy, then distractor response rates are naturally low. If there are few respondents selecting a particular distractor, then the correlation between the selection of the distractor and the total score will be affected by the limited variation due to low selection rates. Nonetheless, the trends in these results are of interest. Closer examination of item characteristics is important in evaluating the overall effect of modification on item discrimination as results varied a great deal across items. Similarly, several items actually became more difficult following modification; such items should be reviewed as well.

Reviewers of an earlier version of this article raised questions about the trend for some items to result in lower overall discrimination in modified form and the possibility that guessing was responsible. The probability of guessing correctly is perhaps the greatest fear preventing more common use of three-option items; however, experimental evidence suggests that the effects of guessing are negligible across subject areas and age groups (Rodriguez, 2005). We relied on this evidence to support the argument that the reduction in overall item discrimination is a function of the slight decrease in difficulty, making the item variance slightly smaller. We also think that this alone is probably not enough to explain the trend. But again, we found distractor discrimination increased across most items. If guessing was uniformly increasing due to elimination of one distractor, the effects would similarly be seen in the distractor statistics. Moreover, we found very small changes in distractor selection rates between original and modified versions. Finally, no changes in test score reliability across original and modified forms also indicates no significant change in random responses or guessing.

Using Distractor Information in Item Reviews

The role of distractors has become more salient as educators demand instructionally relevant information. This means that we need information about what students know and can do as well as information about the misconceptions or errors in problem solving students continue to use. Testing companies are now developing sound principles of item development that focus attention on the contributions of distractors (e.g., King, Gardner, Zucker, & Jorgensen, 2004). Such procedures include distractor analysis to ensure distractors are relatively equally plausible and selected by lower ability test takers, those with misconceptions.

The item, as the building block of a measure, must function in a way that contributes to the overall measure. In his review of item analysis, Livingston (2006) reminds us that when item difficulty is not as expected or not within the target range, given the goals of measurement, we should review distractor performance. This often provides a clue as to why an item may be too easy or too difficult. Similarly, when an item does not achieve the target level of discrimination, an examination of distractor functioning is important. He argued that complete item analysis is important during three stages, including item pretesting, before scoring, and after scores have been reported.

We have decades of evidence of the importance of information found in distractor response patterns. The pattern of incorrect option choice is related to ability in important and informative ways: For high ability test takers, only one or two options are likely to be selected, and for middle ability test takers, other options are more commonly selected (Levine & Drasgow, 1983). Several item response theory (IRT) models have been introduced to evaluate the quality of distractors, recognizing their important role in comprehensive item evaluation (Penfield & de la Torre, 2008; Thissen, Steinberg, & Fitzpatrick, 1989). Distractor discrimination, trace lines, and IRT models examining distractor performance are useful for item design, item analysis, and item modification.

As we observed in Tables 4 and 5, the direction of changes in item discrimination and distractor discrimination are not always consistent; as one improves, the others do not necessarily improve. For most of the items, distractor discrimination improved substantially following modification. In some cases, total item discrimination declined, but in many cases, very little—for those items in Table 4, we observe slight declines in item discrimination, where all final values are at least .35 and in Table 5, the modified item discrimination values were no less than .48, values that contribute to overall score quality. In math, the decline was typically about .05 and in reading it was typically .01. When selecting items of similar content, difficulty, and discrimination, additional information can be obtained by selecting the items with stronger distractor discrimination. Operational test design systems do not routinely include item selection models that accounts for distractor effectiveness. But because of the arguments presented here, we strongly recommend the use of distractor effectiveness in all stages of item development and selection for operational use.

Limitations

There are limitations in this study that are a function of the overall quality of some items. The items employed in this study were selected based on prior item statistics. As we know, item functioning can be a local characteristic (classical test theory statistics are sample specific). There were several items that performed poorly with this particular sample. In mathematics, there were 13 items with discrimination values less than .30 and 3 with values less than .20. In reading, there were 14 items with discrimination values less than .30 and 6 items with values less than .20.

Particularly troubling was one reading item, which in original format had an 18% correct response rate with discrimination of .06; in modified version, this item had a correct response rate of 19% and a discrimination of −.09, certainly not good for any purpose. This item was a vocabulary item employing a word that has several meanings depending on the syllable of emphasis. Although it was used in a sentence, the item required students to identify one of the alternate meanings if the stress was on the opposite syllable. Modification is not necessarily the answer to poorly conceived items. The impact of item-writing quality is potentially much greater than the impact of changes that are made through modification.

One additional limitation to note is that all of the items in the current study were modified in ways that went beyond elimination of the least plausible distractor. Many item stems were edited; had graphics modified, removed, or added; and were reformatted in a number of ways to improve accessibility, as described above. More information on this process can be found in Kettler et al. (2011). Changes to item and distractor statistics must be interpreted with this process in mind.

Conclusion

A reorientation in terms of item development will help. Asking item writers or editors to attend to the distractors requires explicit attention to the attractive aspect of the incorrect options. Asking ourselves: “Is this an effective attractor?” rather than “distractor” will improve our ability to be explicit about the intent of the option and characteristics of the students to whom it attracts. Improvements in item writing and item modifications that strive to make items and tests more widely accessible will contribute to the development of high-quality tests for wider audiences.

Authors’ Note

The positions and opinions expressed in this article are those solely of the author team.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research and/or authorship of this article: The current study was implemented as part of the Consortium for Alternate Assessment Validity and Experimental Studies (CAAVES) project, a multi-state project funded by the U.S. Department of Education (Award to Idaho Department of Education; #S368A0600012).

References

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.

Beddow, P. A., Kettler, R. J., & Elliott, S. N. (2008). *Test Accessibility and Modification Inventory*. Nashville, TN: Vanderbilt University.

Center for Universal Design. (2008). *The principles of universal design*. Retrieved from http://www.ncsu.edu/ncsu/design/cud

Clark, R., Nguyen, F., & Sweller, J. (2006). *Efficiency in learning: Evidence-based guidelines to manage cognitive load*. San Francisco, CA: Pfeiffer.

Ebel, R. L. (1951). Writing the test item. In E. F. Linquist (Ed.), *Educational measurement* (pp. 185-249). Washington, DC: American Council on Education.

Elliott, S. N., Kettler, R. J., Beddow, P. A., Kurz, A., Compton, E., McGrath, D., . . .Roach, A. T. (2010). Effects of using modified items to test students with persistent academic difficulties. *Exceptional Children*, *76*, 475-495.

Haladyna, T. M. (1999). *Developing and validating multiple-choice test items* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.

Haladyna, T. M., & Downing, S. M. (1988, April). *Functional distractors: Implications for test-item writing and test design*. Paper presented at the annual meeting of the AERA, New Orleans, LA.

Haladyna, T. M., & Downing, S. M. (1989a). A taxonomy of multiple-choice item writing rules. *Applied Measurement in Education*, *2*, 37-50.

Haladyna, T. M., & Downing, S. M. (1989b). Validity of a taxonomy of multiple-choice item writing rules. *Applied Measurement in Education*, *2*, 51-78.

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

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*, 309-334.

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

Kettler, R. J., Rodriguez, M. C., Bolt, D. M., Elliott, S. N., Beddow, P. A., & Kurz, A. (2011). Modified multiple-choice items for alternate assessments: Reliability, difficulty, and differential boost. *Applied Measurement in Education*, *24*, 210-234.

King, K. V., Gardner, D. A., Zucker, S., & Jorgensen, M. A. (2004). *The distractor rationale taxonomy: Enhancing multiple-choice items in reading and mathematics*. San Antonio, TX: Pearson. Retrieved from http://www.pearson-assessments.com/research.html

Lazarus, S. S., Thurlow, M. L., Christensen, L. L., & Cormier, D. (2007). *States' alternate assessments based on modified achievement standards (AA-MAS) in 2007* (Synthesis Report 67). Minneapolis: University of Minnesota, National Center on Educational Outcomes.

Levine, M. V., & Drasgow, F. (1983). The relation between incorrect option choice and estimated ability. *Educational and Psychological Measurement*, *43*, 675-685.

Livingston, S. A. (2006). Item analysis. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 421-441). Mahwah, NJ: Lawrence Erlbaum.

Penfield, R. D., & de la Torre, J. (2008). *A new response model for multiple-choice items*. Paper presented at the annual meeting of the National Council on Measurement in Education, New York, NY. Retrieved from http://testing.wisc.edu/research%20papers/DistractorModelNCME2008-1.pdf

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.

Rodriguez, M. C. (2009). Psychometric considerations for alternate assessments based on modified academic achievement standards. *Peabody Journal of Education*, *84*, 595-602.

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 students: Bridging the gaps between research, practice, and policy* (pp. 201-216). New York, NY: Springer.

Tarrant, M., & Ware, J. (2010). A comparison of the psychometric properties of three- and four-option multiple-choice questions in nursing assessments. *Nurse Education Today*, *30*, 539-543.

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.

Author Biographies

Michael C. Rodriguez is the Campbell Leadership Chair in Education and Human Development and a professor of Quantitative Methods in Education at the University of Minnesota. His research interests include item writing and item-response models, with substantive interests in early literacy and youth development. He consults internationally on item writing and test design.

Ryan J. Kettler is an associate professor of School Psychology at Rutgers, The State University of New Jersey. His research interests are within a program on data-based decision making, with active areas in universal screening, inclusive assessment, and teacher effectiveness.

Steve N. Elliott is the Mickelson Foundation Professor in the Sanford School of Social and Family Dynamics at Arizona State University and a Professorial Fellow in the Learning Sciences Institute at the Australian Catholic University. His research focuses on scale development and educational assessment practices with students with disabilities or at risk for educational difficulties.