University of Nebraska - Lincoln
DigitalCommons@University of Nebraska - Lincoln

Faculty Publications in the Biological Sciences | Papers in the Biological Sciences

6-2018

Multiple–True–False Questions Reveal the Limits of the Multiple–Choice Format for Detecting Students with Incomplete Understandings

Brian Couch
University of Nebraska - Lincoln, bcouch2@unl.edu

Joanna K. Hubbard
University of Colorado, Boulder, jhubbard@truman.edu

Chad Brassil
University of Nebraska - Lincoln, cbrassil@unl.edu

Follow this and additional works at: http://digitalcommons.unl.edu/bioscifacpub

 Part of the Biology Commons, Educational Assessment, Evaluation, and Research Commons, Educational Methods Commons, and the Higher Education Commons

Couch, Brian; Hubbard, Joanna K.; and Brassil, Chad, "Multiple–True–False Questions Reveal the Limits of the Multiple–Choice Format for Detecting Students with Incomplete Understandings" (2018). *Faculty Publications in the Biological Sciences*. 685.
http://digitalcommons.unl.edu/bioscifacpub/685

This Article is brought to you for free and open access by the Papers in the Biological Sciences at DigitalCommons@University of Nebraska - Lincoln. It has been accepted for inclusion in Faculty Publications in the Biological Sciences by an authorized administrator of DigitalCommons@University of Nebraska - Lincoln.

University of Nebraska Lincoln
digitalcommons.unl.edu

Published in *BioScience* 68:6 (June 2018), pp 455–463.
doi:10.1093/biosci/biy037
Copyright © 2018 Brian A. Couch, Joanna K. Hubbard, and Chad E. Brassil.
Published by Oxford University Press on behalf of the American Institute of Biological Sciences. Used by permission.

Multiple–True–False Questions Reveal the Limits of the Multiple–Choice Format for Detecting Students with Incomplete Understandings

Brian A. Couch, Joanna K. Hubbard, and Chad E. Brassil

Abstract

By having students select one answer among several plausible options, multiple–choice (MC) questions capture a student's preferred answer but provide little information regarding a student's thinking on the remaining options. We conducted a crossover design experiment in which similar groups of introductory biology students were assigned verbatim questions in the MC format or multiple–true–false (MTF) format, which requires students to separately evaluate each option as either true or false. Our data reveal that nearly half of the students who select the correct MC answer likely hold incorrect understandings of the other options and that the selection rates for individual MC options provide inaccurate estimations of how many students separately endorse each option. These results suggest that MC questions systematically overestimate question mastery and underestimate the prevalence of mixed and partial conceptions, whereas MTF questions enable students and instructors to gain a more nuanced portrait of student thinking with little additional effort.

Keywords: assessment, multiple–choice, multiple–true–false, question format, undergraduate education

Brian A. Couch (bcouch2@unl.edu) is an assistant professor and **Chad E. Brassil** (cbrassil@unl.edu) is an associate professor affiliated with the School of Biological Sciences at the University of Nebraska–Lincoln. **Joanna K. Hubbard** (jhubbard@truman.edu) is an assistant professor affiliated with the Department of Biology at Truman State University, in Kirksville, Missouri.

Instructors use assessment ubiquitously throughout undergraduate science courses to gauge student understanding of important concepts. In addition to measuring student understanding, assessment can serve to communicate course expectations, provide feedback that promotes learning, facilitate student discussion, and empower students to take responsibility for their own learning (Angelo 1998, Black and Wiliam 2009). Assessment prompts come in a variety of formats, and the item types selected for a test, quiz, or other assessment activity can have significant effects on student study behaviors and overall student performance (Bridgeman and Morgan 1996, Stanger-Hall 2012).

Undergraduate teaching and assessment take place in a context of limited resources and practical constraints (Dancy and Henderson 2010, Ebert-May et al. 2011). As a consequence, many instructors choose to administer closed-ended items with predefined response options, allowing for rapid machine grading. The multiple–choice (MC) format, in which students select a single preferred answer from a list of several plausible options, has achieved widespread use in science, technology, engineering, and math (STEM) disciplines, including biology courses. Over 30% of undergraduate STEM instructors report the use of MC exams in all or most of the courses they teach (Hurtado et al. 2012), and this percentage is likely higher among biology instructors teaching large introductory courses that serve as gateways to a major. Indeed, most introductory biology textbooks include supplementary learning programs and test banks that use MC questions. With MC questions, incorrect responses are designed to represent known student misconceptions, and the response rates for each so-called distractor are taken by instructors as indications of the prevalence of the misconception among students. MC questions have thus been employed for diagnostic purposes in the form of in-class clicker questions or within concept inventories, allowing instructors to modify their instruction on the basis of student performance (Mazur 1996, Wood 2004, Caldwell 2007, Libarkin 2008, Adams and Wieman 2011).

As diagnostic probes, assessment items are designed to accurately reveal student thinking regarding a concept or problem, a task that is made difficult by the complex and incoherent structure of student mental models. Biology students often have mixed or partial understandings in which they simultaneously hold correct and incorrect ideas about a particular concept (Nehm and Reilly 2007). For example, students may correctly understand that DNA mutations arise randomly in a population and also incorrectly believe that organisms can induce specific mutations to overcome an environmental stress (Nehm and Schonfeld 2008). The

mixed nature of student thinking presents a problem for the MC format, in which students can only endorse one response option. MC response selections provide information on a student's preferred answer but provide no direct information regarding a student's desire to select the remaining options. As a result, students can select the correct option while simultaneously believing (but being unable to indicate) that some of the remaining distractors are also correct. Conversely, a student may select an incorrect option but still think the correct option is true, albeit to a lesser extent. Therefore, the artificial requirement to only select one answer in the MC format creates a problem for biology instructors, because it can obscure their ability to diagnose student thinking regarding the various answer options. Furthermore, MC questions are prone to test-taking strategies (e.g., option elimination), which tend to be underused by low-performing students and students from underrepresented groups (Kim and Goetz 1993, Ellis and Ryan 2003, Stenlund et al. 2017).

Multiple–true–false (MTF) questions represent an alternative format that retains many of the benefits and grading efficiencies of the MC format. MTF questions consist of a question stem followed by a series of prompts or statements that students evaluate as being true or false. MTF questions function similarly to MC questions in which students "select all that apply," except that MTF questions require marking of both correct and incorrect statements. By having students evaluate each statement, MTF questions have the potential to detect students with mixed conceptions, because student answers to a single question stem can include both correct and incorrect responses (Parker et al. 2012). The MTF format also potentially mitigates test-taking strategies that rely on option comparison. Previous research has shown that MTF questions are adequately familiar to students and that the conceptions detected by MTF questions mirror the response patterns observed in open-ended interviews (Federer et al. 2013). Although MTF questions cannot replace open-ended questions, similar multiple–response questions can have item difficulties similar to free-response (FR) questions (Kubinger and Gottschall 2007) and recapitulate some aspects of FR answers (Wilcox and Pollock 2014). In a previous study comparing MTF and FR formats, we also found that the rate of correct responses to MTF questions correlates with the rates at which students will list the corresponding conceptions in FR answers (Hubbard et al. 2017).

Despite the potential strengths of the MTF format, studies have not directly compared student response patterns between MC and MTF questions. This comparison is important because MC questions are in widespread use, but the degree to which these questions fail to capture the

presence of incomplete understandings remains unclear. The parallel structure of the MC and MTF formats provides an opportunity to investigate this limitation (**Figure 1**); MC questions with one correct response and three distractors are functionally equivalent to MTF questions with one true statement and three false statements, and these questions can be interchanged by restructuring the response input while leaving the question prose the same. When answering such MTF questions, many upper-division biology students who correctly identify the true statement also incorrectly mark at least one of the remaining false statements as true (Couch et al. 2015). Although these results suggest that the MC format may overlook a substantial number of students holding mixed or partial conceptions, these same questions were not answered by students in the MC format, preventing a direct comparison of MC and MTF responses.

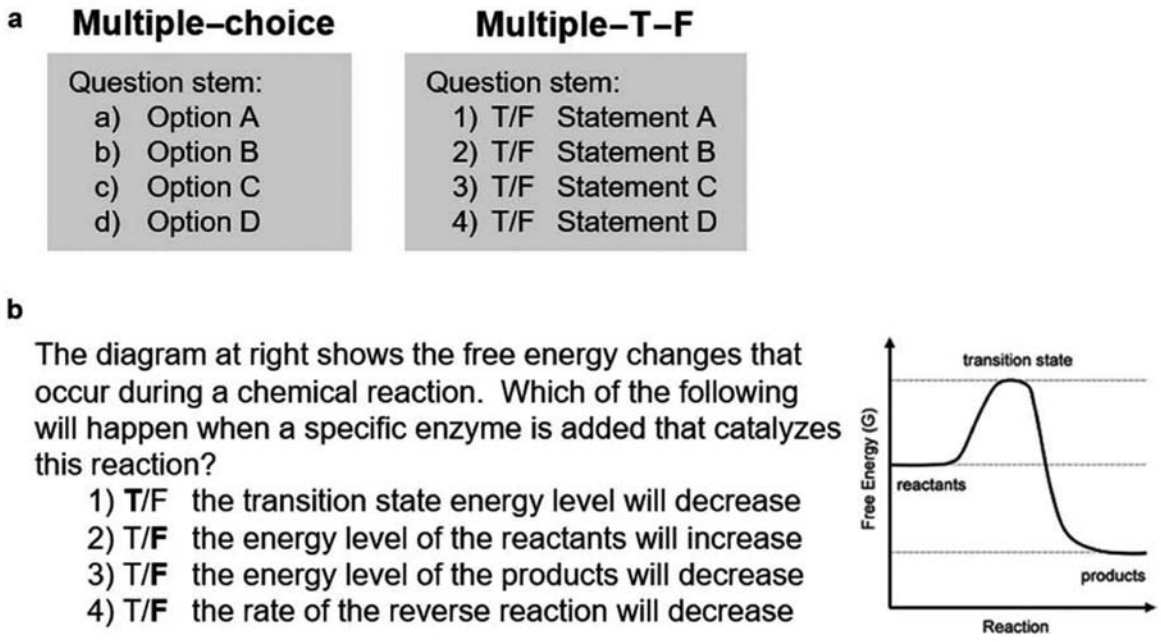

Figure 1. MC and MTF question formats. **(a)** Generic format for MC and MTF questions. For MC questions, students select one response option. For MTF questions, students answer either true or false for each of the four statements. These formats are interchangeable when a question has only one correct or true option or statement, and the various options are nonmutually exclusive. **(b)** Example of an interchangeable MC×MTF question. In the MC format, this question appeared with four response options, with "A" being the correct option. In the MTF format, this question appeared as is shown in the figure, with the first statement being true and the remaining statements all being false.

In this study, we sought to address several research questions related to how response patterns and question scores differ when verbatim questions are posed in either an MC or MTF format. To what extent do the two formats differ in revealing students who have not mastered all the concepts included in a question? How do the rates at which students select a particular MC option compare with the rates at which they will endorse the same option expressed as an MTF statement? How do the question scores and overall scores that students receive differ between the MC and MTF formats? Are findings regarding the two formats on exams recapitulated in the context of clicker questions? By addressing these questions, we sought to provide empirical information to guide instructors as they select and interpret different question formats.

Experimental design

For our main investigation, we implemented a within-subjects controlled experimental design within the four-unit exams of an introductory biology course that enabled us to compare response patterns for verbatim MC and MTF questions administered to similar groups of students. This design was similar to that described previously (Hubbard et al. 2017), but the exams included in these two studies took place in different semesters with different students and different questions. Clicker questions were asked in MC and MTF-like formats throughout the term, so the students had practice with both formats prior to exams.

For each unit exam, we began by generating a question bank consisting of MC and MTF questions, each having four different response options or statements (**Supplemental Figure S1**). The question bank for each exam consisted of 5 control MC questions, 10 experimental MC×MTF questions having one correct or true and three incorrect or false response options, and 9 control MTF questions containing two or three true statements. The answer options for the experimental MC×MTF questions were nonmutually exclusive, meaning that each option could be interpreted independently from the others. We did not include MTF questions with either zero or four true statements because of previous observations that students were reluctant to select those patterns during interviews (Couch et al. 2015), suggesting that their inclusion could have introduced response artifacts.

These questions were used to make two exam versions. Control MC and control MTF questions appeared identically on both versions. Half of the MC×MTF questions appeared in the MC form on Version A and in the MTF form on Version B (i.e., Set 1). The other half of the experimental

questions appeared in the reciprocal arrangement: as MTF questions on Version A and as MC questions on Version B (i.e., Set 2). Importantly, the MTF section had a roughly even balance of questions with one, two, or three true statements to discourage students from biasing their question responses toward a particular pattern (Cronbach 1941). Versions A and B were then distributed to the students in a semi-random fashion on exam day, alternating across auditorium rows. Across the semester, 20 control MC questions, 32 control MTF questions, and 36 experimental MC×MTF questions were included in the final analyses. A total of 249 students were enrolled in the course, and 194 students agreed to have their course data released for research purposes, representing a 78% participation rate. Additional details on exam construction, administration, and processing can be found in the supplemental materials.

Data analyses

For the within-student experimental design, the main effect of question format was robust to chance differences in the particular samples of students taking the two test versions. However, to check for differences, we assessed the equivalence of students taking the different exam versions. We calculated overall student scores for control questions, collected incoming ACT scores for the students taking each version, and analyzed the differences at the student level between versions for each exam with Student's *t*-tests. To compare scores on control questions, we calculated the percentage correct for each MC and MTF question and determined the Pearson's correlation between the percentage correct for each question on the two different versions. The percentage correct for an MTF question equals the average percentage correct for each of the four true–false (T–F) statements.

Apparent mastery was determined for the experimental MC×MTF questions appearing in either the MC or MTF formats. For the MC format, apparent mastery was calculated as the percentage of students who selected the correct option. For the MTF format, apparent mastery was calculated as the percentage of students who provided a fully correct answer in which they answered all four T–F statements correctly (i.e., they answered *true* for the one true statement and *false* for the three false statements). Apparent mastery rates were compared at the question level between the MC and MTF formats using a paired Student's *t*-test.

We further wished to compare the rates at which students would endorse or provide an affirmative marking of the various answer options in the two formats. In the MC format, the endorsement rates were

calculated as the percentage of students who selected each option, and in the MTF format, the endorsement rates were calculated as the percentage of students who marked *true* for each T–F statement. The endorsement rates were then grouped according to whether the underlying option was intended to be correct or true versus incorrect or false. Incorrect or false options were further grouped according to whether the option was the first, second, or third most popular distractor in the MC format. The endorsement rates for the response options were compared between the MC and MTF formats at the question level using a two-way ANOVA (2 formats × 4 options), with question as a random effect, to protect against type I errors. This was followed by a post hoc paired Student's *t*-tests to identify differences in statement endorsement rates between the two formats.

For question-level scores, we calculated the percentage correct for each experimental MC question and the average percentage correct for the four T–F statements constituting each MTF question. For overall scores, we calculated the percentage correct for each student across all experimental MC questions and all experimental MTF statements. We compared question scores and overall scores across the two formats by (a) calculating Pearson's correlations between the two formats, (b) determining whether average scores differed between formats with Student's *t*-tests, and (c) identifying the intersection of the regression line with the one-to-one line for scores in the two formats; the shape of the regression line was determined by comparing the fit of linear and nonlinear models.

Comparison with multiple–choice and multiple–true–false–like clicker questions

To determine whether the results from exams were robust to context, we implemented a follow-up experiment the next year with different assessment stakes and response technique. We used two sections of the same course taught by the same instructor to compare response patterns for 16 new clicker questions asked in either MC or MTF-like formats. MTF-like clicker questions were delivered by having the students manually input all the statements thought to be true on their clicker devices in alphanumeric mode and omit statements thought to be false. For example, the students would type "AC" if they thought statements "A" and "C" were both true. These experimental clicker questions with one correct or true option were embedded within a broader set of MC questions (with a single correct option) and MTF questions with two or three true statements. The experimental questions were displayed

alternately to one section in the MC format and to the other section in the MTF format. A total of 468 students were enrolled in the two course sections, and 405 students agreed to have their course data released for research purposes, representing an 87% participation rate. To determine whether the effect of format was robust across the two contexts (i.e., high-stakes exams and low-stakes clickers), we calculated the fractional relationship of endorsement rates (MTF or MC) in each context (the main data from exams versus the follow-up data from clickers) and compared these relationships with a two-way ANOVA at the question level (2 contexts x 4 options).

Students perform similarly on control questions

The bulk of our analyses focused on the data set coming from exams. For this main investigation, there was no significant difference in overall control question scores or incoming ACT scores across the different versions, suggesting they were taken by groups with similar performance for each exam (**Supplemental Table S1**). At the individual question level, we calculated the percentage correct for control MC and control MTF questions and found a strong correlation between these question scores on the two exam versions (**Supplemental Figure S2**; MC: $r = .95$, $p < .001$; MTF: $r = .96$, $p < .001$). These correlations provide a baseline of the minimal variation expected for identical questions and indicate that other factors, such as question order, did not have major impacts on control question scores.

Multiple–choice questions overestimate mastery of all the response options

Using the experimental MC×MTF exam questions, we estimated the degree to which the different formats revealed complete or incomplete understandings of the various response options. We analyzed this "apparent question mastery" by comparing the percentage of students who answered correctly in the MC format with the percentage of students who gave a fully correct answer in the MTF format, in which they answered all four T–F statements correctly. For most questions, the percentage correct in the MC format exceeded the percentage fully correct in the MTF format (**Figure 2** a). On average, the students answered the MC format correctly 67% of the time, whereas a similar group of students provided a fully correct answer in the MTF format only 36% of the time (**Figure**

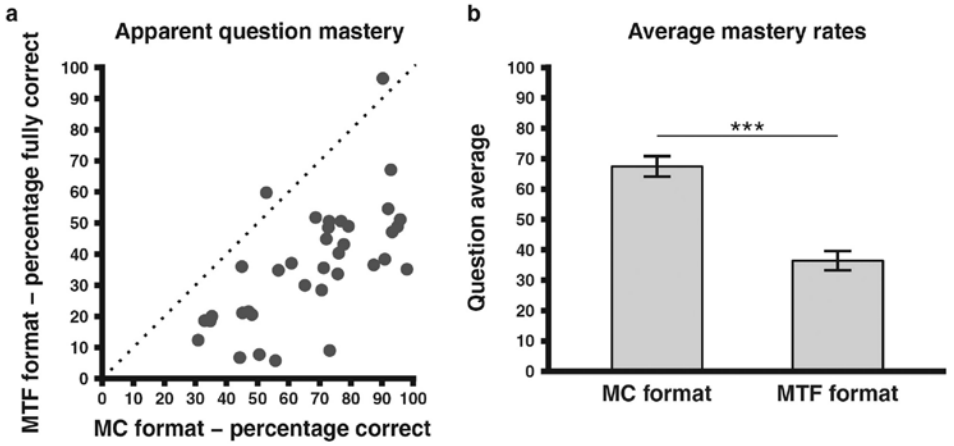

Figure 2. The apparent mastery rates for individual questions. **(a)** The dots represent apparent mastery rates for each MC × MTF question. The *x*-axis gives the percentage of students who selected the correct option in the MC format, and the y-axis gives the percentage of students who gave a fully correct response in the MTF format, in which all four true–false statements were answered correctly. The dotted line represents the one-to-one line in which response options would appear if they had the same apparent mastery rates in the different formats. **(b)** The average apparent mastery rates in the MC and MTF formats. The gray bars represent the average percentage correct in the MC format or percentage fully correct in the MTF format. The error bars represent standard errors of the mean. Paired Student's *t*-test: $t(35) = 11.39$, *** $p < .001$.

2 b; $t(35) = 11.39$, $p < .001$). Thus, although many of the students were given full credit in the MC format, only about half of these students would have been able to demonstrate complete question mastery when asked to evaluate all the response options.

Multiple–choice questions misestimate the independent endorsement rates for each option

To further understand the differences between the MC and MTF formats, we analyzed the endorsement rates for each response option. In the MC format, endorsement of an option was indicated by a student selecting this option. In the MTF format, an endorsement occurred when a student marked *true* for a given T–F statement. We compared the percentage selecting each MC option with the percentage marking *true* for each corresponding T–F statement, and these relationships were separately visualized for the intended correct or true option (**Figure 3** a) and the incorrect or false options (**Figure 3** b–d). For the correct or true option,

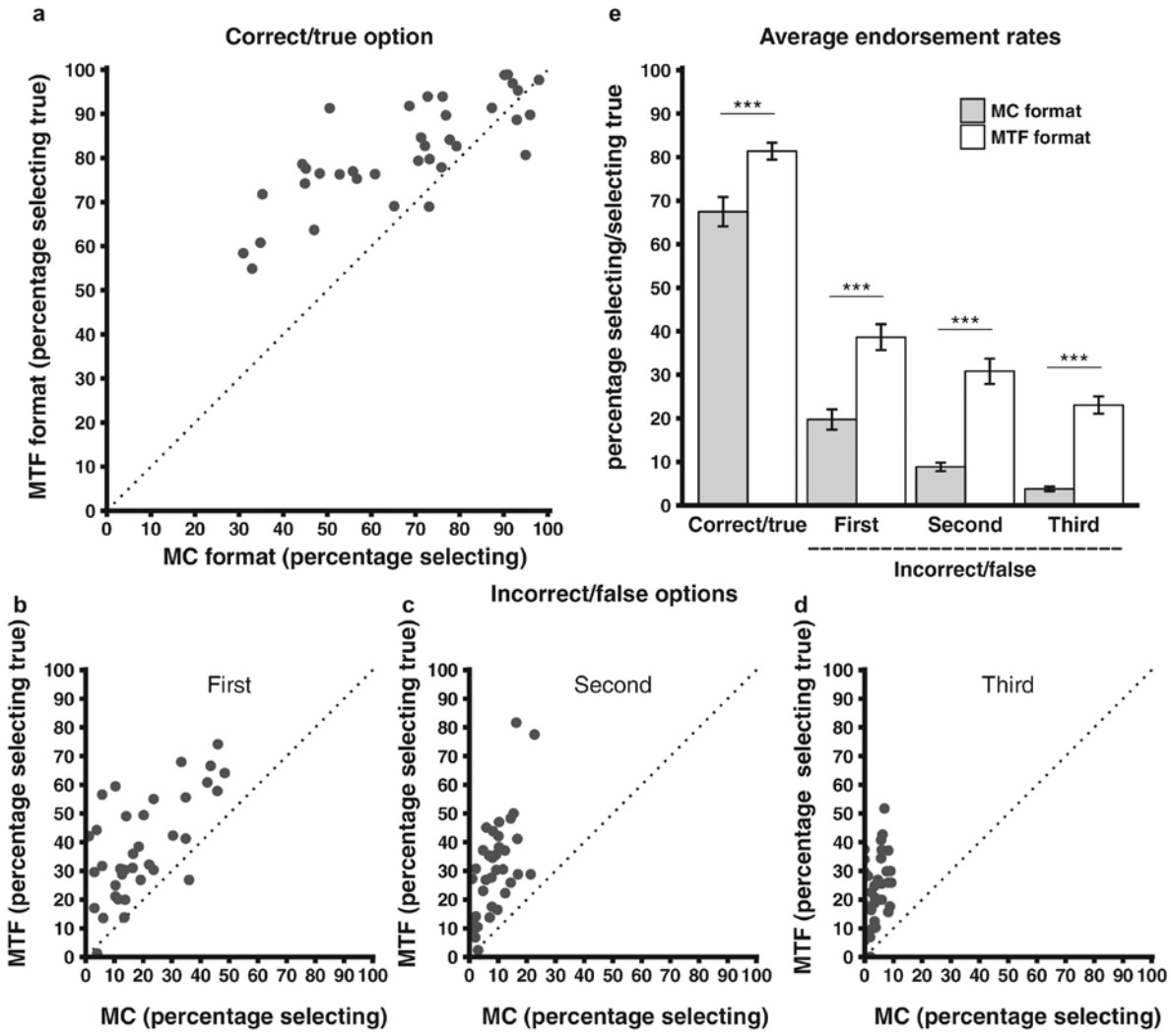

Figure 3. The endorsement rates for individual response options. (**a**–**d**) The dots represent endorsement rates for the response options in each MC×MTF question. The *x*-axis gives the percentage of students who selected the given option in the MC format, and the y-axis gives the percentage of students who marked true for the corresponding T–F statement. The gray dots in **(a)** show endorsement rates for the correct or true option, and the gray dots in **(b)**–**(d)** show endorsement rates for the incorrect or false options, grouped based being the first, second, or third most popular distractor in the MC format. The dotted lines represent the one-to-one lines in which response options would appear if they had the same endorsement rates in the different formats. **(e)** The average endorsement rates for the correct or true option and the three incorrect or false options. The gray bars represent the average percentage of students selecting the given MC option, and the white bars represent the average percentage of students marking true for the corresponding MTF statement. The error bars represent standard errors of the mean. ANOVA: main effect of format, $F(1, 245) = 120.8$, $p < .001$; post hoc paired Student's *t*-tests: correct, $t(35) = -6.33$; first distractor, $t(35) = -8.26$; second distractor, $t(35) = -9.17$, third distractor: $t(35) = -10.46$; all *** $p < .001$.

a higher percentage of students tended to mark *true* in the MTF format than the percentage of students who selected the corresponding MC option. A higher percentage of students also marked each incorrect or false option as *true* in the MTF format than the percentage of students who selected each corresponding MC distractor.

To estimate the magnitude of these differences, we calculated the average endorsement rate for each response option across all MC×MTF questions (**Figure 3** e; ANOVA: main effect of format, $F(1, 245) = 120.8$, $p < .001$). For the correct or true option, we found that roughly 67% of the students selected this option in the MC format, whereas 81% of the students marked *true* for this statement in the MTF format ($t(35) = -6.33$, $p < .001$). For the incorrect or false options, we found that these options were selected at relatively low rates in the MC format (20%, 9%, and 4% averages for the first, second, and third distractors, respectively), but these same statements were endorsed to a greater extent in the MTF format (39%, 31%, and 23% averages for the corresponding false statements; first distractor, $t(35) = -8.26$; second distractor, $t(35) = -9.17$, third distractor: $t(35) = -10.46$; all $p < .001$). Thus, compared with the MTF format, MC questions underestimated both the percentage of students who would have endorsed the correct option as well as the percentage of students who would have incorrectly endorsed each false option.

The effect of question format can be particularly striking when considering individual questions. For the sample question shown in Figure 1, 95% of the students correctly identified the correct option in the MC format, implying that most of the students had mastered the underlying concept that enzymes catalyze reactions by lowering the transition-state energy level. The MC format also provided little indication that the students struggled with the other distractors: only 5% of the students picked any of the other options. In the MTF format, we found that only 49% of the students answered all four T–F statements correctly, suggesting that many of the students still struggled with one or more concepts underlying reaction dynamics. This question along with additional examples in supplemental table S2 demonstrated how the two formats showed some correspondence in overall answer patterns, whereas the MC format obscured finer details in how the students would have separately responded to the various response options.

Multiple–choice and multiple–true–false formats differ in their question scores and overall scores

We further compared how scores differed for individual MC×MTF questions. We calculated the question score for each question (i.e., the percentage correct for MC and the average percentage correct of all four statements for MTF) and analyzed the association between scores in the MC and MTF formats (**Figure 4** a). There was a strong correlation between the percentage correct for questions in the MC and MTF formats (r = .79, p < .001). On average, MC question scores were five points lower than MTF question scores (MC: mean [M] = 67.5, standard error of the mean [SEM] = 3.4; MTF: M = 72.1, SEM = 1.8; $t(35) = -2.05$, p = .048), but differences in the scores for each question depended on the question difficulty, as was indicated by the position of the linear regression line relative to the one-to-one line. For questions of which roughly 75% of the students answered correctly in the MC format, the percentage correct in the MTF format was similar. For questions of which the percentage correct in the MC format was below 75%, the percentage correct in the MTF format tended to be higher (e.g., an MC question with 48% correct had

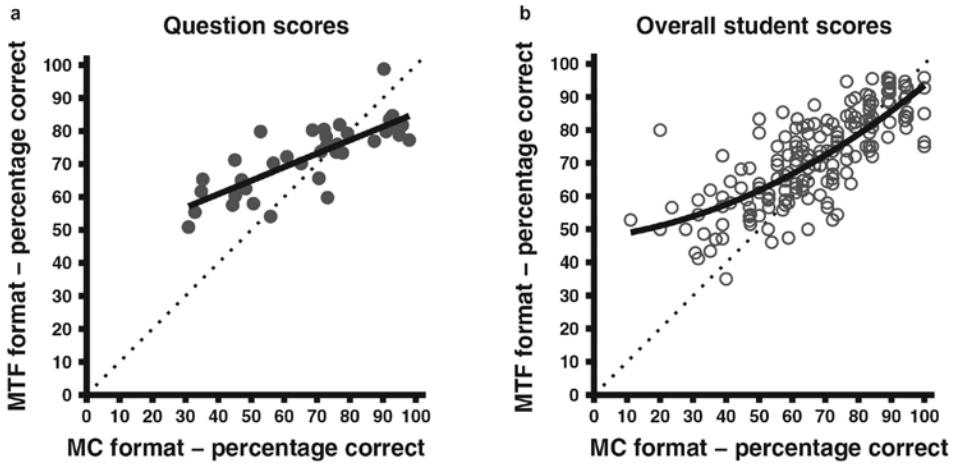

Figure 4. The individual question scores and overall student scores. **(a)** The gray dots represent the percentage correct for each MC×MTF question in either the MC or MTF format (r = .79, p < .001). The percentage correct for a MTF question equals the average percentage correct for each of the four T–F statements. The solid line represents the linear regression line. **(b)** The open gray dots represent the percentage correct across all experimental MC questions and MTF statements for each student (r = .75, p < .001). The solid line represents the quadratic regression line. The dotted lines in both panels represent the one-to-one lines in which questions would appear if they had the same percentage correct in the different formats.

an MTF percentage correct of 63%). Conversely, for questions of which the percentage correct in the MC format was above 75%, the percentage correct in the MTF format tended to be lower (e.g., an MC question with 93% correct had an MTF percentage correct of 83%).

To determine how the MC and MTF formats affected the overall student scores, we separately calculated the percentage correct across all experimental MC and MTF questions for each student (figure 4b). Again, there was a strong correlation between overall scores in the two formats ($r = .75$, $p < .001$), with the average student scores being five points lower in the MC format than in the MTF format (MC: M = 67.1, SEM = 1.3%; MTF: M = 71.8, SEM = 1.0%; $t(192) = -5.47$, $p < .001$). Similar to question scores, the effect of format on overall student performance depended on student ability levels. In this case, however, the relationship was best fit with a nonlinear regression line ($F(1) = 5.61$, $p = .02$ for linear versus nonlinear fit). By comparing the regression line with the one-to-one line, we saw that the low-performing students scored higher on MTF than on MC questions, whereas the high-performing students performed at more similar levels between the two formats.

Differences between the two formats are recapitulated on clicker questions

To determine whether the observed response patterns were robust to context, we compared MC and MTF-like clicker question responses from the follow-up experiment. We observed that the MC format again underestimated the endorsement rates for both correct and incorrect options and that the fractional relationship between the endorsement rates in the two formats was similar across contexts (supplemental figure S3; ANOVA: main effect of context, $F(1, 241) = 0.08$, $p = .78$, interaction between context and option, $F(4, 241) = 0.36$, $p = .84$). There was a significant effect of answer option, which indicated that the fractional relationship differed among the various options (main effect of option, $F(4, 241) = 18.17$, $p < .001$).

Discussion

When we consider the purposes of assessment, the MTF format has certain advantages over the MC format, particularly in the biological sciences, in which students often evaluate different aspects or explanations of a phenomenon. First, although the MC format implicitly rests on a

cognitive model in which student conceptions are coherent at the level of an individual question, the MTF format acknowledges that students typically hold mixed and partial understandings. Second, the MTF format provides a question score that includes partial credit and produces estimations of the extent to which students separately endorse each response option. Finally, the MTF format requires little additional writing or grading effort but can uncover a greater richness in student thinking to inform instructors as they make judgments about student learning.

Our findings suggest that the most important divergence in MC and MTF questions lies at the level of how the two formats convey the level of understanding of the various response options. In particular, our results show that the MC format systematically overestimates the percentage of students with full question mastery and that selection rates in the MC format differ significantly from corresponding evaluations of each statement in the MTF format. Therefore, instructors should keep in mind that correct MC responses do not equate with full question mastery and that students who switch from an incorrect option to the correct option after a period of instruction or peer discussion may still retain underlying misconceptions. Furthermore, instructors should use caution when attempting to use distractor selection rates on MC clicker questions, concept inventories, or other instruments to estimate the frequency of incorrect ideas. Our previous research has shown that students infrequently list incorrect conceptions in their answers to FR questions (Hubbard et al. 2017), so the current findings suggest that the MTF format has a unique capacity to diagnose the presence of latent misconceptions that go underestimated by both the MC and FR formats.

With respect to question scores, we found that for easy MC questions, the requirement to evaluate all options resulted in lower MTF scores. However, for difficult MC questions, partial credit and potential guessing enabled students to achieve higher MTF scores, so instructors should keep in mind that student MTF question scores may overestimate how readily students would be able to identify the correct answer in the MC format. Instructors can address this apparent discrepancy by analyzing individual MTF statement response rates to diagnose student understanding of the various concepts in the question. We also discovered that lower-performing students benefited the most from the MTF format, suggesting that the MTF format can selectively raise exam grades for students at risk of failing a course.

By targeting and revealing a more nuanced portrait of student thinking, the MTF format also has additional benefits related to how assessment supports the learning process (Black and Wiliam 2009). Assessment

can serve to communicate and represent learning expectations, and whereas the MC format conveys a sense that students can achieve success by recognizing the correct answer from a list of options, the MTF format establishes that learning involves evaluating both correct and incorrect ideas (Richardson 1992). Next, MTF questions provide specific information about areas of partial understanding, enabling students and instructors to adjust their practices to promote more complete learning. Finally, in cases in which assessment questions are used to facilitate peer discussion, such as with clicker questions, the MTF format potentially encourages students to discuss all the different options rather than attempting to find one correct answer.

In choosing a question format, instructors aim to gather the most valid and reliable portrait of student understanding available under the particular testing conditions (Crocker and Algina 2006). Many of the validity and reliability advantages of the MTF format stem from a student's ability to answer more MTF statements than MC questions in a given time period. Although the ratio varies based on the length of the question stem and the number of answer choices, students can answer roughly 2.6–3.4 MTF statements in the same time as 1 MC question with 4–5 response options (Frisbie and Sweeney 1982, Kreiter and Frisbie 1989). Thus, students take slightly longer to answer a full MTF question compared with a similar MC question, but each MTF question provides multiple pieces of information on student understanding, whereas an MC question only captures a single student response. This can enable coverage of a broader range of topics and improves an assessment's content validity (i.e., the degree to which an assessment samples across a given domain).

Multiple–true–false question scoring

When using the MTF format, instructors must choose how to score the questions. A main concern is that the relatively high guess rate will introduce noise that undermines the internal reliability of an instrument (i.e., consistency of responses across items). Thus, researchers have developed scoring rules to account for the guess rate by only giving students credit for a question if they exceed the guessing threshold or by applying a penalty for wrong answers. However, these various scoring rules reduce the amount of information in the score, introduce potential artifacts, and provide no benefit to test reliability (Gross 1982, Hsu et al. 1984, Tsai and Suen 1993). Among scoring rules, the rule in which students only earn credit for a question if they answer all four T–F statements correctly has the lowest guess rate, but by reducing all levels of partial understanding

to zero credit, this rule forfeits a significant amount of information on within-question variation and decreases test reliability (Tsai and Suen 1993). Thus, the partial-credit scoring rule used here is suitable for instructional purposes (Siddiqui et al. 2016). Furthermore, this partial-credit scoring rule is easy to calculate and consistent with the underlying premise that MTF questions capture mixed and partial conceptions.

Limitations and considerations for using multiple–true–false questions

Although the data presented here suggest that the MC and MTF formats differ in how they reveal student understanding, it is important to consider potential artifacts introduced by the MTF format. For example, early research indicated that students tend to mark *true* more often than *false* in the MTF format, and this difference is more pronounced when students are asked to evaluate each option as true or false than when directed to only indicate the true options (Cronbach 1941). However, these results contradict more recent data suggesting that students tend to leave alternatives blank on MTF assessments with directions to only indicate the true options (Pomplun and Omar 1997). We aimed to combat potential MTF response tendencies by giving students practice answering MTF questions in class and by including an even balance of questions with one, two, or three true statements. Thus, students would not have received a benefit from systematically guessing *true* or marking certain response patterns on MTF exam questions. Although some MTF response patterns may seem implausible to an expert, previous interview studies have found that students may not view contradictory answers as illogical, in some cases because they have misinterpreted the meaning of statements that include naïve ideas (Federer et al. 2013). Taken together, these limitations underscore the inherent difficulties in measuring student thinking and highlight the need for additional research to understand how faculty interpret and value student responses to closed-ended questions.

During the design process, instructors should consider their assessment objectives and choose formats that meet their needs while taking into account the available resources and administration constraints. Compared with MC questions, our data suggest that the MTF format provides a more complex picture of student thinking regarding the various options while requiring virtually no additional question writing or scoring efforts. There are two cases in which the MC format would still be appropriate. The first situation occurs when the answer options are mutually

exclusive, such that answering one option negates the other options. For example, the MC format would be appropriate for a genetics problem in which students predict the proportion of offspring with a particular genotype. The second situation occurs when the instructor wants students to make comparisons among response options. For example, instructors in clinical medicine might ask students to select the best treatment plan among several viable options (Chandratilake et al. 2011). Nonetheless, when the answer options contain different—even if closely related—conceptions, the MTF format provides the most direct way to assess student understanding of these ideas.

Acknowledgments — We thank Kathleen Brazeal, Tanya Brown, Allison Couch, and Mary Durham for providing feedback on project design and manuscript revisions. This research was supported by an internal award from the University of Nebraska–Lincoln and was classified as exempt from IRB review, Project 14314. Our source data will be uploaded to Dryad once the manuscript is published.

Supplemental material — Supplementary data are available (with subscription) at *BIOSCI* online.

References

Adams WK, Wieman CE. 2011. Development and validation of instruments to measure learning of expert-like thinking. International Journal of Science Education 33: 1289–1312. doi:10.1080/09500693.2010.512369

Angelo TA. 1998. Classroom assessment and research: An update on uses, approaches, and research findings. Jossey-Bass.

Black P, Wiliam D. 2009. Developing the theory of formative assessment. Educational Assessment Evaluation and Accountability 21: 5–31. doi:10.1007/s11092-008-9068-5

Bridgeman B, Morgan R. 1996. Success in college for students with discrepancies between performance on multiple–choice and essay tests. Journal of Educational Psychology 88: 333–340. doi:10.1037/0022-0663.88.2.333

Caldwell JE. 2007. Clickers in the large classroom: Current research and best-practice tips. CBE–Life Sciences Education 6: 9–20. doi:10.1187/cbe.06-12-0205

Chandratilake M, Davis M, Ponnamperuma G. 2011. Assessment of medical knowledge: The pros and cons of using true/false multiple–choice questions. National Medical Journal of India 24: 225–228.

Couch BA, Wood WB, Knight JK. 2015. The Molecular Biology Capstone Assessment: A concept assessment for upper-division molecular biology students. CBE–Life Sciences Education 14 (art. ar10). doi:10.1187/cbe.14-04-0071

Crocker L, Algina J. 2006. Introduction to Classical and Modern Test Theory. Wadsworth.

Cronbach LJ. 1941. An experimental comparison of the multiple–true–false and multiple multiple–choice tests. Journal of Educational Psychology 32: 533–543.

Dancy M, Henderson C. 2010. Pedagogical practices and instructional change of physics faculty. American Journal of Physics 78: 1056–1063. doi:10.1119/1.3446763

Ebert-May D, Derting TL, Hodder J, Momsen JL, Long TM, Jardeleza SE. 2011. What we say is not what we do: Effective evaluation of faculty professional development programs. BioScience 61: 550–558. doi:10.1525/ bio.2011.61.7.9

Ellis APJ, Ryan AM. 2003. Race and cognitive-ability test performance: The mediating effects of test preparation, test-taking strategy use and self-efficacy. Journal of Applied Social Psychology 33: 2607–2629. doi:10.1111/j.1559-1816.2003.tb02783.x

Federer MR, Nehm RH, Beggrow EP, Ha M, Opfer JE. 2013. Evaluation of a new multiple–true–false concept inventory for diagnosing mental models of natural selection. Paper presented at the National Association for Research in Science Teaching; 6–9 April 2013, Rio Grande, Puerto Rico.

Frey BB, Petersen S, Edwards LM, Pedrotti JT, Peyton V. 2005. Item-writing rules: Collective wisdom. Teaching and Teacher Education 21: 357–364.

Frisbie DA, Sweeney DC. 1982. The relative merits of multiple–true–false achievement tests. Journal of Educational Measurement 19: 29–35. doi:10.1111/j.1745-3984.1982.tb00112.x

Gross LJ. 1982. Scoring multiple–true/false tests some considerations. Evaluation and the Health Professions 5: 459–468. doi:10.1177/016327878200500407

Hsu T-C, Moss PA, Khampalikit C. 1984. The merits of multiple– answer items as evaluated by using six scoring formulas. Journal of Experimental Education 52: 152–158.

Hubbard JK, Potts MA, Couch BA. 2017. How question types reveal student thinking: An experimental comparison of multiple–true–false and free-response formats. CBE–Life Sciences Education 16 (art. ar26). doi:10.1187/ cbe.16-12-0339

Hurtado S, Eagan K, Pryor J, Whang H, Tran S. 2012. Undergraduate Teaching Faculty: The 2010–2011 Higher Education Research Institute (HERI) Faculty Survey. HERI, University of California Los Angeles.

Kim YH, Goetz ET. 1993. Strategic processing of test questions: The test marking responses of college students. Learning and Individual Differences 5: 211–218. doi:10.1016/1041-6080(93)90003-B

Kreiter CD, Frisbie DA. 1989. Effectiveness of multiple–true–false items. Applied Measurement in Education 2: 207–216.

Kubinger KD, Gottschall CH. 2007. Item difficulty of multiple–choice tests dependent on different item response formats: An experiment in fundamental research on psychological assessment. Psychological Science 49: 361–374.

Libarkin J. 2008. Concept Inventories in Higher Education Science. National Research Council.

Mazur E. 1996. Peer Instruction: A User's Manual. Addison-Wesley.

Nehm RH, Reilly L. 2007. Biology majors' knowledge and misconceptions of natural selection. BioScience 57: 263–272. doi:10.1641/B570311

Nehm RH, Schonfeld IS. 2008. Measuring knowledge of natural selection: A comparison of the CINS, an open-response instrument, and an oral interview. Journal of Research in Science Teaching 45: 1131–1160. doi:10.1002/tea.20251

Parker JM, Anderson CW, Heidemann M, Merrill J, Merritt B, Richmond G, Urban-Lurain M. 2012. Exploring undergraduates' understanding of photosynthesis using diagnostic question clusters. CBE–Life Sciences Education 11: 47–57. doi:10.1187/cbe.11-07-0054

Pomplun M, Omar H. 1997. Multiple–mark items: An alternative objective item format? Educational and Psychological Measurement 57: 949–962.

Price RM, Andrews TC, McElhinny TL, Mead LS, Abraham JK, Thanukos A, Perez KE. 2014. The Genetic Drift Inventory: A tool for measuring what advanced undergraduates have mastered about genetic drift. CBE– Life Sciences Education 13: 65–75. doi:10.1187/cbe.13-08-0159

Richardson R. 1992. The multiple–choice true/false question: What does it measure and what could it measure? Medical Teacher 14: 201–204. doi:10.3109/01421599209079488

Siddiqui NI, Bhavsar VH, Bhavsar AV, Bose S. 2016. Contemplation on marking scheme for Type X multiple–choice questions, and an illustration of a practically applicable scheme. Indian Journal of Pharmacology 48: 114–121.

Stanger-Hall KF. 2012. Multiple–choice exams: An obstacle for higher-level thinking in introductory science classes. CBE–Life Sciences Education 11: 294–306. doi:10.1187/cbe.11-11-0100

Stenlund T, Eklöf H, Lyrén P-E. 2017. Group differences in test-taking behaviour: An example from a high-stakes testing program. Assessment in Education Principles Policy and Practice 24: 4–20. doi:10.1080/0969594X.2016.1142935

Tsai F-J, Suen HK. 1993. A brief report on a comparison of six scoring methods for multiple–true–false items. Educational and Psychological Measurement 53: 399–404. doi:10.1177/0013164493053002008

Wilcox BR, Pollock SJ. 2014. Coupled multiple–response versus free-response conceptual assessment: An example from upper-division physics. Physical Review Physics Education Research 10 (art. 020124). doi:10.1103/PhysRevSTPER.10.020124 Wood W. 2004. Clickers: A teaching gimmick that works. Developmental Cell 7: 796–798. doi:10.1016/j.devcel.2004.11.004