Journal of Education, Teaching and Social Studies
ISSN 2642-2336 (Print) ISSN 2642-2328 (Online)
Vol. 2, No. 4, 2020
www.scholink.org/ojs/index.php/jetss

Original Paper

An Empirical Investigation of the Fairness of Multiple-Choice Items Relative to Constructed-Response Items on Tests of Students' Mastery of Course Content

Michael Joseph Wise[1*]

[1] Department of Biology, Roanoke College, 221 College Lane, Salem, Virginia 24153, USA

[*] Michael Joseph Wise, E-mail: wise@roanoke.edu

ORCID: https://orcid.org/0000-0003-0091-303X

Received: September 8, 2020 Accepted: September 18, 2020 Online Published: September 28, 2020
doi:10.22158/jetss.v2n4p16 URL: http://dx.doi.org/10.22158/jetss.v2n4p16

Abstract

Although many instructors prefer multiple-choice (MC) items due to their convenience and objectivity, many others eschew their use due to concerns that they are less fair than constructed response (CR) items at evaluating student mastery of course content. To address three common unfairness concerns, I analyzed performance on MC and CR items from tests within nine sections of five different biology courses I taught over a five-year period. In all nine sections, students' scores on MC items were highly correlated with their scores on CR items (overall r = 0.90), suggesting that MC and CR items quantified mastery of content in an essentially equivalent manner—at least to the extent that students' relative rankings depended very little on the type of test item. In addition, there was no evidence that any students were unfairly disadvantaged on MC items (relative to their performance on CR items) due to poor guessing abilities. Finally, there was no evidence that females were unfairly assessed by MC items, as they scored 4% higher on average than males on both MC and CR items. Overall, there was no evidence that MC items were any less fair than CR items testing within the same content domain.

Keywords

constructed response, fairness, gender bias, guessing, liberal arts, multiple choice, test item

1. Introduction

Multiple-choice questions are popular with teachers for a number of reasons: they are efficient, objective, and generally highly reliable tools for assessment of mastery of course content (Haladyna & Rodriguez, 2013; Rodriguez & Albano, 2017). Nevertheless, many teachers remain resistant to using multiple-choice questions (i.e., MC items), unaware that many of their objections have been addressed and debunked by recent research. For instance, simple guidelines have been developed over the past couple of decades to help teachers write clear, high-quality MC items that are free from the writing flaws that have given MC items a reputation for being confusing, ambiguous, or too susceptible to guessing (Haladyna & Downing, 1989; Haladyna, 2004; Frey, Petersen, Edwards, Pedrotti, & Peyton, 2005; Towns, 2014; Rodriguez & Albano, 2017). In addition, several sources have rebuffed the common criticism that MC items simply require memorization and recognition by demonstrating that MC items can be written to target the higher-order cognitive processes of analysis and evaluation (Aiken, 1982; Haladyna, 1997; Morrison & Free, 2001; Tractenberg, Gushta, Mulroney, & Weissinger, 2013; Domyancich, 2014; Scully, 2017).

One concern that has not been adequately addressed, however, is that MC items are in some fundamental ways not as "fair" as questions in which students are required to come up with their own answers (i.e., constructed-response, or CR items) (Alker, Carlson, & Hermann, 1969; Frakes & Lathen, 1985; Feinberg, 1990; Bond, 1995; McCoubrie, 2004; Simkin & Kuechler, 2005; Kubinger & Wolfsbauer, 2010). Unlike traditional psychometric characteristics of test items (e.g., difficulty and discriminability), "fairness" does not have a standard definition or method of quantification (Zieky, 2006; Haladyna & Rodriguez, 2013). Intuitively, however, a test item can be considered "unfair" if students of a particular group are less likely to respond correctly to the item than are students of a different group who have equal knowledge of the construct tested by the item. The goal of the current study is to address three issues related to the fairness of MC items. The first is mainly a concern of teachers: MC items may be less valid than CR items as tests of mastery. The second is more commonly voiced by students: MC items may reward good guessers who have poor mastery and punish bad guessers who have good mastery of the content. The third is a significant sociological concern: MC items may systematically underestimate the mastery of diversity-based groups, particularly females, relative to CR items.

A teacher's main goal in writing any test item is to obtain the most accurate measurement of students' mastery of a concept or skill—i.e., to maximize what is known as the "construct validity" of an item (Reeves & Marbach-Ad, 2016). In a typical classroom setting, true mastery is generally an unknowable quality. Therefore, any test item is at best an attempt to attach a quantitative estimate of mastery in a way that fairly discriminates among students in terms of their varying levels of mastery. It may seem reasonable to expect that a test item for which students have to construct their own answers should give a more valid estimate of mastery than a test item for which students merely pick an answer from a list of options provided to them. The practical concern is that the relative rankings of students' mastery will differ depending on whether the students are tested using CR or MC items. If these rankings differ (i.e., if there is a low correlation between scores on CR and MC items), then these item types are likely not to be equally valid as tests of mastery. However, if students' scores on CR and MC items that test the same constructs are highly correlated, then the two types of items must be very similar in their construct validity (Lukhele, Thissen, & Wainer, 1994; Bacon, 2003). In that sense, a high positive correlation would be strong evidence that MC items are just as fair as CR items in terms of quantifying student mastery. A low correlation, in contrast, would suggest that either the MC items, the CR items, or both possessed inadequate construct validity.

Although students may guess the correct answer to any type of test item, the limited number of options available makes MC items more susceptible to guessing than are CR items (Zimmerman & Williams, 2003; Bar-Hillel, Budescu, & Attali, 2005; Bush, 2015; Campbell, 2015). In particular, unskilled test-writers often unintentionally provide cues that a test-wise student can use to discern the correct response without actually possessing mastery of the tested construct (Dolly & Williams, 1986; Bar-Hillel & Attali, 2002; Attali & Bar-Hillel, 2003). As a result, guessing can lower the discriminability of an item and the reliability of a test. Fortunately, a test-writer can easily learn to avoid cueing in their MC items once they recognize the problem (Haladyna, Downing, & Rodriguez, 2002; Attali & Bar-Hillel, 2003; Schroeder, Murphy, & Holme, 2012; Towns, 2014; Ibbett & Wheldon, 2016; Rodriguez & Albano, 2017). In terms of fairness, the practical issue is not so much that MC items are more susceptible to guessing than are CR items, but that some students may be consistently better guessers than other students. More specifically, students commonly believe that MC items favor test-wise students at the expense of students who know the content better but are not good at guessing (Hoffmann, 1962; Alker et al., 1969). The idea that students vary in test-wiseness is not controversial. The unproven notion is that the good guessers will tend to be the students with low mastery, while the students who have mastered the content better will be poor guessers. This unproven notion is at the heart of the second fairness issue regarding MC items addressed in the current study.

The third fairness issue is important from a sociological standpoint: MC tests may unfairly disadvantage certain groups—particularly those that are traditionally underrepresented in STEM areas (Bell & Hay, 1987; Ben-Shakhar & Sinai, 1991; Bond, 1995; Lissitz, Hou, & Slater, 2012). For instance, it has often been claimed that females’ mastery, competency, or aptitude may be underestimated by a test that uses MC items compared to a test on the same set of constructs that uses CR items (DeMars, 2000; Simkin & Kuechler, 2005; Kuechler & Simkin, 2010). The available evidence addressing this claim is mixed, with variation in outcomes attributed to such factors as the stakes, setting, and subject matter of the tests (Lumsden & Scott, 1987; Feinberg, 1990; Bridgeman & Lewis, 1994; Becker & Johnston, 1999; Beller & Gafni, 2000; DeMars, 2000; Chan & Kennedy, 2002; Stanger-Hall, 2012). Clearly, more empirical data on the issue of gender differences in performance (particularly in classroom settings) would be of great value for teachers who use or are considering the use of MC items in their tests.

To address these three fairness issues, I examined five years of test-performance data from biology courses that I have taught at Roanoke College. Specifically, I analyzed students’ relative performance on MC and CR items within courses to ask three main questions: (1) How strongly are students’ scores on MC and CR items correlated within tests of the same set of constructs? (A low or negative correlation would suggest unfairness in terms of unequal construct validity.) (2) Are some students substantially worse at answering MC items than their performance on CR items would predict? (Such a pattern would suggest MC items are unfair to “bad guessers.”) (3) Do females (or males) perform worse on MC items than their performance on CR items would predict? (Such a pattern would be evidence of gender-based unfairness.)

2. Methods

2.1 The Data Set

The sources of data for this study were nine lecture sections of five different biology courses (Table 1) taught by the author at Roanoke College, a selective liberal arts college of nearly 2,000 undergraduate students in southwestern Virginia, USA. These courses included introductory biology courses for majors, general ecology, and a general-education science course for non-science majors. The number of students per section ranged from 17 to 48, and 18 students completed two different courses in the set. The full dataset included 257 sets of scores (one per student per lecture section) that were calculated from combined items across the tests administered in each course. These tests included a final exam in each course, plus three to five hourly exams for all courses—except for the 2015 sections of BIOL 205, which included one midterm exam and eight half-hour quizzes. (For simplicity, all exams and quizzes will be referred to as “tests.”) In all, 239 different students were part of this study, with 18 students taking two separate courses.

Table 1. Summary Data for the Nine Sections of the Five Courses Included in this Study. The *r* in the Last Column is the Pearson Product-moment Correlation between Students’ Scores on Multiple-choice (MC) and Constructed-response (CR) Items within each Course

Course	Year	Students	Total potential points		*r*
			MC items	CR items	
BIOL 120: Principles of Biology	2013	48	267	500	0.93
BIOL 125: Biodiversity	2012	34	286	366	0.87
BIOL 125: Biodiversity	2013	34	244	440	0.94
BIOL 125: Biodiversity	2015	23	244	395	0.91
BIOL 180: Exploring Biological Diversity	2016	17	294	307	0.92
BIOL 205: General Ecology	2012	34	280	247	0.90
BIOL 205: General Ecology	2015	21	491	408	0.94
BIOL 205: General Ecology	2015	23	491	408	0.81
INQ 251: Bugs in the System	2016	23	170	263	0.90

Each test was made up of a combination of multiple-choice (MC) and constructed-response (CR) items. The different types of items within a test covered the same content domain and set of constructs. MC items constituted about one-third to just over one-half of the total points for the tests in a course (Table 1). On most tests, MC items were counted as two points each, but they were counted as three points on the quizzes and final exam for the 2015 sections of BIOL 205, and one point on the final exam for the 2012 section of BIOL 205 (These differences in points-per-item merely reflect differences in weighting on tests of different lengths, and they have no impact on the inferences of the analyses in this study). Every MC item consisted of five response options, including one correct answer (the key) and four distractors.

The CR items included the following types of questions: fill-in-the-blank; labeling, interpreting and/or drawing diagrams; mathematical problems (e.g., Hardy-Weinberg and Mendelian genetics scenarios); short-answer questions requiring one or two sentences; and essay questions requiring one or two paragraphs. The point values of the CR items varied depending on the length and complexity of the required responses. All test items were written and graded solely by the author of this paper—a fact that allowed for consistency in content coverage, writing style, and grading procedures.

2.2 Statistical Analyses

For each student in each course section, I calculated two standardized performance metrics—one for MC items and one for CR items. Within a course section, I calculated a total for MC points for each student across all tests, then divided this sum by the total possible MC points (Table 1) to obtain a proportional score. I standardized these proportions by subtracting the section's mean score for MC items from each student's score and dividing by the standard deviation. The same procedures were performed on the CR item scores. The means for both MC and CR scores were thus standardized to zero within each section, and students' scores were expressed as the number of standard deviations above or below the mean. This standardization allowed for meaningful comparisons of students' relative performances, as it factored out differences in mean scores between MC and CR items within and between courses.

For each student within each course section, I calculated a single metric (Δ_{M-C}) to indicate each student's relative performance on MC versus CR items by subtracting his/her standardized CR score from his/her standardized MC score. A positive value of Δ_{M-C} would indicate that a student scored relatively better (compared to the section mean) on MC items, while a negative value would indicate relatively better performance on CR items. For example, a student who scored 0.2 standard deviations (std) below the section mean for MC items and 0.3 std above the section mean for CR items would have a Δ_{M-C} score of -0.5 (or -0.2 minus 0.3).

2.2.1 Potential Validity Differences

If MC items and CR items are equally effective in sampling how well students have mastered content, then students' scores should be similar on the two types of test items. In other words, if MC and CR items test mastery of the same material in a consistent manner, then there should be a high positive correlation between students' MC and CR scores within a course. Such a correlation would not provide an absolute estimate of validity, but it would suggest that MC items were just as fair as CR items in their estimates of true mastery. To test this prediction, I calculated Pearson product-moment correlations between students' standardized MC scores and standardized CR scores within each of the nine course sections. I also calculated a single correlation with all 257 pairs of standardized MC and CR scores combined across all nine sections. All statistical analyses reported in this paper were performed using JMP_{IN} v. 4.0.4 (SAS Institute, Cary, North Carolina, USA).

2.2.2 Potential Guessing Bias

I examined the distribution of $\Delta_{\text{M-C}}$ scores in this study to look for evidence of unfairness due to differences in students' guessing abilities on MC questions. Specifically, if MC items were susceptible to good guessers who did not know the content well (as judged by CR items), then there would be an excess of students with higher $\Delta_{\text{M-C}}$ scores than would be expected by random chance. Similarly, if MC items disproportionately tend to trip up students who know the material well, then there would be an excess of students with lower-than-expected $\Delta_{\text{M-C}}$ scores. In contrast, if MC and CR items are equivalent in how well they test mastery (as opposed to guessing ability), then the distribution of $\Delta_{\text{M-C}}$ scores would be expected to follow a normal distribution with a mean of zero and with an absence of extreme (i.e., low-probability) outliers.

To test these predictions, I compared the distribution of $\Delta_{\text{M-C}}$ scores with a normal distribution (having the same mean and std as the actual distribution) using the Shapiro-Wilk goodness-of-fit test. I also performed an outlier analysis by calculating the Mahalanobis distances (MD) for the relationships between the 257 pairs of standardized MC and CR scores. By convention, an MD > 13.82 (the critical value of a chi-square distribution at $P < 0.001$, with df = 2) for any point would indicate that the point was a significant outlier.

The fact that a subset of students took two different courses included in this dataset allowed for an additional way to address whether guessing was a source of unfairness. Specifically, if some students consistently outperform their content knowledge on MC items (indicated by positive $\Delta_{\text{M-C}}$ scores) due to being good at guessing, then that good-guesser trait should be exhibited across courses. Similarly, if other students consistently underperform on MC items (indicated by negative $\Delta_{\text{M-C}}$ scores) due to being bad at guessing, then that bad-guesser trait should be seen across courses as well. Thus, a plot of students' $\Delta_{\text{M-C}}$ scores in one course against their $\Delta_{\text{M-C}}$ scores in a second course should reveal a line with a positive slope if the fairness of MC items is undermined by differential guessing abilities. To test this prediction, I created such a plot and calculated the Pearson correlation coefficient between $\Delta_{\text{M-C}}$ scores in the first and second courses for the 18 students who took two different courses. A significant positive correlation would indicate that performance on MC items were consistently inflated in some students due their being good guessers and deflated in others due their being bad guessers, which would be a sure sign of unfairness.

2.2.3 Potential Gender Bias

To assess potential gender biases in relative performance on different types of test items, I ran analyses of variance (ANOVAs) on MC scores, CR scores, and $\Delta_{M\text{-}C}$ scores using gender as the explanatory variable. I included students' academic standing (i.e., year) as a covariate, allowing for the fact that college experience is likely to explain a substantial amount of the variation in test scores. The academic standings of the students comprised 84 freshmen, 86 sophomores, 59 juniors, and 28 seniors, and the genders comprised 163 females and 94 males. As far as I was aware, the students' identified genders matched their biological genders.

3. Results

3.1 Potential Validity Differences

The Pearson correlation coefficients between MC and CR scores were positive and highly statistically significant ($P < 0.0001$) for all nine sections, ranging from a low of $r = 0.81$ for one section of General Ecology in 2015 to a high of $r = 0.94$ for a second section of General Ecology in 2015 and for Biodiversity in 2013 (Table 1). Across all nine sections together, the correlation between the 257 pairs of standardized MC scores and CR scores was $r = 0.90$ ($P < 0.0001$; Figure 1). In addition, the relative rankings of students within courses were strikingly consistent whether based on MC scores or CR scores, as shown by the clustering of points around straight lines with a slope of one in Figure 2. While there are a few points that deviate from the line, there are no points that indicate a reversal in rankings when scores are based on one type of question versus the other.

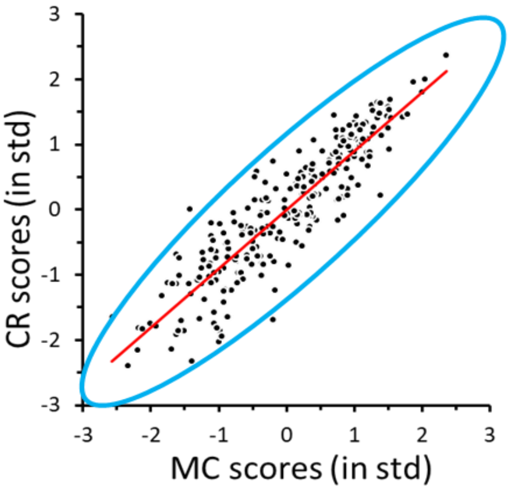

Figure 1. Correlation between Standardized CR Scores and MC Scores for 257 Students across Nine Sections of Five Different Courses (*r* = 0.90; *P* < 0.0001). The Red Line Represents the Regression of CR Scores on MC Scores. The Blue Oval Demarcates the 99% Density Ellipse for the Regression

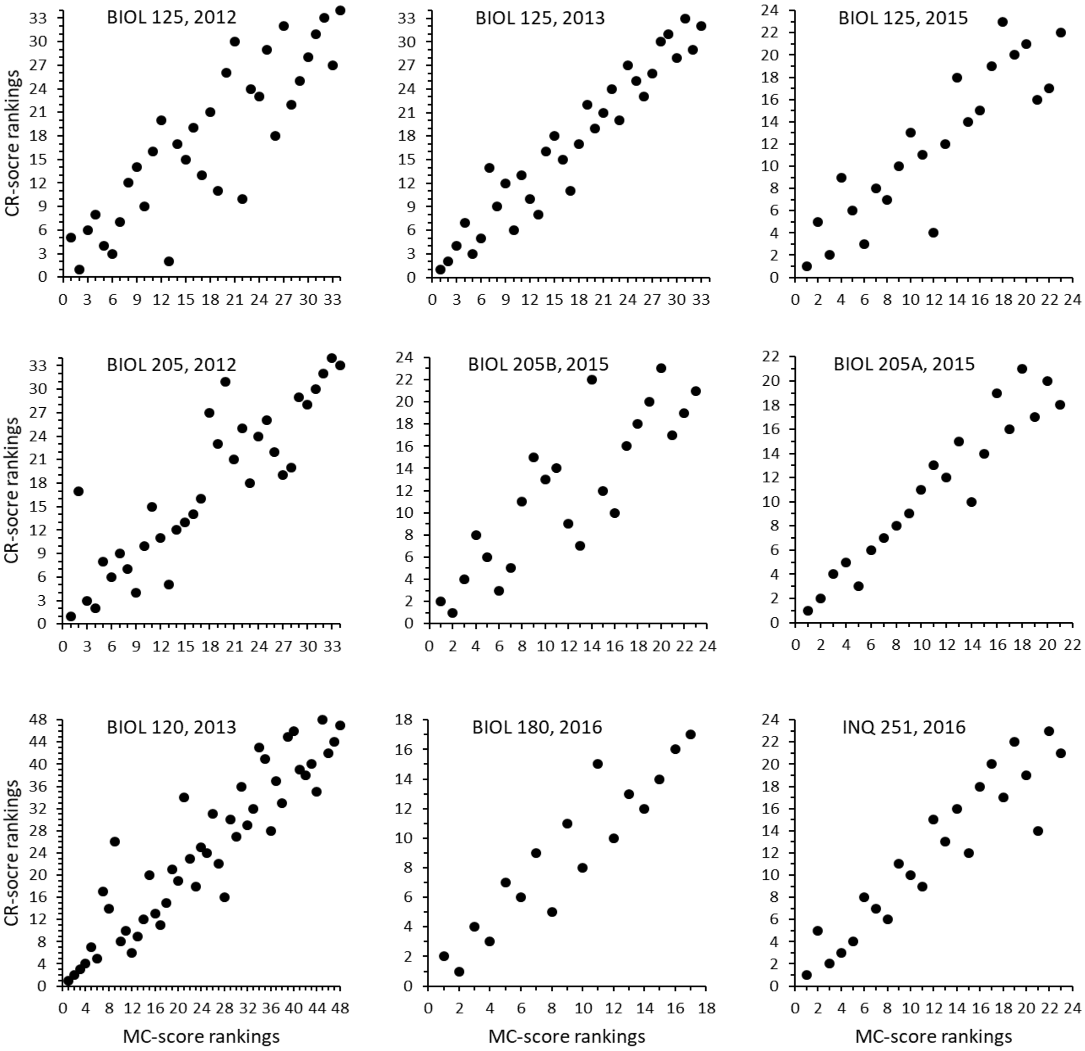

Figure 2. Relationship between Students' Test-score Rankings if Based only on CR Items versus only on MC Items for the Nine Course Sections in this Study. The Rankings are Shown in Order from the Lowest Score (Rank = 1) to the Highest Score in each Course Section

3.2 Potential Guessing Bias

The distribution of the 257 $\Delta_{\text{M-C}}$ scores was very well depicted by a normal distribution with a mean of 0 and standard deviation of 0.43 (Figure 3). Specifically, the Shapiro-Wilk goodness-of-fit test strongly supported the hypothesis that the $\Delta_{\text{M-C}}$ scores were normally distributed (W = 0.99173; P = 0.98). In addition, the 99% density ellipse for the regression of standardized CR scores on MC scores fit tightly around the regression line (Figure 1). The distances between all of the points outside the density ellipse to the ellipse are rather short, indicating that there were no extreme outliers. Furthermore, the maximum Mahalanobis distance for the correlations between the 257 standardized MC and CR scores was 3.55 (which is much less than the critical value of 13.82 for considering a point to be an outlier). The lack of outliers is a strong indication that there were no students who did much better (or worse) on MC items (e.g., through guessing) than their mastery as judged by CR items would predict.

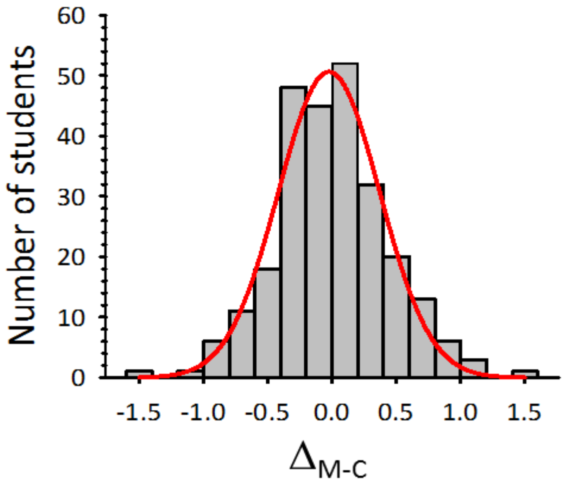

Figure 3. Distribution of $\Delta_{M\text{-}C}$ Scores for 257 Students across Nine Sections of Five Different Courses. The Normal Distribution with the Same Mean and std as this Distribution is Superimposed in Red. The Distribution does not Differ from Normal, as Determined by the Shapiro-Wilk Goodness-of-fit Test

(W = 0.99173; *P* = 0.98)

Students' relative performance on MC and CR items in one course was not a good predictor of their relative performance in a second course. Specifically, the correlation between $\Delta_{M\text{-}C}$ scores in the first and second courses (for the 18 students who took two different courses) was small, negative, and not statistically significant (*r* = -0.19; *P* = 0.46; Figure 4). If being inordinately good or bad at MC items (due to guessing ability that is inversely correlated with mastery) were a trait displayed by students, then there would have been a line with a positive slope approaching a value of one. Because this pattern was not evident, there was no indication that MC items could have been unfair due to differential guessing abilities.

3.3 Potential Gender (and Academic Standing) Bias

Academic standing had a predictable effect on performance (*P* = 0.02; Table 2A, B, Figure 5). Based on Tukey's-HSD test for pairwise comparisons, seniors scored significantly higher than any other academic class on MC items, and significantly higher than freshmen and sophomores on CR items. In addition, juniors scored significantly higher than freshmen on both MC and CR items (Figure 5). However, academic standing had no effect on the *relative* performance on MC vs. CR items (*P* = 0.75; Table 2C).

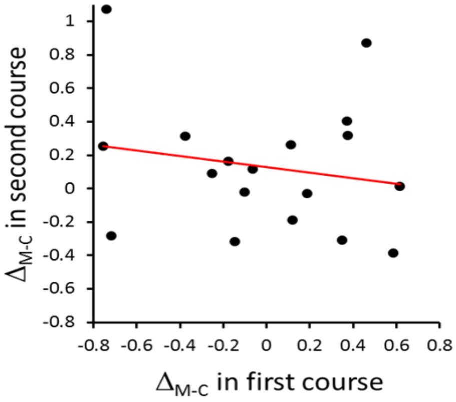

Figure 4. Lack of Correlation between Students' $\Delta_{M\text{-}C}$ Scores in Two Different Courses. The Points Represent 18 Different Students. The Red Line Represents the (non-significant) Regression of Students' $\Delta_{M\text{-}C}$ Scores for the Second Course on the $\Delta_{M\text{-}C}$ Scores for the First Course (*r* = -0.19; *P* = 0.46)

Across all nine sections, females scored slightly higher than males (4% higher in terms of least-squares means) on both MC and CR items ($P < 0.05$; Table 2A, B; Figure 6). Nevertheless, the sexes exhibited no difference in their relative performance on MC vs. CR items, with the mean $\Delta_{M\text{-}C}$ scores for both sexes being essentially zero (mean ± SE = -0.004 ± 0.05 for males and 0.003 ± 0.04 for females). Moreover, this minute difference in $\Delta_{M\text{-}C}$ between males and females was not close to statistical significance ($P = 0.8$; Table 2C).

Table 2. **Summary of ANOVA Results of the Effects of Academic Standing (i.e., Year in School) and Gender on Students' Scores on MC Items, CR Items, and the Standardized Measure of Relative Performance on MC and CR Items ($\Delta_{M\text{-}C}$)**

Source of Variation	df	Mean square	*F*-ratio	*P*-value
A. MC item				
Academic standing	3	0.146099	11.7155	<0.0001
Gender	1	0.082302	6.5997	0.01
Error	252	0.012471		
B. CR-item scores				
Academic standing	3	0.194147	9.6195	<0.0001
Gender	1	0.100461	0.0266	0.03
Error	252	0.020183		
C. $\Delta_{M\text{-}C}$				
Academic standing	3	0.075404	0.3830	0.77
Gender	1	0.072581	0.0569	0.81
Error	252	0.189520		

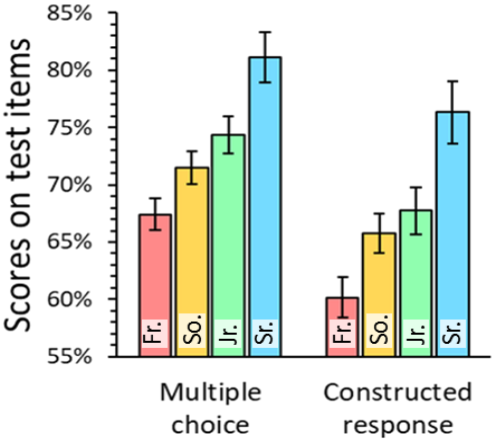

Figure 5. Comparison of Scores for Students of Different Academic Standing on MC and CR Items. Columns and Bars Represent Means ±1 Standard Error. The Dataset Included 84 Freshmen (Red), 86 Sophomores (Yellow), 59 Juniors (Green), and 28 Seniors (Blue)

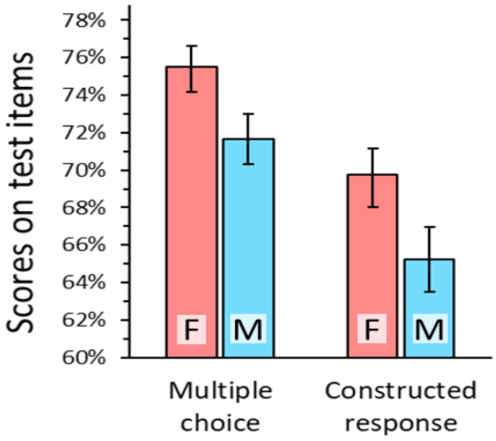

Figure 6. Comparison of Males' and Females' Scores on MC and CR Items. Columns and Bars Represent Means ± 1 Standard Error. The Dataset Included 163 Females (Red, "F") and 94 Males (Blue, "M")

4. Discussion

To appreciate fully the implications and limitations of this study, it is important to acknowledge its retrospective nature. Specifically, none of the items on the tests was written with the intention to address the questions that are the focus of this paper. I designed each test to include different types of test items that together assessed mastery of as complete a sample of constructs as possible. Whether a specific construct was tested with an MC or a CR item was decided organically, based on inspiration while I was writing. For instance, if I could come with a good set of four distractors, then I would be inclined to test a construct with an MC item. The idea to address empirically the issues regarding fairness came only after I had taught all of these courses.

If I had designed the tests *a priori* to investigate the fairness issues that are the focus of this study, I might have used a controlled design in which pairs of MC and CR items were written to test the exact same construct. Such an experimental design would be desirable in terms of obtaining cleaner estimates of parameters like the correlations between scores on MC and CR items. However, tests in real classrooms rarely include multiple items that assess the identical specific construct. Instead, I compared scores on MC and CR items that tested content within the same "domain"—that is, all content covered within a single course. Although my approach may have been statistically messier, it did have the advantage of greater authenticity in the following sense: Because the items on the tests were written in the way that classroom teachers would write their questions (i.e., not as a controlled experiment to test pedagogical hypotheses), the results of this study should be especially relevant to teachers in real classrooms.

4.1 Potential Validity Differences

In this study, students' scores on MC items and CR items on tests within the same content domain were very similar. Specifically, across nine biology courses, the correlation coefficients between scores on MC items and CR items ranged from 0.81 to 0.94, with a mean of 0.90. The high correlation does not by itself address whether the items were effective at estimating mastery. That is, the correlations are not quantifications of construct validity. (The best tool to evaluate true mastery and quantify construct validity is a clinical interview of students—a process which is not plausible for most classes.) What the high correlation does indicate is that the two types of items were nearly equally effective at quantifying mastery, and thus they must at least be very similar in construct validity—whatever the true value of validity may be.

In more concrete terms, equal validity of MC and CR items suggests that a student's grade would be very similar regardless of which type of items made up the tests. In this study, the relative rankings of students were indeed quite similar whether they were based solely on their scores on MC items or solely on their scores CR items. Combined with the high positive correlations between scores on MC and CR items, the consistency in relative grade rankings suggests that a test can be equally fair whether it is made up entirely of MC items, CR items, or a combination of the two types of items. More to the point, this study revealed no indication that MC items were any less fair than CR items to students due to any concerns regarding inadequate construct validity.

Like the current study, a wide range of investigations on a variety of types of tests have found that scores on MC and CR items tend to be positively correlated (Ward, 1982; Frakes & Lathen, 1985; Wainer & Thissen, 1993; Rodriguez, 2003; Bleske-Rechek, Zeug, & Webb, 2007; Bible, Simkin, & Kuechler, 2008; Park, 2010; Pepple, Young, & Carroll, 2010; AAAS, 2011; Tiemeier, Stacy, & Burke, 2011; Hift, 2014; Walke, Kamat, & Bhounsule, 2014). The large magnitude of the positive correlations between MC and CR scores in the current study are particularly striking when one considers the fact that MC and CR items were not explicitly paired by construct, but were written as organic test items to assess classroom mastery. These high positive correlations found in numerous studies would not be possible unless MC and CR items were similar in construct validity. In particular, I submit that if a teacher wants to include MC items on a test to take advantage of their efficiency and objectivity, then the results of the current study can be evinced as support that doing so need not involve a sacrifice in fairness in terms of construct validity.

4.2 Potential Guessing Bias

Even a completely naïve test-taker is expected to answer a certain proportion of multiple-choice items correctly through random guessing. The potential for guessing may thus make a MC item less difficult than a CR item that tests the same construct. However, this fact does not necessarily make MC items less fair. They would only be unfair to the extent that guessing decouples students' performance from students' mastery in a way that is unique to MC items. Specifically, MC items would be unfair if students with low mastery tended to be good at guessing answers to MC items, while students with higher mastery tended to be worse at guessing when they did not know an answer. Such a phenomenon would be evidenced by a cluster of students with higher scores on MC items than their scores on CR would predict (the good guessers) and a cluster of students with lower scores on MC items than their scores on CR items would predict (the bad guessers).

Across the nine courses in this study, the mean score for MC items was five percentage points higher than the mean score for CR items. This difference in difficulty was likely to have been influenced by students with incomplete mastery being more likely to guess a correct response for MC items than to construct a correct written response for CR items. Nevertheless, this study found no evidence that MC items were unfair due to a decoupling of guessing ability from mastery of constructs. Specifically, there were no students with substantially higher (or lower) scores on MC items than their scores on CR items would predict (There were no clusters of scores in the upper left or lower right corner of the plot in Figure 1).

Moreover, the distribution of $\Delta_{\text{M-C}}$ scores was nearly perfectly described by a normal (Gaussian) distribution. If there were a substantial number of students who achieved higher scores on MC items due to guessing than their mastery based on CR items would suggest, then the right-hand tail of the distribution would have been thicker and longer than would be found in a normal distribution. Likewise, if there were a substantial number of students who had relatively higher scores on CR items than on MC items due to being poor guessers, then the left-hand tail would be thicker and longer than expected. The distribution of actual $\Delta_{\text{M-C}}$ scores shows no evidence of either phenomenon. More precisely, the lack of significant Mahalanobis distances indicates that there are no outliers in terms of unexpectedly high or low $\Delta_{\text{M-C}}$ scores.

Data from the set of 18 students who took two different courses included in this study offer another way to examine the issue of guessing abilities. Assume for the sake of argument that some students are consistently better (or worse) at guessing correct responses to MC items than their performance on CR items would predict. Then relative guessing abilities of students should be consistent regardless of which course they are taking. That is, students would have relatively consistent values of $\Delta_{\text{M-C}}$ across courses. The data in the current study provide no evidence of such a pattern. Although the dataset available to examine such a pattern is not large, there is not even a slight indication that being a good guesser at MC items is a trait of students that is negatively correlated with performance on CR items.

It is important to stress that, without a clinical interview of test-takers, there is no way to know which items in this dataset were answered correctly due to guessing by which students. However, the inferences regarding fairness do not rely on the absence of guessing—only that students with poor mastery are not better at guessing correct answers on MC items than are students with higher mastery. The combined evidence of the high positive correlation of students' scores on MC and CR items, the lack of outliers in the distribution of $\Delta_{M\text{-}C}$ scores, and the lack of a correlation of $\Delta_{M\text{-}C}$ scores for students across different courses strongly suggest that the MC items were not any less fair than CR items due to any phenomenon related to guessing.

4.3 Potential Gender Bias

A long-standing criticism of MC items, especially in the context of standardized tests, is that they tend to be unfair to certain diversity-based groups—especially those who are historically underrepresented in STEM fields (Bell & Hay, 1987; Ben-Shakhar & Sinai, 1991; Bond, 1995; Lissitz et al., 2012). In particular, both conventional wisdom and some empirical results suggest that females are disadvantaged by MC items while males are disadvantaged by CR items—especially essay questions (Lumsden & Scott, 1987; Bolger & Kellaghan, 1990; Bridgeman & Lewis, 1994; DeMars, 2000; Kuechler & Simkin, 2010; Stanger-Hall, 2012). Psychological research suggests that personality characteristics that vary between genders (e.g., risk aversion, extroversion, openness to experience, and pessimism) can influence behavior on MC tests (Alker et al., 1969; Ávila & Torrubia, 2004; Kubinger & Wolfsbauer, 2010). In addition, expectations that arise from students' awareness of stereotypes may have effects that disadvantage females on MC tests (Zoller & Ben-Chaim, 1990; Hirschfeld, Moore, & Brown, 1995; Steele, 1997; Halpern, 2004).

Despite expectations, gender discrepancies in performance on MC items have not been universally observed (Bell & Hay, 1987; Feinberg, 1990; Becker & Johnston, 1999; Beller & Gafni, 2000; Chan & Kennedy, 2002; AAAS, 2011; Simkin, Kuechler, Savage, & Stiver, 2011). The contradictory results have led to research on specific factors that may contribute to performance differences between males and females on MC items. For instance, the patterns can vary across subject matter, with a greater female disadvantage on MC items in more quantitative fields (Bell & Hay, 1987; Halpern, 2004). In addition, in an analysis of performance on the International Assessment of Educational Progress (IAEP) mathematics test, Beller & Gafni (2000) found that males' advantage on MC items was only evident on the more difficult items. Similarly, Halpern (2004) noted that the advantage that males have on standardized tests occurs mainly for more unfamiliar material—that is, content that is not strictly tied to in-class learning. The method of scoring can also affect gender differences in performance on MC tests. Specifically, if there is a numeric penalty to discourage guessing, females tend to be disadvantaged, as they are more likely to be risk averse and thus less likely than males to make potentially beneficially choices based on partial knowledge (Bolger & Kellaghan, 1990; Ben-Shakhar & Sinai, 1991).

The question remains whether females are disadvantaged by the types of MC items that are typically asked to test classroom mastery—that is, items that cover material that should be familiar to the test-takers, items that are not excessively difficult, and items that are scored without penalties to discourage guessing. The dataset from the biology courses that make up the current study can be used to address this question in an authentic manner because all of the items were designed to test mastery of content and constructs specifically covered in the course. Furthermore, there were no numeric penalties for guessing on any of the tests.

Across the nine biology courses in the current study, females scored an average of 4% higher than males on both MC and CR items. This result is consistent with the nearly universal pattern of females earning higher grades than males in classroom assessments (Halpern, 2004). More importantly in terms of the focal issue of this study, there was no difference between male and female students in terms of relative performance on MC items compared to CR items. Specifically, the $\Delta_{\text{M-C}}$ was essentially zero for both genders. Therefore, there was no hint of evidence that MC items were unfair to females (or to males) relative to CR items.

Finally, there was a consistent trend toward more-experienced students performing better at both MC and CR items. Specifically, seniors performed significantly better than juniors, sophomores, and freshmen on both types of items. Nevertheless, academic standing did not influence students' relative performance on MC and CR items. Thus it appears that as students gained experience with college courses, their abilities on MC and CR items increased at the same rate—roughly five percentage points per academic year. It is also possible that the mean scores increased with academic standing as the poorer-performing students tended to drop out of the curriculum while the better-performing students remained through their senior year. Either way, the use of MC items was equally fair to students across the range of academic standing.

5. Conclusion

Recent studies have shown that multiple-choice items are efficient, reliable, and objective tools for assessing mastery of content, and they can even be used for testing higher-order cognitive skills. The results of the current study on test items from nine biology courses should help alleviate lingering concerns that multiple-choice items may be less fair to students than constructed-response items. Specifically, the high positive correlations (mean r = 0.90) between students' scores on MC and CR items within courses suggest that the two types of items were essentially equally valid tests of mastery. Second, there was no evidence that any of the 239 students received substantially higher (or lower) overall scores due to being better (or worse) guessers of MC answers than the assessment of their mastery based on CR items suggests they should have received. Third, there was no evidence that MC items unfairly disadvantaged females (or males) relative to their performance on CR items. These results should reassure teachers who use MC items on their tests that they are not unfairly judging student mastery, as well as encourage others who have been reluctant to take advantage of the benefits of MC items on their tests.

One important caveat is that not all MC items are created equal. For instance, MC items that are overly complex, poorly written, or do not have a clear correct response option will not only frustrate students, but they will reduce the validity of the items. In addition, MC items that have unintentional cues can be exploited by test-wise students, such that guessing is more likely to disconnect performance from mastery. Reviews have shown that tests written by teachers, and even commercial test banks, are often replete with MC-item-writing flaws that can lead them to be less valid and less reliable—and thus unfair—as an assessment tool (Hansen & Lee, 1997; Tarrant, Knierim, Hayes, & Ware, 2006; Moncada & Moncada, 2010; Baig, Ali, Ali, & Huda, 2014; Ibbett & Wheldon, 2016; Rush, Rankin, & White, 2016). Fortunately, it is not difficult to learn to recognize these flaws and learn to avoid them in MC-item writing. Many sources are available to help teachers write clear, highly discriminating MC items that can function as fair assessments of student mastery (Haladyna & Downing, 1989; Haladyna, 2004; Frey et al., 2005; Haladyna & Rodriguez, 2013; Towns, 2014; Rodriguez & Albano, 2017; Scully, 2017).

Even though this study suggests that students' relative rankings on tests are likely to be very similar regardless of whether MC or CR items are used, I do not suggest that MC items should be the only assessment tool used in a class, or even on a single test. For one thing, even though MC items can test higher-level thinking, in actual practice they tend to be written in a manner that requires only the lower-level cognitive processes of recognition and recall (Tarrant et al., 2006; Momsen, Long, Wyse, & Ebert-May, 2010; Baig et al., 2014; Rush et al., 2016). The inclusion of CR items can help ensure that tests will include sufficient items that require higher-order cognitive processes. Moreover, CR items are required to test the vital skills of logically formulating and clearly writing arguments (Aiken, 1987). Finally, students may spend less time preparing and study more superficially for MC tests because of the perception that they will merely need to recognize the correct answers provided on the test (Gustav, 1964; Biggs, 1979; Simkin & Kuechler, 2005; Funk & Dickson, 2011; Pinckard, McMahan, Prihoda, Littlefield, & Jones, 2012; Stanger-Hall, 2012). In conclusion, while they cannot achieve all assessment objectives, carefully written MC items constitute a convenient, reliable, and ultimately fair tool in a teacher's toolkit for assessing mastery of course content.

Acknowledgments

I thank S.E. Wise and several anonymous reviewers for constructive comments on the manuscript text.

References

AAAS. (2011). *Vision and Change in Undergraduate Biology Education: A Call to Action*. Washington, D.C.: American Association for the Advancement of Science.

Aiken, L. R. (1982). Writing multiple-choice items to measure higher-order educational objectives. *Educational and Psychological Measurement*, *42*, 803-806. https://doi.org/10.1177/001316448204200312

Aiken, L. R. (1987). Testing with multiple-choice items. *Journal of Research and Development in Education*, *20*(4), 44-58.

Alker, H. A., Carlson, J. A., & Hermann, M. G. (1969). Multiple-choice questions and student characteristics. *Journal of Educational Psychology*, *60*(3), 231-243. https://doi.org/10.1037/h0027615

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. https://doi.org/10.1111/j.1745-3984.2003.tb01099.x

Ávila, C., & Torrubia, R. (2004). Personality, expectations, and response strategies in multiple-choice question examinations in university students: A test of Gray's hypotheses. *European Journal of Personality*, *18*, 45-59. https://doi.org/10.1002/per.506

Bacon, D. R. (2003). Assessing learning outcomes: A comparison of multiple-choice and short-answer questions in a marketing context. *Journal of Marketing Education*, *25*(1), 31-36. https://doi.org/10.1177/0273475302250570

Baig, M., Ali, S. K., Ali, S., & Huda, N. (2014). Evaluation of multiple choice and short essay questions items in basic medical science. *Pakistan Journal of Medical Sciences*, *30*(1), 3-6.

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. https://doi.org/10.1198/000313002623

Bar-Hillel, M., Budescu, D., & Attali, Y. (2005). Scoring and keying multiple choice tests: A case study in irrationality. *Mind & Society*, *56*, 3-12. https://doi.org/10.1007/s11299-005-0001-z

Becker, W. E., & Johnston, C. (1999). The relationship between multiple choice and essay response questions in assessing economics understanding. *The Economic Record*, *75*(231), 348-357. https://doi.org/10.1111/j.1475-4932.1999.tb02571.x

Bell, R. C., & Hay, J. A. (1987). Differences and biases in English language examination formats. *British Journal of Educational Psychology*, *57*, 212-220. https://doi.org/10.1111/j.2044-8279.1987.tb03154.x

Beller, M., & Gafni, N. (2000). Can item format (multiple-choice vs. open-ended) account for gender differences in mathematics achievement? *Sex Roles*, *42*(1/2), 1-21. https://doi.org/10.1023/A:1007051109754

Ben-Shakhar, G., & Sinai, Y. (1991). Gender differences in multiple-choice tests: The role of differential guessing techniques. *Journal of Educational Measurement*, *28*(1), 23-35. https://doi.org/10.1111/j.1745-3984.1991.tb00341.x

Bible, L., Simkin, M. G., & Kuechler, W. L. (2008). Using multiple-choice tests to evaluate students' understanding of accounting. *Accounting Education: an international journal*, *17,* S55-S68. https://doi.org/10.1080/09639280802009249

Biggs, J. (1979). Individual differences in study processes and the quality of learning outcomes. *Higher Education*, *8*, 381-394. https://doi.org/10.1007/BF01680526

Bleske-Rechek, A., Zeug, N., & Webb, R. M. (2007). Discrepant performance on multiple-choice and short answer assessments and the relation of performance to general scholastic aptitude. *Assessment & Evaluation in Higher Education*, *32*(2), 89-105. https://doi.org/10.1080/02602930600800763

Bolger, N., & Kellaghan, T. (1990). Method of measurement and gender differences in scholastic achievement. *Journal of Educational Measurement*, *27*(2), 165-174. https://doi.org/10.1111/j.1745-3984.1990.tb00740.x

Bond, L. (1995). Unintended consequences of performance assessment: Issues of bias and fairness. *Educational Measurement: Issues and Practice*, *14*(4), 21-24. https://doi.org/10.1111/j.1745-3992.1995.tb00885.x

Bridgeman, B., & Lewis, C. (1994). The relationship of essay and multiple-choice scores with grades in college courses. *Journal of Educational Measurement*, *31*(1), 37-50. https://doi.org/10.1111/j.1745-3984.1994.tb00433.x

Bush, M. (2015). Reducing the need for guesswork in multiple-choice tests. *Assessment & Evaluation in Higher Education*, *40*(2), 218-231. https://doi.org/10.1080/02602938.2014.902192

Campbell, M. L. (2015). Multiple-choice exams and guessing: Results from a one-year study of general chemistry tests designed to discourage guessing. *Journal of Chemical Education*, *92*, 1194-1200. https://doi.org/10.1021/ed500465q

Chan, N., & Kennedy, P. E. (2002). Are multiple-choice exams easier for economics students? A comparison of multiple-choice and "equivalent" constructed-response exam questions. *Southern Economic Journal*, *68*(4), 957-971. https://doi.org/10.2307/1061503

DeMars, C. E. (2000). Test stakes and item format interactions. *Applied Measurement in Education*, *13*(1), 55-77. https://doi.org/10.1207/s15324818ame1301_3

Dolly, J. P., & Williams, K. S. (1986). Using test-taking strategies to maximize multiple-choice test scores. *Educational and Psychological Measurement*, *46*(3), 619-625. https://doi.org/10.1177/0013164486463014

Domyancich, J. M. (2014). The development of multiple-choice items consistent with the AP chemistry curriculum framework to more accurately assess deeper understanding. *Journal of Chemical Education*, *91*, 1347-1351. https://doi.org/10.1021/ed5000185

Feinberg, L. (1990). Multiple-choice and its critics. *The College Board Review*, *157*, 12-17.

Frakes, A. H., & Lathen, W. C. (1985). A comparison of multiple-choice and problem examinations in introductory financial accounting. *Journal of Accounting Education*, *3*(1), 81-89. https://doi.org/10.1016/0748-5751(89)90039-0

Frey, B. B., Petersen, S., Edwards, L. M., Pedrotti, J. T., & Peyton, V. (2005). Item-writing rules: Collective wisdom. *Teaching and Teacher Education*, *21*, 357-364. https://doi.org/10.1016/j.tate.2005.01.008

Funk, S. C., & Dickson, K. L. (2011). Multiple-choice and short-answer exam performance in a college classroom. *Teaching of Psychology*, *38*(4), 273-277. https://doi.org/10.1177/0098628311421329

Gustav, A. (1964). Students' preference for test format in relation to their test scores. *The Journal of Psychology*, *57*, 159-164. https://doi.org/10.1080/00223980.1964.9916685

Haladyna, T. M. (1997). *Writing Test Items to Evaluate Higher Order Thinking*. Needham Heights, MA: Allyn & Bacon.

Haladyna, T. M. (2004). *Developing and validating multiple-choice test items* (3rd ed.). New York, NY: Routledge.

Haladyna, T. M., & Downing, S. M. (1989). A taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education*, *2*(1), 37-50. https://doi.org/10.1207/s15324818ame0201_3

Haladyna, T. M., & Rodriguez, M. C. (2013). *Developing and Validating Test Items*. New York: 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. https://doi.org/10.1207/S15324818AME1503_5

Halpern, D. F. (2004). A cognitive-process taxonomy for sex differences in cognitive abilities. *Current Directions in Psychological Science*, *13*(4), 135-139. https://doi.org/10.1111/j.0963-7214.2004.00292.x

Hansen, J. D., & Lee, D. (1997). Quality multiple-choice test questions: Item-writing guidelines and an analysis of auditing testbanks. *Journal of Education for Business*, *73*(2), 94. https://doi.org/10.1080/08832329709601623

Hift, R. J. (2014). Should essays and other open-ended-type questions retain a place in written summative assessment in clinical medicine? *BMC Medical Education*, *14*(249). https://doi.org/10.1186/s12909-014-0249-2

Hirschfeld, M., Moore, R. L., & Brown, E. (1995). Exploring the gender gap on the GRE subject test in economics. *Journal of Economic Education*, *26*(1), 3-15. https://doi.org/10.1080/00220485.1995.10844852

Hoffmann, B. (1962). *The Tyranny of Testing*. New York: The Crowell-Collier Press.

Ibbett, N. L., & Wheldon, B. J. (2016). The incidence of clueing in multiple choice testbank questions in accounting: Some evidence from Australia. *e-Journal of Business Education & Scholarship of Teaching*, *10*(1), 20-35.

Kubinger, K. D., & Wolfsbauer, C. (2010). On the risk of certain psychotechnological response options in multiple-choice tests: Does a particular personality handicap examinees? *European Journal of Psychological Assessment*, *26*(4), 302-308. https://doi.org/10.1027/1015-5759/a000040

Kuechler, W. L., & Simkin, M. G. (2010). Why is performance on multiple-choice tests and constructed-response tests not more closely related? Theory and an empirical test. *Decision Sciences Journal of Innovative Education*, *8*(1), 55-73. https://doi.org/10.1111/j.1540-4609.2009.00243.x

Lissitz, R. W., Hou, X., & Slater, S. C. (2012). The contribution of constructed response items to large scale assessment: Measuring and understanding their impact. *Journal of Applied Testing Technology*, *13*(3), 1-50.

Lukhele, R., Thissen, D., & Wainer, H. (1994). On the relative value of multiple-choice, constructed response, and examinee-selected items on two achievement tests. *Journal of Educational Measurement*, *31*(3), 234-250. https://doi.org/10.1111/j.1745-3984.1994.tb00445.x

Lumsden, K. G., & Scott, A. (1987). The economics student reexamined: Male-female differences in comprehension. *Research in Economic Education*, *18*(4), 365-375. https://doi.org/10.1080/00220485.1987.10845228

McCoubrie, P. (2004). Improving the fairness of multiple-choice questions: A literature review. *Medical Teacher*, *26*(8), 709-712. https://doi.org/10.1080/01421590400013495

Momsen, J. L., Long, T. M., Wyse, S. A., & Ebert-May, D. (2010). Just the facts? Introductory biology courses focus on low-level cognitive skills. *CBE-Life Sciences Education*, *9*(winter), 435-440. https://doi.org/10.1187/cbe.10-01-0001

Moncada, S. M., & Moncada, T. P. (2010). Assessing student learning with conventional multiple-choice exams: Design and implementation considerations for business faculty. *International Journal of Education Research*, *5*(2), 15-29.

Morrison, S., & Free, K. W. (2001). Writing multiple-choice test items that promote and measure criticial thinking. *Journal of Nursing Education, 40*(1), 17-24.

Park, J. (2010). Constructive multiple-choice testing system. *British Journal of Educational Technology*, *41*(6), 1054-1064. https://doi.org/10.1111/j.1467-8535.2010.01058.x

Pepple, D. J., Young, L. E., & Carroll, R. G. (2010). A comparison of student performance in multiple-choice and long essay questions in the MBBS stage 1 physiology examination at the University of the West Indies (Mona Campus). *Advances in Physiological Education*, *54*, 86-89. https://doi.org/10.1152/advan.00087.2009

Pinckard, R. N., McMahan, C. A., Prihoda, T. J., Littlefield, J. H., & Jones, A. C. (2012). Short-answer questions and formula scoring separately enhance dental student academic performance. *Journal of Dental Education*, *76*(5), 620-634. https://doi.org/10.1002/j.0022-0337.2012.76.5.tb05296.x

Reeves, T. D., & Marbach-Ad, G. (2016). Contemporary test validity in theory and practice: A primer for discipline-based education researchers. *CBE--Life Sciences Education*, *15*(spring), rm1. https://doi.org/10.1187/cbe.15-08-0183

Rodriguez, M. C. (2003). Construct equivalence of multiple-choice and constructed-response items: A random effects synthesis of correlations. *Journal of Educational Measurement*, *40*(2), 163-184. https://doi.org/10.1111/j.1745-3984.2003.tb01102.x

Rodriguez, M. C., & Albano, A. D. (2017). *The College Instructor's Guide to Writing Test Items: Measuring Student Learning*. New York, NY: Taylor & Francis Group.

Rush, B. R., Rankin, D. C., & White, B. J. (2016). The impact of item-writing flaws and item complexity on examination item difficulty and discrimination value. *BMC Medical Education*, *16*, 250. https://doi.org/10.1186/s12909-016-0773-3

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

Scully, D. (2017). Constructing multiple-choice items to measure higher-order thinking. *Practical Assessment, Research & Evaluation*, *22*(4).

Simkin, M. G., & Kuechler, W. L. (2005). Multiple-choice tests and student understanding: What is the connection? *Decision Sciences Journal of Innovative Education*, *3*(1), 73-97. https://doi.org/10.1111/j.1540-4609.2005.00053.x

Simkin, M. G., Kuechler, W. L., Savage, A., & Stiver, D. (2011). Why use multiple-choice questions on accounting certification examinations. *Global Perspectives on Accounting Education*, *8*, 27-46.

Stanger-Hall, K. F. (2012). Multiple-choice exams: An obstacle for higher-level thinking in introductory science classes. *CBE--Life Sciences Education*, *11*, 294-306. https://doi.org/10.1187/cbe.11-11-0100

Steele, C. M. (1997). A threat in the air: How stereotypes shape intellectual identity and performance. *American Psychologist*, *52*(6), 613-629. https://doi.org/10.1037/0003-066X.52.6.613

Tarrant, M., Knierim, A., Hayes, S. K., & Ware, J. (2006). The frequency of item writing flaws in multiple-choice questions used in high stakes nursing assessments. *Nurse Education Today*, *26*, 662-671. https://doi.org/10.1016/j.nedt.2006.07.006

Tiemeier, A. M., Stacy, Z. A., & Burke, J. M. (2011). Using multiple-choice questions written at various Bloom's taxonomy levels to evaluate student performance across a therapeutics sequence. *Innovations in Pharmacy*, *2*(2), Article 41. https://doi.org/10.24926/iip.v2i2.224

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

Tractenberg, R. E., Gushta, M. M., Mulroney, S. E., & Weissinger, P. A. (2013). Multiple choice questions can be designed or revised to challenge learners' critical thinking. *Advances in Health Science Education*, *18*, 945-961. https://doi.org/10.1007/s10459-012-9434-4

Wainer, H., & Thissen, D. (1993). Combining multiple-choice and constructed-response test scores: Toward a Marxist theory of test construction. *Applied Measurement in Education*, *6*(2), 103-118. https://doi.org/10.1207/s15324818ame0602_1

Walke, Y. S. C., Kamat, A. S., & Bhounsule, S. A. (2014). A retrospective comparative study of multiple choice questions versus short answer questions as assessment tool in evaluating the performance of the students in medical pharmacology. *International Journal of Basic & Applied Pharmacology*, *3*(6), 1020-1023. https://doi.org/10.5455/2319-2003.ijbcp20141212

Ward, W. C. (1982). A comparison of free-response and multiple-choice forms of verbal aptitude tests. *Applied Psychological Measurement*, *6*(1), 1-11. https://doi.org/10.1177/014662168200600101

Zieky, M. J. (2006). Fairness reviews in assessment. In S. M. Downing, & T. M. Haladyna (Eds.), *Handbook of Test Development* (pp. 359-376). Mahwah, NJ: Lawrence Erlbaum Associates.

Zimmerman, D. W., & Williams, R. H. (2003). A new look at the influence of guessing on the reliability of multiple-choice tests. *Applied Psychological Measurement*, *27*(5), 357-371. https://doi.org/10.1177/0146621603254799

Zoller, U., & Ben-Chaim, D. (1990). Gender differences in examination-type preferences, test anxiety, and academic achievements in college science education—A case study. *Science Education*, *74*(6), 597-608. https://doi.org/10.1002/sce.3730740603