CBE—Life Sciences Education
Vol. 14, 1–12, Summer 2015

Article

Examining the Impact of Question Surface Features on Students' Answers to Constructed-Response Questions on Photosynthesis

Michele Weston,* Kevin C. Haudek,† Luanna Prevost,‡ Mark Urban-Lurain,* and John Merrill§

*Center for Engineering Education Research, †Department of Biochemistry and Molecular Biology, and §Department of Microbiology and Molecular Genetics, Michigan State University, East Lansing, MI 48824; ‡Department of Integrative Biology, University of South Florida, Tampa, FL 33620

Submitted July 3, 2014; Revised January 17, 2015; Accepted January 27, 2015
Monitoring Editor: Janet Batzli

One challenge in science education assessment is that students often focus on surface features of questions rather than the underlying scientific principles. We investigated how student written responses to constructed-response questions about photosynthesis vary based on two surface features of the question: the species of plant and the order of two question prompts. We asked four versions of the question with different combinations of the two plant species and order of prompts in an introductory cell biology course. We found that there was not a significant difference in the content of student responses to versions of the question stem with different species or order of prompts, using both computerized lexical analysis and expert scoring. We conducted 20 face-to-face interviews with students to further probe the effects of question wording on student responses. During the interviews, we found that students thought that the plant species was neither relevant nor confusing when answering the question. Students identified the prompts as both relevant and confusing. However, this confusion was not specific to a single version.

INTRODUCTION

National calls for biology and science, technology, engineering, and mathematics (STEM) education reform are shifting the focus of STEM instruction from a collection of facts to a focus on key concepts or principles in the disciplines (American Association for the Advancement of Science, 2011; Achieve, 2013). This is an attempt to help students see connections between different science disciplines and to draw connections between units/chapters within a course. A common fundamental principle in these calls is tracing matter through biological processes. However, substantial research in biology education has documented that undergraduates in introductory-level courses have a number of alternative conceptions about cellular respiration and photosynthesis that are grounded in a failure to trace matter through biological systems (Wilson *et al*., 2006; Hartley *et al*., 2011; Maskiewicz *et al*., 2012; Parker *et al*., 2012, 2013). Instructors need assessments that can reveal the nature of student understanding in order to improve and evaluate student learning.

CBE Life Sci Educ June 1, 2015 14:ar19
DOI:10.1187/cbe.14-07-0110
Address correspondence to: John Merrill (merrill3@msu.edu).

CONSTRUCTED-RESPONSE ASSESSMENT

Assessment is an important classroom practice. The information gained from assessment allows instructors to make important and timely instructional and curricular decisions (Pellegrino *et al*., 2001; Pellegrino, 2006). Ultimately, the goal of assessment is to improve student learning (Wiggins, 1998; Pellegrino *et al*., 2001; Pellegrino, 2006). Assessments that reveal student conceptual understanding are especially important and useful for this goal in undergraduate STEM education (Smith and Tanner, 2010). Assessments can be

broadly classified as closed form (e.g., multiple choice, true/false), in which students must select an appropriate answer, or open form (e.g., essay, interview, short answer), in which students must construct their own answers. These open-form, or constructed-response questions, give faculty greater insight into student understanding than do multiple-choice questions (Birenbaum and Tatsouka, 1987). Additionally, constructed-response assessments may reveal deeper understanding and represent a more authentic task (Kuechler and Simkin, 2010). In contrast, there is evidence that closed-form assessments, such as multiple-choice items, invoke memorization strategies and influence student study habits, encouraging surface-level learning (Biggs, 1973; Stanger-Hall, 2012).

Recent work in undergraduate biology education provides evidence that students can select a correct answer on a multiple-choice assessment yet still exhibit mixed models of understanding (i.e., including both correct and incorrect ideas) or be unable to supply scientifically accurate explanations for their choices (Nehm and Schonfeld, 2008; Lyons *et al*., 2011; Haudek *et al*., 2012). In particular, Parker *et al*. (2012) investigated students' explanations of the source of plant biomass and found that students were less likely to identify a correct source of matter in a constructed-response version of a question as compared with a multiple-choice question. In addition, the researchers were able to identify mixed-model answers, responses that included both correct and incorrect sources of biomass, in student responses to the constructed-response version. Both these findings further support the benefits of using constructed-response items as opposed to closed-form type items (Parker *et al*., 2012). In this paper, we continue to examine the impact of formative constructed-response assessment and how item features of these assessments affect student explanations.

ITEM FEATURE EFFECTS

An important consideration for constructed-response questions is their item validity, or the degree to which the students' cognitive processes elicited by the question match those intended by the question writer (Pollitt *et al*., 2008). Item validity is required in order to use responses to make accurate inferences about student understanding (Pollitt *et al*., 2008). Sometimes students respond to constructed-response questions in unexpected ways, which could indicate a problem with their interpretation of the question and thus the item's validity (Sweiry, 2013). One possible cause of misinterpreting the question is its context. When answering a question with a detailed context, students may expect that the answer requires their everyday knowledge or preconceptions about the context rather than the disciplinary knowledge that the question was intended to elicit (Sweiry, 2013). Another important factor influencing student responses is their expectations of what is required in a complete answer based on cues at the question level or even the sentence level (Crisp *et al*., 2008). Schurmeier *et al*. (2010) studied alterations in question wording and tasks and found that even small changes could influence the content of student responses. They suggested that this may be due to changing the cognitive load for students. When evaluating item validity, it is important to study the effects of question context and wording.

Instructors often desire multiple versions of questions to prevent memorization during pre- and posttesting and student question sharing over time. From a learning perspective, students can benefit from answering questions across multiple cases, because it requires them to transfer their knowledge across situations (National Research Council, 1999). For these reasons, it is worthwhile to develop constructed-response questions that can be used interchangeably without expecting significant differences in the content of student responses.

PREVIOUS WORK ON ITEM FEATURES IN BIOLOGY

Previous research has shown that novices can have difficulty applying the appropriate conceptual knowledge to solve a problem (Chi *et al*., 1981; Baxter and Glaser, 1998). Instead of identifying the core concepts that characterize an assessment task, novices may focus on surface features of the problem (Chi *et al*., 1981; Baxter and Glaser, 1998). Surface features are defined by Chi *et al*. (1981) as objects that the question refers to as well as discipline-specific terms. In their work with biology students, Smith *et al*. (2013) used a card-sorting strategy to distinguish how novices and faculty categorize biology assessment questions. Novices (nonmajors in a laboratory course) sorted questions based on surface features such as the type of organism, whereas faculty sorted them based on "deep features," or core concepts. Nehm and Ha (2011) found that undergraduates' responses to constructed-response questions about natural selection were significantly affected by two surface features: whether the question addressed trait gain or trait loss and whether it made evolutionary comparisons between or within species. Later, Nehm *et al*. (2012a) found that undergraduate students were influenced by other surface features of questions about evolution. For example, students constructed explanations using different concepts in response to questions with organisms from different kingdoms or questions with different traits of interest. Similar results have been found when examining student responses to genetics questions, in that students were less likely to identify mutations as a source of new alleles in response to a question version about bacteria than one about animals (Prevost *et al*., 2013). So far, little work has been done that investigates the influence of question surface features on student ability to trace matter in photosynthesis questions.

PREVIOUS USE OF A CONSTRUCTED-RESPONSE QUESTION ON PHOTOSYNTHESIS

One item that has been used and revised in our research using constructed-response questions is on photosynthesis (Lyons *et al*., 2011; Weston *et al*., 2012). Hereafter, this question will be referred to as the photosynthesis question. Originally, this question was developed by the Diagnostic Question Cluster (DQC) project (Parker *et al*., 2012). It was used in the DQC as a multiple-choice question and as a constructed-response question. The multiple-choice version was assessed for content validity by disciplinary experts and construct validity through student interviews (Parker *et al*., 2012). The constructed-response version asks,

> Each Spring, farmers plant ~5–10 kg of dry seed corn per acre for commercial corn¡ production. By the Fall, this same acre of corn will yield ~4–5 metric tons of dry harvested corn. Explain this huge increase in mass.

This question asks about plant growth at the macroscopic level. However a correct explanation requires students to trace the path of matter from a molecular input of photosynthesis to the final form of organic material in the plant. Thus, while they trace the path of matter, students must also move between scales.

In the past, we have asked this photosynthesis question with multiple species: a maple tree, corn plants, radish plants, and peanut plants. We used the different versions in pre- and postinstruction assignments and for different exam forms. We chose these four species, because we assumed they were all familiar to students. But throughout our work with the photosynthesis question, we encountered several instances of students including species-specific ideas in their responses, which raised concerns that the scientific content of their responses was influenced by these surface features of the question.

This project investigates what impact, if any, the surface features have on student responses. For example, when conducting student interviews using the corn version of the photosynthesis question, Parker *et al*. (2012) interviewed a student who said that corn is not a photosynthesizing plant, because it is yellow, not green. In written responses to the peanut version of the question, we noticed that some students accounted for the addition of biomass through cell division causing one peanut to turn into many peanuts. In addition, both questions ask about commercial plant production and could be influenced by the students' preconceptions about the agriculture industry. For instance, in a written response, one student talked about corn production requiring commercial fertilizers and "an application of an N, P, L fert [*sic*] in order to produce profitable yields." This student's response appears to include information that might otherwise not be included if the question were about noncommercial plants. We chose to compare responses to two of the species versions for which we had evidence of students bringing species-specific preconceptions to their answers: corn and peanut plants.

In addition to species differences, we were interested in investigating whether the order of prompts influenced the content of student responses. We noted previously that many responses to the original constructed-response question included a source of matter but not a process involved in adding biomass to plants (Lyons *et al*., 2011). Students may not have known that photosynthesis is the process involved in adding biomass to plants, or alternatively, they may not have thought that a complete answer requires a process. This difference is important to instructors, who are interested in assessing their students' understanding of photosynthesis. One explanation for the incomplete responses could be the vagueness of the prompt "explain." Pollitt *et al*. (2008) studied student responses to 2000 questions from a large-scale test and found the prompt "explain" to be particularly troublesome, because it has multiple meanings and can be answered at multiple depths. To guide student responses and be more specific about what the question requires, we revised the question's "explain" prompt to be: "Where did the huge increase in biomass come from and by what process?" When we added this prompt, we found that it led to an increase in the number of students who mentioned a process in their answer but a decrease in the number who included a source of biomass (Weston *et al*., 2012). Other research has shown question-order effects can happen within a single item and may arise because the content elicited by the two prompts is similar (Schuman and Presser, 1981). It is possible that the change we noticed was because students find the prompts redundant and only answer the one prompt, which they feel encapsulates the other (Schuman and Presser, 1981). An alternative explanation is that students may forget to address the first prompt and only respond to the last prompt. In this study, to see whether the order that the prompts appear in the question stem changes how students respond, we tested two versions of the question, each with a different prompt first.

This project investigated the effects on student responses of changing two surface features of the photosynthesis question: the plant species and the order of prompts.

RESEARCH QUESTIONS

1. Does changing the surface features of this photosynthesis question stem result in significant changes to student responses?
   - A. Does changing the species of plant in the question stem influence the concepts present in student responses?
   - B. Does changing the order of prompts in the question stem influence the concepts present in student responses?
2. What parts of this photosynthesis question stem do students find important when they are answering it?

METHODS

Four Versions of the Question Stem

We used four different versions of the photosynthesis question (Table 1). Two of the versions asked about corn plants as the species and the other two versions asked about peanut plants. Two of the versions had the order of prompts "What process adds this huge increase in biomass?" and then "and where does the biomass comes from?" The other two versions had the reverse order of prompts, "Explain where the huge increase in biomass comes from" and then

Table 1. Four versions of the photosynthesis question[a]

		Order of prompt		
		Process, where (PW)	Where, process (WP)	Total
Species	Corn (CO)	Version 1 (V1) $n = 83$	Version 2 (V2) $n = 77$	160
	Peanut (PE)	Version 4 (V4) $n = 70$	Version 3 (V3) $n = 92$	162
	Total	153	169	322

[a]This table includes the different versions of the question stem, the order of the prompts, the species of plant identified in each version, and the number of students responding to each of the four versions of the question stem. Total $n = 322$.

"and by what process." Because of the class size in this study, we constrained our investigation to two species so that we could do a cross-over design with our second surface-feature variable, the order of prompt, and have a sufficient *n* in each group to maintain statistical power. The general format of the question is shown below along with the two order-of-prompts variations:

> Question text: "Each Spring, farmers plant ~5-10 kg of species per acre for commercial species production. By the Fall, this same acre of species will yield ~4-5 metric tons of dry harvested species." order of prompts.
>
> Order of prompts: process, where. "Explain what process adds this huge increase in biomass and where the biomass comes from."
>
> Order of prompts: where, process. "Explain where this huge increase in biomass comes from and by what process."

Written Homework Responses

We gathered responses to the four versions of the question in an introductory molecular and cell biology course for biology majors at a large public university. Each student was randomly assigned one of the four versions of the question stem as a homework question in an online course-management system. The question was given postinstruction following a cell metabolism unit that covered photosynthesis and cellular respiration. This question was part of a homework assignment for which the points awarded comprised <1% of the students' overall course grade. Students were given credit for any reasonable attempt to answer and were not graded on the content of their answers. We received responses from 326 out of 468 enrolled students. Four students were removed from the analysis, because we did not have complete demographic data for them (see Supplemental Table S1). We used student ethnicity, gender, course grade, and cumulative grade point average (GPA) in our demographic analysis. We chose to include this information to test whether our homework responders and interviewed students were representative of the class. In addition, instructors can use this information to compare their student population with the sample used in this study.

Text Analysis of Homework Responses

Our approach to analyzing the students' constructed responses takes advantage of automated text analysis (TA) that can rapidly summarize keywords and concepts from large data sets of student writing in response to constructed-response questions (Haudek *et al.*, 2012; Nehm *et al.*, 2012b; Prevost *et al.*, 2012). Previous work has shown these techniques to be reliable and to provide good insight into student thinking (Haudek *et al.*, 2012, Nehm *et al.*, 2012b). We used IBM SPSS Text Analytics for Surveys version 4 and Modeler 14.2 (SPSS, 2010b, 2011) for computerized analysis. The TA features of these programs are similar in that they extract *terms*—words and phrases—from written data. These terms are then aggregated into conceptual TA *categories* that are specified by the user (hereafter, titles of conceptual TA categories are italicized throughout the paper). The goal of creating conceptual TA categories is that all terms within a single category could be interchanged without fundamentally altering the meaning of a student's response to the specific question in consideration, while encompassing an idea that has disciplinary relevance. For example, the category *sugars* includes responses that mention glucose, sugar, cellulose, starch, glyceraldehyde-3-phosphate, and any relevant synonyms and/or misspellings. Initially, we chose the categories that are relevant to common correct ideas and alternative conceptions in the students' writing and then added categories for less common ideas until we established a comprehensive group of categories reflecting the concepts in this set of responses.

Human Scoring with a Photosynthesis Rubric

In addition to using computerized TA, we created an analytical photosynthesis scoring rubric (PS rubric) based on well-documented alternative conceptions in photosynthesis (Table 2; Eisen and Stavy, 1988; Köse, 2008; Parker *et al.*,

Table 2. Distribution of responses in the PS rubric[a]

Scientific or alternative concept	Examples[b]	Corn[c]	Peanut[d]	WP[e]	PW[f]
Correct process	Photosynthesis, Calvin cycle	53%	59%	57%	55%
Incorrect process	Light reactions alone, respiration, cell division	39%	35%	36%	39%
Correct source of biomass	CO_2, carbon from the atmosphere	36%	23%	30%	29%
Incorrect source of biomass	Sunlight as mass, oxygen, ATP	21%	22%	25%	29%
Water as source	Water	6%	7%	5%	9%
Nutrients from the soil	Nutrients from the soil, minerals, fertilizer	29%	24%	17%	26%
Correct product	Glucose, sugar	23%	20%	26%	16%
Incorrect product	CO_2, ATP, energy	6%	4%	5%	5%

[a]The PS rubric includes eight scientific and alternative concepts typically found in student responses. It is scored dichotomously based on the presence or absence of each concept. Each response can fall into 0–8 of the individual scientific or alternative concepts.

[b]The PS rubric identifies concepts in context. For example, it is possible for carbon dioxide to be used correctly as a source of biomass by one student and incorrectly as a product of photosynthesis by another.

[c]$n = 160$.

[d]$n = 162$.

[e]$n = 169$.

[f]$n = 153$.

2012). The PS rubric scores for eight scientific and alternative concepts (hereafter, titles of concepts from the rubric will be underlined throughout the paper; see Supplemental Table S2 for the complete PS rubric with student examples). Responses are scored dichotomously and receive a "1" for the presence of the scientific or alternative concept, or a "0" for its absence. Each response is scored independently for each of the scientific and alternative concepts and therefore can fall into 0–8 of the individual scientific or alternative concepts.

Three disciplinary experts independently scored the 322-response data set using the PS rubric. The scorers then met to resolve any scoring disagreements and arrive at a consensus score for those responses. Any necessary revisions to the rubric were also discussed.

Interviews and Coding for Consistency

Approximately 4–6 wk after instruction, we conducted 20 student interviews. The purpose of these interviews was to investigate how closely students' verbal responses resemble their written responses to this question and to ask about the photosynthesis question stem itself. The course instructor emailed the students, asking them to participate in face-to-face interviews to give feedback about the course. Only homework responders were solicited. Each student received a payment of $20 for participating in interviews. We chose students who responded on a first-come-first-served basis until we had five students from each question-stem version. The group of interviewed students had a statistically significant higher course grade average (3.03 vs. 2.40, $p < 0.008$) than the rest of the homework responders (see Supplemental Table S1).

Interviews were conducted by four interviewers who followed the same interview protocol (see the Supplemental Material for the complete interview protocol). The protocol was similar in structure to other clinical interview protocols (Haudek *et al*., 2012; Parker *et al*., 2012). First, we asked students to respond out loud to the same photosynthesis question stem they had answered in their homework. After their initial verbal responses, we showed students their original written homework responses and asked them to reconcile any differences. In the second part of the interview, we asked the students questions about how they interpreted the question prompt itself. Third, we asked clarifying and/or follow-up questions about the content of the students' interview responses to the photosynthesis question. All selected students provided consent to be interviewed and recorded, and the interviews were subsequently transcribed for analysis.

Two disciplinary experts independently scored each student's initial verbal response to the question using the PS rubric. Any disagreements were discussed and consensus was reached about a proper score. We also developed a coding scheme to characterize interviews based on whether the student's response stayed consistent throughout the interview or whether new ideas about plant biomass were uncovered. If the student seemed to modify his or her initial answer by adding or removing ideas through the interview probing, then the interview was coded as "inconsistent." If no ideas were added to or removed from his or her answer through the interview probing, then the interview was coded as "consistent." Two disciplinary experts independently read a full transcript of each interview and coded the interview as "consistent" or "inconsistent." After coding independently, the experts discussed any scoring disagreements and came to a consensus.

Data Analysis

We analyzed the content of written student responses in two ways: via TA and via human scoring with the PS rubric. We used the results of these analyses as variables to determine whether there were differences in responses based on the form of the question the students received. The dependent variables were the question surface features (species and order of prompts), while the independent variables were the 25 TA categories and the eight PS rubric concepts. We used univariate tests with a false discovery rate (FDR) correction that adjusts the significance level for multiple tests. We used an FDR correction, because it controls the expected proportion of incorrectly rejected null hypotheses (Benjamini and Hochberg, 1995). This correction works by ranking the tests based on their p values and then multiplying the desired alpha level for each test by its proportion of the ranks producing an overall false rejection rate at the desired alpha level (Benjamini and Hochberg, 1995), in this case $\alpha = 0.05$.

Logistic regression analysis was run using the JMP nominal logistic model script, loading species, order of prompts, and the cross between species and order of prompts as construct model effects, and whether the observation was in the TA category or PS rubric concept as the response variable.

A chi-square test was used to determine whether the likelihood of a student's response belonging to each TA category or rubric concept was independent of the version he or she received. These 33 tests were done independently. A general assumption of the chi-square test is that expected values for any cell are greater than five. For 14 of the variables, this did not hold true (*ATP*, *carbon*, *chlorophyll*, *electron transport chain*, *fertilizer*, *green*, *I don't know*, incorrect product, incorrect source of biomass, *nitrogen*, *oxygen*, *respiration*, *self-pollination*, and *storage*). Therefore, we used Fisher's exact test rather than the chi-square tests for those TA categories and PS rubric concepts. The Fisher's exact test has more power than a chi-square test when there are low expected values, because the significance is measured exactly, not approximated (Agresti, 1992).

We used Mann-Whitney *U*-tests to compare the distribution of responses across TA categories and PS rubric concepts. We used whether students were interviewed or not interviewed as the grouping variable. We chose the Mann-Whitney test as an alternative to the Student's t test, because the Mann-Whitney test is nonparametric and the data were not normally distributed (McKnight and Najab, 2010). The FDR correction was also applied to this test.

Logistic regression, chi-square tests, and Fisher's exact tests were done in JMP version 11.1.1 (SAS Institute, 2014). The demographic analysis (independent sample t tests of mean course grade and cumulative GPA compared interviewed and not-interviewed groups and responders with nonresponders) and Mann-Whitney *U*-tests were done in IBM SPSS Statistics version 21 (SPSS, 2010a).

RESULTS

Research Question 1. Does Changing the Surface Features of This Photosynthesis Question Stem Result in Significant Changes to Student Responses?

Results for Student Written Responses. We investigated whether the surface features used in the different question-stem versions influenced the content of student responses. Using TA, we examined the concepts used by students when constructing their answers. We created 25 categories to capture scientific and alternative conceptions from student responses. One category called *I don't know* was created for students who expressed this in their responses. The two most common categories (*photosynthesis* and *growth*) were the same for all four versions of the question stem, regardless of species or order of prompts (see Supplemental Figure S1). About half of the categories are commonly used (present in more than 10% of the responses) in all versions of the questions. However, none of the specific categories we created to capture previously observed species language (*cell division*, *nitrogen*, and *green*) appeared in more than 10% of the responses.

We also compared responses across different species and order-of-prompts versions of the question stem by using the results of scoring with the PS rubric (Table 2). The PS rubric looks for eight scientific and alternative concepts in student writing, and many of the concepts encompass multiple TA categories. This gives the rubric a larger grain size for analyzing responses. Responses were scored by experts for each of the eight scientific and alternative concepts in the PS rubric. Examples of student responses and how they are classified by TA categories and scoring with the PS rubric are given in Table 3.

From our scoring with the PS rubric, we found that more than 50% of the responders were scored for correct process, in this case photosynthesis (see Table 2). Incorrect processes, such as respiration and cell division, were detected in more than 35% of student responses and are captured by the alternative concept incorrect process. Students also included content that addressed the other part of the question prompt, which asks where the biomass comes from. Around 30% of the responders included a correct source of biomass, a little less than 30% included an incorrect source of biomass, and ~20% were scored for nutrients from the soil. Nutrients from the soil is designed to identify students who write about nutrients from the soil as a source of plant biomass, which is a previously researched alternative concept (Eisen and Stavy, 1988). Few students wrote about products of photosynthesis, but those who did were more likely to be scored for correct product than for incorrect product.

We used logistic regression, chi-square tests, and Fisher's exact tests to test for significance in differences of distributions of responses across PS rubric concepts and TA categories for our two surface features, species and order of prompts. We chose logistic regression, because it will predict the outcome of a binary response variable, that is, whether the response is in or out of a TA category or PS rubric concept, based on multiple predictor variables. The logistic regression can also be used to look for interaction effects between two predictor variables, in this case 1) species and 2) order of prompts (Jaccard, 2001). The results of the logistic regression indicated there were no interaction effects between the predictor variables for any of the PS rubric concepts or TA categories. Therefore, we were able to eliminate the interaction term and use chi-square tests.

Table 3. TA categories and PS rubric concepts for three example responses

Response	TA categories	PS rubric concepts
From seedlings the plants reproduce (divide) their cells at an incredible rate. If given the proper nutrients the plants will continuously reproduce cells and eventually produce peanuts.	*Cell division* and *organic or inorganic substance*	Incorrect process and nutrients from the soil
The increase in biomass comes from photosynthesis. This is because photosynthesis allows plants to use light energy and carbon dioxide and convert it to carbohydrates (sugars). As these sugars accumulate it allows the plants to grow larger and larger.	*Photosynthesis*, *carbon dioxide*, *sugars*, *growth*, and *solar radiation*	Correct process, correct source, and correct product
Dry seed corn undergoes photosynthesis. The light reactions and Calvin Cycle create sugar, which enters cellular respiration, and all of these processes causes [*sic*] the seeds to grow. The increase in biomass comes from these processes.	*Photosynthesis*, *respiration*, *sugars*, and *growth*	Correct process, incorrect process, and correct product

We used a chi-square test to see whether differences across the four versions of the question are due to chance or some other factor. For those TA categories and PS rubric concepts that had expected values less than five, we used the Fisher's exact test rather than the chi-square test. We also adjusted the significance level for 33 tests using an FDR correction for multiple tests. None of the differences across question-stem versions with different species and orders of directives were significant at the alpha levels set by the FDR correction.

In addition to testing for differences in individual categories and PS rubric concepts across versions of the question stem with different surface features, we also compared the total number of TA categories per response to see whether students used a different number of ideas when responding to different prompts. Figure 1A shows that the distribution of the total number of categories is similar across the two order-of-prompts versions. Figure 1B shows that the distribution of the total number of categories in responses to the corn version is similar to the distribution for peanut version, but the corn version has some responses with more categories. Box-and-whisker plots display the distribution of the total number of

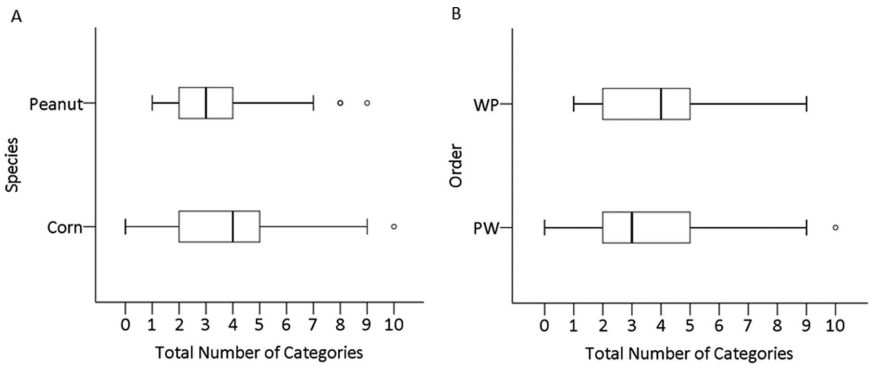

Figure 1. Distribution patterns of TA categories by (A) species and (B) order of prompts. $n = 322$.

categories for responses to each version (Figure 1). In a box-and-whisker plot, the left edge of a box represents the 25th percentile, and therefore the lowest 25% of responses fall to the left of that line. The right edge of the box represents the 75th percentile, and therefore the top 25% of responses fall to the right of that line. The thick line inside the box represents the median. The "T" bars, or whiskers, extending from the box show the distribution of data within the 95% confidence interval. Any values outside of the 95% confidence interval are considered outliers and are shown as circles on the graph.

We then used the TA and PS rubric scoring results to compare the concepts the interview group ($n = 20$) wrote about in their homework responses with the rest of the responders ($n = 302$). We used Mann-Whitney *U*-tests for this analysis and the FDR correction for multiple tests. The results showed that the distribution of responses were significantly different for only one TA category, *seed dispersal*, and for none of the PS rubric concepts. This category, *seed dispersal*, was created to help capture the idea observed in previous semesters that plants gain biomass by dispersing their seeds throughout the field, with each growing into a new plant that has biomass. One student from the interview group wrote about seed dispersal in his homework response, whereas no one in the not-interviewed group mentioned it. These results show that the content of written responses (as measured by TA categories and the PS rubric) were nearly identical for the group of interviewed students when compared with the noninterviewed students.

Results for Student Interviews. Each of the 20 interviews began with asking the student to verbally answer the same version of the photosynthesis question that he or she had been given on the homework. The interviewers waited for the students to finish answering and did not ask probing questions during this part of the interview. The purpose of this part of the interview was to compare how students responded in writing with how they answer verbally during an interview. We used both TA and the PS rubric to identify scientific and alternative conceptions in this part of the interview transcripts to compare the results with each student's written homework response. We found that six out of eight of the scientific and alternative concepts scored by the PS rubric showed agreement above 75% between the homework and interview. That means that, for each of those concepts, a student was scored the same (either 0 or 1) on both on the homework and the interview. For example, this student's homework response would be considered in agreement with his interview response for the scientific concept correct process, because both the homework and the interview were scored as a 1. All student responses are verbatim:

> Student 1 homework: Photosynthesis is the cause of this increase in mass. The corn uses the energy from the sun for the light reactions in photosynthesis. These reactions store the sun's energy in the form of ATP and NADPH which are used in the Calvin cycle. This energy is used to from [*sic*] glucose from carbon dioxide which is then used by the plant. The increase in biomass is a result of the buildup of glucose in the plant.
>
> Student 1 interview: I think when I answered this the first time I said that photosynthesis is obviously the main cause, the reason being that the plants would absorb carbon dioxide from the atmosphere in the form of CO_2. And then through the different cycles with photosynthesis, like carbon fixation, the carbon dioxide would be turned into sugars and so that would be the increase in mass. It's from the sugars that are being produced by the corn, which help it grow and get bigger.

Nutrients from the soil and water as source were the two alternative concepts that did not have high agreement between the homework and the interview. For nutrients from the soil, four responses added this alternative concept and four responses lost it in the interview. After reviewing the responses, there do not seem to be any other commonalities in these responses.

As for the alternative concept, water as source, five students were scored 1 on the homework only, and one student was scored 1 on the interview only. Four out of five of the student responses that were scored 1 on the homework only for water as source talk about water as an input to photosynthesis. However, during the interview, these students seemed more uncertain about the photosynthetic process, and only two of them were able to explain photosynthesis and its inputs. This may be a possible explanation for why water as source was less common during the interview than on the homework. For example, one student mentioned water being necessary for photosynthesis on the homework but then struggled with identifying a process during the interview:

> Student 2 homework: The biomass comes from the growth of the plant. The process of photosynthesis makes food for the plant. Like any other species when an organism is given food it will grow. The seeds must be given water and sunlight so that it [*sic*] can actually

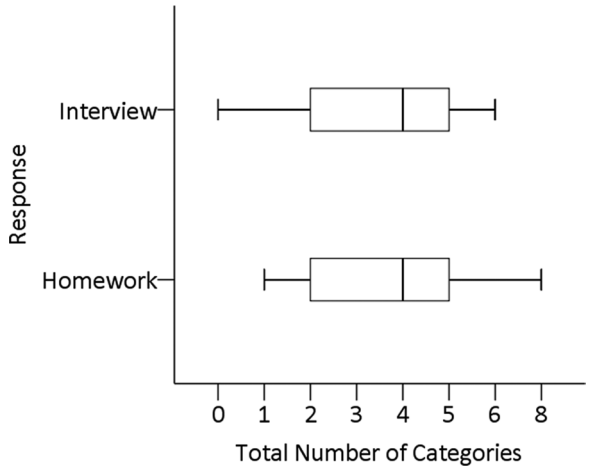

Figure 2. Distribution of total number of TA categories for responses to the homework and during the interview. $n = 20$.

> begin the process of photosynthesis with the light reactions. The dark reactions do not necessarily need light but it uses the products form [*sic*] the light reactions to start. If there was not photosynthesis you would just eat the peanut seeds and not actually peanuts.
>
> Student 2 interview: I know he, the instructor, talked about how plant … I know plant seeds, they grow; but it's through the nutrients and Calvins … all that stuff … photosynthesis stuff and Kreb's, Calvin, cellular respiration. And then I guess they just grow. I've not a clue.

When we used TA to compare how individual students responded to the homework and the interview, we found that students showed consistency above 75% between their homework and interview responses for 20 of the 25 categories. Two of the five categories with agreement below 75% are aligned with the PS rubric agreement results. The category *water* corresponds to the alternative concept water as source, and the category *in/organic substance* often corresponds to the alternative concept nutrients from the soil. Two other categories with poor agreement between homework and interview, *energy* and *growth*, are not fundamental to the scoring for any single PS rubric concept; therefore they are not well aligned to any of the eight PS rubric concepts.

We compared the total number of categories in students' homework responses with the number of categories in their interview responses. Figure 2 shows the distribution of the total number of TA categories for two groups, the written homework responses and the interview responses. The homework responses and interview responses have the same median and similar distributions. One student could not answer the question during the interview, so his interview response had no scientific content and was only in the category *I don't know*.

The majority of students showed agreement between their homework and interview responses, as measured by PS rubric scoring (Table 4) and TA categorization (see Supplemental Table S3). Also, the overall number of TA categories did not change from the homework to the interview (Figure 2).

We also coded interviews for whether or not the student modified his or her initial interview answer throughout the course of the interview. Modifying the initial interview response could include adding or removing a concept. We found that 14 students were consistent throughout the interview, meaning they did not substantially modify their explanation. Six students were coded as inconsistent. These students were an even mix of those whose initial responses had correct ideas, incorrect ideas, and mixed ideas. For some of these students, their inconsistency stemmed from the fact that they "readopted" ideas from their homework when shown their original homework responses.

Research Question 2. What Parts of This Photosynthesis Question Stem Do Students Find Relevant When They Are Answering It?

Toward the end of each interview, we asked the student which parts of the question stem are the most relevant to answering it, which parts are confusing, and then followed up by asking him or her to explain why. We found that none of the 20 students chose the species (corn or peanut) as something that was the most relevant for them in answering the question or as something that was confusing. In contrast, nine students identified the prompts as relevant to answering the question, and several found some aspect of the prompts confusing. Students also chose parts of the question stem that include experimental and control variables, such

Table 4. Distribution of responses from interviewed students across scientific and alternative conceptions on the homework and interview[a]

Scientific and alternative conception	Agreement		Disagreement		Percent agreement
	Present in homework and interview	Absent in homework and interview	Present in homework only	Present in interview only	
Correct process	9	7	3	1	80
Incorrect process	2	13	4	1	75
Correct source	4	13	0	3	85
Incorrect source	0	17	2	1	85
Water as source[b]	4	10	5	1	70
Nutrients from the soil[b]	3	9	4	4	60
Correct product	2	13	5	0	75
Incorrect product	0	18	0	2	90

[a]The following categories did not have any responses and were removed from the table: *chlorophyll*, *fertilizer*, *green*, *I don't know*, *nitrogen*, and *self-pollination*. $n = 20$.

[b]These alternative concepts did not have agreement between the homework and interview for at least 15 students.

as the amount of time that passed, the change in biomass, and the fact that the same acre of corn was measured before and after the growing season, as parts of the stem that are relevant to answering it.

Of those students who found the prompts both relevant and confusing, two said the prompts were confusing because they both seem to be asking for the same thing. Three other students indicated that the prompt that asks for a process was confusing. One student's explanation of what he finds confusing about the prompt asking for a process is shown below:

> Student 7: I guess it would be if they're asking for like one specific process here, it's just one thing, or if they're asking for just like whole steps of processes that leads to it. That would probably be it.

Other parts of the question stem that confused students were the words "dry harvested corn" and "biomass," because they did not know what they meant.

In summary, none of the 20 students who were interviewed chose the plant species as something that was the most relevant to them in answering the question or as a part of the question that was confusing. In contrast, a total of nine students chose one or both of the prompts as the most relevant parts of the question, and five students found the prompts confusing. Two of the students who chose the prompts as relevant also chose them as confusing.

DISCUSSION

This report analyzed both written and verbal student responses to different versions of a photosynthesis question. The versions varied in plant species and orders of prompts. We used the results of computerized TA and expert rubric scoring to compare the content of student written responses to these different versions of the question stem (research question 1). In addition, we confirmed these findings using student interviews (research question 1) and further used these interviews to investigate how students interpret the question stem itself (research question 2).

TA categories and scientific and alternative concepts from the PS rubric can be used to characterize responses in distinct ways. TA categories have a finer grain size and represent the majority of content ideas in student writing. The categories provide a comprehensive summary of conceptual ideas in individual responses and across an entire data set. Concepts found in the PS rubric represent a larger grain size and scientific and alternative concepts that are of interest for student understanding of photosynthesis. Because there may be many permutations of the scientific and alternative concepts, one PS concept can encompass multiple TA categories.

Research Question 1. Does Changing the Surface Features of This Photosynthesis Question Stem Result in Significant Changes to Student Responses?

Species. There was no difference in the scientific or alternative concepts in responses to question versions with different species, as measured by the distribution of responses in any TA category or PS rubric Also, changing this surface feature did not influence the number of ideas, as measured by TA categories, used by students in their responses either. This is seemingly in contrast to other studies that have identified surface features as relevant variables in the way students respond to questions (Nehm and Ha, 2011; Prevost *et al*., 2013; Smith *et al*., 2013).

We found that questions with familiar, but different, species did not result in more variable alternative conceptions from students, as was seen by Nehm *et al*. (2012a) when they altered other item features. Our interviews offer some insight. Because students chose the prompts and not the species as relevant to answering the question, we can see that they were unlikely to be focused on the species as they formulated their answers. Nehm *et al*. (2012a) chose item features based on their familiarity to students in order to assess a range of questions with differing familiarity. This differs from the stimulus of this report, in that we investigated the prevalence and influence of species-specific ideas that we had observed in student responses, despite the familiarity of both species. We found that, for our questions, previously observed preconceptions about corn and peanut plants were not prevalent in the population and did not significantly influence student responses.

In our analysis, some TA categories were designed to look for species-specific preconceptions. The category *cell division* was designed to determine whether the concept of cell division was more common in responses to the peanut version. We found that equal distributions of responses (28%) to the corn and peanut versions were distributed in the category *cell division*. The category *nitrogen* was used to look for a pattern of associating nitrogen-fixing processes with peanut plants. In this data set, only three responses were in the *nitrogen* category, and all three listed it along with other nutrients that the plant takes in through the soil. These three students all received the peanut version, but after reviewing these responses, it seems as though they were describing nitrogen as a required nutrient for plant growth, not the nitrogen-fixing process. Another category informed by our previous research was *green.* During an interview in a previous semester, a student said that he was not sure whether or not corn undergoes photosynthesis because it is yellow, not green. Because students have differing familiarity with plant species, we used the TA category *green* to see whether there was a disproportionate number of students talking about one of the species being green. While ~7% more of the corn responses than the peanut responses were in the category *green*, the difference was not significant. In the end, none of the TA categories we created to capture unusual ideas uncovered during previous use of this question was significantly influenced by the question version. Therefore, it is unlikely the surface features of the question itself are triggering these unusual ideas in student thinking in any consistent way.

Order of Prompts. Our results showed that there was no significant difference in the content of responses to versions of the question with different orders of prompts nor was there a difference in the number of ideas used by students in their responses. This is consistent with work done by Schuman and Presser (1981), in which some of the questions they studied did not have order effects, even though they asked about similar content. In the past, we noticed that adding

a second prompt, "What process adds this huge increase in biomass?," led to an increase in the number of students who mentioned a process in their answer but a decrease in the number who mentioned a source of biomass or an answer to the prompt "Explain where this huge increase in biomass comes from?" (Weston *et al*., 2012). Our results show that this occurrence is not likely to be due to the order of the two prompts. Our interviews may offer some insight. We found that a few students view the prompts as redundant, so they may think that when they answer with a process, a source or input is implied. This is a special type of item-ordering effect wherein one question is believed to be more broad and encapsulate the other (Schuman and Presser, 1981). However, we need to conduct more detailed and focused interviews in order to test this hypothesis.

Implications. These findings are important in our work with constructed-response questions, because they allow us to make more valid inferences about student understanding across these question versions. We now have evidence to allow the interchangeable use of familiar species in the question with at least some populations of students, without influencing the content of their responses. These results help meet a call for developing variations of questions that are shown to maintain validity (Nehm *et al*., 2012a). For example, Smith *et al*. (2009) used similar questions that addressed the same genetics content, in order to compare how students answer clicker questions individually with how they answer them in peer groups. Multiple question versions are needed by instructors as well, for example, when using pre- and posttesting to measure the effect of instructional interventions or for giving multiple forms of an exam in a fair way. Items within the diagnostic question clusters are examples of questions with surface features that could possibly be varied as we varied the photosynthesis question (Wilson *et al*., 2006; Parker *et al*., 2012). In part, this may be due to the similarity of the question task (i.e., tracing matter or energy through an organism). However, as other reports have documented item-feature effects in evolutionary explanations (Nehm and Ha, 2011), it is not clear to what extent content domain or question task may interact with specific item features. This research area is open to further exploration.

Our results also show that the order of the two prompts in this question does not significantly influence the content of responses. That means we can continue using the two-part version of the photosynthesis question, which elicits more complete responses than the original. Pollitt *et al*. (2008) call for more detailed prompts than "explain" alone, and these results will be useful, in that they show that the order of two-part prompts does not affect validity. Also, the methods that we used in comparing the content of responses to the four questions in a cross-over design will continue to be useful to our work validating constructed-response questions.

Research Question 2. What Parts of This Photosynthesis Question Stem Do Students Find Relevant When They Are Answering It?

During the interviews, we asked each student which parts of the question stem are the most relevant to answering it and whether any parts are confusing. We found that students thought the plant species was neither relevant nor confusing to them when answering the question. In contrast, they identified the prompts as both relevant and confusing. However, this confusion was not specific to a single version. The interviews showed that, while students could identify the parts of the question that give them directives, several students were unsure of what the question was asking. This confusion seems to be due to students being unsure of what constitutes a scientific "process" in the context of the question. Research has shown that students' understanding and explanations of scientific processes may change as they move through a learning progression for carbon-transforming processes (Parker *et al*., 2013). One explanation for students' confusion about the word "process" may be due to differences in their progress along this learning progression, particularly for students who are in its early stages.

Implications. It is encouraging that students identified the prompts as relevant, as opposed to the species of plant, as this should focus them on the information the question was designed to assess. It is important in evaluating the item validity, and thus the inferences that we make about student responses, that the question cue students properly about what is expected in their responses (Crisp *et al*., 2008). Therefore, it is unlikely that unexpected responses are due to a mismatch between the students' and writer's ideas of relevant parts of the question.

However, despite being able to identify the question prompts as the relevant parts for answering the question, students still had trouble defining a scientific "process." This could indicate that the task is not defined completely to students, which is a requirement for constructed-response questions (Liu, 2010). While we were able to support the validity of the item when students picked the prompts and not the species as relevant to answering it, we uncovered the word "process" as a possible source of confusion to some students. Complete answers to many of our constructed-response questions require a process, so students' poor understanding of this term could account for unexpected responses. In the future, we plan to further investigate student understanding of scientific processes and its effect on their responses.

Validating the Photosynthesis Question with Interviews. We found the content of written responses of the interviewed students to be nearly identical to that of the rest of the class, as shown by the results of our Mann-Whitney *U*-tests. Therefore, the photosynthesis constructed-response question is likely to elicit student understanding equally well for the groups of students in the class who were interviewed and not interviewed, although it is important to note that the interview group had higher course grade averages.

In comparing the interviewed students' written homework responses with their verbal interview responses, we found that the majority of students showed agreement for most PS rubric concepts (Table 4) and TA categories (see *Results for Student Interviews* and Supplemental Table S3). This further bolsters previous findings that student writing is well-aligned to interviews and is a better measure of student thinking than multiple-choice items (Haudek *et al*., 2012; Beggrow *et al*., 2013). We also found that the two PS rubric concepts that did not have greater than 75% agreement (water as source and nutrients from the soil) were aligned with the disagreements we found using TA categories. This result lends validity to both methods of analysis, in that they are finding similar constructs in student responses. Finally,

having a validated constructed-response question about photosynthesis allows continued investigation into how students trace matter through biological processes, without the need for in-depth interviews. Students' written responses matched their verbal responses closely in both content (TA categories) and expert scoring (PS rubric). This finding allows the use of student writing about tracing matter in photosynthesis to be used more broadly to investigate new research questions about student understanding of these fundamental principles.

In addition to comparing students' homework responses with their initial responses during the interview, we also wanted to see whether we would uncover additional ideas in the interviews through probing. When we coded the interviews for consistency, we found that a large majority of students did not modify their answers over the course of the interview. Those who did modify their answers tended to add an idea from their homework responses after seeing them during the interview. Therefore, some students did not add "novel" ideas at the time of the interview but instead were reminded of a concept that they had talked about previously in their homework. These few students who were coded "inconsistent" may represent students with a less-stable mental model of photosynthesis. These students may have "bits" of knowledge about photosynthesis but have not yet connected them in meaningful ways. In addition, they are probably less likely to have a framework constrained by scientific principles for thinking about problems (Parker *et al*., 2012), as evidenced by their tendency to pick up and discard ideas when discussing their responses.

Limitations. Students included in this study came from a single public university and from an introductory majors biology course. We recognize that geographic and cultural variables, as well as other demographics, most likely influence familiarity with given species.

Conclusions from our face-to-face interviews may be limited in that the subsample of students who volunteered for interviews had a higher course grade average than those students who did not volunteer for the interview or did not answer the homework (see Supplemental Table S1). Therefore, interview data are representative of higher-performing students in this course. However, as we have noted before, there is only one significant difference (*seed dispersal*) in the content of interviewed students' homework and noninterviewed students' homework as measured by TA categories and scientific or alternative concepts from the PS rubric. Similarly, the homework responders are more representative of higher-performing students within the course and the university than students who did not respond. Therefore, some conclusions about student understanding about photosynthesis may not extend to students who did not respond.

Finally, this study focused on our manipulations of one question, which limits the generalizability of our results to other questions.

ACKNOWLEDGMENTS

The authors thank and acknowledge the contributions of Joyce Parker (for help with item creation, rubric development, and interview protocol), Alexander Lyford and Jennifer Kaplan (for help with statistical testing and discussions about results), and Emily Norton-Henry and Matthew Berry (for help in expert scoring and conducting interviews). The work presented here was supported by the National Science Foundation (DUE 1022653). Any opinions, findings and conclusions expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. Data collection for this project was approved by the Michigan State University Institutional Review Board (x10-577). Analytic resources used for lexical analysis that are described in this report, as well as information about the Automated Analysis of Constructed Response research project and help with TA, can be found at: www.msu.edu/~aacr.

REFERENCES

Achieve (2013). K-LS1. From Molecules to Organisms: Structures and Processes, Next Generation Science Standards: For States, By States. www.nextgenscience.org/sites/ngss/files/Appendix-L_CCSS%20Math%20Connections%2006_03_13.pdf.

Agresti A (1992). A survey of exact inference for contingency tables. Stat Sci *7*, 131–153.

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

Baxter GP, Glaser R (1998). Investigating the cognitive complexity of science assessments. Educ Meas *17*(3), 37–45.

Beggrow EP, Ha M, Nehm RH, Pearl D, Boone WJ (2013). Assessing scientific practices using machine-learning methods: how closely do they match clinical interview performance? J Sci Educ Technol *23*, 160–182.

Benjamini Y, Hochberg Y (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Roy Stat Soc B *57*, 289–300.

Biggs J (1973). Study behaviour and performance in objective and essay formats. Aus J Educ *17*, 157–167.

Birenbaum M, Tatsouka KK (1987). Open-ended versus multiple-choice response formats—it does make a difference for diagnostic purposes. Appl Psychol Meas *11*, 385–395.

Chi MT, Feltovich PJ, Glaser R (1981). Categorization and representation of physics problems by experts and novices. Cogn Sci *5*, 121–152.

Crisp V, Sweiry E, Ahmed A, Pollitt A (2008). Tales of the expected: the influence of students' expectations on question validity and implications for writing exam questions. Educ Res *50*, 95–115.

Eisen Y, Stavy R (1988). Students' understanding of photosynthesis. Am Biol Teach *50*, 208–212.

Hartley LM, Wilke BJ, Schramm JW, D'Avanzo C, Anderson CW (2011). College students' understanding of the carbon cycle: contrasting principle-based and informal reasoning. BioScience *61*, 65–75.

Haudek KC, Prevost LB, Moscarella RA, Merrill J, Urban-Lurain M (2012). What are they thinking? Automated analysis of student writing about acid-base chemistry in introductory biology. CBE Life Sci Educ *11*, 283–293.

Jaccard J (2001). Interaction Effects in Logistic Regression, vol. 135, Thousand Oaks, CA: Sage.

Köse S (2008). Diagnosing student misconceptions: using drawings as a research method. World Appl Sci J *3*, 283–293.

Kuechler WL, Simkin MG (2010). Why is performance on multiple-choice tests and constructed-response tests not more closely related? Theory and an empirical test. Des Sci J Innovative Educ *8*, 55–73.

Liu X (2010). Essentials of Science Classroom Assessment, Thousand Oaks, CA: Sage.

Lyons C, Jones S, Merrill J, Urban-Lurain M, Haudek KC (2011). Moving across scales: using lexical analysis to reveal student reasoning

about photosynthesis. Paper presented at the National Association of Research on Science Teaching, held April 3–6, 2011, Orlando, FL.

Maskiewicz AC, Griscom HP, Welch NT (2012). Using targeted active-learning exercises and diagnostic question clusters to improve students' understanding of carbon cycling in ecosystems. CBE Life Sci Educ *11*, 58–67.

McKnight PE, Najab J (2010). Mann-Whitney U test. In: Corsini Encyclopedia of Psychology, Vol. 2, ed. IB Weiner and EW Craighead, Hoboken, NJ: John Wiley & Sons, 960.

National Research Council (1999). How People Learn: Brain, Mind, Experience, and School, Washington, DC: National Academies Press.

Nehm RH, Beggrow EP, Opfer JE, Ha MS (2012a). Reasoning about natural selection: diagnosing contextual competency using the ACORNS instrument. Am Biol Teach *74*, 92–98.

Nehm RH, Ha MS (2011). Item feature effects in evolution assessment. J Res Sci Teach *48*, 237–256.

Nehm RH, Ha MS, Mayfield E (2012b). Transforming biology assessment with machine learning: automated scoring of written evolutionary explanations. J Sci Educ Technol *21*, 183–196.

Nehm RH, Schonfeld IS (2008). Measuring knowledge of natural selection: a comparison of the CINS, an open-response instrument, and an oral interview. J Res Sci Teach *45*, 1131–1160.

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 Sci Educ *11*, 47–57.

Parker JM, de los Santos EX, Anderson CW (2013). What learning progressions on carbon-transforming processes tell us about how students learn to use the laws of conservation of matter and energy. Educ Química *24*, 399–406.

Pellegrino JW (2006). Rethinking and Redesigning Curriculum, Instruction and Assessment: What Contemporary Research and Theory Suggests, A Paper Commissioned by the National Center on Education and the Economy for the New Commission on the Skills of the American Workforce, Washington, DC: National Center on Education and the Economy.

Pellegrino JW, Chudowsky N, Glaser R (2001). Knowing What Students Know: The Science and Design of Educational Assessment, Washington, DC: National Academies Press.

Pollitt A, Ahmed A, Baird J, Tognolini J, Davidson M (2008). Improving the Quality of GCSE Assessment, A Report Commissioned by the Qualifications and Curriculum Authority, Carrickfergus, UK.

Prevost LB, Haudek KC, Merrill J, Urban-Lurain M (2012). Deciphering student ideas on thermodynamics using computerized lexical analysis of student writing. Paper presented at the American Society for Engineering Education Annual Conference, held June 13, 2012, San Antonio, TX.

Prevost LB, Knight J, Smith M, Urban-Lurain M (2013). Student writing reveals their heterogeneous thinking about the origin of genetic variation in populations. Paper presented at the Annual Meeting of the National Association for Research in Science Teaching, held April 6–9, 2013, Rio Grande, PR.

SAS Institute (2014). JMP 11 Basic Analysis, Cary, NC.

Schuman H, Presser S (1981). Questions and Answers in Attitude Surveys: Experiments on Question Form, Wording, and Context, New York: Academic.

Schurmeier KD, Atwood CH, Shepler CG, Lautenschlager GJ (2010). Using item response theory to assess changes in student performance based on changes in question wording. J Chem Educ *87*, 1268–1272.

Smith JI, Combs ED, Nagami PH, Alto VM, Goh HG, Gourdet MA, Hough CM, Nickell AE, Peer AG, Coley JD (2013). Development of the biology card sorting task to measure conceptual expertise in biology. CBE Life Sci Educ *12*, 628–644.

Smith JI, Tanner K (2010). The problem of revealing how students think: concept inventories and beyond. CBE Life Sci Educ *9*, 1–5.

Smith MK, Wood WB, Adams WK, Wieman C, Knight JK, Guild N, Su TT (2009). Why peer discussion improves student performance on in-class concept questions. Science *323*, 122–124.

SPSS (2010a). IBM SPSS Statistics 21 (version 21.0.0), Chicago, IL.

SPSS (2010b). SPSS Text Analytics for Surveys (version 4.0), Chicago, IL.

SPSS (2011). IBM SPSS Modeler (version 14.2), Chicago, IL.

Stanger-Hall KF (2012). Multiple-choice exams: an obstacle for higher-level thinking in introductory science classes. CBE Life Sci Educ *11*, 294–306.

Sweiry E (2013). A framework for the qualitative analysis of examinee responses to improve marking reliability and item and mark scheme validity. Paper presented at the 39th Annual Conference of the International Association for Educational Assessment, held October 20–25, 2013, Tel Aviv, Israel.

Weston M, Haudek KC, Prevost LB, Lyons C, Urban-Lurain M, Merrill J (2012). How do biology undergraduates "explain" photosynthesis? Investigating student responses to different constructed response question stems. Paper presented at the NARST Annual International Conference, held March 25–28, 2012, Indianapolis, IN.

Wiggins G (1998). Educative Assessment: Designing Assessments to Inform and Improve Student Performance, San Francisco, CA: Jossey-Bass.

Wilson CD, Anderson CW, Heidemann M, Merrill JE, Merritt BW, Richmond G, Sibley DF, Parker JM (2006). Assessing students' ability to trace matter in dynamic systems in cell biology. Cell Biol Educ *5*, 323–331.