# How Question Types Reveal Student Thinking: An Experimental Comparison of Multiple-True-False and Free-Response Formats

**Joanna K. Hubbard, Macy A. Potts, and Brian A. Couch***
School of Biological Sciences, University of Nebraska–Lincoln, Lincoln, NE 68588

## ABSTRACT

**Assessments represent an important component of undergraduate courses because they affect how students interact with course content and gauge student achievement of course objectives. To make decisions on assessment design, instructors must understand the affordances and limitations of available question formats. Here, we use a crossover experimental design to identify differences in how multiple-true-false (MTF) and free-response (FR) exam questions reveal student thinking regarding specific conceptions. We report that correct response rates correlate across the two formats but that a higher percentage of students provide correct responses for MTF questions. We find that MTF questions reveal a high prevalence of students with mixed (correct and incorrect) conceptions, while FR questions reveal a high prevalence of students with partial (correct and unclear) conceptions. These results suggest that MTF question prompts can direct students to address specific conceptions but obscure nuances in student thinking and may overestimate the frequency of particular conceptions. Conversely, FR questions provide a more authentic portrait of student thinking but may face limitations in their ability to diagnose specific, particularly incorrect, conceptions. We further discuss an intrinsic tension between question structure and diagnostic capacity and how instructors might use multiple formats or hybrid formats to overcome these obstacles.**

## INTRODUCTION

In response to national calls for transformations in science, technology, engineering, and mathematics (STEM) teaching (National Research Council [NRC], 1999; President's Council of Advisors on Science and Technology, 2012), undergraduate STEM instructors have incorporated more student-centered instruction as a means to improve student learning (DeAngelo *et al.*, 2009; Hurtado *et al.*, 2012; Eagan *et al.*, 2014). As instructors make changes to their course activities, they must also consider the manner in which their assessments reveal student understandings. Assessments represent a fundamental component of any course because they allow students to interact with material, provide instructors and students with information regarding student thinking, and produce scores reflecting student performance (NRC, 2003; Black and Wiliam, 2009; Brame and Biel, 2015). Assessments also play an integral role in facilitating iterative learning cycles and guiding teaching transformation efforts. Just as individual students can use assessments to refine their own understandings, instructors can use assessments to identify and address widely held misconceptions among students, evaluate specific learning activities, and make decisions about their instructional practices (Tanner and Allen, 2004). To maximize the impact of these instructional efforts, instructors must be able to critically interpret and apply results from their various assessments.

**Kathryn E. Perez,** *Monitoring Editor*
Submitted December 5, 2016; Revised February 1, 2017; Accepted February 27, 2017
CBE Life Sci Educ June 1, 2017 16:ar26
DOI:10.1187/cbe.16-12-0339
*Address correspondence to: Brian A. Couch (bcouch2@unl.edu).

Assessments employ a variety of question formats that generally fall into the categories of closed-ended and open-ended (Martinez, 1999; Goubeaud, 2010; Kuechler and Simkin, 2010). Closed-ended formats include multiple-choice, true–false, matching, card sorting, and ordering; while open-ended formats include short answer, fill-in-the-blank, free-response, concept mapping, and diagramming. Each format has different affordances and limitations, so instructors face the task of weighing the ability of different formats to assess student thinking against the time and resources required to develop, administer, and score the assessments. The multiple-choice (MC) and free-response (FR) formats, in particular, have been used and researched in many undergraduate classrooms. MC questions consist of a question stem and a series of response options, with one correct option among several incorrect options, or distractors. At the other end of the spectrum, FR questions solicit written answers to open-ended prompts. MC questions are widely used in introductory STEM courses (DeAngelo *et al.*, 2009; Hurtado *et al.*, 2012; Stanger-Hall, 2012) and standardized tests, because they are easily administered and can be machine graded, while the FR format requires more grading effort, may suffer from grading inconsistencies, and depends on student writing skills (Case and Swanson, 1993; Martinez, 1999). Nevertheless, many argue that FR questions more authentically capture student knowledge because students construct their own answers rather than selecting an answer from among several possible options (Milton, 1979; Birenbaum and Tatsuoka, 1987; Martinez, 1999; Haudek *et al.*, 2012). Some researchers have argued that carefully constructed MC questions can assess higher-level thinking in a way similar to FR questions (Simkin and Kuechler, 2005; Kuechler and Simkin, 2010), but the time and effort needed to develop such questions may negate the time saved by machine grading.

Beyond providing a numeric standing on an overall scale, assessments play a critical role in diagnosing the degree to which students hold correct and incorrect understandings of course concepts. Students may hold partial conceptions characterized by a lack of knowledge or mixed conceptions, wherein they simultaneously hold correct and incorrect ideas regarding a particular concept. Mixed conceptions can be seen with the concept of natural selection, for which students may have a correct understanding regarding the outcome (i.e., beneficial traits become more predominant in the population) but incorrectly identify desire or need as a mechanism that produces new alleles or traits in a population (Nehm and Reilly, 2007; Nehm and Schonfeld, 2008). Previous studies using concept assessments have revealed that partial and mixed (rather than complete) understandings are widespread among undergraduate biology students, even for graduating seniors (Couch *et al.*, 2015).

The utility of an assessment can be partly judged by the degree to which it can detect the presence of correct and incorrect understandings. MC questions suffer the limitation that student selection of a particular answer provides little indication regarding their thinking on the remaining options (Parker *et al.*, 2012). Indeed, we have found that roughly half of the students who select the correct MC option would endorse an additional incorrect option if given the opportunity (B.A.C., unpublished data). These data suggest that selection of the correct MC answer should not be equated with complete understanding of all the answer options. Conversely, FR questions have the potential to fully reveal student understandings, because student answers can include any combination of correct and incorrect ideas. However, despite this potential, student answers may not provide sufficient information to confirm the presence of such conceptions. For example, students may avoid writing about areas in which they have uncertainty, instead choosing to elaborate on areas of greater confidence. Furthermore, students may misinterpret the FR prompt and provide a response that only tangentially relates to the target answer. In each of these cases, the FR answer provides insufficient information to fully diagnose the extent to which students hold specific ideas.

Multiple-true-false (MTF) questions represent a promising alternative to the MC and FR formats (Parker *et al.*, 2012). Much like the MC format, MTF questions consist of a question stem followed by several answer options. With MTF questions, however, students must select true or false for each answer option, rather than identifying the best answer. The MTF format functions similarly to MC questions in which students "select all that apply," except that the MTF format requires affirmative marking of both true and false statements. As a consequence of their closed-ended structure, MTF questions maintain the ease of machine grading and avoid the potential issues related to scoring consistency and student writing ability associated with open-ended formats. By having students evaluate different answer options, the MTF format also allows an instructor to probe student agreement with specific correct and incorrect ideas. While MTF questions may not collect the full range of ideas produced in response to an FR question, a comparison of the FR format with a related multiple-select format found that 74% of FR answers contained one or more elements aligned with a closed-ended answer option (Wilcox and Pollock, 2014). Thus, while the MTF format cannot recapitulate all aspects of FR questions, it holds promise as a machine-gradable mechanism for determining the degree to which students hold correct and incorrect ideas.

Despite the potential of MTF questions to assess student thinking with minimal grading resources, few studies have directly compared the MTF and FR formats. Here, we report results from an experimental comparison to understand differences in how MTF and FR questions reveal student thinking. To illustrate these differences, we focus on overall response patterns, rather than elaborating on the specific content of student conceptions. In comparing student responses to MTF and FR exam questions, we sought to answer several research questions:

- How do student responses to MTF statements compare with the rates at which the corresponding correct and incorrect conceptions are included in FR answers?
- To what extent can FR questions compel students to provide answers that address the ideas represented in MTF statements?
- How do responses to MTF and FR questions relate to a student's overall exam performance?
- To what extent are partial and mixed conceptions inferred from MTF and FR answers based on information included or not included in a student's answer?
- How do student scores on MTF and FR questions compare?

In answering these questions, we sought to better understand how these two question formats function, so that practitioners might better understand the inherent tendencies of each format and improve their interpretation and use of assessment results.

## METHODS

### Experimental Rationale and Context

We conducted this study in two sections of an introductory biology course both taught by one of the authors (B.A.C.) at the University of Nebraska–Lincoln. This is the first course in a two-semester introductory series that serves as the gateway for a diverse array of life sciences majors. The course was divided into four units covering discrete content areas (i.e., macromolecules and cells, cellular respiration and photosynthesis, cell cycle and genetics, and inheritance and gene expression). The course employed a variety of pedagogical strategies, including pre-class assignments, in-class clicker questions and group activities, and post-class homework quizzes. Importantly, students were exposed to MTF and FR question formats throughout the term through clicker questions and postclass homework quizzes. Students answered 43 clicker questions posed as either MC or MTF with a single correct statement, 16 MTFs with two correct statements, and 17 MTFs with three correct statements (76 clicker questions in total). For nearly all questions, students voted individually, discussed responses with their peers, and voted again, after which the instructor or students provided explanations regarding the correct and incorrect answers (Crouch and Mazur, 2001; Vickrey *et al.*, 2015). Students typically answered two FR questions on each weekly homework assignment. For these questions, the instructor posted an example of a correct answer and a scoring rubric on the course Web page, but students did not receive personalized feedback. Consequently, students had practice with both question formats before the experimental treatment that took place during the course's four unit exams. A total of 468 students were enrolled in the two sections of the course, and 405 students agreed to have their course data released for research purposes, representing 87% of total enrollment (see Table 1 for demographic information).

**TABLE 1. Student demographics**

| Demographic categories | *n* | % |
|---|---|---|
| Gender | | |
| Female | 238 | 59 |
| Male | 167 | 41 |
| Race/ethnicity | | |
| Non-URM (white, Asian, international) | 352 | 87 |
| Underrepresented minority (URM) | 45 | 11 |
| Generation status | | |
| Continuing generation | 281 | 69 |
| First generation | 124 | 31 |
| High school location | | |
| Urban or other | 267 | 66 |
| Rural | 135 | 33 |
| Major | | |
| Life sciences | 253 | 62 |
| Other STEM | 20 | 5 |
| Non-STEM | 82 | 20 |
| Undeclared | 43 | 11 |
| Class rank | | |
| First-year | 189 | 47 |
| Sophomore | 103 | 25 |
| Junior | 58 | 14 |
| Senior | 23 | 6 |
| Other | 32 | 8 |

### Identification of Student Conceptions

Our experimental strategy built on a rationale that has served as a basis for developing questions on concept assessments (Adams and Wieman, 2011). When developing concept assessments, developers often begin by administering an open-ended question to students in the context of a semistructured interview or written assignment to identify the various conceptions that students hold. Student-generated correct and incorrect conceptions are then translated into MC or MTF answer options, and these closed-ended questions are administered as a means to estimate the extent to which these various ideas are held by students (e.g., Bretz and Linenberger 2012; Couch *et al.*, 2015; Newman *et al.*, 2016). Thus, in many cases, a single question can be articulated in a manner suitable for either a closed-ended or open-ended format. In a closed-ended format, student thinking is inferred based on selection of the various response options included by the instructor. In an open-ended format, student thinking is inferred based on the presence and correctness of these same conceptions in addition to other conceptions that students may include in their responses.

To develop questions that could appear in either an MTF or FR format, we began by analyzing responses to open-ended questions administered to students as part of homework quizzes and exams during a previous term. These questions covered a range of content representative of the breadth of each unit. For each question, 100 responses were open-coded by two independent coders to identify the various conceptions students included in their answers, and the frequency of each conception was tallied across student responses. Once a concept appeared in more than 10 student responses, it was marked as a "common conception," and additional occurrences were not counted. The coders then compared and discussed their list of concepts and created a compiled list of distinct conceptions along with wording variations for each conception found in student responses.

### Exam Construction and Administration

For each unit exam, we began by generating a question bank consisting of eight control MCs, 10 control MTFs, and eight experimental MTF × FR questions that could appear in either the MTF or FR format (Figure 1A). All MC and MTF questions included four different response options or statements. In writing questions, we took care to consider the length, structure, and wording of question stems and response options, in accordance with established item-writing guidelines (Frey *et al.*, 2005). The experimental MTF × FR questions were developed based on the open-ended questions from the previous year, but modified such that students with access to materials from the previous year could not simply memorize responses. For most MTF × FR questions, the MTF and FR question stems appearing

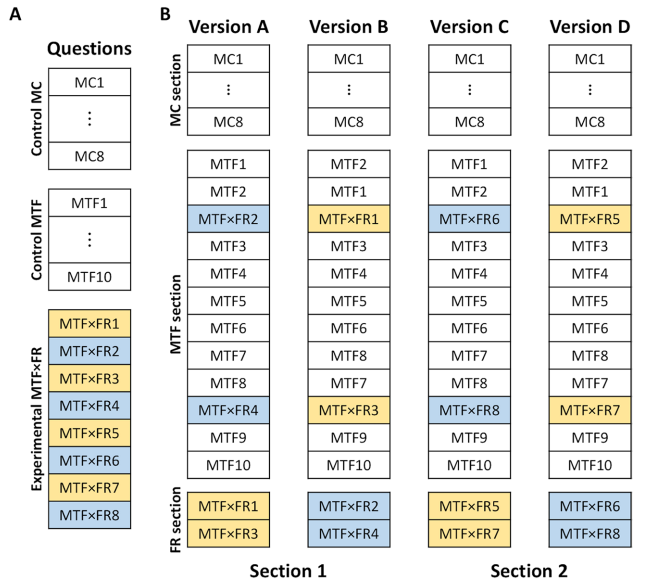



**FIGURE 1. Overview of crossover experimental design. (A) Before each exam, a question bank was generated containing eight control MC questions, 10 control MTF questions, and eight MTF × FR questions. (B) These questions were used to make four exam versions, with two versions used in each course section. Control questions were identical on all four versions. Two of the MTF × FR questions appeared in the MTF format on version A and in the FR format on version B. Conversely, two additional MTF × FR questions appeared in the FR format on version A and in the MTF format on version B. The remaining four questions followed the same pattern for versions C and D used in the second course section.**

on the exam were nearly identical in wording, with minor adjustments to sentence syntax. For the MTF form, six of the most common conceptions from previous student responses were adapted into true–false (T/F) statements. These were narrowed down to the final four based on how well they fit with the question prompt and the desired balance of true and false statements. In a few cases, FR question stems included slight modifications or additional wording to provide clarity or scaffolding to direct student responses. Two authors (J.K.H. and B.A.C.) discussed and refined each question before finalizing each exam (see Supplemental Material 1 for an example of the question development process).

Questions from the question bank were used to make four exam versions, two per course section (Figure 1B). Control MC and MTF questions appeared with identical wording across all versions. Two of the MTF × FR questions appeared in the MTF form on version A and in the FR form on version B. Two additional experimental questions appeared in the reciprocal arrangement, as FR questions on version A and as MTF questions on version B. The other four experimental questions followed the same pattern across versions C and D. Finally, the order of adjacent control questions was inverted in some cases to further minimize potential copying between students, but general topic flow was maintained for each exam version so that the question order mirrored the order of topics covered during the previous unit.

Each exam version ultimately contained eight MC questions, 12 MTF questions, and two FR questions used as data for the current study.[1] 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). The two versions (A and B in section 1 and C and D in section 2) were distributed to students in a semirandom manner on exam day, such that exam versions alternated across auditorium rows. Across the four unit exams, this crossover design yielded a total of 32 experimental MTF × FR questions and 128 associated statements/conceptions that were answered by half of the students in a course section in the MTF format and the other half of the students from the same course section in the FR format. Students had 50 min to complete the first three unit exams and 120 min to complete the fourth unit exam along with a cumulative final. The cumulative final did not contain any questions included in this study. Students recorded their answers to closed-ended questions on Scantron sheets that were scored by the institutional testing center. Students wrote their FR answers on pages in the test booklet that were initially scored for an exam grade and later coded for the experimental comparison. The total number of students taking each exam version ranged from 80 to 114 students (Table 2).

## Data Processing

A coding rubric was created for each FR question based on the four statements included in the MTF format. The goal of the coding rubric was to score FR answers based on whether a student would likely have provided a correct or incorrect answer for each corresponding MTF statement. An "unclear" code was applied when the coder could not determine how a student would have responded to the corresponding MTF statement, either because the FR answer was ambiguous or because the student did not address the concept. FR answers that were entirely blank were coded as unclear for all four statements; on average, only 2% of answers were entirely blank per question, ranging from 0% to 8%.

We have adopted specific conventions to describe MTF and FR structures. We use the term "question" to refer to an entire question (i.e., an MTF prompt with four associated T/F statements or an FR prompt). We use the terms "statement" and "conception" to refer to an individual T/F statement or the corresponding FR conception, respectively. Finally, we use "answer" to describe an answer to an entire question (i.e., all four MTF statement selections or the full FR explanation) and "response" to refer to a response to an individual T/F statement or FR conception.

A minimum of 30% of the answers for each FR question were cocoded by two raters. Raters began by cocoding a batch of 15 FR answers. If interrater agreement exceeded 80% for all four statements, then the two raters cocoded an additional batch. If the raters achieved 80% agreement for the second batch, then one rater coded the remaining FR answers. If interrater agreement fell below 80% at any point, then the raters would come to consensus and cocode additional answers until agreement was achieved on two consecutive

[1]An additional two MC and two MTF questions were also included on each exam but were not included in the current data set, because they were part of a different study.

TABLE 2. Student performance on control questions across different exam versions[a]

| | Exam 1 | | | | Exam 2 | | | | Exam 3 | | | | Exam 4 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Version[b] | A | B | C | D | A | B | C | D | A | B | C | D | A | B | C | D |
| *n* | 110 | 114 | 88 | 87 | 114 | 107 | 92 | 82 | 109 | 105 | 85 | 84 | 108 | 107 | 80 | 82 |
| % correct | 73.4 | 75.6 | 76.1 | 75.6 | 75.2 | 76.9 | 78.8 | 74.8 | 77.0 | 74.5 | 75.2 | 77.1 | 67.3 | 67.8 | 63.7 | 69.5 |
| SD | 14.0 | 13.3 | 12.6 | 13.9 | 16.0 | 14.7 | 14.4 | 13.9 | 14.0 | 14.8 | 14.2 | 13.8 | 17.4 | 16.6 | 16.5 | 16.8 |
| ANOVA | $F(3, 395) = 0.83$, $p = 0.48$ | | | | $F(3, 391) = 1.44$, $p = 0.23$ | | | | $F(3, 379) = 0.80$, $p = 0.50$ | | | | $F(3, 373) = 1.71$, $p = 0.17$ | | | |

[a]Student performance did not differ across exam versions for any of the four unit exams.
[b]Versions A and B were given in section 1 and versions C and D were given in section 2 for each of the four unit exams.

batches (Turpen and Finkelstein, 2009). Following individual coding, an additional 10 FR answers were cocoded to confirm interrater reliability. Average interrater agreement was 91% across all conceptions.

### Statistical Analyses

Exam versions were distributed semirandomly for each exam, such that the group of students that took version A on the first exam was somewhat different from the group of students that took version A on subsequent exams. Therefore, we used four one-way analyses of variance (ANOVAs; one for each exam) to compare student performance on control questions that appeared identically across all four versions of each exam. For control question, students were given 1 point for each MC question answered correctly and 0.25 points for each MTF statement answered correctly.

To compare MTF × FR responses, we calculated the correlation between the two formats for the percentage of students with correct responses and the percentage of students with incorrect responses. We also used a Mantel-Haenszel test to determine whether question format impacted the relative proportion of correct and incorrect responses between formats. This test determines whether two dichotomous variables (i.e., question format and response) are independent, while accounting for repeated measurements across questions.

We also determined the relationship between question scores for the two formats using two different scoring rules. In the first comparison, MTF and FR questions were both scored with a "partial" scoring model in which students received credit for each statement that they addressed correctly. In the second comparison, MTF questions were scored with an "all-or-nothing" scoring rule in which students only received credit for an entire question if they addressed all associated statements correctly, and these MTF scores were compared with FR partial-credit scores. Unclear responses in the FR format were counted as incorrect for all question score calculations. Correlation analyses were used to compare question score relationships. All analyses were completed in R, version 3.2.4 (R Core Team, 2016).

## RESULTS

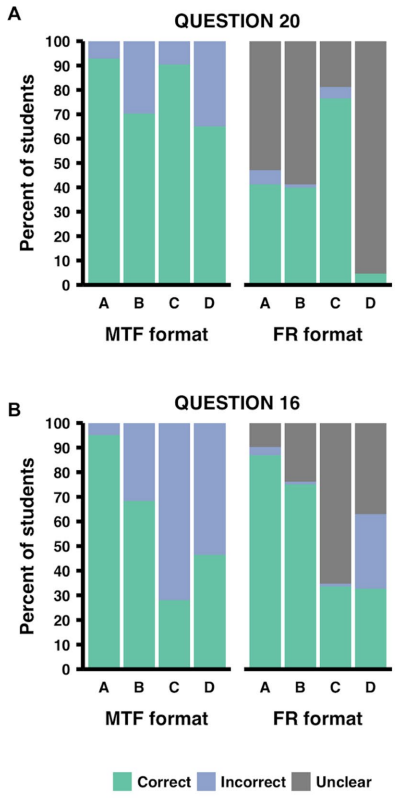
We first compared performance on control questions across the four exam versions to determine whether the versions were taken by comparable groups of students. There were no significant differences across versions for performance on control questions for each exam (Table 2 and Supplemental Figure 1; exam 1: $F(3, 395) = 0.83$, $p = 0.48$; exam 2: $F(3, 391) = 1.44$, $p = 0.23$; exam 3: $F(3, 379) = 0.80$, $p = 0.50$; exam 4: $F(3, 373) = 1.71$, $p = 0.17$). These data establish that the four exam versions were taken by equivalently performing groups of students.

### Example Questions Reveal Common MTF and FR Answer Tendencies

We next analyzed the experimental MTF × FR questions to understand how student responses compared between the two formats. While the response patterns varied across questions, we highlight two comparisons in Figure 2 that illustrate common features present in many questions (results for all 32 questions are provided in Supplemental Figures 2–5).

Question 20 asked students to describe how meiosis, gamete formation, and fertilization can increase genetic diversity within a population (Figure 2A). In the MTF format, students readily identified that recombination can generate chromosomal diversity (statement A, 93% correct), while fewer students recognized independent assortment of homologous chromosomes during meiosis as a mechanism that increases genetic diversity (statement B, 70% correct). For the two statements related to fertilization, most students recognized that fertilization produces new combinations of alleles in offspring (statement C, 91% correct), while fewer students answered correctly regarding the misconception that fertilization adds new alleles to a population (statement D, 64% correct). In the FR format, roughly half of students listed recombination or independent assortment as contributing to genetic diversity (concept A, 41% correct and concept B, 40% correct, respectively). Students prominently cited that fertilization produced new allele combinations (concept C, 77% correct), but very few students confirmed or refuted the misconception regarding fertilization adding new alleles to the population (concept D, 5% correct). In many cases, the FR answer did not provide sufficient information to infer how students would have responded to the corresponding MTF statement, and students only rarely provided information suggesting that they would have responded to the corresponding MTF question incorrectly.

Question 16 asked students to describe the flow of matter and energy within a closed system containing a plant and a few crickets (Figure 2B). In the MTF format, nearly all students recognized that photosynthesis in the plant provided the crickets with oxygen (statement A, 95% correct), while fewer students agreed that photosynthesis in the plant also served as a source of chemical energy for the crickets (statement B, 68% correct). Students performed poorly on the statement addressing the misconception that plants do not perform cellular respiration (statement C, 28% correct). Students also struggled with the notion that the system will be sustainable when the rates of photosynthesis and cellular respiration are in balance (statement D, 46%




**MTF:** Which of the following describe how the events of meiosis, gamete formation, and fertilization increase genetic variation within a population?

A. **T**/F recombination, or crossing over, creates novel arrangements of alleles on a chromosome
B. **T**/F homologous chromosomes assort independently during meiosis
C. **T**/F fertilization results in new combinations of alleles in offspring
D. T/**F** fertilization adds new alleles to the population

**FR:** Describe how the events of meiosis, gamete formation and fertilization increase genetic variation within a population.

**MTF:** A large plant is placed into a sealed glass chamber containing a few crickets that feed on the plant and an ample amount of water. The chamber is illuminated continuously by a light source, allowing the plant to grow. Which of the following describe the flow of matter and energy that occurs in the enclosed system?

A. **T**/F photosynthesis by the plant provides the crickets with a source of oxygen
B. **T**/F photosynthesis by the plant provides the crickets with a source of chemical energy
C. T/**F** without crickets, there would be no mechanism to release $CO_2$ into the atmosphere of the system
D. T/**F** the system will be sustainable if the rate of photosynthesis is greater than the rate of all cellular respiration

**FR:** A large plant is placed into a sealed glass chamber containing a few crickets that feed on the plant and an ample amount of water. The chamber is illuminated continuously by a light source, allowing the plant to grow. Describe the flow of matter and energy that occurs in the enclosed system. Under what conditions will this system be sustainable for a long period of time?

**FIGURE 2. Examples of student responses to two MTF × FR questions. Bars represent the proportion of students providing correct, incorrect, or unclear responses for questions in the MTF or FR formats. Data are shown for (A) question 20 and (B) question 16. Wording of the question in both formats is provided, and correct MTF responses are shown in bold and underlined. The MTF statements have been reordered from the exam for the purpose of presentation. Comparisons of student responses for all questions are provided in Supplemental Figures 2–5.**

correct). In the FR format, nearly all students indicated that plants release oxygen that is consumed by the crickets (concept A, 87% correct), and many students indicated that the crickets obtain chemical energy by feeding on the plant (concept B, 75% correct). While some students mentioned that plants perform cellular respiration or release carbon dioxide, a large fraction did not provide sufficient information to determine their thinking on this concept (concept C, 65% unclear). For the last concept (D), an additional line was added to the FR question prompt to direct students to describe the conditions under which the system would be sustainable. This prompt guided many students to provide either correct (33%) or incorrect (30%) responses related to the balance of photosynthesis and respiration rates, while a substantial fraction of students still provided unclear responses (37%).

### Correct, but Not Incorrect, Response Rates Correlate between MTF and FR Questions

To understand how statement response rates compare across all questions, we examined the correlation between the percentage of students who provided correct and incorrect responses for both formats. At the statement level, the percentage of students providing a correct response in the MTF format correlated significantly with the percentage of students providing a correct response in the FR format (Figure 3A; $r = 0.46$, $t_{126} = 5.80$, $p < 0.001$). Students rarely performed better in the FR format, as most of the data points fell below the one-to-one line. Alternatively, we did not detect a significant correlation between the percentage of students providing incorrect responses in the MTF and FR formats (Figure 3B; $r = 0.15$, $t_{126} = 1.74$, $p = 0.08$). Students rarely wrote incorrect responses in the FR format, even when a notable percentage of students selected incorrect responses in the MTF format.

### The Proportion of Correct to Incorrect Responses Differs between MTF and FR Questions

To understand broader patterns in MTF and FR answers, we next compared the average proportion of students providing correct and incorrect responses in each format across all the experimental questions (Figure 4A). In the MTF format, students responded correctly 72% of the time and incorrectly 28% of the time. In the FR format, students gave responses that included correct and incorrect conceptions at lower rates, with 38% of students providing correct conceptions and 7% listing incorrect conceptions, and the *relative* proportion of correct to incorrect responses was also influenced by question format (Mantel-Haenszel $\chi^2 = 205.8$, $p < 0.001$). In the FR format, ~55% of students provided unclear responses that did not adequately address the conceptions given in the MTF statements. This large percentage of unclear responses explains the lower combined percentage

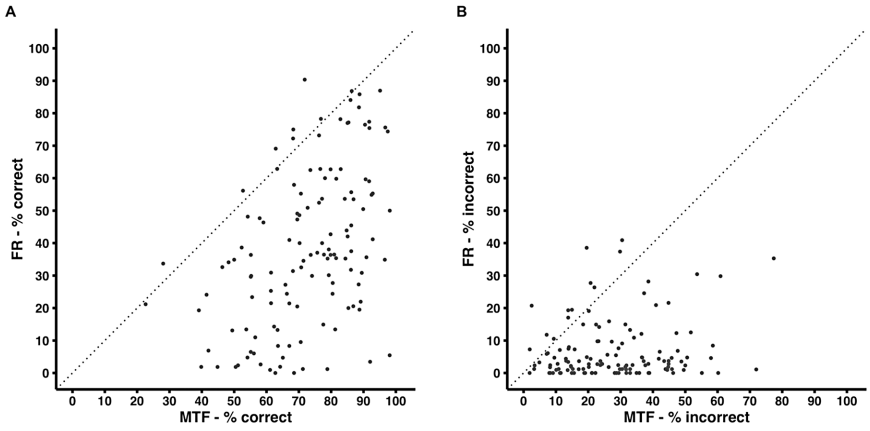



FIGURE 3. Scatter plots showing the correlation between the percentage of (A) correct responses and (B) incorrect responses for questions in the MTF and FR formats. Points represent the 128 statements/conceptions. Dotted lines represent the one-to-one line where data points would fall if they were equivalent in the two formats.

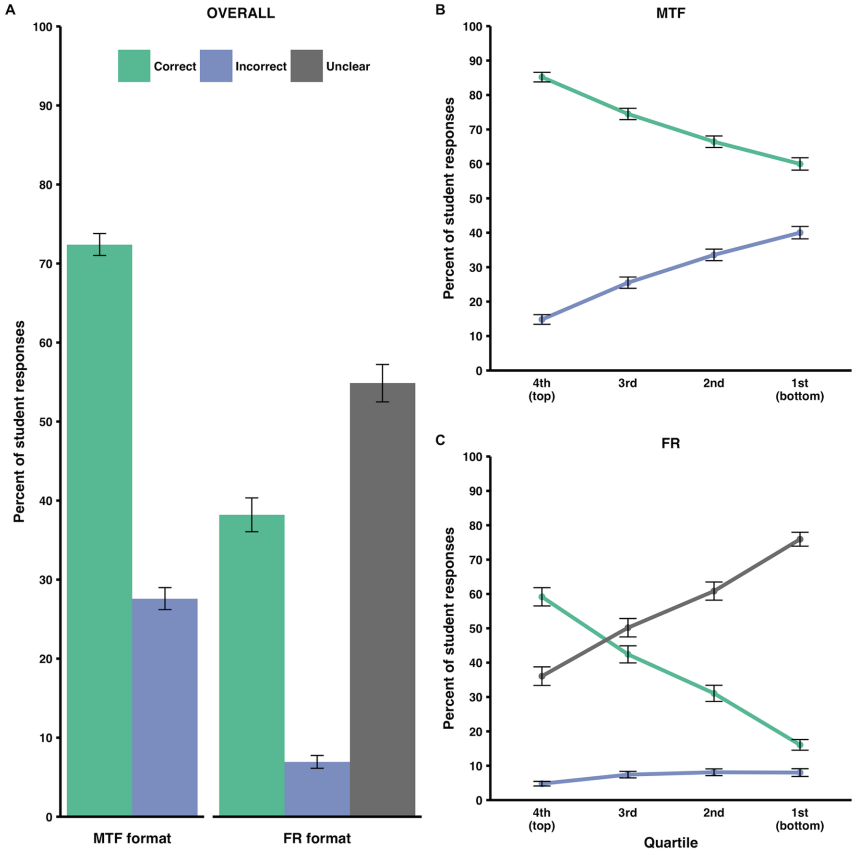



FIGURE 4. Average proportion of students providing correct, incorrect, or unclear answers for questions in the MTF or FR formats. (A) Bars represent proportion of student responses for all students. (B and C) Points represent the proportion of student responses for students in each quartile based on overall exam performance for the entire term; the fourth quartile contains students with the highest 25% of overall course grades, and the first quartile contains students with the lowest 25% of overall course grades. Error bars represent standard errors across statements/concepts ($n = 128$ statements).

of correct and incorrect responses compared with the MTF format.

### Lower-Performing Students Are Characterized by More Unclear FR Answers

We wanted to further explore how MTF and FR answers differed among high- and low-performing students, so we compared response rates across students divided into quartiles based on overall exam performance. The fourth (top) quartile contained students with the highest 25% of overall exam scores, and the first (bottom) quartile contained students with the lowest 25% of overall exam scores. For the MTF format, the percentage of students responding correctly steadily decreased across quartiles, while the percentage of students responding incorrectly increased proportionately, as expected (Figure 4B). For the FR format, the percentage of correct conceptions also decreased across quartiles, but the number of incorrect conceptions remained low across all four quartiles (Figure 4C). Instead, students with overall scores in the bottom quartiles were characterized by an increase in unclear responses.

### MTF and FR Questions Infer Mixed and Partial Conceptions to Different Extents

We next considered the answers students gave for full questions to further understand how each question revealed student thinking (Figure 5). For MTF questions, 33% of students gave fully correct answers in which they answered all four T/F statements correctly, 66% revealed mixed conceptions characterized by some combination of correct and incorrect responses, and only 1% answered all four statements incorrectly. FR questions produced a broader array of patterns based on the combination of correct, incorrect, and unclear conceptions included in students' answers. Very few students (9%, Figure 5, teal segment) gave fully correct answers in which they correctly addressed all four conceptions, and almost no students (<1%, Figure 5, blue segment) incorrectly addressed all four conceptions. The majority of FR answers consisted of a combination of correct, incorrect, and unclear responses. Among these answers, a relatively small number could be classified as mixed conceptions consisting of correct and incorrect conceptions (13%, Figure 5, salmon and green segments, respectively). Conversely, a substantial number of

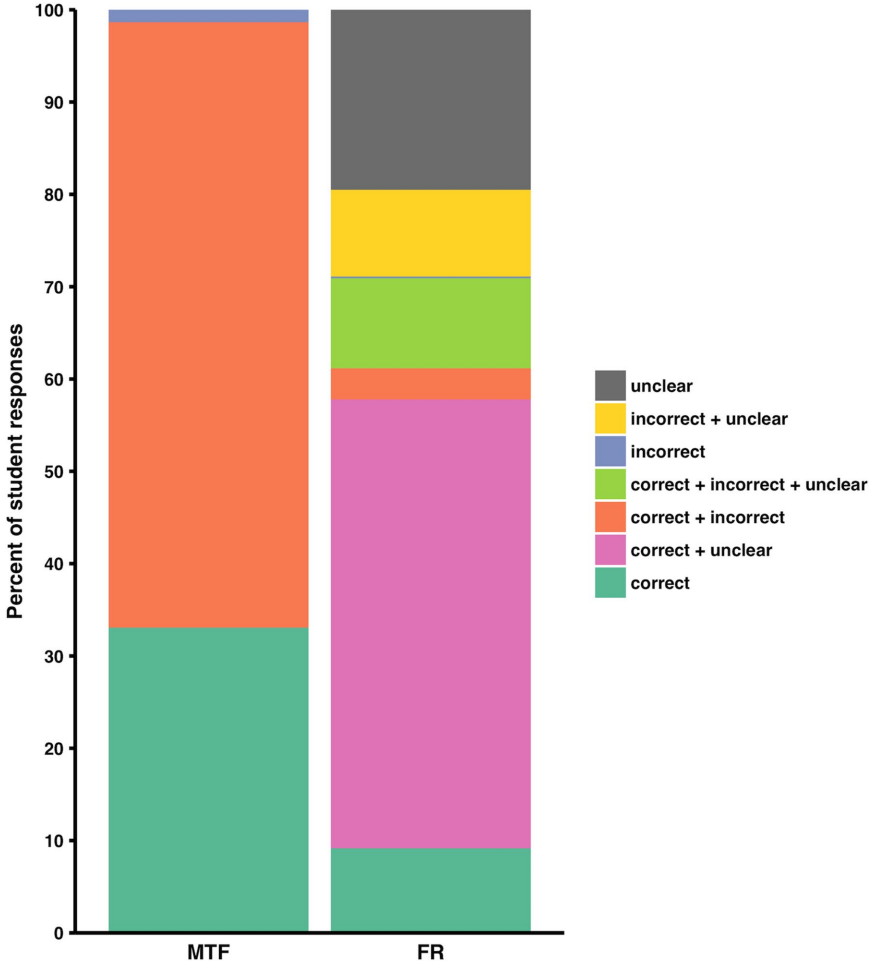



**FIGURE 5. Distribution of various answer types at the question level. Bars represent various combinations of correct, incorrect, and unclear responses for full questions in the MTF and FR formats, according to the legend shown at right (MTF: *n* = 3108 question responses; FR: *n* = 3114 question responses).**

answers could be characterized as partial understandings, because they consisted of correct and unclear conceptions and, in some cases, a combination of mixed and unclear conceptions (58%, Figure 5, pink and green segments, respectively). Finally, a subset of students provided answers that had no correct elements and thus could not be considered mixed or partial for the given conceptions (29%, Figure 5, yellow and gray segments, respectively). Thus, while putative mixed conceptions predominated for MTF questions, FR answers were more commonly characterized by partial conceptions, with true mixed conceptions being much less common.

### MTF and FR Scores Correlate

We also compared the scores that students would have received in each format.[2] In the first comparison, answers in both formats were scored using a partial-credit scoring model where students received partial credit for each correct statement and no credit for any incorrect or unclear statements (Figure 6A). Under this scoring model, MTF scores had an average of 72%, while FR scores had a much lower average of 38%. Partial-credit scores between the two formats were significantly correlated ($r = 0.516$, $p = 0.003$). In the second case, answers to MTF questions were scored using an all-or-nothing scoring model wherein students only received credit for an entire question if they answered all four associated T/F statements correctly (Figure 6B). A comparison of the MTF all-or-nothing scores to FR partial-credit scores was made based on previous findings that these scores can yield similar results (Kubinger and Gottschall, 2010). MTF all-or-nothing scores averaged 33%, which is closer to the 38% average for FR questions. MTF all-or-nothing scores also correlated significantly with FR scores ($r = 0.504$, $p = 0.003$). Thus, while both MTF scoring models produce similar correlations with FR questions, MTF all-or-nothing scoring yielded absolute scores that better approximated FR scores than MTF partial scoring.

## DISCUSSION

Instructors make many choices when designing their courses, including decisions related to assessing student understanding of course content. The format, design, and implementation of these assessments represent important decisions, because they can impact how students interact with course material and what information instructors can gain regarding student learning (NRC, 2003; Stanger-Hall, 2012; Momsen *et al.*, 2013; Brame and Biel, 2015). Recognizing that assessment formats each have different affordances and limitations, we compared student responses to MTF and FR questions to help practitioners develop deeper insights on how to translate assessment results into diagnostic information that can be used to guide instruction.

### MTF Responses Can Predict Correct, but Not Incorrect, FR Conceptions

Across all the T/F statements, we found a moderate correlation between response rates for correct conceptions. We also found on average that MTF correct responses occurred at much higher rates than corresponding FR correct conceptions. Indeed, MTF correct rates almost always exceeded the rate at which students included corresponding FR correct conceptions. These results suggest that, while MTF format can provide some information on the frequency of correct FR conceptions, the MTF correct response rate will generally overestimate the proportion of students who will write the corresponding correct FR conception. This difference likely results from MTF questions providing specific prompts that enable some students to answer

[2]The actual scores students received on MTF and FR questions were based on a partial credit scoring rule. FR exam scoring was more lenient than the specific coding rubric used here for research purposes.

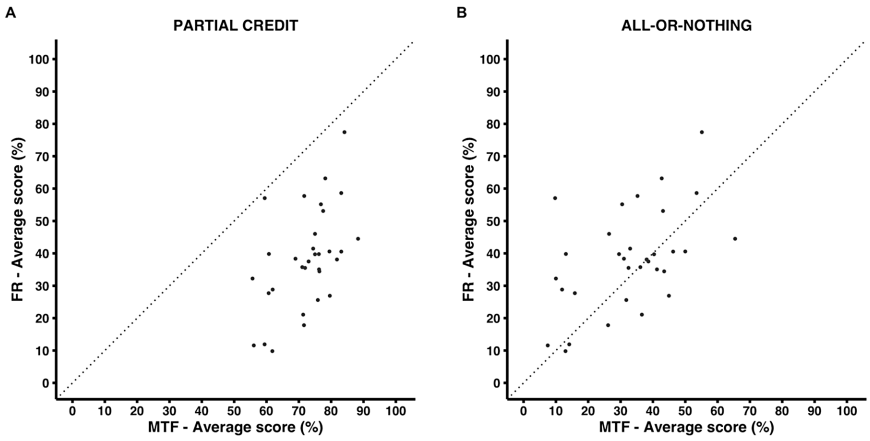



**FIGURE 6. Scatter plots showing the correlation between the average score for questions in the MTF and FR formats. (A) Comparison of the MTF and FR formats scored using a partial-credit model that gives credit for each correct selection/conception. (B) Comparison of the MTF format (scored with an all-or-nothing model that gives credit for a question only when all four accompanying statements are answered correctly) to the FR format (scored using the partial-credit model). For FR questions, unclear responses are counted as incorrect for scoring purposes. Points represent the 32 MTF × FR questions. Dotted lines represent the one-to-one line where data points would fall if they were equivalent in the two formats.**

correctly, even if they would not have generated the correct conception in an open-ended format or do not fully understand the underlying concept.

However, we found no significant correlation between response rates for incorrect conceptions. While students responded incorrectly in the MTF format 28% of the time, students infrequently expressed the corresponding incorrect conception in the FR format (7% of responses). These results suggest that MTF questions provide little information regarding the likelihood of a student volunteering an incorrect conception in the FR format, likely stemming from the fact that students rarely include explicitly incorrect conceptions aligned with the MTF statements in their FR answers. Thus, MTF questions may be uniquely suited for identifying incorrect conceptions, because they elicit responses to all included statements.

### FR Answers Include Many Unclear Answers, Particularly for Lower-Performing Students

We also sought to better understand the extent to which FR question stems can compel students to address specific conceptions in their answers. In this case, we built the FR scoring rubric around the four specific MTF statements, but this rubric could potentially have included any number of other conceptions related to the question stem. We found that more than half of the time, there was not enough information included in a student's FR answer to definitively diagnose specific conceptions. In some cases, this phenomenon stemmed from the fact that the corresponding MTF statement was designed to ask about a specific misconception and students could answer the FR question completely without addressing this conception. For example, students could answer question 20 correctly without including the misconception that fertilization adds new alleles (corresponding to MTF statement D). In other cases, adequate question scaffolding was present to induce at least some students to provide an answer with specific diagnostic information, but this information rarely allowed us to diagnose a specific incorrect conception. For example, some students stated for question 16 that plants can produce carbon dioxide through cellular respiration, but very few students stated that plants do not perform cellular respiration (corresponding to MTF statement C). The extent of unclear responses highlights both the difficulty of drafting FR question stems that can adequately prompt students to address a range of specific ideas and the limitations inherent in interpreting student open-ended responses (Criswell and Criswell, 2004). However, when assessing procedural knowledge (i.e., algebraic calculation), FR questions are uniquely able to reveal a broad range of mistakes made by students (Birenbaum and Tatsuoka, 1987), and unclear responses may be less common. While the current study focused on conceptual understandings, further work will help reveal the extent to which the observed patterns occur for diagnosing science process skills and other important competencies (American Association for the Advancement of Science, 2009).

To further understand FR answer tendencies, we analyzed responses separately for students in each quartile based on their total exam scores. As expected, we found that the correct response rate decreased across quartiles, with correct responses occurring 59% of the time for the top quartile and 16% for the bottom quartile. Surprisingly, the proportion of students providing incorrect conceptions did not change across the quartiles. Instead, a decrease in overall exam performance coincided with an increase in unclear responses, climbing to 76% of all responses for the bottom quartile. As a consequence, FR questions became increasingly less diagnostic for lower-performing students, because their answers contained very little information that could be definitively interpreted as either correct or incorrect with respect to specific MTF conceptions. These results also suggest that lower-performing students may struggle with understanding question stems and targeting their answers around specific conceptions valued by the instructor.

### MTF and FR Questions Imply Different Kinds of Incomplete Understanding

In both question formats, the majority of student answers reflected some type of incomplete understanding of various conceptions related to a question. Thus, we sought to further characterize these incomplete understandings in order to decipher student answer patterns. For MTF questions, among students who did not answer all four T/F statements correctly, almost all appeared to have mixed understandings, because their answers contained evidence of both correct and incorrect conceptions. For FR questions, students showed a much broader array of different understandings. Among students not addressing all four FR conceptions correctly, many students were classified as having partial understandings consisting of

correct and unclear conceptions, while only a small number revealed true mixed understandings with both correct and incorrect conceptions (Figure 5). Thus, taking question results at face value, a main difference between the two formats lies in the manner in which they reveal incomplete understandings.

### A Comparison of MTF and FR Question Scores

In addition to the type of diagnostic information provided by question format, students and instructors may also be concerned with the relative scores produced by each format and the consequent effect on student grades. We compared MTF scores calculated with partial or all-or-nothing methods with FR scores calculated with a partial scoring rule based on the coding rubric. We found correlations between the question formats for both comparisons, but we determined that MTF partial scoring produces much higher scores than the FR format, while MTF all-or-nothing scoring yields scores on par with the FR format. These results agree with previous findings that MTF questions, when scored based on the correctness of an *entire* question, can approximate the difficulty of FR questions (Kubinger and Gottschall, 2010; Kuechler and Simkin, 2010). Conversely, when comparing graduate student performance on equivalent FR and MC questions with multiple correct answers, others have found that these two formats were most comparable with a partial-credit scoring method (Kastner and Stangl, 2011). These results suggest that the correspondence of scores across two question formats may depend on additional factors, such as the student population and question content.

The alignment of MTF all-or-nothing scores to FR partial scores is difficult to interpret, because these scoring schemes measure student thinking in fundamentally different ways. MTF all-or-nothing scoring estimates the number of students with question mastery (i.e., complete understanding), while FR partial scoring gauges the average degree of understanding regarding the question. The score alignment seemingly stems from a coincidental intersection of the difficulty of the two scoring schemes. While the MTF format includes specific cues that help students address various conceptions, the all-or-nothing scoring scheme increases the overall difficulty of MTF questions (i.e., produces lower scores) to be more comparable to the difficulty of an FR question that lacks such cues. Moreover, we caution instructors about grading using MTF all-or-nothing scoring for several reasons. First, this scoring rule produces very low scores and may incite discontent among students. Second, a justification for using MTF questions lies in their ability to probe different conceptions related to a particular scenario. Collapsing these different dimensions into a single binary score for a question contradicts the underlying logic for the question format. Third, while MTF all-or-nothing scoring and FR partial scoring may produce similar scores, they represent distinctly different evaluations of student performance, with one identifying students who have demonstrated a complete understanding, while the other gives students credit for each correct conception. However, MTF all-or-nothing scoring may be suitable in cases in which instructors wish to specifically assess mastery (e.g., clinical practice).

### MTF and FR Answers Reflect Fundamental Tensions in Question Structure and Prompting

The advent of concept inventories has elevated the role of assessments in diagnosing specific conceptions held by students for the purposes of improving instruction (e.g., Wright and Hamilton, 2008; Shi *et al.*, 2010; Couch *et al.*, 2015; Newman *et al.*, 2016). As questions are increasingly used for formative purposes, it becomes important that instructors understand how different question formats shape student responses. We propose that the differences observed between MTF and FR results stem from fundamental tensions between the amount of structure and prompting in a question and the nature of the ambiguity resulting in student responses.

At one end of the spectrum, MTF questions represent a highly structured, closed-ended format wherein students provide simple evaluations of predefined biological conceptions. In addition to the obvious grading efficiencies, the direct prompting offered by this format enables instructors to probe a broad range of conceptions, including some conceptions that students may not readily offer in response to FR questions. As a result, instructors may learn that a substantial number of students demonstrate an incorrect conception in the MTF format that would not have been detected readily in an FR answer (Figure 3B). For example, roughly one-third of students indicated that fertilization adds new alleles to a population in the MTF format, while almost no students mentioned this misconception in their FR answers.

While MTF questions have the ability to probe specific conceptions, student answers to MTF questions can stem from a variety of different thought processes. For example, students may hold a priori conceptions that align well with a given statement, so their responses accurately capture their underlying thinking. In other cases, students may hold understandings that align poorly with a particular statement, so they engage in educated reasoning that leads to their answer selection. Finally, students with little understanding regarding a statement might guess, but their answers will likely still reflect a predisposition toward superficial question features (Cronbach, 1941; Ebel, 1978). In the absence of student interviews or response modeling, instructors have a limited capacity to resolve these different reasoning processes from raw student data.

At the other end of the spectrum, FR questions represent an open-ended format with less structured prompting for student answers. This aspect of FR questions enables students to compose a vast range of different answers and describe their understandings in their own words (Birenbaum and Tatsuoka, 1987; Martinez, 1999; Criswell and Criswell, 2004). For FR questions, student responses can stem from a variety of thought processes. Students will presumably list any a priori conceptions they feel address the question, and these conceptions may or may not align with specific conceptions on a grading rubric. Students with little understanding regarding a question will struggle to formulate a valid response and will be unlikely to guess correctly. In each of these cases, the instructor has a greater capacity to resolve underlying reasoning processes from raw student responses.

A challenge with interpreting FR answers, however, lies in deciphering and categorizing student understandings. We found that student FR answers were often unclear with respect to specific conceptions. Of course, some of this uncertainty can be attributed to limitations in question design, as students may not have been adequately prompted to address specific conceptions. However, we argue that this problem extends beyond poor question design for three reasons. First, even in cases in which we provided additional and specific

prompting, we still struggled to resolve student thinking regarding a particular conception. Second, there were many cases in which a majority of students addressed a particular conception, but some students provided ambiguous responses, suggesting that even reasonably well-targeted question wording will not elicit corresponding responses from all students. Finally, our finding that lower-performing students give more unclear responses indicates that question interpretation depends on underlying student understanding, and a certain amount of understanding may be necessary for students to know that an answer should address particular conceptions (Criswell and Criswell, 2004). While this is likely a skill that students can develop with practice, we did not see the proportion of unclear responses decrease throughout the term. However, students in this study received minimal feedback on their FR homework and exam answers. If students receive feedback on the correctness of their answers and the overall quality of their written responses, we might expect improvements in their ability to construct complete and correct FR answers (Brame and Biel, 2015). Moreover, engaging students in the process of providing feedback through self- and peer review can yield greater improvement and promote self-reflection (Willey and Gardner, 2009). More work is needed to identify effective mechanisms for training students to provide quality FR answers and to better understand how this important skill changes as students advance through their academic careers.

### Affordances and Limitations of MTF and FR Questions

In considering the output of MTF and FR formats, we recognize that one of the biggest differences lies in how these questions characterize incomplete understandings. Our findings suggest that the prompting provided by MTF statements causes many students to display mixed understandings consisting of correct and incorrect conceptions. However, instructors should adjust their interpretations to acknowledge that these same conceptions may not have been explicitly articulated in an FR answer, likely because students can use educated reasoning and guessing in their MTF response process. Thus, the MTF format runs the risk of identifying misconceptions that may not have been prominent in students' minds. Conversely, the FR format reveals fewer students with mixed and more students with partial understandings. While this may reflect deficiencies in student mental models or question targeting, instructors should also recognize that students may hold additional incorrect conceptions that are not readily detected by the FR format.

We propose that the tendencies discovered here reflect a broader dilemma in question structure and design. As questions become more structured, such as with MTF questions, instructors have a greater capacity to restrict and grade student responses, but they lose the ability to understand the thought processes underlying student responses. As questions become less structured, such as with FR questions, instructors have a greater capacity to decipher student thinking, but they lose the ability to discern particular conceptions within a student answer, because a student may have either misunderstood the targeting of the question or the student provided an answer in which the specific conception was indiscernible. This dilemma can be partially addressed by adapting the question formats to fall more in the middle of the structure spectrum. For example, MTF statements might be followed by an open-ended probe for students to explain their reasoning (Haudek *et al.*, 2012), or FR questions could be divided into more explicit prompts designed to elicit specific conceptions (Urban-Lurain *et al.*, 2010). Instructors can also design assessments that include multiple question formats to take advantage of the affordances of closed-ended and open-ended formats.

In considering the trade-offs inherent to question design, instructors may wish to consider how the resolution of their questions relates to the manner in which student results will be used to inform instructional practices. For example, instructors may prioritize closed-ended questions on summative assessments (e.g., unit exams) when the main purpose is to quantify student understanding rather than to provide detailed feedback to improve student learning. Instructors may also use open-ended and closed-ended questions at different points in a learning cycle. For example, they may use FR questions to elicit a range of student answers, determine the extent to which students can generate certain conceptions, and help students gain practice in writing open-ended responses. Conversely, because MTF questions can better detect the presence of incorrect conceptions, instructors may wish to use this format on formative assessments (e.g., clickers or quizzes) to identify and subsequently address misconceptions. Instructors can also use an interplay between question formats to help guide student learning. For example, instructors can use student answers to FR questions to develop MTF statements to be used either in the same course or a future iteration of the course (see Supplemental Material 1).

While a fundamental goal of assessments is to measure how well students have achieved course learning goals and to inform instructional practices, summative assessments can also be used to promote good study habits. In anticipation of an exam with FR questions, students have been shown to adopt more active study strategies that promote deep learning, while in anticipation of MC questions, students adopt more passive strategies that promote surface learning (Stanger-Hall, 2012; Momsen *et al.*, 2013). Further research is needed to determine how students prepare for MTF exam questions. When answering MTF questions, students must evaluate each statement independently, rather than simply recognizing one correct answer from a list of plausible options. Consequently, similar to FR questions, MTF questions may encourage students to use study strategies that promote deep learning. Understanding how students perceive different formats, such as MTF questions, also represents an important consideration to assessment design. Students perceive FR questions to be more difficult but to also provide a better representation of their knowledge compared with MC questions (Zeidner, 1987; Struyven *et al.*, 2002; Tozoglu *et al.*, 2004). This perception of difficulty can lead to increased test anxiety and influence student performance (Meichenbaum, 1972; Crocker and Schmitt, 1987). However, the response option structure of MTF questions may alleviate some of the anxiety associated with the open-ended FR format, while still providing greater resolution of student thinking than the MC format (Martinez, 1999). Future research in this area will help instructors understand how students interact with various question formats and how these formats might be used to improve student learning.

## ACKNOWLEDGMENTS

We thank Jacob Morehouse, Anh Nguyen, and Lamar Surrett for assistance with coding student responses. We also thank Kathleen Brazeal, Tanya Brown, Mary Durham, and the University of Nebraska–Lincoln (UNL) discipline-based education research community for critical research discussions. This work was supported by an internal award from UNL. This research was classified as exempt from Institutional Review Board review, project ID 14314.

## REFERENCES

Adams, W. K., & Wieman, C. E. (2011). Development and validation of instruments to measure learning of expert-like thinking. *International Journal of Science Education*, *33*, 1289–1312.

American Association for the Advancement of Science. (2009). *Vision and change in undergraduate biology education: A call to action*, Washington, DC.

Birenbaum, M., & Tatsuoka, K. K. (1987). Open-ended versus multiple-choice response formats—it does make a difference for diagnostic purposes. *Applied Psychological Measurement*, *11*, 385–395.

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability*, *21*, 5–31.

Brame, C. J., & Biel, R. (2015). Test-enhanced learning: the potential for testing to promote greater learning in undergraduate science courses. *CBE—Life Sciences Education*, *14*, es4.

Bretz, S. L., & Linenberger, K. J. (2012). Development of the enzyme-substrate interactions concept inventory. *Biochemistry and Molecular Biology Education*, *40*, 229–233.

Case, S. M., & Swanson, D. B. (1993). Extended-matching items: a practical alternative to free-response questions. *Teaching and Learning in Medicine*, *5*, 107–115.

Couch, B. A., Wood, W. B., & Knight, J. K. (2015). The molecular biology capstone assessment: a concept assessment for upper-division molecular biology students. *CBE—Life Sciences Education*, *14*, ar10.

Criswell, J. R., & Criswell, S. J. (2004). Asking essay questions: answering contemporary needs. *Education*, *124*, 510–516.

Crocker, L., & Schmitt, A. (1987). Improving multiple-choice test performance for examinees with different levels of test anxiety. *Journal of Experimental Education*, *55*, 201–205.

Cronbach, L. (1941). An experimental comparison of the multiple true-false and multiple multiple-choice tests. *Journal of Educational Psychology*, *32*, 533–543.

Crouch, C. H., & Mazur, E. (2001). Peer instruction: ten years of experience and results. *American Journal of Physics*, *69*, 970–977.

DeAngelo, L., Hurtado, S., Pryor, J. H., Kelly, K. R., & Santos, J. L. (2009). *The American college teacher: National norms for 2007–2008*. Los Angeles: Higher Education Research Institute, University of California–Los Angeles.

Eagan, M. K., Berdan Lozano, J., Aragon, M. C., Suchard, M. R., & Hurtado, S. (2014). *Undergraduate teaching faculty: The 2013–2014 HERI faculty survey*. Los Angeles: Higher Education Research Institute, University of California–Los Angeles.

Ebel, R. L. (1978). The ineffectiveness of multiple true-false test items. *Educational and Psychological Measurement*, *38*, 37–44.

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.

Goubeaud, K. (2010). How is science learning assessed at the postsecondary level? Assessment and grading practices in college biology, chemistry and physics. *Journal of Science Education and Technology*, *19*, 237–245.

Haudek, K. C., Prevost, L. B., Moscarella, R. A., Merrill, J., & Urban-Lurain, M. (2012). What are they thinking? Automated analysis of student writing about acid-base chemistry in introductory biology. *CBE—Life Sciences Education*, *11*, 283–293.

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

Kastner, M., & Stangl, B. (2011). Multiple choice and constructed response tests: do test format and scoring matter? *Procedia—Social and Behavioral Sciences*, *12*, 263–273.

Kubinger, K. D., & Gottschall, C. H. (2010). Item difficulty of multiple choice tests dependant on different item response formats—an experiment in fundamental research on psychological assessment. *Psychology Science*, *49*, 361–374.

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*, 55–73.

Martinez, M. E. (1999). Cognition and the question of test item format. *Educational Psychologist*, *34*, 207–218.

Meichenbaum, D. H. (1972). Cognitive modification of test anxious college students. *Journal of Consulting and Clinical Psychology*, *39*, 370–380.

Milton, O. (1979). Improving achievement via essay exams. *Journal of Veterinary Medicine Education*, *6*, 108–112.

Momsen, J., Offerdahl, E., Kryjevskaia, M., Montplaisir, L., Anderson, E., & Grosz, N. (2013). Using assessments to investigate and compare the nature of learning in undergraduate science courses. *CBE—Life Sciences Education*, *12*, 239–249.

National Research Council (NRC). (1999). *Transforming undergraduate education in science, mathematics, engineering, and technology*. Washington, DC: National Research Council.

NRC. (2003). *Evaluating and improving undergraduate teaching in science, technology, engineering, and mathematics*. Washington, DC: National Research Council.

Nehm, R. H., & Reilly, L. (2007). Biology majors' knowledge and misconceptions of natural selection. *BioScience*, *57*, 263–272.

Nehm, R. H., & Schonfeld, I. S. (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.

Newman, D. L., Snyder, C. W., Fisk, J. N., & Wright, L. K. (2016). Development of the central dogma concept inventory (CDCI) assessment tool. *Cell Biology Education*, *15*, ar9.

Parker, J. M., Anderson, C. W., 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.

President's Council of Advisors on Science and Technology. (2012). *Engage to excel: Producing one million additional college graduates with degrees in science, technology, engineering, and mathematics*, Washington, DC: Office of Science and Technology.

R Core Team. (2016). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. www.r-project.org/

Shi, J., Wood, W. B., Martin, J. M., Guild, N. A., Vicens, Q., & Knight, J. K. (2010). A diagnostic assessment for introductory molecular and cell biology. *CBE—Life Sciences Education*, *9*, 453–461.

Simkin, M. G., & Kuechler, W. L. (2005). Multiple-choice tests and student understanding: what is the connection? *Decision Sciences Journal of Innovative Education*, *3*, 73–98.

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.

Struyven, K., Dochy, F., & Janssens, S. (2002). Students' perception about assessment in higher education: a review. Joint Northumbria/Earli SIG Assessment and Evaluation Conference, held August 28–30, 2002, University of Northumbria at Newcastle.

Tanner, K., & Allen, D. (2004). Approaches to biology teaching and learning: from assays to assessments—on collecting evidence in science teaching. *Cell Biology Education*, *3*, 197–201.

Tozoglu, D., Tozoglu, M. D., Gurses, A., & Dogar, C. (2004). The students' perceptions: essay versus multiple-choice type exams. *Journal of Baltic Science Education*, *2*, 52–59.

Turpen, C., & Finkelstein, N. D. (2009). Not all interactive engagement is the same: variations in physics professors' implementation of peer instruction. *Physical Review Special Topics–Physics Education Research*, *5*, 1–18.

Urban-Lurain, M., Moscarella, R. A., Haudek, K. C., Giese, E., Merrill, J. E., & Sibley, D. F. (2010). Insight into student thinking in STEM: lessons learned from lexical analysis of student writing. Paper presented at the National Association for Research in Science Teaching Annual International Conference, held March 2010 in Philadelphia, PA.

Vickrey, T., Rosploch, K., Rahmanian, R., Pilarz, M., & Stains, M. (2015). Research-based implementation of peer instruction: a literature review. *CBE—Life Sciences Education*, *14*, es3.

Wilcox, B. R., & Pollock, S. J. (2014). Coupled multiple-response versus free-response conceptual assessment: an example from upper-division physics. *Physical Review Special Topics—Physics Education Research*, *10*, 1–11.

Willey, K., & Gardner, A. (2009). Improving self- and peer assessment processes with technology. *Campus-Wide Information Systems*, *26*, 379–399.

Wright, T., & Hamilton, S. (2008). Assessing student understanding in the molecular life sciences using a concept inventory. *ATN Assessment 8*, 216–224.

Zeidner, M. (1987). Essay versus multiple-choice type classroom exams: the student's perspective. *Journal of Educational Research*, *80*, 352–358.