

2020

A Comparison of the Two-Option Versus the Four-Option Multiple-Choice Item: A Case for Fewer Distractors

Allan Bateson

Office of Training, Education & Development, U.S. Food & Drug Administration

William R. Dardick

Department of Educational Leadership, George Washington University

Follow this and additional works at: <https://scholarworks.bgsu.edu/pad>



Part of the [Human Resources Management Commons](#), [Industrial and Organizational Psychology Commons](#), and the [Other Psychology Commons](#)

Recommended Citation

Bateson, Allan and Dardick, William R. (2020) "A Comparison of the Two-Option Versus the Four-Option Multiple-Choice Item: A Case for Fewer Distractors," *Personnel Assessment and Decisions*: Vol. 6 : Iss. 3 , Article 5.

DOI: <https://doi.org/10.25035/pad.2020.03.005>

Available at: <https://scholarworks.bgsu.edu/pad/vol6/iss3/5>

This Measurement and Measures is brought to you for free and open access by the Journals at ScholarWorks@BGSU. It has been accepted for inclusion in Personnel Assessment and Decisions by an authorized editor of ScholarWorks@BGSU.

A COMPARISON OF THE TWO-OPTION VERSUS THE FOUR-OPTION MULTIPLE-CHOICE ITEM: A CASE FOR FEWER DISTRACTORS

Allan Bateson¹ and William R. Dardick²

1. Office of Training, Education & Development, U.S. Food & Drug Administration

2. Department of Educational Leadership, George Washington University

ABSTRACT

KEYWORDS

multiple choice tests,
alternate choice format,
test reliability, plausible
distractors

Multiple choice test items typically consist of the key and 3-4 distractors. However, research has supported the efficacy of using fewer alternatives. Haladyna and Downing (1993) found that it is difficult to write test items with more than one plausible distractor, resulting in items with a correct answer and one alternative, also known as the alternate choice (AC) format. We constructed two 32-item tests; one with four alternatives (MC4) and one with two (AC), using an inter-judge agreement approach to eliminate distractors. Tests were administered to 138 personnel working for a U.S. Government agency. Testing time was significantly less and scores were higher for the AC test. However, score differences disappeared when both forms were corrected for guessing. There were no significant differences in test difficulty (mean p -values). The corrected KR-20 reliabilities for both forms, after applying the Spearman-Brown formula, were AC = .816 and MC4 = .893. We discuss the results with respect to the resources spent writing and reviewing test items, and in more broadly sampling a content domain using the AC format due to reduced testing times.

The multiple-choice item has been dominant throughout the testing industry and education over the past century. Tests constructed of multiple choice questions (MCQs) are generally judged easier to administer and score than tests using less objective formats. Also, test developers long thought that having more distractors (incorrect options) was preferred in order to reduce the possibility that test takers could guess the correct answers to test items (Haladyna & Downing, 1989; Owen & Froman, 1987; Sax & Reiter, 1980). However, over the past 50 years there has been much theoretical discussion, and empirical demonstration, extolling the merits of the three-option MCQ (Costin, 1970, 1972; Delgado & Prieto, 1998; Ebel, 1969; Grier, 1975; Lord, 1944; Rodriguez, 2005; Rogers & Harley, 1999; Sidick et al., 1994; Straton & Catts, 1980; Trevisan et al., 1994; Tversky, 1964). The result of this research points to the superiority of the three-option MCQ from two important perspectives: (a) based on a comparison of the psychometric properties, and (b) a reduction in the resources required in the test development and administration process; it takes less time to develop distractors and less time for test takers to complete tests with three versus four response options.

The purpose of this research was to explore the impact

of a further reduction to the number of response options, eliminating two distractors from a set of four-option, multiple-choice items, resulting in test items with two response options. Although the two-option format, also known as alternate choice (AC), has been of some interest (Rodriguez, 2005), Downing (1992) commented on the lack of research on the AC format and called for more to determine the potential utility of the format. Given the resources (time and money) required to write additional distractors and the difficulty in writing additional, plausible ones, we wanted to explore the possibility of developing and using tests with AC format items (Haladyna & Downing, 1993; Sidick et al., 1994).

Review of the Literature on the Number of Response Options

Despite the significant body of research supporting the case for three-option items, the four- and five-option MCQ remains the prevailing choice for high stakes (e.g., credentialing and education) testing. This has been the

Corresponding author:

Allan Bateson

Email: allan.bateson@fda.hhs.gov

case despite research showing no significant differences in item discrimination, item difficulty, or test reliability for tests employing the three-option versus either the four or five-option MSQ formats (Costin, 1970, 1972; Delgado & Prieto, 1998; Owen & Froman, 1987; Schneid et al., 2014; Sidick et al., 1994). Results from Owen and Froman (1987) showed no differences in item discrimination, item difficulty, or test scores for the three versus five-option MSQ tests. Sidick et al. (1994) found no practical differences in psychometric properties for employment tests consisting of either three- or five-option items. Costin (1970) found that mean discrimination indices were actually higher for the three-option than for the four-option item test measuring student knowledge of psychology. In a meta-analysis covering eighty years of research on multiple-choice items, Rodriguez (2005) reported increases in both item discrimination and reliability for three-option versus four-option MC tests.

We offer that there are at least two reasons for the persistent use of the four- and five-option MCQ tests despite the three-option format showing similar, or better, psychometric results. The first is that, all things being equal in terms of distractor performance, more distractors result in less opportunity to guess the correct answer to items, lower the average p -value of the test (increases the difficult level), and increase item and score reliabilities. The second reason has to do with the greater population of test takers more typical of high-stakes testing. Greater numbers may equate to a greater number of personnel available to participate in test-item writing, item reviewing, and pilot testing, where non-performing distractors may be identified in pilot testing and replaced prior to test use in a high-stakes environment. The first author has done testing in both a large organization ($N = 800,000$) and a much smaller one ($N = 1500$) and found it much easier to get ongoing SME assistance in the numbers required, across the testing process, from the larger one.

Also, it may be easier to set/adjust cut scores empirically in larger organizations rather than by using a judgment method such as Angoff (1971). The smaller number of personnel available for test preparation prior to operational use may preclude activities such as large-scale pilot testing and require organizations to use judgment approaches for setting cut scores. Personnel time away from their jobs and budgetary constraints are additional concerns for both smaller public- and private-sector organizations.

Several studies have addressed the issue of the time savings in test administration with fewer alternatives. Costin (1972) stated that students can more quickly complete items with three options than those with four options. This makes sense given that reading and evaluation time should be less. Aamodt and McShane (1992) conducted a meta-analysis of the impact of several test item characteristics on test scores and test completion times. They found that

testing time was significantly less for three-option than four-option tests. Where Aamodt and McShane analyzed test completion time, Schneid et al. (2014), using a computerized testing approach, were able to collect data on time to complete each item in a pharmacology exam. The authors found that students answered three-option MCQs on average five seconds faster (36 versus 41 seconds) than four- or five-option items. Testing time, whether per item or the entire test, is important in that significantly reduced testing times for tests with fewer options provides opportunities to either reduce testing time or to increase the length of the test while keeping administration time constant. Increasing test length with the same testing time may result in increased test reliability and validity, as well as a more thorough sampling of the content domain. In addition to the potential reduction in time needed for testing, Sidick et al. (1994) stated that test development time would also be less with fewer distractors, thereby saving both time and money. As a result, efficiencies may be realized in both test development and administration by using fewer options.

Although the debate continues with respect to the three-option MCQ, some attention has also been given to the two-option item, also referred to as alternate choice (AC; Downing, 1992; Ebel, 1982; Haladyna et al., 2002). Although there has been some debate regarding whether the AC item is really any different than the true-false (TF) item, Ebel (1982) proposed the AC format over the TF format for several reasons. He stated that AC items might be less ambiguous in that they provide specific options (best choice) rather than making judgments regarding absolute certainty. He also stated that AC items could be used to measure higher level cognitive processes, an ongoing issue with respect to MCQs in general.

Downing (1992) concurred with Ebel and stated that the AC item should be considered as a multiple-choice item with two options. He argued that there appear to be no real differences between AC items and those with more options in terms of mean item discrimination. AC items do appear to be easier, but that could be due to the higher probability of guessing the correct answer. He commented on the lack of research on the AC format, questioned why more had not been done to date, and called for more research to determine the potential utility of the format for high-stakes credentialing exams (e.g., licensing and certification).

Another reason for interest in the AC format is that item writers often find it difficult to write multiple, equally plausible distractors for MCQs. Haladyna and Downing (1993) state that writing such distractors may be the most difficult part of writing test items. The authors conducted a distractor analysis to determine the number of effectively performing distractors per test item, across several different tests, and found the number of such distractors to be one. They also found the number of distractors per item to be unrelated to item difficulty. Both of these results add sup-

port for research to explore the characteristics and functionality of the AC item.

Purpose of This Research

The purpose of this study was to explore the characteristics of AC items and the AC test in relation to the four-option alternative (MC4). We examined the psychometric characteristics of two tests with the same stem for all MCQs but with either two (AC) or with four response options. These characteristics include test scores, testing time (start to completion), internal consistency reliabilities (KR-20), and item difficulties (p -values). We also investigated the impact of guessing on test scores, given that the potential for guessing would be higher, resulting in higher scores for the AC format test.

METHOD

Participants

The test takers for the study were FDA investigators attending a 3-week training course. Data were collected from four different course offerings, each delivered in different parts of the U.S. The participants were relatively new investigators who had been working to complete their new investigator training. The content of the test items for the two forms was in food good manufacturing practices (GMPs). Although all participants may not have attended national training in this content area, the investigators in this course would all have gained some experience conducting food GMP inspections at their home offices prior to attending. GMP food inspections would be a typical starting point for these investigators as food inspections constitute approximately two-thirds of all inspections conducted by the FDA.

Test Development

Test items were selected from a bank of questions covering GMPs for food preparation and storage. All items in the bank were four-option MCQs written by experienced investigators with the assistance of a contractor. Fifty items were randomly selected from the bank and served as the foundation from which the final tests were constructed. The MC4 and AC tests consisted of the same set of items but with two alternatives eliminated from each four-option item to construct the AC test.

Two alternatives from each item were eliminated using a judgmental process. Four experienced investigators were individually presented with the set of 50 items, and correct answers, and instructed to select the one best alternative to the correct answer. Items were retained where there was agreement from three out of the four judges. The final test forms consisted of 32 items with identical stems. We did not add any additional items to keep administration time to around 30 minutes (the amount of time we were allotted for administration by personnel responsible for the course).

Items were presented in the same order on both forms. The correct answers and distractors were distributed equally across the alternative options in both forms.

Few studies have used judges to reduce the number of options for test items, with most studies eliminating non-performing alternatives that were selected by test takers less than 5% of the time (Crehan et al., 1993; Landrum et al., 1993; Sidick et al., 1994; Williams & Ebel, 1957). Stratton and Catts (1980) used judgments by economics teachers on one test form to eliminate the distractor that they determined students would know was wrong. On another form they randomly eliminated a distractor, an approach they deemed emasculating to the items in their discussion as it negatively impacted test reliability. Cizek, Robinson, and O'Day (1998) used two different approaches to determine which options to eliminate; one was by a panel of content experts trained in item construction and review, and the second was by using the results of item analyses (item difficulty and item discrimination data). Trevisan et al. (1994) used what they called an incremental approach. They started by constructing two-option MCQs (alternate choice items) and then added additional distractors to create items with more options. Rodriguez (2005) conducted a meta-analysis focusing on the optimal number of alternatives for multiple-choice items. The studies they included used one of four-option deletion methods: (a) random deletion, (b) deletion of ineffective distractors, (c) deleting the most effective distractor, and (d) adding distractors.

However, no studies used the approach of having independent judges select the one best alternative from four-option MCQ items to create an AC format. This judgmental approach is important because it addresses the capability of item writers to construct MCQ items with distractors that are all equally plausible. Haladyna and Downing (1993) stated that writing distractors takes time and may be the most difficult part of item writing. Although it is fairly straightforward for item writers to construct an initial, plausible distractor, it is very difficult to construct additional distractors that are all equally plausible to the test taker who does not know the answer to the item.

The authors conducted their study to answer one question regarding the frequency of occurrence of effective and ineffective distractors for the different testing programs addressed in their paper. They conducted a descriptive analysis of the distractors and found the majority of distractors did not perform well. In fact, the number of effective distractors was basically one, providing some support to further investigate the viability of the two-option AC item format. Our judges in effect just reversed the process of writing the key (correct answer) and then their first (and probably best) distractor. They were provided complete four-option items with the key and then selected the best distractor from the three alternatives to include in the AC test, thus eliminating the two distractors they judged to be

least plausible. Rogers and Harley (1999) argue that test-wise test takers do much the same thing; they eliminate the least plausible distractors where they do not immediately know the answer to the item.

Procedures

Tests were administered to FDA investigators on the last day of Week 2 of a 3-week-long class. The content of the tests for this study was not related to the content of the course they were taking. They were told that the organization was trying to get information on different test formats and item types in an effort to make testing as effective and efficient as possible. The testing was self-paced with an instruction sheet as the first page indicating how long the activity would take, what information to provide before and after completing the test, and how to mark their answers. The test forms were interleaved into one stack and were thus distributed randomly, one at a time, to each student as they were sitting in their seats for the course.

RESULTS

The tests were completed by 138 investigators in their first year on the job ($M = 8.15$ months, $s = 2.96$). Seventy persons completed the AC form, and 68 completed the MC4 test.

We conducted independent means t -tests to identify differences between the AC and MC4 test item formats. Table 1 provides descriptive statistics, t -values, significance levels (p -values), and 95% confidence intervals (CIs). The test for differences in mean scores showed the mean AC score ($p < .001$) was significantly higher than for the MC4 format. The CI indicates a 95% probability that the true effect size lies between the lower and upper bounds of the interval. The size of the effect is sufficiently large where the spread between the lower and upper CI excludes 0, as it does in this case. Both p -values (point estimates) and confidence intervals (interval estimates) add to the body of evidence regarding the research findings and have been provided for all comparisons (Greenland et al., 2016; Wasserstein & Lazar, 2016). Average item difficulty values (P -values) were also calculated for both forms. The MC4 test ($M = 0.67$) was more difficult than the AC test ($M = 0.79$; the average p -values just represent the test results as mean % correct in addition to the raw scores).

One criticism of multiple-choice tests is that test takers may guess correctly where they don't know the answer versus having to construct their response with a test format such as short answer (Frary, 1988). It may also be that the greater difficulty level of the MC4 test is due in part to the greater number of options to be considered. If we assume that a test taker either knows the answer to a test question or guesses randomly to questions they miss, on a test with four-option questions (such as the MC4 test in this paper),

test takers would be expected to correctly guess one question for every three guessed incorrectly (25% of the time). Guessing on the AC test would be much higher (50% of the time). The two tests have the same stems, the items are presented in the same order, the key (correct answer) is the same, but the AC test has one-half of the response options of the MC4 test. The higher mean score on the AC test, and the corresponding mean P -value, should be directly affected by the higher probability of guessing correctly. To test this assumption, we first applied a correction for guessing formula to all scores on both forms to account for the probabilities of guessing (Frary, 1988). The correction for guessing formula is

$$\text{Corrected Score} = \# \text{ Correct} - \# \text{ Incorrect} / (\# \text{ of response options per item} - 1).$$

For the 32-item tests in this study, with a score of 26, the calculation for the MC4 test would be

$$\text{Corrected Score} = 26 - 6 / (4 - 1) = 24$$

and for the AC test would be

$$\text{Corrected Score} = 26 - 6 / (2 - 1) = 20.$$

We then analyzed the corrected mean scores (AC = 19.06, MC4 = 19.31) with an independent means t -test and found they were not statistically different (see Table 1). This result supports our proposition that the number of response options directly impacts test scores and item difficulty. It also seems likely that reliability is directly impacted by not only increasing the number of items on a test but also by increasing the number of response options per item.

We also analyzed the differences in the amount of time required by test takers to complete the two forms. The t -test was significant, with test takers requiring on average 24% less time to complete the AC test (see Table 1). This result is consistent with previous research (Aamodt & McShane, 1992; Costin, 1972). The AC test should be completed faster based on the number of options test takers would need to read and consider before answering (Schneid et al., 2014). If we were to add more items to the AC test (to 42 items), the longer AC test and the 32-item MC4 test would be completed in approximately the same amount of time.

The KR20 reliability for the MC4 test form was higher than for the AC format. Given that test takers required 24% less time to complete the AC format test, we applied the Spearman-Brown formula (Crocker & Algina, 2008) to estimate the increase in AC form reliability if we added 24% more items to the AC format test. The reliability of the AC format test increased from 0.46 to 0.53, closer to that of the MC4 test (0.57).

We estimated the values for tests containing the number

TABLE 1.
Descriptive Statistics and Item Format Comparisons

Statistic	AC	MC4	<i>t</i>	95% CI
Test scores				
<i>M</i>	25.53	22.49	5.64 ($p < .001$)**	1.98, 4.11
<i>s</i>	2.59	3.67		
<i>N</i>	70	68		
<i>KR20</i>	0.46	0.57		
<i>P</i> -values				
<i>M</i>	0.79	0.67	0.30 ($p = 0.765$)**	-1.95, 1.44
<i>s</i>	0.18	0.28		
Corrected score				
<i>M</i>	19.06	19.31	-5.15 ($p < .001$)*	-6.37, -2.84
<i>s</i>	5.18	4.89		
Test time (in minutes)				
<i>M</i>	14.82	19.43	-5.15 ($p < .001$)*	-6.37, -2.84
<i>s</i>	4.78	5.77		
<i>KR20</i> ***	0.53	0.57		

Note. Mean test scores are number correct on each 32-item test. *($df = 131$), **($df = 136$). ****KR20* recalculated for the AC test with an increase in items by 24% to equate to the MC4 testing time.

of items typical for certification tests taken by our investigators (100 items per test). Applying the Spearman-Brown formula once again, the reliability estimates for both forms were AC = 0.73 and MC4 = 0.81. The issue of these somewhat low reliabilities will be addressed in the discussion section below.

DISCUSSION

The purpose of this study was to explore the characteristics of AC items and the AC test in relation to the four-option alternative (MC4). Although there has been some research on the AC format, Downing (1992) argued that more was needed to determine the utility of the format.

We employed a judgment process that required independent interjudge agreement to remove two distractors from the four-option test items (a qualitative approach). It is important to note that the strategy we used has not been used before. The approach used most often relies on data from item analysis to eliminate non-functioning distractors (a quantitative approach). We are not suggesting our method as a recommended approach for item writing, because it would eliminate the purpose of cutting item writing time, but simply as a better way of simulating what would happen if item writers only wrote two options. Additional research could help determine how well SME item writers compare with empirical methods for identifying non-functioning distractors. If item writers compared favorably to empirical analysis and were able to develop a single functioning distractor using the AC format, we would be able to spend less time writing and reviewing items.

The results were predictable with respect to test scores and testing time. Scores were significantly higher and testing time significantly less for the AC compared to the MC4 test. We thought that AC test scores might be higher based on a higher probability of guessing the correct answer. When corrected for guessing, score differences disappeared, leaving only the difference in testing time. The correction for guessing is used for formula scoring, a test-scoring strategy where test takers are instructed not to answer questions unless they are reasonably sure of the answer, because they will be penalized for answering incorrectly (Rowley & Traub, 1977). It also assumes that test takers blindly select options when they don't know the answer to a question.

The application of the formula shows the impact that fewer distractors alone has on test scores—fewer distractors results in a greater chance of guessing the correct answer. We are not proposing this actually be done but just trying to see if guessing explains the differences between the AC and other numbers of response options. Also, formula scoring is not used in organizations today so the ease in guessing does reduce the utility of AC items. We also understand that test takers do not blindly guess when they don't know the answers to questions but use other strategies such as waiting to answer a question because other questions in the test might cue the answer.

Another means of addressing the higher scores with fewer distractors is in setting passing scores. When using a judgmental approach such as a modified Angoff (Angoff, 1971), item raters could be instructed to set the *p*-values (items with higher values are easier) greater than or equal to 0.50, whereas for MC4 items that value would be 0.25. The

result may be higher passing scores for AC tests. Additional research is needed to determine whether the difference in instructions on setting the floor for p -values would result in higher passing scores for AC versus tests with more distractors and thus similar mean scores compared to MC4 tests.

For testing time, Sidick et al. (1994) addressed it as an opportunity to increase the length of the test while keeping testing time constant, thereby more fully sampling the content domain. We used the Spearman-Brown formula to estimate the increase in reliability from adding additional AC items while keeping testing time constant. Although reliability did improve somewhat, it was still lower than for the MC4 test (AC = .53, MC4 = .57). Because most of our certification tests consist of 100 items, we again estimated the potential increase in reliability (AC = .73, MC4 = .81). The reliabilities for both formats are still low for tests used to make decisions about individuals. One possible explanation for the low internal consistency reliabilities for both forms is that job content is multidimensional. Test items were selected at random from the test bank; there were not enough items from any duty area to construct a test of reasonable length.

One possible approach to address this would be to increase test length and develop reliable subscales to capitalize on the multidimensionality of the content domain. A longer test would increase reliability, and test takers could be given more specific feedback on their performance in addition to total score. More research is needed, not only with increased sample sizes but also with scale development and in different content areas.

There are also trade offs to be considered when increasing the number of test items. Where organizations are concerned about reducing testing time, the AC format may not be a viable option given the lower reliability. However, if the concern is more thoroughly sampling a specific content domain and keeping testing time at current levels, AC tests may be an option.

Anecdotally, many students remarked that the test was fair. They were referring to the AC format, and they didn't know the MC4 test was being completed by half of the class. We propose that perceptions of fairness might be higher for the AC format. We already addressed previous research that found students preferred tests with fewer options (Owen & Froman, 1987). We have seen no research comparing the number of response options with respect to perceived fairness. Research might be warranted given the generally negative perceptions of testing in general.

The practitioner issues driving this study emphasize the difficulty in developing plausible distractors and the time devoted to testing are meaningful issues in the applied world of testing and credentialing. We believe the AC format provides the potential for considerable cost and time savings for organizations doing test development and administration. It takes more time to write items with four

options than for two. We have observed that item writers often struggle to write those additional, hopefully equally plausible distractors, an observation shared by many others (Haladyna & Downing, 1993; Sidick et al., 1994). Indeed, Haladyna and Downing (1988) argued that it is probably not worth the time and effort required to develop additional distractors. We see this as another area where additional research is needed. What is the average time required to write an item stem, key, and one terrific, plausible distractor; and then what is the additional time spent trying to develop additional, equally plausible distractors that may not be effective?

We also propose developing test items without requiring a set number of alternatives for all items on the test. The result would be a test with items having different numbers of distractors (a mixed response-option test). Item writing instructions would be to write as many plausible distractors as possible but to stop when you can't write any more, even if you don't have three. We think this approach also merits further investigation.

Although the results in this paper did not make a convincing case for the blanket adoption of the AC format as opposed to the three- and four-options formats, we have tried to address some of the psychometric and practical issues related to reducing the number of response options. We hope that there will be more research interest in alternatives to the three- and four-option tests and expect that there will be. After all, Haladyna and Downing (1993) found the number of plausible distractors in their study to be "approximately one" (p. 1008).

REFERENCES

- Aamodt, M., & McShane, T. (1992). A meta-analytic investigation of the effect of various test item characteristics on test scores and test completion times. *Public Personnel Management*, 21(2), 151-160.
- Angoff, W. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.) *Educational measurement* (2nd ed., pp. 508-600). American Council in Education.
- Cizek, G., Robinson, K., & O'Day, D. (1998). Nonfunctioning options: A closer look. *Educational and Psychological Measurement*, 58(4), 605-611.
- Costin, F. (1970). The optimal number of alternatives in multiple-choice achievement tests: Some empirical evidence for a mathematical proof. *Educational and Psychological Measurement*, 30, 353-358.
- Costin, F. (1972). Three-choice versus four-choice items: Implications for reliability and validity of objective achievement tests. *Educational and Psychological Measurement*, 32, 1035-1038.
- Crehan, K., Haladyna, T., & Brewer, B. (1993). Use of an inclusive option and the optimal number of options for multiple-choice items. *Educational and Psychological Measurement*, 53, 241-247.

- Crocker, L. & Algina, J. (2008). *Introduction to classical & modern test theory*. Cengage Learning.
- Delgado, A., & Prieto, G. (1998). Further evidence favoring the three-option items in multiple-choice tests. *European Journal of Psychological Assessment, 14*(3), 197-201.
- Downing, S. (1992). True-false, alternate-choice, and multiple-choice items. *Educational Measurement: Issues and Practice, 11*, 27-30.
- Ebel, R. (1969). Expected reliability as a function of the number of choices per item. *Educational and Psychological Measurement, 29*, 565-570.
- Ebel, R. (1982). Proposed solutions to two problems of test construction. *Journal of Educational Measurement, 19*, 267-278.
- Frary, R. (1988). Formula scoring of multiple-choice tests (correction for guessing). *Educational Measurement: Issues and Practice, Summer*, 33-38.
- Greenland, S., Senn, S., Rothman, K., Carlin, J, Poole, C., Goodman, S., & Altman, D. (2016). Statistical tests, P-values, confidence intervals, and power: A guide to misinterpretations. *American Statistician, Online Supplement to the ASA Statement on Statistical Significance and P-values*, 1-12.
- Grier, J. (1975). The number of alternatives for optimum test reliability. *Journal of Educational Measurement, 12*(2), 109-113.
- Haladyna, T., & Downing, S. (1988). Functional distractors: Implications for test-item writing and test design. Paper presented at the annual meeting of the American Educational Research Association, New Orleans LA.
- Haladyna, T., & Downing, S. (1989). A taxonomy of multiple-choice item writing rules. *Applied Measurement in Education, 1*, 37-50.
- Haladyna, T., & Downing, S. (1993). How many options is enough for a multiple-choice test item? *Educational and Psychological Measurement, 53*, 999-1009.
- Haladyna, T., Downing, S., & Rodriguez, M. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education, 15*(3), 309-334.
- Landrum, R., Cashin, J., & Theis, K. (1993). More evidence in favor of three-option multiple-choice tests. *Educational and Psychological Measurement, 53*, 771-778.
- Lord, F. (1944). Reliability of multiple-choice tests as a function of the number of choices per item. *Journal of Educational Psychology, 35*(3), 175-180.
- Owen, S., & Froman, R. (1987). What's wrong with three-option multiple choice items? *Educational and Psychological Measurement, 47*, 513-522.
- Rodriguez, M. (2005). Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. *Educational Measurement: Issues & Practice, 24*(2), 3-13. DOI: 10.1111/j.1745.3992.2005.0006.x
- Rogers, W. & Harley, D. (1999). An empirical comparison of three- and four-choice items and tests: Susceptibility to testwise-ness and internal consistency reliability. *Educational and Psychological Measurement, 59*(2), 234-247.
- Rowley, G., & Traub, R. (1977). Number-right scoring, and test-taking strategy. *Journal of Educational Measurement, 14*(1), 15-22.
- Sax, G., & Reiter, P. (1980). Reliability and validity of two-option multiple-choice and comparably written true-false items (Report No. 143). U.S. Department of Education.
- Schneid, S., Armour, C., Park, Y., Yudkowsky, R., & Bordage, G. (2014). Reducing the number of options on multiple-choice questions: response time, psychometrics, and standard setting. *Medical Education, 48*, 1020-1027. Doi: 10.1111/medu.12525\
- Sidick, J., Barrett, G., & Doverspike, D. (1994). Three-alternative multiple choice tests: An attractive option. *Personnel Psychology, 47*(4), 829-835.
- Straton, R., & Catts, R. (1980). A comparison of two, three and four-choice item tests given a fixed total number of choices. *Educational and Psychological Measurement, 40*, 357-365.
- Trevisan, M., Sax, G., & Michael, W. (1994). Estimating the optimum number of options per item using an incremental option paradigm. *Educational and Psychological measurement, 54*(1), 86-91.
- Tversky, A. (1964). On the number of alternatives at a choice point. *Journal of Mathematical Psychology, 1*, 386-391.
- Wasserstein, R., & Lazar, N. (2016). The ASA's statement on p-values: context, process, and purpose. *American Statistician*. DOI: 10.1080/00031305.2016.1154108.
- Williams, B., & Ebel, R. (1957). The effect of varying the number of alternatives per item on multiple-choice vocabulary test items. In E. M. Huddleston (Ed.), *Fourteenth yearbook of the National Council on Measurements Used in Education* (pp. 63-65). National Council on Measurements Used in Education.

RECEIVED 10/19/18 ACCEPTED 02/07/20