Characterizing Students' Ideas about the Effects of a Mutation in a Noncoding Region of DNA

Scott A. Sieke,[†‡] Betsy B. McIntosh,[†‡] Matthew M. Steele,[§] and Jennifer K. Knight[‡*]
[†]Department of Molecular, Cellular, and Developmental Biology, University of Colorado–Boulder, Boulder, CO 80309; [§]CREATE for STEM Institute, Michigan State University, East Lansing, MI 48824

ABSTRACT

Understanding student ideas in large-enrollment biology courses can be challenging, because easy-to-administer multiple-choice questions frequently do not fully capture the diversity of student ideas. As part of the Automated Analysis of Constructed Responses (AACR) project, we designed a question prompting students to describe the possible effects of a mutation in a noncoding region of DNA. We characterized answers from 1127 students enrolled in eight different large-enrollment introductory biology courses at three different institutions over five semesters and generated an analytic scoring system containing three categories of correct ideas and five categories of incorrect ideas. We iteratively developed a computer model for scoring student answers and tested the model before and after implementing an instructional activity designed to help a new set of students explore this concept. After completing a targeted activity and re-answering the question, students showed improvement from preassessment, with 64% of students in incorrect and 67% of students in partially incorrect (mixed) categories shifting to correct ideas only. This question, computer-scoring model, and instructional activity can now be reliably used by other instructors to better understand and characterize student ideas on the effects of mutations outside a gene-coding region.

INTRODUCTION

In high-enrollment biology courses, student understanding is often measured through multiple-choice questions, as grading written answers presents a significant time and resource burden for instructors (Simkin and Kuechler, 2005). Multiple-choice questions can also be used effectively for rapid in-class feedback to both instructors and students and can be paired with peer instruction and technological tools such as clickers to create an effective active-learning environment (Smith *et al.*, 2009; Hubbard and Couch, 2018). However, from an assessment perspective, multiple-choice questions offer a relatively limited view of student conceptual understanding, because they force students to select the one answer with which they most strongly agree, but do not necessarily include or allow them to express all of their ideas on a particular topic (Birenbaum and Tatsuoka, 1987; Kuechler and Simkin, 2010; Couch *et al.*, 2015). Students can also hold both scientific and naive conceptions simultaneously, shifting toward only holding scientific conceptions as they become more expert-like; such nuances cannot be captured in multiple-choice answers (Opfer *et al.*, 2012). Additionally, multiple-choice questions tend to test a lower cognitive level of understanding (Bloom, 1956). Students answering multiple-choice questions are more likely to use convergent thought processes to arrive at a single correct answer rather than employ divergent thought processes to think of all possible solutions to a problem (van den Bergh, 1990; Danili and Reid, 2006). Furthermore, students may use study strategies that privilege memorization and surface-level learning rather than pursuing a deep understanding of biological content when they know they will be asked to answer low cognitive-level questions typical of the multiple-choice format (Ward *et al.*, 1980;

Ross Nehm, *Monitoring Editor*
Submitted Sep 6, 2018; Revised Feb 5, 2019; Accepted Feb 5, 2019
CBE Life Sci Educ June 1, 2019 18:ar18
DOI:10.1187/cbe.18-09-0173

[†]These authors contributed equally to this article.
*Address correspondence to: Jennifer Knight (Jennifer.Knight@colorado.edu).

Martinez, 1999; Stanger-Hall, 2012). Although it is possible to write multiple-choice questions that access higher levels of critical thinking such as analysis and evaluation, instructors do not always expend the extra effort required to do so (Simkin and Kuechler, 2005). Several modifications on the multiple-choice format, such as two-stage and multiple true/false questions, are better at revealing "mixed" ideas that contain both correct and incorrect ideas, even among relatively advanced students (Couch *et al.*, 2015, 2018). However, the high guess rate inherent in this multiple true/false format (50%) and the forced nature of selecting an answer rather than constructing one's own answer still result in a relatively surface-level diagnosis of student ideas (Couch *et al.*, 2015). Alternatively, open-ended or constructed-response questions prompt students to respond in their own words rather than selecting an answer that most closely resembles their thinking (Martinez, 1999). Such questions can better reveal when students only partially understand a concept and when they hold both correct and incorrect ideas simultaneously (Hubbard *et al.*, 2017). Students themselves are also aware that constructed-response questions allow them to better communicate depth of knowledge, and they report altering their study strategies to focus more on conceptual application than knowledge-level recall when they know their exams will use this format (Scouller, 1998; Martinez, 1999; Simkin and Kuechler, 2005). Despite these advantages, resource and time limitations often prevent instructors from using constructed-response questions to understand student thinking, in both formative and summative assessment.

Recently, computer assisted–scoring models have begun to offer an economical and effective tool to combat the limitations of constructed-response questions. A variety of such computer assisted–scoring models exist, such as SPSS Text Analysis, in which text from student answers is extracted using large libraries of terms developed by the researchers (Nehm and Haertig, 2012; Prevost *et al.*, 2016). Another tool, Summarization Integrated Development Environment analytics (Kang *et al.*, 2008), identifies patterns in student responses and uses these patterns to categorize the student-submitted responses into categories developed by the research team (Nehm *et al.*, 2012). All computer assisted–scoring techniques rely on humans first scoring a large number of student answers, ideally reflecting a wide range of student ideas. Computer assisted–scoring models are then constructed using the human-scored data to create categories of student responses. Once a computer assisted–scoring model is functional, instructors can administer the developed question to students using existing course management or other software, upload deidentified student answers, and receive rapid and detailed feedback informing them about the ideas represented in their students' answers. This method circumvents reading hundreds of answers and provides information to instructors quickly enough for them to use student ideas as formative feedback. The computer scoring of answers is not intended to be used for grading, as even under the best circumstances it cannot yet achieve a human-level of grading accuracy; rather, it is intended to provide feedback to instructors, who can then share this feedback with students (Ha *et al.*, 2011). The Automated Analysis of Constructed Responses (AACR) group has developed questions and computer assisted–scoring models on a diverse set of topics (Ha *et al.*, 2011; Haudek *et al.*, 2011, 2012; Nehm and Haertig, 2012; Nehm *et al.*, 2012; Weston *et al.*, 2015; Prevost *et al.*, 2016). The set of AACR questions for which computer models are currently available comprise topics commonly covered in introductory biology, biochemistry, chemistry, and statistics (see https://msu.edu/~aacr). In this paper, we respond to a need for additional questions regarding topics in introductory biology.

Ideas relating to the processes of cellular transcription and translation, as well as the effect that mutations can have on these processes, are difficult for students to grasp (Marbach-Ad, 2001; Fisher, 2006; Duncan and Reiser, 2007; Wright *et al.*, 2014; Haskel-Ittah *et al.*, 2018). To address this, Prevost *et al.* (2016) developed a series of questions and a computer model based on a multiple-choice question from the Genetics Concept Assessment (Smith *et al.*, 2008) to score student ideas about the consequences of a nonsense mutation. While more than half of the students surveyed correctly reported that such a mutation will stop translation, many students incorrectly indicated that the effect on translation was due to an earlier halting of replication, transcription, or both. To further explore these misconceptions, Pelletreau *et al.* (2016) developed an activity to help students understand the possible effects of mutations on these processes. They administered the stop-codon mutation questions developed by Prevost *et al.* (2016) before and after this activity as a pre/posttest and showed that students from multiple institutions with different instructors and different classroom norms improved their understanding of the effect of a nonsense mutation after the activity. Thus, there is clearly value in developing and using questions with computer-assisted scoring to characterize the efficacy of instructional activities on student learning.

While misconceptions about nonsense mutations are prevalent, the question developed here specifically builds on a previous research finding from the Introductory Molecular and Cell Biology Assessment (IMCA) that students struggle to understand the roles of promoter and coding regions of DNA and how to distinguish them from noncoding and nonregulatory regions (Shi *et al.*, 2010). A question in the IMCA asks about the effect of replacing the coding sequence of a bacterial gene with the coding sequence of a similar human gene, while preserving the bacterial promoter. This change creates a human gene whose expression could be driven by the bacterial promoter. However, more than 70% of students tested answered the question incorrectly; almost half of these students thought that the amino acid sequence would consist of a "hybrid" bacterial gene sequence. This suggests that students misunderstood the role of the promoter as a regulatory sequence, believing instead that the promoter sequence contributes to the actual gene product.

In this study, we developed a constructed-response question about the effect of an insertion mutation before the promoter region of a gene, characterized more than 1000 student answers, developed and tested the viability of a computer assisted–scoring model, and used this question to measure the impact of an activity designed to help teach this concept.

METHODS

Question Development

We designed the question for this study using published guidelines for developing constructed-response questions, including minimizing jargon as much as possible in the prompt; using

wording that elicits answers of several words to several sentences in length, rather than single-word answers; and using wording to avoid responses with chains of reasoning (Haudek *et al.*, 2011; Nehm and Haertig, 2012). We pilot tested the value of a draft version of this question by giving it to students from three introductory biology classes and found that it elicited an appropriate range of correct and incorrect ideas (unpublished data) that could be used to further explore student understanding of this concept. We then clarified the wording and formatting of the question based on feedback from 10 faculty associated with the AACR project and 10 interviews with students enrolled in an introductory biology course. A draft version of this question showed the sequence of both strands of DNA and was universally described as confusing in interviews with students. Accordingly, we chose to show only a single strand of DNA and to use the common descriptor of "coding strand," as that term could be correctly defined and used by interviewees. We also chose to use the word "promoter" rather than a more general descriptor, because students are commonly taught the definition of a promoter sequence and correctly identified that the promoter was critical for the binding of RNA polymerase to the DNA strand. The final version of the "noncoding mutation" (NCM) question is shown below.

> The following is a eukaryotic DNA sequence. The coding sequence of the gene is in bold and italicized, and the promoter is underlined.
>
> DNA 5′ T G*A A G G A A T TA T A A T A C G A C C … ***A T G A T G T A C G C A T AAAC G T*** 3′
>
> A mutation occurs in which a base (T) is inserted into the DNA sequence after the G, at the position marked with an asterisk, before transcription begins. How will this alteration influence the mRNA produced?

Data Collection

We collected a total of 1127 student responses from students enrolled in eight different large introductory biology courses at three institutions over five semesters. The concepts of mutations and their potential effects on transcription and translation were taught in four of these courses (introductory biology and genetics courses); students in the other four courses were also enrolled in introductory biology courses in which they learned about mutations and genes in a more general sense, not explicitly being taught the content of this question. We took this approach purposefully, so that we could collect a wide range of ideas and better characterize student thinking on this topic. All students received completion extra credit for answering the question through online course management software as part of a normal homework assignment. For an additional set of students, we administered the question pre- and postinstruction, separated by an in-class activity designed to help teach these concepts. For these students, we matched individual student answers pre- and postactivity ($n = 259$).

For all students, identifiers were removed by course instructors before sharing data with research personnel. This research was approved by the University of Colorado Human Subjects Institutional Review Board (protocol 0610.10).

Characterizing Student Answers

Two authors (S.A.S. and J.A.K.) constructed an initial scoring system by categorizing answers from 301 students, working together to iteratively group student ideas into categories. Computer assisted–scoring models require large numbers of human-scored student answers in each of the chosen categories to provide examples of the language diagnostic of category membership, so we did not include categories for infrequent ideas (those that occur less than 2% of the time) in the scoring system. Further, computer assisted–scoring models are inefficient at recognizing categories that require inferential interpretation on the part of the reader, so student responses that alluded to ideas without explicit statement were not included as members of a given category. As we coded additional sets of student answers from different institutions, collected at different points during introductory courses, we continually modified the categories to reflect the full set of student ideas. To ensure interrater reliability, three of the authors (S.A.S., B.B.M., and J.A.K.) individually coded the same 21 student answers and agreed upon 91% of items graded across eight scoring categories. Each of authors then characterized 220 answers individually. All answers were then sequentially checked by the other two authors, and all disagreements were resolved. The final eight categories of student answers, with definitions and examples, are shown in Table 1.

Developing a Computer Assisted–Scoring Model

We generated a computer assisted–scoring model by using a machine-learning text classification scheme to assign nonexclusive categories to student writing (see Aggarwal and Zhai, 2012, and the references therein). Using this method, each individual student response became a "document" and each category in the scoring model became a "class." The system then made predictions on whether each given document (student answer) was a member of each class (scoring category). To generate these predictions, we used an ensemble of eight individual machine-learning algorithms, one for each of our desired categories of student ideas (Jurka *et al.*, 2012). Each individual algorithm was trained using the 1127 hand-scored student responses, resulting in the production of a set of scoring models capable of generating category predictions for new unscored student writing. For this question, we produced a total of 64 scoring models during the training phase (eight individual machine learning–algorithm scoring models for each of the eight categories). Each individual model returned a prediction of the probability that each student answer was a member of each category. We then combined the predictions of the set of individual algorithms using a naive Bayes optimal classifier stacking routine (Mitchell, 1997) to produce a single prediction for each category. A 10-fold cross-validation was performed to evaluate the performance of the scoring model. Each of these validations was assembled by sampling the full training set such that the class distribution of each validation matched the class distribution of the full sample and each response was used exactly once across the 10 validations. The Cohen's kappa (κ) values generated by this cross-validation process were used as the primary metric to evaluate the interrater reliability between human scoring and computer scoring. For each of the scoring categories shown in Table 2, κ reports the level of agreement between each scorer (human vs. computer), taking into account

TABLE 1. Category of student ideas[a]

Category	Definition	Example student answer
Transcription Unaffected	The mutation does not affect the process of transcription.	"The mutation will have no alteration to the mRNA produced."
Outside Promoter Region	The mutation occurs outside of the promoter region; the location of the mutation is in a noncoding region.	"Because the mutation occurred before the promoter sequence and outside of the coding region…"
Enhancer Region	The mutation, if in an enhancer region, could have an effect on the amount of mRNA produced or on the rate of transcription.	"It is possible, however, that this mutation could affect an enhancer which would only affect the speed of the production of mRNA."
mRNA Different Composition	The mRNA sequence is different, longer, shorter, changed, altered, mutated, incorrect, or interrupted; codons are altered, affected, shifted, or changed.	"The insertion of T will be read by the RNA polymerase and ultimately make the RNA 1 base pair longer than it would have been without the mutation."
Function Disrupted	The mutation will cause a change in the function of either the mRNA or the protein product resulting from this gene.	"The mRNA will be totally random and will not be able to make a functioning protein."
Frameshift	The mutation is or causes a frameshift.	"The mRNA will be frameshifted by one base."
Protein or Translation Affected	The mutation will change the amino acid or protein sequence, or translation is affected in some general way, apart from enhancer effects.	"This mutation will be made up of different amino acids."
Stop Codon	Translation will be terminated, the protein will be truncated, or a stop codon will be produced as a consequence of the mutation.	"This mutation will create a stop codon that will truncate the protein significantly."

[a]Student ideas were coded into eight categories. The table shows each category name, a definition of the answers that fall into that category, and an example student answer. Blue shading denotes correct ideas, and red shading denotes incorrect ideas.

random chance (Cohen, 1960). Typically, a value of 1 indicates perfect agreement, 0.81–0.99 indicates almost perfect agreement, 0.61–0.80 indicates substantial agreement, and 0.41–0.60 indicates moderate agreement (McHugh, 2012). Also shown in Table 2 are the frequency of answers in each category and the precision and recall values for each category. "Precision" is defined as the fraction of predicted positives that match the human reference score (the ratio of true positives to predicted positives), and "recall" is defined as the fraction of human reference positives that match the predicted score (the ratio of true positives to human reference positives). The precision and recall provide an additional helpful diagnostic tool to understand the limitations of lower-performing rubric bins (those with lower Cohen's kappa). Here, we note that the lower-performing rubric bins also have low recall values, or high false-negative rates, indicating that the scoring models are more likely to underscore than overscore the relevant rubric bins. There are many possible reasons for this behavior, with the most likely being that the training data set is insufficient to fully describe all possible ways student writing could display the concepts indicative of rubric bin membership.

It is important to note that the computer model can only score the question for which it was designed, not related or transfer questions. A new model has to be generated to score each new question. In addition, the performance of the scoring models described here is not independent of the student population used to produce them. Although we assembled a training set representative of the students enrolled in introductory biology and genetics courses at several institutions, students at institutions with different demographics or in more advanced courses will likely display writing content and styles not found in our training set. As a result, the scoring of student writing from such populations may not perform as well when processed with these models as the results presented here.

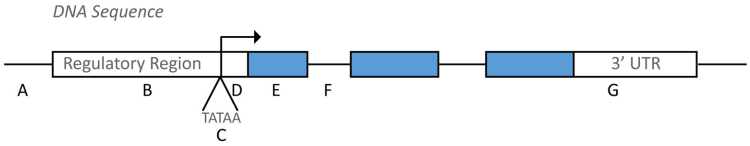
To improve the accuracy of computer scoring for categories in which the κ value was initially low, we wrote 73 additional "mock" student answers, resulting in a total data set of 1200 answers for the computer model. These mock responses

TABLE 2. Model statistics[a]

Category	Frequency	Cohen's kappa	Precision	Recall
Transcription Unaffected	0.53	0.82	0.913	0.916
Outside Promoter Region	0.42	0.83	0.912	0.890
Enhancer Region	0.10	0.82	0.930	0.756
mRNA Different Composition	0.29	0.59	0.785	0.621
Function Disrupted	0.09	0.67	0.875	0.572
Frameshift	0.22	0.88	0.917	0.889
Protein/Translation Affected	0.19	0.71	0.879	0.668
Stop Codon	0.05	0.70	0.788	0.651

[a]Cohen's kappa values range from high (0.81 and above) to moderate (0.4–0.6; Cohen, 1960; McHugh, 2012).

A) Using the sequence diagrammed below: (1) Explain how a single base deletion mutation at each of the regions indicated could affect the process of transcription and the amount of mRNA made. (2) Explain how this single base deletion could affect the process of translation and/or the amount or sequence of the protein product by filling out the table below.



	Transcription begins?	Amount or sequence of mRNA different? (explain)	Translation begins?	Amount or sequence of protein different? (explain)
A	Yes	No	Yes	No
B	Yes	Yes, could be increased or decreased	Yes	No, unless you get more/less protein because had more/less mRNA

B) You discover a single base deletion in region B of this DNA sequence. Regarding transcription, this mutation will likely:

A) Prevent transcription initiation at the TATAA box.
B) Result in an alteration to the mRNA sequence.
→ C) Result in an increase or decrease in the amount of mRNA transcribed.
D) Have no effect on transcription or the mRNA sequence.

FIGURE 1. In-class activity in which students explore the effects of single-base insertions in different noncoding and coding regions of DNA. (A) Students worked with their peers to determine how a single-base insertion into the DNA sequence in regions A through G would affect the initiation and products of transcription and translation. This figure shows the first part of this activity with suggested responses; see Supplemental Figure 1 for the complete activity. (B) Following time to work on the worksheet, the whole class responded to clicker questions to check understanding and prompt whole-class discussion. An example of one of these clicker questions is shown; see Supplemental Figure 2 for the complete list of clicker questions.

included specific phrasings and vocabulary to help train the computer model. For the remainder of the paper, when describing student ideas, we used only the 1127 student-submitted answers and excluded the mock answers written by the research team. The complete computer assisted–scoring model is publicly available for use through the AACR project website (https://apps.beyondmultiplechoice.org/AutoReport). The input required to use this model is a sheet with student answers stored as plain text in a single column (.xlxs or .csv). This computer model rapidly analyzes the student-submitted answers and then generates an instructor report.

Developing a Classroom Learning Activity

We designed an in-class activity to address several areas of student confusion regarding the effects of mutations in different gene regions. The activity, which included a handout (Supplemental Figure 1) and a series of clicker questions (Supplemental Figure 2), provided students with practice differentiating between mutations that affect transcription and those that affect translation. The handout asks students to work through solutions on their own or in small groups of two to three, filling in a table to determine the effects of a DNA base deletion in different gene regions (Figure 1A and Supplemental Figure 1). The questions ask about transcription and translation separately, encouraging students to consider whether these processes would initiate normally, and whether the amount or sequence of mRNA or protein produced would be different. After students worked on the activity for approximately 10 minutes, they answered the set of clicker questions (Figure 1B and Supplemental Figure 2), which are intended to spot-check student understanding of the activity and allow for further discussion. We captured students' initial ideas by having them answer the NCM question on a homework assignment immediately before the in-class activity and their postactivity ideas by having them answer the same question again on the next homework assignment following the activity, both for extra credit. Overall, 308 students provided preactivity ("pre") responses and 285 provided postactivity ("post") responses, for a total of 258 matched students who responded both pre and post. We then tracked individual student responses to measure changes in student thinking as a result of the in-class activity (Supplemental Table 1).

Statistical Analysis

We illustrated the data as a network diagram (Figure 2) to impart the relative frequencies of each category of student ideas, as shown by the diameter of each circle, and the frequency at which two categories co-occur in a student's answer, as shown by the width of the connecting lines. We calculated the percent co-occurrence by using the frequency that any two codes were present in a particular answer when either one or the other code was present. Co-occurrences of less than 5% were not shown, and co-occurrences were grouped into 5% bins. The Sankey diagram (Figure 3) illustrates the transitions of individual students from fully incorrect, mixed, or fully correct ideas before a classroom activity, to fully correct, mixed, or fully incorrect ideas after the activity. We used a chi-square test to compare the transition of student ideas pre- to postactivity. All statistical analyses, the creation of the network and Sankey diagrams, and the interrater reliability between human and computer coding were accomplished using RStudio v. 3.4.3 and the following packages: crossprod, igraph, networkD3, stats, and carat.

RESULTS

Student Responses Demonstrate a Range of Ideas

Student responses to the NCM question fall into eight distinct categories, illustrating three correct and five incorrect categories of student thought (Table 1), with human-computer scoring accuracies ranging from Cohen's kappa 0.59 (moderate) to 0.88 (near perfect; Cohen, 1960; Table 2). In general, the computer model was more accurate at scoring correct categories than incorrect categories. Correct student ideas ranged from simple ("the mutation will have no alteration to the mRNA produced": Transcription Unaffected) to more explanatory ("mutation occurred before the promoter sequence and outside of the

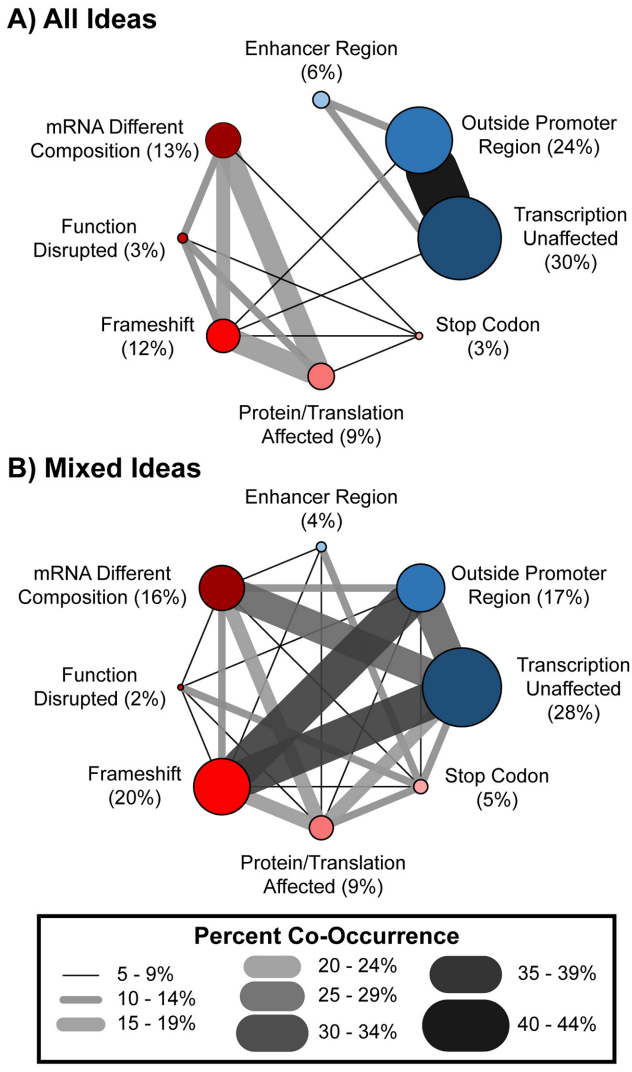


FIGURE 2. Frequency and co-occurrence of student ideas. The diameter of the circles indicates the frequency at which each category occurred relative to the total number of student ideas, and the shaded lines represent the percent of student responses in which any two categories were stated together in the same answer. This co-occurrence is calculated as the frequency at which an individual category is represented, divided by the total number of times each of these two categories is used. This way, co-occurrence functions as a measurement of how "linked" two ideas are, normalized by their frequency of category occurrence. Data were plotted from (A) the whole data set of student ideas and (B) only those 11% of responses that contained mixed ideas. Correct categories are shown in shades of blue, while incorrect categories are shown in shades of red.

coding region before the promoter": Outside Promoter Region). Some answers also included the qualifying idea that, if the mutation happened to occur in a regulatory region, the rate of transcription could be altered (coded as Enhancer Region). Only 1.4% of student answers included ideas that were not captured in the eight categories. Many of these answers were off-topic or incoherent, so we did not create an additional rubric category to represent them. Two correct ideas were also represented: that a mutation might not have an effect if it were present in an intron and that the cell may repair the mutation before transcription begins. As both of these ideas were rare, and are captured by the larger rubric category of Transcription Unaffected, we also did not create additional rubric categories for them.

Incorrect student ideas were more diverse than correct ideas. Because the computer model relies on consistent terminology as one means to identify categories, the varied use of student vocabulary challenges the development of a computer-scoring model. This is seen in the lower kappa values, which compare human and computer scoring, for incorrect versus correct categories (Table 1, red vs. blue categories). For example, student answers in the mRNA Different Composition category describe that the mRNA will be different, longer, shorter, changed, altered, mutated, incorrect, or interrupted, in addition to stating that codons can be altered, affected, shifted, or changed. In contrast, in the category Protein or Translation Affected, for which there is a higher human to computer-scoring agreement, students state that the mutation will change the amino acid or protein sequence or that translation is affected in some general way. In the category Frameshift, student use of only one or two descriptive words ("frameshift" or "shift") leads to a high scoring agreement.

In addition to looking at the individual ideas in student answers, we also categorized their entire responses as wholly correct—containing only correct ideas, wholly incorrect—containing only incorrect ideas, or mixed—containing both correct and incorrect ideas (Table 3). Wholly correct answers (~53%) contained one or more ideas that fell into correct categories exclusively. For example, a student may simply state, "There will not be a change" (one idea), or may include all three correct ideas, as shown in the table. Wholly incorrect answers (~36%) ranged from statements that "the mRNA would contain an additional base" (one idea) to answers that described a cascade of negative effects, for example, "if there is a frameshift, the mRNA will have a different composition, resulting in an altered protein." Mixed answers (~11%) contained one or more correct and incorrect ideas, representing the types of mixed understanding that are difficult to capture using forced-response multiple-choice questions.

Students Often Express Multiple Ideas in One Response

In addition to describing the frequencies of individual ideas, we can also visualize how often students link different ideas together. In a single answer, students generally tend to combine correct ideas with other correct ideas (Figure 2A, lines between blue nodes) or incorrect ideas with other incorrect ideas (Figure 2A, lines between red nodes). Figure 2A shows each coded category as a circle, sized to represent its relative frequency, out of the total number of student ideas. Lines between circles represent the co-occurrence of ideas, with the thickness of each line representing the frequency with which the two categories co-occurred when either category was present. In correct responses, for example, ~30% of student responses contained the idea of Transcription Unaffected, and ~24% of student responses contained the idea that the mutation occurred Outside Promoter Region; in ~42% of answers in which a student mentioned one of these two idea categories, the student mentioned both. In incorrect answers, students described the presence of a Frameshift in ~12% of all responses, mRNA Different Composition in ~13% of responses, and Protein/Translation Affected in ~9% of

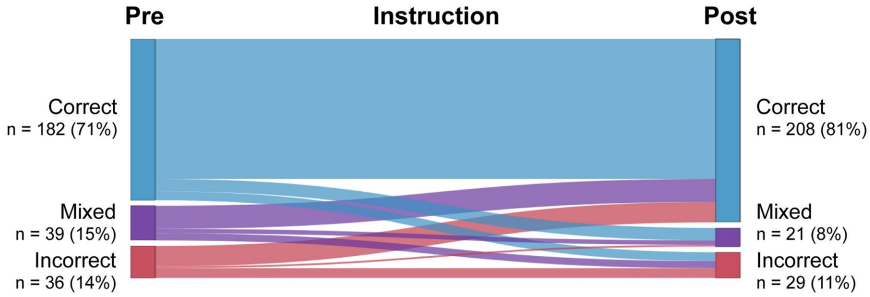


FIGURE 3. Categorization of student ideas before and after in-class instruction. The Sankey diagram shows the number of students with wholly correct, wholly incorrect, or mixed (containing both correct and incorrect) ideas when responding to the AACR question preinstruction (left) and postinstruction (right). Line thickness represents the number of students whose answers fall into each category.

responses. Similar to co-occurrence of correct ideas, three incorrect ideas are closely linked. The ideas mRNA Different Composition and Protein/Translation Affected co-occur in ~22% of responses, mRNA Different Composition and Frameshift co-occur in 18% of responses, and Frameshift and Protein/Translation Affected co-occur in ~21% of all student responses (Figure 2A). This co-occurrence can also be visualized by normalizing the fraction to the overall number of observed student ideas (the AACR report generated from the website for faculty uses this visualization technique; see Supplemental Figure 3A).

In ~11% of student answers, the ideas expressed fall into the "mixed" category, containing both correct and incorrect ideas. To better visualize the correct and incorrect ideas that students hold simultaneously, we generated a network diagram containing only the 340 statements found in mixed-idea responses (Figure 2B). Ideas that frequently occur together in mixed-idea answers are Frameshift and Transcription Unaffected, as well as Frameshift and Outside Promoter Region. Frameshift co-occurs with the correct ideas Outside Promoter Region and Transcription Unaffected more than 30% of the time. For example, "The mRNA will be one nucleotide longer after that point [the insertion], but will not alter transcription [*sic*]." Another high co-occurrence exists between mRNA Different Composition and Transcription Unaffected, where 26% of student mixed-idea answers contained both ideas.

Student Ideas Become More Expert-like in Response to an In-Class Activity

We used the information about the nature of correct and incorrect student ideas to develop a classroom activity and then test the efficacy of this activity. The activity asks students to consider the effects of a mutation at seven different regions of a DNA sequence and to describe the possible effect on the processes of transcription and translation and on the sequences of their products: mRNA and amino acids (Figure 1; complete activity shown in Supplemental Figure 1). We collected and matched answers to the NCM question from 258 students who completed the question both pre- and postactivity in order to chart how students' answers change after experiencing an activity designed around these concepts. The computer assisted–scoring model performed similarly on this smaller data set: scoring of the preinstruction responses was marginally better on average than the postinstruction responses (0.12 SEκ—the standard error difference in Cohen's kappa), but by rubric category, no single pre/post difference rose to the level of statistical significance (maximum standard error difference 0.52 SEκ). We present the human-scored results of student answers here. Before the activity, ~71% of students expressed only correct ideas; an additional ~15% expressed mixed ideas, and ~14% expressed only incorrect ideas. After completing the activity in class, ~81% of students expressed only correct ideas (Figure 3). This increase in the number of correct student responses corresponds to a decrease in the number of both mixed and incorrect ideas. Approximately 67% of students with mixed answers and 64% of students with incorrect answers expressed completely correct ideas after the activity, significantly more than moved toward incorrect ideas ($p < 0.001$ via chi-square test). A small percentage of students

TABLE 3. Examples of student answers coded into holistic and analytic categories[a]

Holistic category	Answer	Analytic categories represented			Student responses (%)
Correct	*"This alternation will not affect the production of the mRNA since it doesn't affect the promoter sequence nor the gene sequence itself.* The polymerase will bind after the insertion. **It is possible, however, that this mutation could affect an enhancer which would only affect the speed of the production of mRNA."**	*Transcription Unaffected*	Outside Promoter Region	**Enhancer Region**	53
Incorrect	*"The addition of a base (T) will cause a frameshift mutation,* in which it will change the codons on the mRNA strand following the inserted base. **This will most likely produce a non-functioning protein."**	mRNA Different Composition	**Function Disrupted**	*Frameshift*	36
Mixed	*"It will lead to a frameshift mutation* sinceit is before the promoter region, but sinceit is before the promoter region **it would not affect the gene or the gene product."**	**Transcription Unaffected**	Outside Promoter Region	*Frameshift*	11

[a]Using a holistic model, student responses were categorized as fully correct, fully incorrect, or mixed (containing at least one idea from both a correct category and an incorrect category). This table shows an example of a student answer in each of these three broad holistic categories, with the analytic categories displayed in the next three columns. Individual ideas within an answer are differentiated here by italics, underlining, and bold type. This table also shows the percent of total student answers ($n = 1127$) that fall into these three broad holistic categories.

who initially expressed mixed or correct ideas preactivity shifted to incorrect ideas postactivity. Most students who went from initially correct to mixed (eight out of 14) said that transcription will not be affected but that mRNA will have a different sequence. None said that the resulting mRNA or protein will be nonfunctional, and only one student described that the mutation would result in a stop codon. On the other hand, the 10 students who moved from correct to incorrect had ideas spanning all incorrect categories.

Similar to the set of answers used to train the computer model, answers from students in this specific course expressed many ideas concurrently in their responses. Preactivity, many students linked the correct ideas in the Transcription Unaffected category, present in ~37% of student responses, with those from the Outside Promoter Region category, present in ~32% of student responses, with a ~45% co-occurrence of these two ideas (Figure 4A). Students also linked incorrect ideas: ~38% of student responses were categorized as both Frameshift and Different Sequence, and ~35% of responses were categorized as Frameshift and Function Disrupted. We also observed "mixed" ideas of ~5–9% co-occurrence between Frameshift and Transcription Unaffected, Frameshift and Outside Promoter Region, Frameshift and Enhancer Region, and mRNA Different Composition and Transcription Unaffected.

As a consequence of the activity, students changed how they answered the NCM question. Postactivity, students significantly increased their use of Enhancer Region, which co-occurred with Outside Promoter Region in ~19% of student answers and Transcription Unaffected in ~24% of student answers (Figure 4B and Supplemental Figure 3B). When only correct student answers were examined, ~7% contained the enhancer region idea preactivity, while ~52% contained this idea postactivity. We also observed a decrease from pre- to postactivity in the number of individual incorrect categories and their co-occurrences. For example, the category mRNA Different Composition decreased from ~10 to ~6%, Frameshift decreased from ~9 to ~4%, and their co-occurrence decreased from ~38 to ~15%. Finally, the co-occurrence of both correct and incorrect ideas, which characterizes mixed-idea responses, decreased. Only mRNA Different Composition and Transcription Unaffected had a more than 5% co-occurrence, our threshold for visualization (Figure 4B).

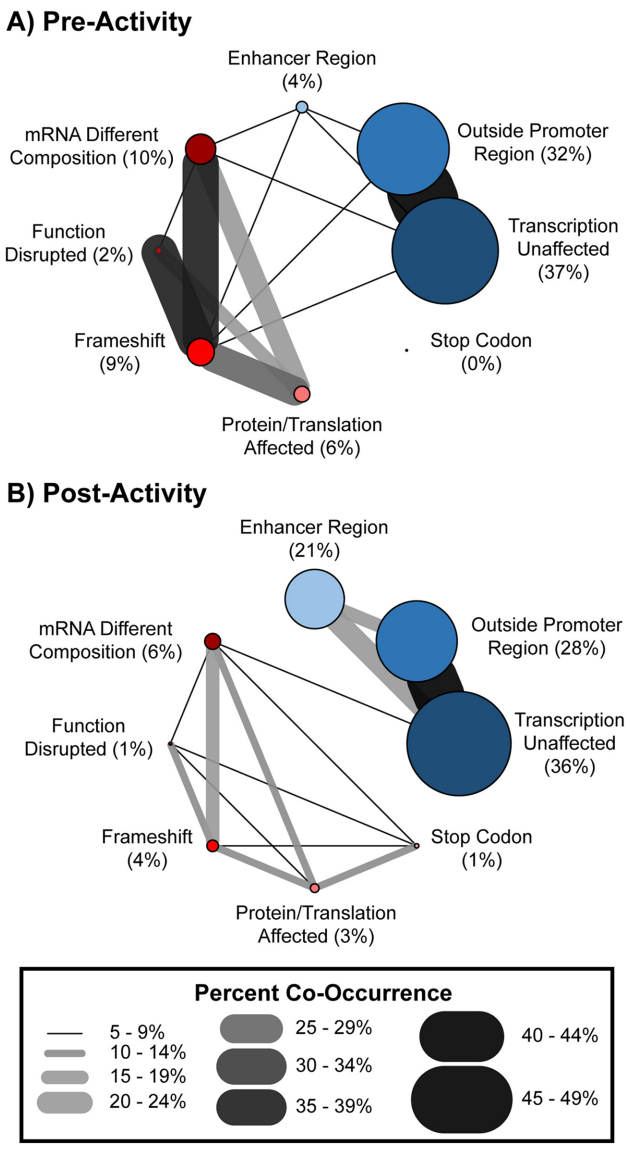


FIGURE 4. Frequency and co-occurrence of student ideas (A) preactivity and (B) postactivity. In each panel, correct categories are shown in shades of blue, while incorrect categories are shown in shades of red. The diameters of the circles indicate the frequency with which each category occurred relative to the total number of student ideas in the data set, and the shaded lines represent the percent of student responses in which any two categories were stated together in the same answer. The co-occurrence is the frequency at which an individual category is represented, divided by the total number of times each of these two categories is used.

Students Maintain Their Understanding on an Exam Question

We measured whether students maintained their understanding of this topic by looking at matched student performance on a single multiple-choice exam question similar to the NCM question (Supplemental Figure 4). Two hundred fifty-six students answered both the postactivity NCM question and the exam question. Of the 206 students who answered the postactivity NCM question with only correct ideas, ~91% went on to answer the exam question correctly. Of the 21 students who answered the NCM question with mixed ideas, ~90% answered the exam question correctly. Of the 29 students who answered the NCM question with only incorrect ideas, ~62% answered the exam question correctly (Supplemental Table 2).

DISCUSSION

The constructed-response question described in this paper, about a single-base insertion before the promoter region of a gene, reveals a variety of student incorrect and mixed ideas about how mutations affect transcription and translation and can be efficiently analyzed using a computer-scoring model to facilitate quick feedback to both instructors and students.

The Computer-Scoring Model Accurately Categorizes Student Answers

We used extensive iterative analysis to create a computer assisted–scoring model with generally very good predictive

value. With the exception of the category mRNA Different Composition, all kappa values for human-computer scoring were at the level of substantial or near-perfect agreement (Table 1). This level of agreement, although not appropriate for individually grading students, is accurate enough to support the use of the question for formative assessment to look at overall trends of student answers and co-occurrences of ideas (Ha *et al.*, 2011; Haudek *et al.*, 2012; Prevost *et al.*, 2016).

Incorrect Student Answers Suggest a Failure to Understand Components of the Process of Transcription

Even before high school, students often have difficulty in understanding genetic mechanisms (Bolger *et al.*, 2012; Haskel-Ittah *et al.*, 2018). Haskel-Ittah *et al.* (2018) showed that nearly a third of students entering high school–level biology classes demonstrate a nonmechanistic misunderstanding of "genes" and "traits," referring to traits as "inside of genes" or "genes and traits are the same thing." Further, undergraduates in biology courses still demonstrate difficulty with understanding the underlying mechanisms of transcription and translation (e.g., Duncan and Reiser, 2007; Southard *et al.*, 2016). Wright *et al.* (2014) found that 36% of students described transcription as "a chemical conversion of DNA into RNA," and 16% stated that "RNA existed before the process of transcription began." In addition, Southard *et al.* (2016) found that only 69% of introductory- and upper-level biology undergraduates were able to describe "mechanism-appropriate entities" such as RNA polymerase being used for transcription. Such fundamental misunderstandings are likely to lead to many of the incorrect ideas seen in the student answers presented here. Students who are unable to describe the mechanism or components of transcription might fail to recognize the importance of the promoter region or imagine that RNA polymerase transcribes in both directions from the location of the promoter region. Similarly, if students view the process of DNA → RNA as a chemical conversion, they are likely not considering the positioning of the promoter region, resulting in a prediction that the RNA will contain an extra base no matter where a nucleotide in inserted. Additionally, if students hold the conception that RNA exists before the process of transcription begins, they might reasonably predict that any DNA insertion will be present in the RNA, regardless of location. Once students imagine that an extra base in present in the mRNA, their conclusion that the extra base causes a frameshift mutation, leading to a truncated or otherwise nonfunctional protein, is logically consistent. Many wholly incorrect answers in our data set followed this general chain of reasoning, in which the original incorrect assumption was the addition of the base in the mRNA product. Thus, although student responses may express a variety of incorrect ideas, there is an internal logical consistency in their models.

Another component of student thinking that may lead to errors is a tendency to link words together, thus developing an immediate association of ideas. For example, Haudek *et al.* (2012) observed that strong word associations between the terms "amino" and "acid" led to students believing that functional amino groups have acidic properties. An example of such word association coloring conceptual understanding has also been observed with the terms "stop" and "codon." Students link the definition of the common term "stop" to "stop codon," and thus imagine that such a codon stops all cellular processes, including DNA replication, RNA transcription, and translation (Prevost *et al.*, 2016). In the current study, students linked the terms "frameshift" and "deletion/insertion," leading to the idea that a nontriplet insertion or deletion *anywhere* in the genome is defined as a frameshift mutation, rather than the true definition, which requires this alteration to occur *only* in a coding region.

In mixed-idea student answers, we saw two common patterns: a correct description of why the *process* of transcription would remain unaffected combined with either the incorrect idea that the mRNA would have a different sequence or the idea that the mutation would result in a frameshift. These mixed-idea answers demonstrate that, even when students correctly state that alterations to the DNA outside the promoter and coding regions do not impact transcription, they may still visualize that the insertion affects the mRNA sequence, likely for one of the reasons described earlier. With regard to the idea that this insertion will lead to a frameshift, we postulate that students likely have an incorrect definition of a frameshift mutation. Students commonly learn about frameshift mutations in conjunction with insertion or deletion mutations, and almost always in the context of the coding region of a gene. If students have not explicitly studied noncoding regions, it is not surprising that they associate an insertion of a single base or any nontriplet addition or deletion to a DNA sequence, regardless of whether the sequence will be transcribed and translated with a frameshift mutation. A typical example of such a student answer is: "It will lead to a frameshift mutation since it is before the promoter region, but since it is before the promoter region it would not effect [*sic*] the gene or the gene product." Thus, the coding system and analysis reveals combinations of student ideas that may otherwise remain hidden.

An In-Class Activity Helps Students Move Toward More Expert-like Answers

To help students grapple with the roles of different gene regions, we designed an in-class activity that directed students to predict the possible effects of mutations in different gene regions. Because NCM question uses the scenario of a single base insertion, we used a conceptually similar setup, a single base deletion, in the activity, allowing students to practice applying knowledge on an easily transferable scenario (Bransford and Schwartz, 1999). Before participating in this pre/postassessment and activity, students had learned about different kinds of mutations and different gene regions and their roles in transcription and translation. Accordingly, a relatively high percentage of students expressed wholly correct ideas when answering the question preactivity (~71%) compared with the population of students used to generate the scoring system (~53%). Nevertheless, many students who had incorrect or mixed answers shifted to wholly correct answers. In addition, postactivity, many more students included the more sophisticated idea that, if the mutation were to occur in an enhancer region, there could be an effect on transcription rate (~4% preactivity to ~21% postactivity). This suggests that students deepened their understanding of the complexity of gene expression as a result of instruction.

Students also generally maintained their understanding of this concept at least until the unit exam that followed instruction. More than 90% of students who correctly answered the postactivity NCM question subsequently answered a similar multiple-choice exam question correctly. Very few students (less than 10%) who answered the post-NCM question with correct ideas answered the exam question incorrectly. Additionally, many of the students who had answered the post-NCM question incorrectly subsequently answered the exam question correctly. This suggests that the in-class work students do to understand this concept may help them build enough understanding to correctly answer the exam question, even if they do not immediately answer the extra-credit follow-up post question correctly. Students may also study the concept further after class if they realize they still do not understand. Thus, the activity and the NCM question, if challenging to the students, may raise their metacognitive awareness and stimulate them to engage in different learning strategies to master the concept (Dye and Stanton, 2017).

Limitations

Because of the nature of computer-assisted scoring, the scoring rubric developed here can only be used to gather information from students on this specific question. The question can help diagnose different ideas that students hold regarding the roles of different regions of DNA, the impact of mutations on the initiation of transcription, and their understanding of certain principles such as frameshifts and how RNA sequences relate to DNA sequences. However, the answers cannot be used to infer student understanding on other related topics, such as the impact of mutations in different locations of a DNA sequence, or the depth of their understanding of the many facets of gene regulation.

Suggestions for Instruction

This activity aims to help students by prompting them to identify and explain the downstream effects of a deletion mutation occurring at varying locations along the gene region. Visualizations can help students understand difficult mechanisms that are too small to see or occur too rapidly to observe directly (Offerdahl *et al.*, 2017), such as the processes of transcription and translation. In addition to encouraging students to visualize the gene regions, the activity also asks students to consider the effects of these mutations on transcription and translation *separately*. Considering the effects on transcription and translation individually may help students separate the mechanisms of the two processes, which are often conflated (Southard *et al.*, 2016). Engaging with a visual aid (in this case, a diagram) may encourage students to draw their own representations to predict possible outcomes, a technique shown to have a positive impact on student learning (Quillin and Thomas, 2015). An extension to this activity could also include an incomplete or unlabeled diagram with instructions for the student to complete the visual representation of a gene and its possible regulatory sequences.

The student answers collected and analyzed here provide evidence that students struggle with biology terminology and have difficulty understanding the roles of different gene regions as influenced by a single base mutation. The NCM question can be used easily by instructors in several different ways to help students master these concepts. The question can be used pre- and postinstruction using online submission tools, such as learning management systems or Google forms, to measure whether students build a more complete understanding after instruction, and they can be used in combination with the in-class activity we designed to facilitate such instruction. With the computer assisted–scoring model, categorizing student ideas into the three correct and five incorrect categories, in addition to more holistically grouping student answers as wholly correct, incorrect, or mixed, can be done quickly and with large numbers of students. Importantly, if instructors wish to share information about student incorrect ideas as part of the learning process, sample student answers, proportions of student answers in each category, and co-occurrences of ideas are all easily visualized using the automatically generated instructor report accessible by uploading student answers at https://apps.beyondmultiplechoice.org/AutoReport. Instructor reports also show an estimate of the probability that the computer scoring of each individual student answer would agree with a human scorer for each category. In addition, the reports include co-occurrence Web diagrams similar to those shown in Figures 2 and 4, which instructors may also find useful for diagnosing and addressing mixed ideas.

In using the in-class activity, we suggest that instructors allow students to explicitly discuss the roles of different gene regions with one another. Explaining *why* mutations do or do not affect the processes of transcription and translation and comparing the processes to the products may be a key step in building understanding (Smith *et al.*, 2009; Knight *et al.*, 2015). We also suggest that instructors encourage students to work on problems involving DNA and RNA sequences before or after this activity, so students can make molecular-level connections with the concepts of sequence alterations (Wright *et al.*, 2017). Finally, given the relatively high proportion of incorrect definitions of "frameshift" in student answers, instructors should engage students in thinking about the many examples of mutations that do not result in a frameshift, such as an insertion in an intron. Instructors can also help students practice this concept through clicker questions, drawing exercises, or other types of in-class activities, so students can recognize that the definition of frameshift dictates that the disrupted sequence must be in the reading frame of the gene.

This question and accompanying activity may also be useful as a follow-up to the three stop-codon mutation questions also generated by the AACR group (Prevost *et al.* 2016), which revealed that many students misunderstand the meaning and consequences of nonsense mutations. Some students think (preinstruction) that such mutations stop the process of replication, and a larger portion think such a mutation stops transcription, which then results in a stop in translation. Both of these ideas can be repaired with an activity designed to be used along with the AACR question to help teach these concepts (Pelletreau *et al.*, 2016). Because the stop-codon mutation questions and the current NCM question all address the potential effects of mutations on transcription and translation, the two sets of questions and activities can be used within the same unit to highlight incorrect student ideas and engage students in learning these challenging concepts. The two activities together prepare students to think about how mutations affect transcription

and translation differently depending on whether they occur in a regulatory region, a coding sequence, or a noncoding, nonregulatory region.

While creating constructed-response questions requires more investment than multiple-choice questions, such questions increase the cognitive level of the learning task. The question we have shared here, as well as others in the existing AACR database, will help instructors learn more about their students' thinking and use this information to improve student understanding.

ACKNOWLEDGMENTS

We are grateful to Mark Urban-Lurain, Kamali Sripathi, John Merrill, Alex Lyford, Jennifer Kaplan, Luanna Prevost, and the rest of the AACR team of researchers for their help in this work. This work was funded by the National Science Foundation (DUE 1323022).

REFERENCES

Aggarwal, C. C., & Zhai, C. (2012). A survey of text classification algorithms. In *Mining text data* (pp. 163–222). Boston, MA: Springer. https://doi.org/10.1007/978-1-4614-3223-4_6

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*(4), 385–395. https://doi.org/10.1177/014662168701100404

Bloom, B. S. (Ed.). (1956). *Taxonomy of educational objectives: The classification of educational goals* (1st ed.). New York: Longmans, Green.

Bolger, M. S., Kobiela, M., Weinberg, P. J., & Lehrer, R. (2012). Children's mechanistic reasoning. *Cognition and Instruction*, *30*(2), 170–206. https://doi.org/10.1080/07370008.2012.661815

Bransford, J. D., & Schwartz, D. L. (1999). Rethinking transfer: A simple proposal with multiple implications. *Review of Research in Education*, *24*(1), 61–100. https://doi.org/10.3102/0091732X024001061

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46. https://doi.org/10.1177/001316446002000104

Couch, B. A., Hubbard, J. K., & Brassil, C. E. (2018). Multiple–true–false questions reveal the limits of the multiple–choice format for detecting students with incomplete understandings. *BioScience*, *68*(6), 455–463. https://doi.org/10.1093/biosci/biy037

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*(1), ar10. https://doi.org/10.1187/cbe.14-04-0071

Danili, E., & Reid, N. (2006). Cognitive factors that can potentially affect pupils' test performance. *Chemistry Education Research and Practice*, *7*(2), 64–83. https://doi.org/10.1039/B5RP90016F

Duncan, R. G., & Reiser, B. J. (2007). Reasoning across ontologically distinct levels: Students' understandings of molecular genetics. *Journal of Research in Science Teaching*, *44*, 938–959.

Dye, K. M., & Stanton, J. D. (2017). Metacognition in upper-division biology students: Awareness does not always lead to control. *CBE—Life Sciences Education*, *16*(2), ar31. https://doi.org/10.1187%2Fcbe.16-09-0286

Fisher, K. M. (2006). A misconception in biology: Amino acids and translation. *Journal of Research in Science Teaching*, *22*(1), 53–62. https://doi.org/10.1002/tea.3660220105

Ha, M., Nehm, R. H., Urban-Lurain, M., & Merrill, J. E. (2011). Applying computerized-scoring models of written biological explanations across courses and colleges: prospects and limitations. *CBE—Life Sciences Education*, *10*(4), 379–393. https://doi.org/10.1187/cbe.11-08-0081

Haskel-Ittah, M., Yarden, A., & Spell, R. (2018). Students' conception of genetic phenomena and its effect on their ability to understand the underlying mechanism. *CBE—Life Sciences Education*, *17*(3), ar36. https://doi.org/10.1187/cbe.18-01-0014

Haudek, K. C., Kaplan, J. J., Knight, J., Long, T., Merrill, J., Munn, A., … Urban-Lurain, M. (2011). Harnessing technology to improve formative assessment of student conceptions in STEM: Forging a national network. *CBE—Life Sciences Education*, *10*(2), 149–155. https://doi.org/10.1187/cbe.11-03-0019

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*(3), 283–293. https://doi.org/10.1187/cbe.11-08-0084

Hubbard, J. K., & Couch, B. A. (2018). The positive effect of in-class clicker questions on later exams depends on initial student performance level but not question format. *Computers & Education*, *120*, 1–12. https://doi.org/10.1016/j.compedu.2018.01.008

Hubbard, J. K., Potts, M. A., & Couch, B. A. (2017). How question types reveal student thinking: An experimental comparison of multiple-true-false and free-response formats. *CBE—Life Sciences Education*, *16*(2), ar26. https://doi.org/10.1187/cbe.16-12-0339

Jurka, T. P., Collingwood, L., Boydstun, A. E., Grossman, E., & van Atteveldt, W. (2012). *RTextTools: Automatic text classification via Supervised Learning version 1.4.2 from CRAN*. Retrieved March 13, 2018, from https://rdrr.io/cran/RTextTools/

Kang, M., Chaudhuri, S., Joshi, M., & Rosé, C. P. (2008). SIDE: The Summarization Integrated Development Environment. In *Proceedings of the 46th annual meeting of the Association for Computational Linguistics on Human Language Technologies: Demo session* (pp. 24–27). Stroudsburg, PA: Association for Computational Linguistics. Retrieved April 23, 2019, from http://dl.acm.org/citation.cfm?id=1564144.1564151

Knight, J. K., Wise, S. B., Rentsch, J., & Furtak, E. M. (2015). Cues matter: Learning assistants influence introductory biology student interactions during clicker-question discussions. *CBE—Life Sciences Education*, *14*(4), ar41. https://doi.org/10.1187/cbe.15-04-0093

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

Marbach-Ad, G. (2001). Attempting to break the code in student comprehension of genetic concepts. *Journal of Biological Education*, *35*(4), 183–189. https://doi.org/10.1080/00219266.2001.9655775

Martinez, M. E. (1999). Cognition and the question of test item format. *Educational Psychologist*, *34*(4), 207–218. https://doi.org/10.1207/s15326985ep3404_2

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica*, *22*(3), 276–282.

Mitchell, T. M. (1997). *Machine learning* (pp. 174–176). Boston: McGraw-Hill.

Nehm, R. H., Ha, M., & Mayfield, E. (2012). Transforming biology assessment with machine learning: Automated scoring of written evolutionary explanations. *Journal of Science Education and Technology*, *21*(1), 183–196. https://doi.org/10.1007/s10956-011-9300-9

Nehm, R. H., & Haertig, H. (2012). Human vs. computer diagnosis of students' natural selection knowledge: Testing the efficacy of text analytic software. *Journal of Science Education and Technology*, *21*(1), 56–73. https://doi.org/10.1007/s10956-011-9282-7

Offerdahl, E. G., Arneson, J. B., Byrne, N., & Brickman, P. (2017). Lighten the load: Scaffolding visual literacy in biochemistry and molecular biology. *CBE—Life Sciences Education*, *16*(1), es1. https://doi.org/10.1187/cbe.16-06-0193

Opfer, J. E., Nehm, R. H., & Ha, M. (2012). Cognitive foundations for science assessment design: Knowing what students know about evolution. *Journal of Research in Science Teaching*, *49*(6), 744–777. https://doi.org/10.1002/tea.21028

Pelletreau, K. N., Andrews, T., Armstrong, N., Bedell, M. A., Dastoor, F., Dean, N., … Smith, M. K. (2016). A clicker-based case study that untangles student thinking about the processes in the central dogma. *CourseSource*, *3*. https://doi.org/10.24918/cs.2016.15

Prevost, L. B., Smith, M. K., & Knight, J. K. (2016). Using student writing and lexical analysis to reveal student thinking about the role of stop codons in the central dogma. *CBE—Life Sciences Education*, *15*(4), ar65. https://doi.org/10.1187/cbe.15-12-0267

Quillin, K., & Thomas, S. (2015). Drawing-to-Learn: A framework for using drawings to promote model-based reasoning in biology. *CBE—Life Sciences Education*, *14*(1), es2. https://doi.org/10.1187/cbe.14-08-0128

Scouller, K. (1998). The influence of assessment method on students' learning approaches: Multiple choice question examination versus assignment essay. *Higher Education*, *35*(4), 453–472. https://doi.org/10.1023/A:1003196224280

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*(4), 453–461. https://doi.org/10.1187/cbe.10-04-0055

Simkin, M. G., & Kuechler, W. L. (2005). Multiple-choice tests and student understanding: What is the connection? *Decision Sciences Journal of Innovative Education*, *3*(1), 73–98. https://doi.org/10.1111/j.1540-4609.2005.00053.x

Smith, M. K., Wood, W. B., Adams, W. K., Wieman, C., Knight, J. K., Guild, N., & Su, T. T. (2009). Why peer discussion improves student performance on in-class concept questions. *Science*, *323*(5910), 122–124. https://doi.org/10.1126/science.1165919

Smith, M. K., Wood, W. B., & Knight, J. K. (2008). The Genetics Concept Assessment: A new concept inventory for gauging student understanding of genetics. *CBE—Life Sciences Education*, *7*(4), 422–430. https://doi.org/10.1187/cbe.08-08-0045

Southard, K., Wince, T., Meddleton, S., & Bolger, M. S. (2016). Features of knowledge building in biology: Understanding undergraduate students' ideas about molecular mechanisms. *CBE—Life Sciences Education*, *15*(1), ar7. https://doi.org/10.1187/cbe.15-05-0114

Stanger-Hall, K. F. (2012). Multiple-choice exams: An obstacle for higher-level thinking in introductory science classes. *CBE—Life Sciences Education*, *11*(3), 294–306. https://doi.org/10.1187/cbe.11-11-0100

van den Bergh, H. (1990). On the construct validity of multiple-choice items for reading comprehension. *Applied Psychological Measurement*, *14*(1), 1–12. https://doi.org/10.1177/014662169001400101

Ward, W. C., Frederiksen, N., & Carlson, S. B. (1980). Construct validity of free-response and machine-scorable forms of a test. *Journal of Educational Measurement*, *17*(1), 11–29.

Weston, M., Haudek, K. C., Prevost, L., Urban-Lurain, M., & Merrill, J. (2015). Examining the impact of question surface features on students' answers to constructed-response questions on photosynthesis. *CBE—Life Sciences Education*, *14*(2), ar19. https://doi.org/10.1187/cbe.14-07-0110

Wright, L. K., Catavero, C. M., & Newman, D. L. (2017). The DNA triangle and its application to learning meiosis. *CBE—Life Sciences Education*, *16*(3), ar50. https://doi.org/10.1187/cbe.17-03-0046

Wright, L. K., Fisk, J. N., Newman, D. L., & Campbell, A. M. (2014). DNA → RNA: What do students think the arrow means? *CBE—Life Sciences Education*, *13*(2), 338–348. https://doi.org/10.1187/cbe.cbe-13-09-0188