Journal of Effective Teaching in Higher Education

Randomize it: Fair Procedures when Constructing Multiple-choice Test-Keys

Dane Christian Joseph, George Fox University, djoseph@georgefox.edu

Abstract. Multiple-choice testing is a staple within the U.S. higher education system. From classroom assessments to standardized entrance exams such as the GRE, GMAT, or LSAT, test developers utilize a variety of validated and heuristic-driven item-writing guidelines. One such guideline that has been given recent attention is to randomize the position of the correct answer throughout the entire answer key. Doing this theoretically limits the number of correct guesses that test-takers can make and thus reduces the amount of construct-irrelevant variance in test score interpretations. This study empirically tested the strategy to randomize the answer-key. Specifically, a factorial ANOVA was conducted to examine differences in General Biology classroom multiple-choice test scores by the interaction of method for varying the correct answer's position and student ability. Although no statistically significant differences were found, the paper argues that the guideline is nevertheless ethically substantiated.

Keywords: multiple-choice, guessing, procedural fairness, testing, factorial ANOVA

Introduction

The purpose of this study was to empirically compare three strategies for varying the position of correct answers in multiple-choice test keys and to assess their impact on test scores. Students arrive in U.S. higher education institutions already familiar with the multiple-choice (MC) assessment format. MC tests are in many ways more convenient for test-makers than their constructed or written-response counterparts (Haladyna, 2004; Haladyna & Rodriguez, 2013). They can be used in both low and high stakes testing environments, are applicable across a large number of subject matter areas, and are relatively easy to administer, score, and interpret when diligently developed (Downing, 2002a; Drummond, Sheperis, & Jones, 2016; Chappuis & Stiggins, 2017).

The decision to administer MC assessments has a mixed history in the educational psychology literature. Mingo, Chang, and Williams (2018) reported that 161 students in an undergraduate educational psychology course least preferred MC but most preferred constructed-response (CR) or essay types out of ten assessment-format choices. On the other hand, Parmenter's (2009) review of studies with undergraduate business students finds they prefer MC assessments even though the students admitted to being more enthusiastic when given essays. Parmenter also suggests that decreasing budgets and increasing class sizes are forcing many professors who teach larger courses to incorporate MC at the expense of CR options; thus, instructors "prefer" MC as a grading-efficient convenience.

Yet other research suggests that some students might prefer MC assessments as they believe they can both: (1) rely on cognitive strategies such as recall and recognition to assist them in selecting the correct answer for particularly difficult test content and (2) get more readily available feedback because of the ease of MC scoring (Mullet, Butler, Verdin, Von Borries, & Marsh, 2014). While MC items are by nature easier to score, they are notoriously difficult to design well and all too often contain low-quality items such as items that only target students' recall of facts and not their ability to understand or apply abstract concepts.

On the other hand, high-quality MC items increase the likelihood that test score interpretations will accurately reflect examinees' domain-specific knowledge of a subject. There are massive consequences attached to test score interpretations (Roediger & Marsh, 2005). For instance, standardized test performance may be proportional to the funding that schools receive (Tindal, 2002). Good test performance might open up the possibility of earning academic scholarships to increase accessibility as well as affordability and offset personal costs (Cohn, Cohn, Balch, & Bradley Jr., 2004). Conversely, poor test performance might limit accessibility and affordability options.

A good deal of psychometric research has demonstrated several contributing factors to test score variance beyond superficial proxies such as student ability or prior achievement. Examples include the quality of the examinees' nutrition prior to taking a test (Figlio & Winicki, 2005); school socioeconomic situation, classroom environment and available resources (Aikens & Barbarin, 2008); and knowledge of test-wise strategies that can be deployed to attain successful guesses (Supon, 2004). Haladyna and Rodriguez (2013) also posit that item quality can influence an item's power to discriminate between students who truly know the correct answer from those who do not. If an item has low discrimination, students without true knowledge of the item's answer may be able to exploit some of the item's features to guess the correct response. Thus, if item quality is dependent on item-writing ability, the latter can also impact test score variance.

As a result of items being too difficult or too easy, guessing contributes construct-irrelevant variance (CIV) to the test score. CIV is a part of the score that does not represent the examinee's true knowledge of item(s) but rather something else, such as their ability to successfully infer and guess correct items (Downing, 2002b; Haladyna & Downing, 2004). This lowers test score reliability and hence, interpretations from one stakeholder (person or group) or test-attempt to another (Bar-Hillel, Budescu, & Attali, 2005). Nolen, Haladyna, and Haas (1992) warn against test score interpretations that lack validity due to CIV. Consequences include inaccurate formative or summative feedback to students to guide their learning as well as large-scale costs such as indefensible admissions' standards.

Rationale for this Study

Researchers have proposed and defended item-writing guidelines to assist item developers because high quality MC items increase the likelihood that precise and accurate test scores will be the result (Haladyna, 2004). Although roughly thirty MC item-writing guidelines (Haladyna & Downing, 1989a, 1989b; Haladyna & Rodriguez, 2013; Haladyna, Downing, & Rodriguez, 2002) have consistently been touted to increase item quality, the empirical validity of some of these guidelines remains in question. One guideline lacking empirical attention suggests varying the position of correct answers among options. The rationale for this approach is to limit the number of successful guesses that examinees can make as a result of response set, the systematic or predictable patterns in item lists.

Three methods that are used to vary the position of correct answers are randomized, arbitrary, and balanced (Attali & Bar-Hillel, 2003). Arbitrary methods are not purely random, although they may resemble it. Human beings are prone to creating patterns. They tend to systematically but biasedly over-place desired items in middle or edge-averse locations as compared to the ends or edges (Ayton & Falk, 1995; Bar-Hillel & Attali, 2002; Christenfeld, 1995; Falk, 1975; Rubinstein, Tversky, & Heller, 1996). Examinees can then benefit from this bias within a testing scenario either by purposefully guessing middle/edge-averse options or because they themselves are also biased to choosing such positions.

In a balanced key, the correct answer is intentionally placed in each possible option an equal number of times. When keys are balanced, students may use elimination strategies such as the *underdog* strategy to successfully guess (Bar-Hillel & Attali, 2002). To do this, examinees answer all the questions on the test as best as they can; count the frequency of each position among the answers; select the position with the lowest frequency (the *underdog* position); eliminate any clear distractors (incorrect answers); and assign the *underdog* to all as yet unanswered items. The higher a student's ability, or the more knowledge a student possesses of the test items, the higher the number of correct answers that can be attained before having to guess. Thus, the greater the potential for *underdog* strategy success.

Bar-Hillel and Wagenaar (1991) discuss the differences between random and non-random processes. True randomization results from a process that uses a purely random device to assign the position of correct answers; examples are die, unbiased coins, or computer programs. Each possible outcome has an equal chance of occurring on any given turn. Yet, the pattern is unpredictable to the human brain. Given that game theory suggests that one can do no better against a random move other than to play randomly; at best, students can split even by using a purely random counter-move (Attali & Bar-Hillel, 2003; Rubinstein et al., 1996). Because examinees cannot purely randomize and because one can assume they are not allowed to utilize pure randomized devices in assessments, they cannot employ successful guessing strategies for a randomized answer-key.

Upon examining the face validity of each method, randomization seems to be the most effective method to make the answer-key pattern unpredictable to examinees and hence, the most difficult for them to exploit through successful guesses. Although randomized outcomes will balance out in the long run, the pattern on any single trial is completely unpredictable to humans. Thus, no successful guessing strategies can be employed on a single-trial randomized answer-key. This was the theoretical basis for the present study.

Methodology

Three research questions were investigated:

1. To what extent does the method of answer-key assignment affect the examinees' performance on a general biology classroom multiple choice test?
2. To what extent do examinees' test scores differ across test formats for combined ability groups?
3. To what extent do high-ability examinees' test scores differ across test formats?

Because student ability factors into the success of the underdog strategy in balanced keys, a proxy for ability was explored as a second independent factor to interact with the first factor, test format. This student ability was operationalized as the amount of knowledge that students would possess on the subject matter and measured as their cumulative grade-point average (CGPA).

Design and Analysis

An experiment was conducted using a between-subjects' factorial ANOVA design. This design was appropriate to analyze the significance of mean differences on the dependent variable (DV) between the groups or levels of the independent variables (IV) or factors (Mertler & Vannatta, 2016). The manipulated factor in the study was the method for varying the location of correct answers: randomized, arbitrary, or balanced. Each student was assigned only one method condition. A second non-manipulated IV was the examinee's proximal ability or knowledge using CGPA. CGPAs were collapsed into high, medium, and low non-contiguous groupings. The DV was the total test score. Individual items were scored dichotomously as either 0 (incorrect) or 1 (correct); whereas, test score was aggregated on a continuous level.

Beyond the ANOVA results, item difficulty and discrimination analyses were conducted to assess the quality of item responses along with Cronbach alphas used to assess the consistency of item responses in each test format. Item difficulty was assessed through the use of item proportions, a classical test theory approach that looks at the average proportion of correct answers over a test domain. Item discrimination was computed in the form of point biserial correlation coefficients.

Sample and Participants

Participants came from a large land-grant research based university located in the Pacific Northwest region of the United States that had an approximate annual enrollment of just under 20,000 undergraduate students. The majority of participants were freshmen and sophomores. A few juniors and seniors were also enrolled. The sampling frame was a total of 540 students from 15 sections of 36 students in each, and 369 students provided informed consent to have their test scores analyzed in this study.

This N was necessary in order to have adequate power for the analysis of between-group differences after a power analysis was conducted using an alpha level set at .05 with desired power recommended at .80 (Cohen, 1988). For large effects, total sample size requirements were n = 133 and n = 107 for interaction and main effects, respectively as well as cell sample size requirements of n = 15 and n = 12 for interaction and main effects, respectively. The course professor taught all course sections in combined lectures with teaching assistants supervising each section's lab meetings.

Instrumentation and Administration

The test instrument was a 100-level General Biology MC exam. The course instructor developed the items based on his experience in writing and using classroom MC tests as well as his expertise in the subject matter given his PhD in Biology. The conventional MC test included 50 items and was to be taken in a 60-minute testing period per psychometric recommendations of one item per minute (Burton, 2006). Each item was worth one point for a total possible test score of 50 points. Each test item had five options, A through E. The researcher then collected the developed instrument and conducted an informal proofreading and screening of the items for style and format concerns surrounding item clarity, grammar, punctuation, and spelling.

Following this, each of the three test formats were created. Each format had the same exact items in the same exact order. They only differed in where the position of the correct answer was located among the five options. Items were included at the end of the test to ascertain examinees' gender, race/ethnicity, and age, in order to determine the representativeness of the sample to the university population. Students also provided their CGPAs.

Microsoft Excel was used to develop the randomized and balanced keys. A random number generating function was employed to assign the position of the correct option for each item in the randomized version. An assignment of correct positions was similarly used for the balanced format by specifying that each possible option represent the correct answer exactly 10 times, i.e., 10 correct answers in position A, 10 in position B, etc. It is important to note that the distribution of correct positions did not turn out to be equally represented throughout. This was because of a conflict among other item-writing guidelines that Haladyna and Downing recommended.

Specifically, the correct answer-position on one of the items had to be relocated post hoc because the guideline for ordering numerical values was violated. In other words, the possible answers were numbers that needed to be ordered by size across the options. Given that the correct answer was the smallest number, it should have been placed in position A to satisfy this item-writing guideline. Since this failed to occur (the correct answer was originally placed in position D), the researcher switched the two locations.

The course instructor developed the arbitrary key by assigning the position of the correct answer among the options. The distribution of correct options for each test format is found in Table 1 below.

Table 1: Distribution of Correct Options for Each Test Format

	Test Format		
Option	Randomized	Arbitrary	Balanced
A	.24	.18	.22
B	.10	.22	.20
C	.32	.18	.20
D	.14	.16	.18
E	.20	.26	.20

The test was administered on days and times during which the course sections typically met for lectures. Each student randomly received one of the three test formats as opposed to randomly varying the formats by whole sections. Its rationale was to reduce systematic error and increase the power of the design to detect treatment effects (Lipsey, 1990; *underdog* & Sax, 1990). Subjects were only informed that unnecessary guessing would tend to lower their total test score.

Results

Cronbach alphas were equal to .80, .69, and .72 for the randomized, arbitrary, and balanced test formats, respectively. This indicated fair reliability in the responses within each test format. Descriptive data of students' demographics for race/ethnicity and gender showed that the sample was indeed representative of the university population. Data were then screened for missing or erroneous values and to ensure that the ANOVA assumptions would be fulfilled. A low score of 23 and a high score of 42 were observed. Some outliers were deleted from the final dataset after examination of box plots. Two outliers were deleted from the analysis because they did not fit the distribution of scores as both revealed extremely low raw scores (9 and 14) for two students who self-reported extremely high CGPAs.

To separate the ability groupings non-contiguously, the researcher used cutoffs that produced three roughly equivalent sample sizes for high-ability (CGPA = 3.7–4.0), medium-ability (3.0–3.3), and low-ability (2.3–2.7). These were arbitrarily chosen to increase cell sample size per ability-grouping while keeping the groups as distinct as possible (see Table 2).

Table 2: ***Cross-tabulation of Cell and Group Sample Sizes for Student Ability x Test Format***

			Test Format		
		Randomized	Arbitrary	Balanced	Total
	High	19	18	15	52
Student Ability	Medium	18	20	19	57
	Low	19	17	18	54
Total		56	55	52	163

Note that although 369 students consented to participate, the noncontiguous grouping strategy cut the analytical sample by more than half. Although this reduced cell and group sample sizes, conditions of statistical power for large effects were still satisfied given the power analysis results in the "Sample and Participants" section above. Descriptive statistics for this analytical group can be found in Table 3, along with the average proportions of correct answers for each test form.

Table 3: Cell Sample Sizes, Mean Test Scores, Standard Deviations, and Mean Correct-Item-Proportions for Test Formats and Ability Groups in ANOVA Analysis

Test Form	N	M	S	P
Randomized				
L	19	29.11	6.09	0.58
M	18	31.44	4.23	0.60
H	19	37.74	3.12	0.75
C	56	32.79	5.88	0.62
Arbitrary				
L	17	29.41	4.98	0.56
M	20	31.45	4.50	0.61
H	18	35.83	3.68	0.71
C	55	32.25	5.08	0.61
Balanced				
L	18	28.61	5.77	0.57
M	19	30.79	4.21	0.60
H	15	36.87	4.20	0.73
C	52	31.79	5.80	0.61

Note: L = low ability, M = medium ability, H = high ability, C = combined ability

Examination of group scores by histograms revealed normality. Levene's test of equality of variances found no statistically significant differences, indicating homogeneity of variances across groups, F(8, 154) = 1.42, p = .19. Neither interaction nor main effect results from Table 4 below revealed statistically significant differences in test scores by the ability groups. The interaction of test format by ability on test score was not statistically significant, F(2, 154) = .32, p = .72, partial eta squared < .01. The main effect of test format on test score was also not statistically significant, F(4, 154) = .35, p = .84, partial eta squared < .01. Although not a research question of interest, the main effect of ability level was statistically significant as expected, F(2, 154) = 39.38, p < .01, partial eta squared = .33.

Table 4: Two-way ANOVA Summary Table of Interaction and Main Effects

Source	df	Sum of Squares	Mean Squares	F-ratio	F-prob	Effect Size
Between treatments	8	1750.79	218.84			
Ability Level	2	1685.02	842.51	39.38	< .01	.33
Test Format	2	13.71	6.85	.32	.72	.00
Ability Level x Test Format	4	30.22	7.55	.35	.84	.00
Within Treatments	154	3294.65	21.39			
Total	163	174979.00				

Discussion

For the purposes of the following discussion, it is assumed that the vast majority of the examinees would have attempted at least some guesses. Based on the theoretical rationale, it was expected that the combined ability randomized group mean would be lower than either of the combined ability arbitrary or combined ability balanced group means. This expectation was not supported by this study's findings. Guessing test-takers are only successful in the randomized format in the following conditions: 1) they have relatively high ability to minimize the number of guesses needed and be able to use a randomized guessing device on the few items left over; 2) they are able to use a successful guessing strategy such as edge-aversion to better effect than students would on the arbitrary format, assuming the randomized key resembled an arbitrary key; 3) they are able to use a successful guessing strategy such as the **underdog** than students would in a balanced format, assuming the randomized key resembled a balanced key.

Use of randomized devices was strictly prohibited from the test; therefore, option 1 above is unlikely and unfeasible. Results from Table 1 do not confirm that the randomized key was edge-averse; options A and E appear more times (0.24 and 0.20, respectively) than their neighboring options B and D (0.10 and 0.14, respectively). Middle-bias seemed to be present since option C appeared more than any other option (0.32). However, inferences from the distribution of this study's randomized key can only be made about this particular randomization outcome. In other words, another randomization trial would likely have produced a different distribution.

Furthermore, students were not told which method of distributing correct answers pertained to their individual tests. Even more vexing is the fact that the arbitrary key produced by the instructor was atypical and revealed no significant edge-aversion. An edge-averse strategy and option 2 above was therefore also unlikely and unfeasible given this reasoning.

Table 1 also shows an unbalanced randomized key. Although randomized keys are expected to balance out in the long run, they are not expected to be uniformly distributed on the vast majority of single trials (Attali & Bar-Hillel, 2003). To do so would require a large number of items for each test version. It follows that single-trial randomized keys might be equated more often than not with some form of a non-balanced key. The results of this study also showed that combined ability group mean test scores from the non-perfectly balanced keys (i.e., the arbitrary and randomized formats) did not significantly differ from the almost perfectly balanced key format. Therefore, option 3 above is unlikely and unfeasible.

Several items revealed severe problems based on the item difficulty and item discrimination results. Three items had almost uniformly distributed proportions (for difficulty) across options A-E, revealing possible examinee confusion over which options were plausible or implausible distractors. Several items also had poor point-biserial correlation coefficients. This means, for example, that examinees across different ability levels would answer the item correct roughly the same number of times. Hence, such an item is described as non-discriminating. For those unfamiliar with psychometric theory, discriminating items are both necessary and desirable from a test interpretation perspective. This revealed poor item design as one would expect that on average higher ability examinees would answer an item correctly more so than lower ability examinees.

Most noteworthy was the fact that even the high-ability students scored on average 0.75, 0.71, and 0.73 of the items correctly for the randomized, arbitrary, and balanced formats, respectively. This equates to the ‘best achieving’ students attaining C-grades across each format. Thus, it is plausible that the test might just have been too difficult for everyone, regardless of ability level.

Some of the design choices in this study were also suboptimal in retrospect. One such example was the decision to operationalize ability through an achievement proxy as imperfect as CGPA. First-year students in their spring semesters would have only had one semester’s worth of courses, typically introductory or survey courses, compared with second- or third-year students who had several semesters’ worth of courses of supposedly increasing cognitive demand.

Some of the analytical choices were also questionable. While classical test theory is certainly rigorous and widely used in its own right, item-response theory is the current dominant and more accurate approach to examining item difficulty and discrimination (Markus & Borsboom, 2013; Price, 2017). Finally, beyond the answer-key varying strategy, no other key patterns were examined. Yet, more recent research has identified important issues in how patterns such as sequences—whether long runs or palindromes—can trick or confuse examinees (Lee, 2018). Tests should never confuse the examinee as this can also create CIV.

Implications and Recommendations

This section describes some pertinent implications and recommendations for classroom practitioners by discussing practical ways that ethical principles, such as equity, impact item development and guessing. Kane (2013) offered possible reasons for compromised validity due to CIV, including the value frameworks from which stakeholders operate. For example, increased accountability from administrative sources might propel test-makers to create lower difficulty (easier) items in order to inflate test scores; conversely, test-takers might adopt cheating or other less desired behavioral strategies to attain higher scores. Thus, values are a key component to item development, testing, and assessment culture. As such, the ethics of item-development procedures must be addressed if tests are to serve honorably the purposes for which they are designed.

Based on this study's results it might seem that more empirical attention is warranted to support the decision to randomize classroom MC tests, and that until more validity evidence is provided, test-makers may as well arbitrarily place the position of correct answers in MC items without much concern for test-takers to benefit from successful guessing. But this is far from the truth. If the item difficulty and item discrimination results have indeed revealed anything substantial, it is that questions of fairness are not solely about the distribution of outcomes but also the procedures used to generate them. This means that item quality is just as important as test score distribution.

Hence, item-writer training is vital to ensuring that the process of item (stem and options) creation is fair. In that sense, the process can attend to equity issues such as student-access to test-wise strategy training. Because test-wise strategy training can result in higher scores (Markus & Borsboom, 2013; Supon, 2004), some students with the resources to access such training can gain differential score advantages that have nothing to do with subject-matter knowledge of the test items but result from their strategic guessing skills. While it would be easy to then recommend that all teachers prepare their students with test-wise strategy training, those with the economic resources to get more advanced or individualized preparation will still differentially benefit over students without such economic footing.

Even this approach might be forgetting that the point of testing and assessment is to provide both the instructor and student with performance feedback to direct future learning and instruction endeavors. As such, perhaps it would be better—as one reviewer recommended—to incorporate low-stakes MC assessments throughout instruction as a retrieval practice exercise. This increases compatibility across what is learned and what is assessed and helps to level the playing field for students.

For a balanced test with a reasonable number of items, the guessing strategy's success still requires a good deal of knowledge of the other items to minimize the required number of guesses. But for an arbitrary test, no such assumption is necessary. And those who know to "guess C" (or at least stay away from the edges) would have an advantage over the odds of guessing correctly. In some practically significant ways then, the empirical results of this study were always going to be a moot point. Instead, what is important is that test-makers and item-writers follow theoretically sound guidelines when producing testing artifacts. When these are empirically supported, all the better.

Yet in special circumstances where certain item-writing guidelines are not empirically supported, such as that observed in this study, a reasonable approach to test development should be adopted. Fair procedures such as randomizing the answer key ensures that students with limited social and economic capital, or diminished test-wise strategy training, are on equal footing with others before the test administration. Leveling the playing field with sound item-writing guidelines before test administration is the only way to equitably minimize CIV and also ensure that student ability and knowledge makes the only difference on the outcome. Randomization is the best and fairest method in this case.

Note: I wish to thank and acknowledge the editors and reviewers for their time and invaluable feedback to improving the presentation of the study and its potential for impact.

Conflicts of Interest

The author declares that there is no conflict of interest regarding the publication of this article.

References

Aikens, N. L., & Barbarin, O. (2008). Socioeconomic differences in reading trajectories: The contribution of family, neighborhood, and school contexts. *Journal of Educational Psychology*, *100*, 235–251. http://dx.doi.org/10.1037/0022-0663.100.2.235

Attali, Y., & Bar-Hillel, M. (2003). Guess where: The position of correct answers in multiple-choice test items as a psychometric variable. *Journal of Educational Measurement*, *40(2)*, 109–128.

Ayton, P., & Falk, R. (1995). Subjective randomness in hide-and-seek games. Paper presented at the 15th bi-annual conference on Subjective Probability, Utility, and Decision Making, Jerusalem, Israel.

Bar-Hillel, M., & Attali, Y. (2002). Seek whence: Answer sequences and their consequences in key-balanced multiple-choice tests. *The American Statistician*, *56(4)*, 299–303.

Bar-Hillel, M., Budescu, D., & Attali, Y. (2005). Scoring and keying multiple-choice tests: A case study in irrationality. *Mind & Society*, *4(1)*, 3–12.

Bar-Hillel, M., & Wagenaar, W.A (1991). The perception of randomness. *Advances in Applied Mathematics*, *12(4)*, 428–454.

Burton, R.F. (2006). Sampling knowledge and understanding: How long should a test be? *Assessment & Evaluation in Higher Education*, *31(5)*, 569–582.

Chappuis, J., & Stiggins, R.J. (2017). *An introduction to student-involved assessment FOR learning* (7th ed.). Hoboken, NJ: Pearson Publications.

Christenfeld, N. (1995). Choices from identical options. *Psychological Science*, *6*, 50–55.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New York, NY: Routledge Publishing.

Downing, S.M. (2002a). Construct-irrelevant variance and flawed test questions: Do multiple- choice item-writing principles make any difference. *Academic Medicine*, *77* (10), S103–S104.

Downing, S.M. (2002b). Threats to the validity of locally developed multiple-choice tests in medical education: Construct-irrelevant variance and construct underrepresentation. *Advances in Health Sciences Education*, *7*, 235–241.

Drummond, R.J., Sheperis, C.J., & Jones, K.D. (2016). *Assessment procedures for counselor and helping professionals* (8th ed.). Hoboken, NJ: Pearson Publications.

Falk, R. (1975). *The perception of randomness*. Unpublished doctoral dissertation (in Hebrew, with English abstract), Hebrew University, Jerusalem, Israel.

Figlio, D.N. & Winicki, J. (2005). Food for thought: The effects of school accountability plans on school nutrition. *Journal of Public Economics*, *89*(2), 381–394.

Haladyna, T.M. (2004). *Developing and validating multiple-choice test items* (3rd ed.). Mahwah, NJ: Lawrence Earlbaum Associates.

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

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

Haladyna, T.M., & Downing, S.M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice*, *23* (1), 17–27.

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

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*(3), 309–334.

Kane, M. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1–73.

Lee, C.J. (2018). The test-taker's fallacy: How students guess answers on multiple-choice tests. *Journal of Behavioral Decision Making*, 1–12. doi:org/10.1002/bdm.2101.

Lipsey, M.W. (1990). *Design sensitivity: Statistical power for experimental research*. Newbury Park, CA: Sage Publications.

Parmenter, D. (2009). Essay versus multiple-choice: Student preferences and the underlying rationale with implications for test construction. *Academy of Educational Leadership Journal*, *13*(2), 57–71.

Markus, K., & Borsboom, D. (2013). *Frontiers of test validity theory: Measurement, causation and meaning (Multivariate applications book series)*. New York, N.Y.: Routledge Publishing.

Mertler, C.A., & Vannatta, R.V. (2016). *Advanced and multivariate statistical methods: Practical application and interpretation*. New York, NY: Routledge Publishing.

Mingo, M., Chang, A., & Williams, H. (2018). Undergraduate students' preferences for constructed versus multiple-choice assessment of learning. *Innovative Higher Education*, *43*(2), 143–152.

Mullet, H.G., Butler, A.C., Verdin, B. Von Borries, R. & Marsh, E.J. (2014). Delaying feedback promotes transfer of knowledge despite student preferences to receive feedback immediately. *Journal of Applied Research in Memory and Cognition*, *3*(3), 222–229.

Nolen, S.B., Haladyna, T.M., & Haas, N.S. (1992). Uses and abuses of achievement test scores. *Educational Measurement: Issues and Practices*, *11*, 9–15.

Price, L. (2017). *Psychometric methods: Theory into practice (Methodology in the social sciences)*. New York, NY: The Guilford Press.

Roediger, H.L., & Marsh, E.J. (2005). The positive and negative consequences of multiple-choice testing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31 (5)*, 1155–1159.

Rubinstein, A., Tversky, A., & Heller, D. (1996). Naïve strategies in competitive games. In W. Albers, W. Guth, P. Hammerstein, B. Moldovanu, & E. van Damme (Eds.), *Understanding strategic interaction*. New York, NY: Springer-Verlag.

Supon, V. (2004). Implementing strategies to assist test-anxious students. *Journal of Instructional Psychology*, *31*(4), 292–296.

Tindal, G. (2002). Large-scale assessments for all students: Issues and options. In G. Tindal & T.M. Haladyna (Eds.), *Large-scale assessment for all students: Validity, technical adequacy, and implementation* (pp. 1–22). Mahwah, NJ: Lawrence Earlbaum Associates.

Trevisan, M.S., & Sax, G. (1990). Reliability and validity of multiple-choice examinations as a function of the number of options per item and student ability. *Paper presented at the Annual Meeting of the American Educational Research Association* (74th Boston, MA. April 16-20).