Automated Writing Assessments Measure Undergraduate Learning after Completion of a Computer-Based Cellular Respiration Tutorial

Juli D. Uhl,[†]* Kamali N. Sripathi,[‡] Eli Meir,[§‖] John Merrill,[¶] Mark Urban-Lurain,[†] and Kevin C. Haudek[†#]

[†]CREATE for STEM Institute, Michigan State University, East Lansing, MI 48824; [‡]UC Davis Genome Center, Biomedical Engineering, Davis, CA 95616; [§]SimBiotic Software, Inc., Missoula, MT 59807; [¶]Department of Microbiology and Molecular Genetics, Michigan State University, East Lansing, MI 48824; [#]Department of Biochemistry and Molecular Biology, Michigan State University, East Lansing, MI 48824

ABSTRACT

The focus of biology education has shifted from memorization to conceptual understanding of core biological concepts such as matter and energy relationships. To examine undergraduate learning about matter and energy, we incorporated constructed-response (CR) questions into an interactive computer-based tutorial. The objective of this tutorial is to teach students about matter and energy and help dispel common misconceptions through the context of cellular respiration. We used a constructed-response classifier (CRC) tool to categorize ideas in responses to three CR questions and measure changes in student thinking about cellular respiration. Our data set includes 841 undergraduates from 19 geographically diverse institutions including two-year colleges, primarily undergraduate institutions, and research-intensive colleges and universities. We found students from all institution types included more scientific ideas in CRs post-tutorial. Students used an average of 2.1 ideas in CRs and frequently used both scientific and developing ideas. We found this mixed thinking persisted after the tutorial regardless of institution type. Students' multiple-choice (MC) selections were correlated with their CRs, but CRs revealed more mixed thinking than would be inferred from MC responses. Our study shows a CRC tool can measure student learning after a computer-based tutorial and provides more complete information than MC responses.

INTRODUCTION

Recent science education reforms at the K–12 (National Research Council [NRC], 2012) and undergraduate levels (American Association for the Advancement of Science [AAAS], 2011) focus on concepts and competencies important to multiple biology fields, and more broadly to all science, technology, engineering, and mathematics (STEM) disciplines. As educational foci shift, instructors and students increasingly interact with computer-based learning tools, including online simulations and tutorials. Instructors and instructional designers must assess student learning to measure effectiveness of computer-based learning. We are interested in the intersection of computer-based learning and automated assessment tools as a potential way for instructors and students to assess learning. We used an automated Constructed-Response Classifier (CRC) tool developed by the Automated Analysis of Constructed Response (AACR) research project, to measure changes in student thinking about a core biological concept as a result of completing an interactive computer-based tutorial developed by SimBiotic Software. To do so, we included three constructed-response (CR) questions, in which students must respond to a prompt by writing their answer in their own

John Coley, *Monitoring Editor*

Submitted Jun 25, 2020; Revised Dec 24, 2020; Accepted Apr 27, 2021

CBE Life Sci Educ September 1, 2021 20:ar33

DOI:10.1187/cbe.20-06-0122

[‖]Conflict of interest statement: Eli Meir is the Director of Research at SimBiotic Software, the company that developed the Cellular Respiration Explored tutorial and participated in the study design for this work. This work should not be construed as promotion of a product to the exclusion of other similar products.

*Address correspondence to: Juli D. Uhl (uhljuli@msu.edu).

words in the tutorial as paired pre- and post-tutorial assessments. Because students often learn by adding new scientific ideas without removing nonscientific ideas, we characterized student thinking in terms of combinations of ideas in their responses. To determine whether this tutorial was beneficial for all students, we compared learning among students enrolled in 19 institutions representing three institutional types. Because student thinking often mixes scientific and nonscientific ideas, we sought to identify ideas other than those in students' multiple-choice (MC) responses by comparing student CRs to their MC selections.

BACKGROUND

In focusing on core concepts and competencies, the goal is for all students to learn to make sense of the plethora of biological information that floods their everyday lives, such as discussions of genetically modified foods and the effects of drugs on cellular components. One core concept defined by the AAAS (2011) is "pathways and transformations of matter and energy." This concept requires students to understand matter- and energy-transforming processes at multiple scales, such as the cellular processes that create and break down metabolites, and how specific elements such as carbon, nitrogen, and phosphorus cycle through organisms and ecosystems.

Students have great difficulty understanding and applying these processes, and their difficulties at all levels of education have been extensively documented. First, students are often confused by the vocabulary used to describe these processes, due to the overlap with colloquial language. For example, students often confuse the colloquial and scientific definitions of respiration and other scientific processes (Bell, 1985; Anderson *et al.*, 1990; Driver *et al.*, 1994). Similarly, students exhibit confusion about the term "energy" due to its use in everyday applications (Anderson *et al.*, 1990; Hartley *et al.*, 2012; Jin *et al.*, 2013; Opitz *et al.*, 2017). Hartley and colleagues (2012) also found that students believe that matter and energy are essentially interchangeable in biological contexts through erroneous application of the physical equation $E = mc^2$. Students also exhibit more complicated misunderstandings regarding matter and energy transformations. Jin and colleagues (2013) described a learning progression that characterizes development of students' thinking about these transformations in socio-ecological systems. With respect to matter transformation, students at less-sophisticated levels could not reliably identify that mass can be lost as gas. Students with more sophisticated understanding of matter were able to understand that all three phases are made of matter and were able to apply the law of conservation of matter. With respect to energy transformations, the authors found increasing sophistication of students' explanations of energy. Student explanations ranged from naïve reasoning that only living organisms possess energy to more sophisticated explanations describing chemical bonds and heat energy.

Research has also focused on characterizing the different ideas that students have about matter- and energy-transforming processes. Wilson and colleagues (2006) collected undergraduates' ideas about matter transformation in various contexts, such as human weight loss and mass change in plant photosynthesis, through student essays and interviews and developed MC questions and distractors based on these responses. Their MC options reflected student ideas that were correct (such as correct, but sometimes separate, identification of the products and processes of weight loss) or incorrect (such as that mass is vaguely converted into energy). Sripathi and colleagues (2019) built upon the work of Wilson and colleagues (2006) by characterizing ideas that occur in undergraduates' CRs. The authors identified three commonly occurring ideas in student CRs, which they labeled "Scientific," because these ideas are essential for a scientific, molecular and/or physiological description of human weight loss. These three ideas were 1) the Correct Molecular Product of CO_2; 2) descriptions of the physiological Exhalation process; and 3) the correct Molecular Mechanism of cellular respiration or similar processes. They also described five "Developing" ideas, so called because these ideas, while not entirely incorrect, would need additional context from Scientific ideas to demonstrate a complete understanding of weight loss. Examples of Developing ideas are if the students used the idea of Matter Converted into Energy or if students included informal descriptions of How to Lose Weight, such as calorie output exceeding calorie input or exercise. Sripathi and colleagues used these two types of ideas to develop descriptive models for student written descriptions: Scientific descriptions included only Scientific ideas, while Developing descriptions contained only Developing ideas. Mixed descriptions included at least one Scientific and at least one Developing idea and were quite common in student responses. The work by Sripathi and colleagues (2019) highlighted the complicated ways that students can combine both correct and incorrect ideas about human weight loss mechanisms.

The persistent challenge of learning biology core concepts, including matter and energy, has led to development of a variety of instructional methods, both computer-based and in-class activities. Many of these have been published and demonstrate learning gains (e.g., Bentley and Connaughton, 2017; Bergan-Roller *et al.*, 2017; Freeman *et al.*, 2017; Goff *et al.*, 2018). For these and any instructional method, especially during a time in which education is rapidly adopting online and computer-based learning, it is essential to assess student learning. One cannot otherwise make claims about the utility of the learning activity. For complex concepts, the instructor or instructional designer must carefully consider how students learn and what they wish to assess. Students often do not learn by replacing Developing ideas with Scientific ones all at once; rather students add new ideas onto their existing ideas (Vosniadou, 2012). It is essential to use assessments that elicit a detailed picture of student thinking and can detect changes in the use of these ideas as expertise develops (Opfer *et al.*, 2012). Two commonly used assessment types are described here.

MC questions are highly constrained in their possible responses and are common in large-enrollment introductory courses and computerized assessments (Scalise and Gifford, 2006). While MC questions are favored because they can be rapidly graded, there are drawbacks: MC questions may cause students to rely on memorization (Stanger-Hall, 2012) or use test-taking strategies (Kim and Goetz, 1993). MC questions may also overestimate student understanding (Nehm and Schonfeld, 2008), fail to capture mixed thinking (Brassil and Couch, 2019), or cause students to consider the incorrect alternatives as correct in later testing (Roediger and Marsh, 2005). Thus, MC questions may not provide a complete picture of student thinking.

In contrast, CR questions, which require students to answer in their own words, can better measure complex student thinking patterns and elicit students' own ideas (Nehm and Haertig, 2012; Nehm *et al.*, 2012; Meir *et al.*, 2019). CR questions may be better suited to target authentic competencies and practices of constructing scientific arguments and explanations (NRC, 2014). Well-written CR questions can characterize mixed student thinking (Haudek *et al.*, 2012; Hubbard *et al.*, 2017). Student misconceptions can be better diagnosed by scoring CRs than MC responses (Birenbaum and Tatsuoka, 1987). Within biology contexts, misconceptions identified by student responses to a CR instrument about natural selection were more consistent with student interviews than an MC instrument (Nehm and Schonfeld, 2008). In the context of photosynthesis, text and concept analysis of student CRs aligned to verbal responses in student interviews (Weston *et al.*, 2015). Such results suggest that CRs align with student thinking during interviews and thus provide a richer picture of student thinking necessary for assessing instructional effectiveness and meaningful learning.

Instructors are often reluctant to use CR questions due to the time and cost to reliably grade and interpret responses (Nehm *et al.*, 2012; Nehm and Haertig, 2012). Automated categorization methods aim to overcome these limitations. For example, machine learning algorithms can predict human scores for undergraduate written responses in the subject areas of biology and chemistry with high measures of interrater reliability (IRR; e.g., Urban-Lurain *et al.*, 2009; Ha *et al.*, 2011; Haudek *et al.*, 2012; Moharreri *et al.*, 2014; Prevost *et al.*, 2016). Automated analysis of student writing and associated predicted scores or categorization of ideas was successfully applied to develop and monitor the effects of teaching interventions (Pelletreau *et al.*, 2016). When incorporated into computerized learning modules, automated feedback to question formats with levels of constraint intermediate between MC and fully constructed, such as fill-in-the-blank, matching, or labeling questions (for other types, see Scalise and Gifford, 2006), can be applied to assist in student learning (Meir *et al.*, 2019; Zhu *et al.*, 2020).

As part of assessing student learning, it is essential to consider whether the lesson or computer-learning module is effective for all students. Most educational research studies draw students from research-intensive colleges and universities (RICUs; Schinske *et al.*, 2017). However, as many as half of all students complete at least part of their education at a two-year college (TYC; National Academies of Sciences, Engineering, and Medicine, 2016). Additionally, approximately one-third of college students are enrolled in primarily undergraduate institutions (PUIs) such as baccalaureate colleges or master's colleges and universities (Carnegie Classification of Institutions of Higher Education, 2018). These three broad categories of institutions likely serve different student demographics; on average, TYCs enroll a more diverse and older student population than 4-year institutions such as RICUs and PUIs (Hussar *et al.*, 2020). Thus, comparing learning of students from different institutional types helps evaluate whether an instructional tool benefits all students.

RESEARCH QUESTIONS

This study examines how undergraduate student thinking changes in response to an interactive, computer-based tutorial. The tutorial was developed by SimBiotic Software and focuses on the core concept of matter and energy transformations during cellular respiration, framed in the context of exercise (Kim *et al.*, 2014). To assess student learning, we used the CRC tool developed by the AACR research group to automatically categorize ideas students include in their CRs pre- and post-tutorial (beyondmultiplechoice.org). Our data were collected from students from 19 institutions, including TYCs, PUIs, and RICUs. By relating the ideas contained in student CRs with their MC selections and institutional types, we investigate the following research questions:

Research Question 1. How do student descriptions about cellular respiration change after completion of an interactive computer-based tutorial focused on cellular respiration?

Research Question 2. Do learning gains vary among students from different institution types?

Research Question 3. How do ideas included in student CRs correspond to their MC selections?

RESEARCH METHODS

Data Collection

Student responses were collected and de-identified from consenting students in classes that opted to use a research version of SimBio's Cellular Respiration Explored tutorial. Instructors were invited to opt into the study by emails to current SimBio users and two webinar recruitment sessions. Invitations to the webinar were sent via email to a SimBio mailing list. The email invited recipients to a 30-minute webinar hosted by SimBio, with two authors (K.C.H., E.M.) as presenters. At the webinar, the authors presented the Cellular Respiration Explored tutorial and basics of the CRC tool. The presentation also briefly outlined expectations for instructors participating in the pilot study (e.g., assign questions pre- and post-instruction). Interested participants were directed to a URL to sign up to use this special version of Cell Respiration Explored free of charge in their courses. Instructors were provided with documentation so they could use the CRC tools to generate their own reports from their students' CRs about cellular respiration before and after completion of the tutorial.

Nineteen classes from a variety of institutional types and geographic locations were involved in the research study. Class sizes ranged from seven to 344 students (mean class size: 50; SD: 72). We grouped students based on Carnegie Classifications (Carnegie Classifications of Institutions of Higher Education, 2018) into three institutional types for this study: TYCs include institutions designated as community colleges, PUIs include public and private master's and baccalaureate institutions, and RICUs include public and private doctoral-granting universities and colleges. Of the 998 students involved in the study, we included in the present analysis 841 students who fully completed both the pre- and post-tutorial CR assessments used to examine student thinking (Table 1). Data collection was completed by the SimBiotic Company and data analysis completed by researchers at Michigan State University. The study was reviewed and approved by the New England Institutional Review Board (IRB no. 120160152) and designated exempt by Michigan State University's Institutional Review Board (IRB x10-577).

CR Items and Automated Analysis Models

Three CR questions were incorporated into a research version of SimBio's computer-based Cellular Respiration Explored tutorial (Kim *et al.*, 2014). For details on the tutorial, including screenshots, see Supplemental Material and Supplemental Figure 1. This version of the tutorial began with the three CR questions as a pre-tutorial assessment and included 51 feedback questions throughout the tutorial (both MC and intermediate-constraint formats) and a post-tutorial 10-question MC quiz, followed once more by the three CR questions. Two of the CR questions assess conceptual understanding of the core biological concept of transformation of matter and energy (AAAS, 2011). These questions were originally developed as part of a diagnostic question cluster to assess student ability to trace matter and energy across scales (Wilson *et al.*, 2006). We refer to the first question as the Weight Loss question, "You have a friend that lost 15 lbs. on a diet. Where did the mass go?" The second we refer to as the Energy from Glucose question, "You eat a sweet and juicy grape. Explain how a molecule of glucose from that grape can be used to move your little finger." We also included a third question, titled the Enzyme Binding question, which asks "Enzymes help in chemical reactions in living organisms. How would a molecular biologist explain the mechanism that helps an enzyme to bind to its correct substrate and reduces the possibility of incorrect interactions?" The Enzyme Binding question was included as a control, as it targets the structure and function core concept (AAAS, 2011), which is not included in the Cellular Respiration Explored tutorial.

Each of the three CR questions has an associated automated categorization model developed by the AACR research group. These and other models are available for use at beyondmultiplechoice.org (Beyond Multiple Choice, n.d.). We used these three models to characterize ideas included in student responses to each of the CR questions, pre- and post-tutorial. All three models are based on analytic scoring rubrics, wherein each category captures one idea and student responses may contain zero, one, or more ideas and ideas can co-occur. The automated categorization model for the Weight Loss question captures a total of eight ideas. For this work, the Carbon Alone category was not included, as this represents a very rare case of student language (see Sripathi *et al.*, 2019). The automated categorization model for the Weight Loss question used in this work captures seven ideas (Table 2). The Cohen's kappa measure of IRR between human codes and predictions made by the Weight Loss automated categorization model range from 0.700 to 0.976, which is considered substantial to almost perfect agreement (Cohen, 1960; Landis and Koch, 1977). The automated categorization model for the Energy from Glucose question captures five ideas in student responses (Supplemental Table 1) and predicts human scores for each category with IRR ranging from 0.364 to 0.861. The automated categorization model for the Enzyme Binding question is capable of capturing 11 distinct ideas in student responses (Supplemental Table 2) and predicts human scores for each category with IRR ranging from 0.684 to 0.955.

For the purposes of the present analysis of responses to the Weight Loss CR question, we adopt the language from Sripathi *et al.* (2019), who defined a Scientific idea as being characteristic of a molecular or physiological description of the processes and products of cellular respiration. A Developing idea is defined as an idea that represents a misconception (i.e., Matter Converted to Energy), a partially correct idea, or informal ideas about dieting (i.e., How to Lose Weight). In application of human codes, Sripathi *et al.* (2019) considered the rubric categories General Metabolism and Molecular Mechanism as mutually exclusive. However, the automated categorization model is not programmed to use rules that consider any pair of categories as mutually exclusive, although it is possible that these rules may be implicitly "learned" by the algorithms during the training process. Thus, for this work, we consider any pair of categories as potentially co-occurring.

Data Analysis

We used the AACR CRC tool to assign a score of presence (1) or absence (0) for each of the ideas in students' pre- and post-tutorial responses for each of the three CR questions. When responses were categorized as 0 for all ideas by the CRC tool, student response text was not relevant to any of the categories and was classified as "none."

Research Question 1. To analyze overall trends in student responses to all three questions, we tallied the total number of responses containing each idea pre- and post-tutorial. For responses to the Weight Loss question, we took advantage of the paired nature of the data to further categorize whether students changed the ideas they included in responses after completing the tutorial. Students who did not include a given idea in either their pre- or post-instruction response were categorized as "idea never used." Students who included a given idea pre-instruction, but not post-instruction were categorized as "idea removed." Students who included a given idea in both their pre- and post-instruction responses were categorized as "idea maintained." Students who did not include a given idea pre-instruction and included it in their post-instruction responses were categorized as "idea added." To compare the proportion of each analytically categorized idea in student responses to the Weight Loss CR question between pre- and post-tutorial, we performed McNemar's test of correlated proportions using SPSS v. 24.

To analyze student thinking as revealed by student responses to the Weight Loss item, we applied three descriptive thinking models (Scientific, Developing, or Mixed) described by Sripathi *et al.* (2019). The Scientific descriptive model is defined as a response that includes one or more Scientific ideas and no Developing ideas. The Mixed descriptive model is defined as a student response that includes one or more each of Scientific and Developing ideas. The Developing descriptive model is defined as a student response that includes one or more Developing ideas and no Scientific ideas. We identified the descriptive model for each pre- and post-tutorial student response so that we could track and visualize change in student thinking with a Sankey diagram (Bogart, n.d.).

Research Question 2. To analyze data for research question 2, we compared ideas categorized in student responses among students from different institutional classifications (Table 1). To compare the number of Scientific and Developing ideas across the three institutional types, we performed an analysis of variance (ANOVA) with permutation testing (see LaFleur and Greevy, 2009), using R version 3.5.2 (R Core Team, 2020)

TABLE 1. Number of classes and students (in parentheses) who responded to all three CR questions both pre- and post-tutorial by institution type

	TYC	PUI	RICU	Total
Public	3 (69)	5 (134)	4 (405)	12 (608)
Private	—	3 (78)	4 (155)	7 (233)
Total	3 (69)	8 (212)	8 (560)	19 (841)

implemented in R Studio v.1.1.463 (RStudio Team, 2020) and the ri2 package developed by (Coppock, 2020).

Research Question 3. To analyze data for research question 3, we compared ideas and descriptive thinking models in post-tutorial CR Weight Loss responses to selections made to a related MC question in the Cellular Respiration Explored post-tutorial quiz. The MC question targets molecular understanding of cellular respiration through understanding how weight loss occurs and states, "Some students have the misconception that during cellular respiration, the matter in glucose is somehow turned into energy. Consider that when we exercise, we burn glucose and also lose mass. Why does this happen?" Students could select from among four options, including a correct response and three distractors. Trained coders for the Weight Loss CR question used the Weight Loss rubric to assign each of the MC options to a single rubric category. Among the MC options, the correct response—"Our cells convert glucose into CO_2 and water, which are eliminated from our bodies when we exercise."—was coded as Correct Molecular Products; two distractors—"Our cells convert the mass in glucose into energy, which is weightless." and "Our cells convert the mass in glucose into energy that is used during exercise. Losing that energy reduces our mass."—were coded as Matter Converted to Energy; and one distractor—"Our cells use up the potential energy stored in glucose and losing that energy during exercise reduces our mass."—was coded as General Metabolism. We grouped student written responses to the corresponding CR question based on the coded categories of their MC selections. Responses from the two Matter Converted to Energy distractors were combined. To measure how frequently CR categories co-occurred with MC options, we calculated phi coefficients using SPSS v. 24.

RESULTS

Research Question 1. How Do Student Descriptions about Cellular Respiration Change after Completion of an Interactive Computer-Based Tutorial Focused on Cellular Respiration?

We used the CRC tool at beyondmultiplechoice.org to generate predictions for student CRs before and after completion of the Cellular Respiration Tutorial; two CR questions assess the concept of matter and energy and the other the concept of structure and function (per AAAS, 2011). Based on the content of the tutorial, we hypothesized that student CRs would include more Scientific ideas related to matter and energy but would be largely unchanged about structure and function after the tutorial.

To test this, we examined the percentage of student responses including the categories for each of the CR questions pre- and post-tutorial (Figure 1). In pre-tutorial responses to the Weight Loss question, the CRC tool predicted most ideas to occur in 20% or fewer of the student responses, with the exception of the documented misconception that Matter Is Converted to Energy (see Wilson *et al.*, 2006), which occurred in 41% of student responses. We see a similar trend in responses to the Energy from Glucose question; most ideas occur in fewer than 20% of responses, with the exception of the matter to energy misconception captured by the Sugar Converted to Energy category, which occurred in 37% of student responses. Thus, many students began the tutorial with this misconception.

In post-tutorial responses to the two CR questions targeting the concept of matter and energy, ideas related to the misconception that Matter or Sugar Is Converted to Energy were included about half as often as in pre-tutorial responses (Figure 1, A and B). Additionally, inclusion of Scientific ideas increased after the tutorial; for the Weight Loss question, the two most frequently observed categories included the Scientific ideas Correct Molecular Products and Exhalation. Likewise, in post-tutorial responses to the Energy from Glucose question, the most frequently observed category captures the Scientific

TABLE 2. Rubric categories and descriptions for the automated categorization model used to identify ideas in student responses to the Weight Loss question[a]

Rubric category	Brief description	Cohen's kappa[b]
Correct Molecular Products[c]	Responses in this category include the idea that the products of cellular respiration, primarily carbon dioxide in any form (e.g., CO_2, carbon dioxide) are the result of mass loss.	0.976
Exhalation[c]	Responses in this category include the idea that excess mass is exhaled or exits the body.	0.892
Molecular Mechanism[c]	Responses in this category include the idea that mass loss occurs due to correct molecular processes (e.g., cellular metabolism, beta oxidation), or describe these processes in specific detail.	0.775
General Metabolism	Responses in this category include the idea that mass loss occurs due to some kind of molecular conversion, even if it is only partially correct.	0.700
Matter Converted to Energy	Responses in this category include the idea that mass loss occurs through vague conversions from matter to energy.	0.827
Excretion	Responses in this category state that the mass is excreted out of the body. Responses must specifically indicate the physiological process of excretion by explicitly using the term "excreted" or similar or indicating physiological waste in their responses.	0.832
How to Lose Weight	Responses in this category include ideas about societal discussions of weight loss, such as "calories in" greater than "calories out" or exercise.	0.806

[a]Adapted from Sripathi *et al.* (2019).
[b]Cohen's kappa (Cohen, 1960) measure of IRR between human scorers and CRC predictions.
[c]Scientific ideas as defined by Sripathi *et al.* (2019).

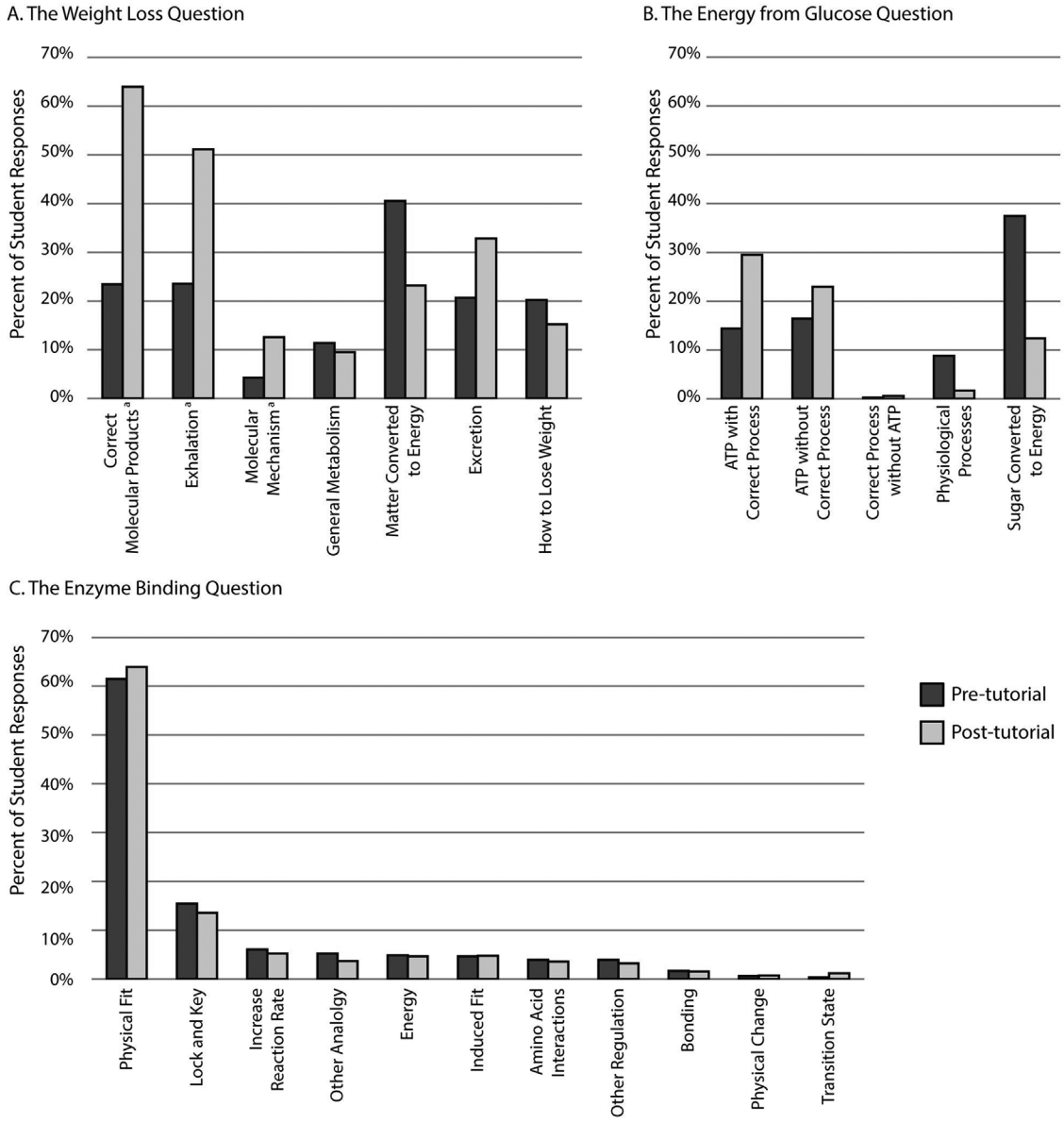

FIGURE 1. Ideas included in student responses to the three CR questions. Percent of pre- and post-tutorial student responses that include the ideas categorized by the automated predictive models for the three CR questions. (A) Weight Loss and (B) Energy from Glucose, which are ordered from most to least scientific, and (C) Enzyme Binding, which is ordered by occurrence. *n* = 841. [a]Scientific ideas as defined by Sripathi *et al.* (2019).

idea of ATP with Correct Process. Thus, in ideas related to the concept of matter and energy, we see an overall increase in Scientific ideas and a decrease in misconceptions after completing the tutorial.

We examined the responses to the Enzyme Binding question for differences between student pre- and post-tutorial descriptions of the concept of structure and function. In responses to the Enzyme Binding question, the most common category is Physical Fit both pre- and post-tutorial. Neither this nor any other category had an obvious difference of ideas included in pre- and post-tutorial responses (Figure 1C). We take these differences in student responses to the matter and energy questions but not to a structure and function question as evidence that student thinking about matter and energy changes in response to the tutorial. Next, we look more closely at student thinking about matter and energy by examining student responses to the Weight Loss question.

To examine changes in student thinking as a result of the tutorial, we compared paired pre- and post-tutorial responses from the same student. Each student was categorized based on the occurrence of the ideas in pre- and post-tutorial responses as: never used pre- or post-tutorial, removed after tutorial, maintained both pre- and post-tutorial, or added after the tutorial (Figure 2). Six of seven categorized ideas had strong evidence of changed proportions between their pre- and post-tutorial responses (McNemar test of correlated proportions, with p values < 0.005). The only idea that did not change from pre- to post-tutorial was General Metabolism ($\chi^2 = 1.974$, $p = 0.160$).

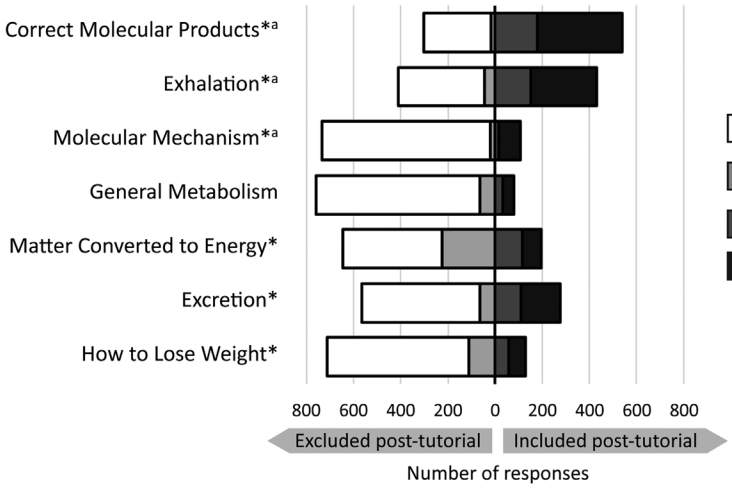

FIGURE 2. Ideas about weight loss change after completion of the tutorial. Compared with their pre-tutorial responses, students add Scientific ideas and the Excretion idea and remove the ideas How to Lose Weight and Matter Is Converted to Energy from their responses after completing the tutorial. $n = 841$. *$p < 0.005$, McNemar test of correlated proportions. [a]Scientific ideas as defined by Sripathi *et al.* (2019).

Research Question 2. Do Learning Gains Vary Among Students from Different Institution Types?

Because we had responses from a variety of institutions, we wanted to determine how students from varying institution types learned after the tutorial. We hypothesized that students from all institutional types would benefit similarly from the tutorial. We divided the responses into three groups—students at TYCs, PUIs, and RICUs—and investigated the numbers and types of ideas used by each student group. We found that students from all three institution types used similar numbers of Developing ideas both pre- and post-tutorial and similar numbers of Scientific ideas pre-tutorial but differed in number of Scientific ideas post-tutorial (Table 3). Despite the apparent difference in Scientific idea inclusion post-tutorial, the magnitude of the difference was small and similar to other magnitudes in difference ($\eta_p^2 = 0.01$). Overall, we conclude that students from all three institution types use similar numbers of Scientific and Developing ideas before and after the tutorial.

To compare student learning across institution types, we compared the difference between the number of Scientific and Developing ideas in student responses from pre- to post-tutorial. Students from all institution types tend to add Scientific ideas (Figure 4A) and remove or maintain Developing ideas (Figure 4B). While there is strong evidence that there is a

Because it appears that four of the categories with changed proportions primarily result from students adding ideas, we examined the average normalized gain in these ideas (Hake, 1998). We found medium gains in two ideas, Correct Molecular Products (<g> = 0.53) and Exhalation (<g> = 0.36), and small gains in adding the Scientific idea Molecular Mechanism (<g> = 0.09) and the Developing idea Excretion (<g> = 0.15). Normalized gains together with the pre/post data shown in Figures 1 and 2, suggest that students tend to add all three Scientific ideas and the Developing idea of Excretion and remove misconceptions as a result of the tutorial.

To further examine student thinking before and after completion of the tutorial, we applied the student descriptive models proposed by Sripathi *et al.* (2019). Student thinking may include a combination of Scientific and Developing ideas (Mixed descriptive model), fully Scientific, or fully Developing descriptive models. Taking advantage of the paired nature of these data, we tracked changes in individual student descriptive models after completion of the tutorial (Figure 3). The most common descriptive model pre-tutorial was Developing (55%), with Scientific, Mixed, and responses categorized as No descriptive model occurring with similar frequency (15%, 14%, 17%, respectively). Post-tutorial, the two most common descriptive models were Scientific (33%) and Mixed (39%). The largest increase in these categories came from students who used Developing descriptions pre-tutorial. More than half of the students who included only Developing ideas pre-tutorial added at least one Scientific idea to their responses post-tutorial. Finally, we note that the frequency of the Mixed descriptive model post-tutorial indicates that many students added Scientific ideas to their responses without removing Developing ideas.

In summary, we found that thinking about matter and energy became more Scientific after the tutorial. These results suggest that student learning about energy and matter as evidenced by changed ideas identified with a CRC tool is directly related to completion of the tutorial.

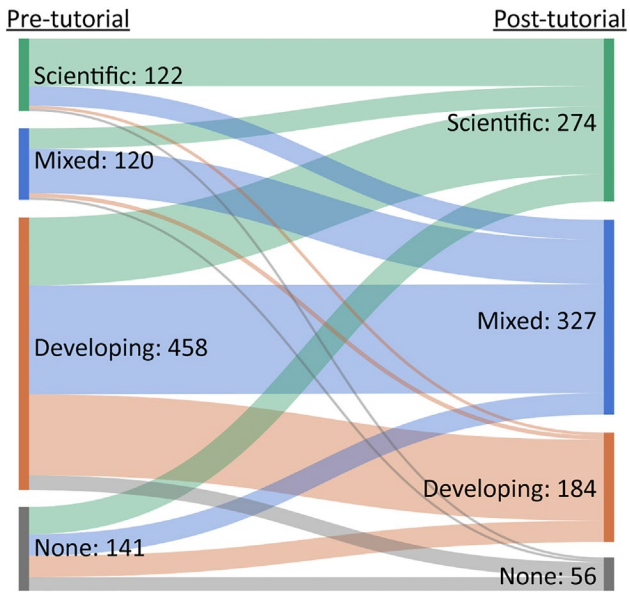

FIGURE 3. Student descriptive models change after completing the tutorial. Paired student responses categorized into descriptive models according to Sripathi *et al.* (2019). Most student responses are classified as a Developing descriptive model pre-tutorial and post-tutorial most are the Mixed or Scientific descriptive model. $n = 841$.

TABLE 3. Students from three institutional types include similar average numbers of ideas included in their responses to the Weight Loss question

	TYC $n = 69$		PUI $n = 212$		RICU $n = 560$		$F(2, 848)$	η_p^2
	M	SD	M	SD	M	SD		
Pre-tutorial Scientific Ideas	0.48	0.815	0.66	0.949	0.46	0.824	4.20	0.01
Post-tutorial Scientific Ideas	1.12	0.932	1.17	0.897	1.34	0.964	5.78*	0.01
Pre-tutorial Developing Ideas	0.99	0.696	0.81	0.792	0.97	0.790	3.38	0.01
Post-tutorial Developing Ideas	0.96	0.736	0.81	0.823	0.79	0.776	2.62	0.01

*$p < 0.05$, ANOVA.

difference in gains in Scientific ideas between three institutional types, $F(2, 838) = 12.46$, $p = 0.0005$, the effect size is small; $\eta_p^2 = 0.03$. There is some evidence that there is a difference in change of Developing ideas between three institutional types, $F(2, 838) = 4.39$, $p = 0.0601$, and again the effect size is small; $\eta_p^2 = 0.01$. It appears that RICU students added more Scientific ideas than either PUI or TYC students. To see whether a single RICU class with high student enrollment was responsible for this apparent difference in Scientific ideas added, we looked at course-level results and found that all RICU courses show similar patterns of gains in Scientific ideas (unpublished data). We conclude that the apparent larger gain in Scientific ideas for RICU students does not represent a large difference in true effect and is likely a combination of the slightly lower number of pre-tutorial Scientific ideas with the slightly higher post-tutorial Scientific ideas in these student descriptions.

To further compare student thinking across institution types, we examined student descriptive models in responses to the Weight Loss question across all three institution types (Supplemental Table 3) and found the most common descriptive model used by students from any of the three institution types in pre-tutorial responses was Developing. For all students, the most common post-tutorial descriptive model was Mixed. These trends suggest that students learn similarly across institution types.

Research Question 3. How Do Ideas Included in Student CRs Correspond to Their MC Selection?

The Cellular Respiration Explored tutorial includes a MC post-tutorial quiz relating to the concepts learned in the tutorial. This quiz was not included as a pre-assessment in the tutorial; thus, we cannot calculate learning gains based on the MC quiz. However, students performed well on the quiz, with a median score of 80% (range 0–100%). After this quiz, students were prompted to complete the CR questions. One of the MC questions assesses understanding of the relationship between glucose and energy during weight loss, similar to concepts assessed by the Weight Loss CR question. We examined the categories predicted in each student's Weight Loss CR based on the student-selected MC option. Although students responded to the MC question first, most included more than one idea in their written responses (Figure 5). We found that students who made the correct MC selection included an average of 2.2 ideas and students who selected one of the distractors or who made no selection included an average of 1.7–1.8 ideas.

Next, we asked whether the CR ideas were correlated with students' MC selections (Table 4), beginning with students who selected the correct MC option. We found that the MC option representing Correct Molecular Products was positively and significantly related to the Scientific ideas Correct Molecular Products and Exhalation and the Developing idea of Excretion. The correct MC choice was negatively associated with the Developing idea of Matter Converted to Energy, and only 19% of students making the correct MC selection included this idea (see Supplemental Table 4). We found the inverse in CRs from students who selected the MC option representing the Matter Converted to Energy idea. We found a positive relationship with the Matter Converted to Energy idea in students' CRs, and a negative relationship with the Developing idea of Excretion and the Scientific ideas Correct Molecular Products and Exhalation.

Importantly, the MC option selected was not a perfect indicator of ideas included in CRs. In CRs from students who selected the MC option representing the General Metabolism idea, there was a negligible, nonsignificant relationship with the same idea. These students' CRs were positively associated with the Developing Matter Converted to Energy idea and negatively associated with the Scientific ideas of Correct Molecular Products and Exhalation.

Moreover, some students who selected the correct MC option wrote responses to the subsequent CR question that contained Developing ideas, like Matter Converted to Energy (19%) or How to Lose Weight (15%; Supplemental Table 4). Likewise, some students who selected the MC option representing the Matter Converted to

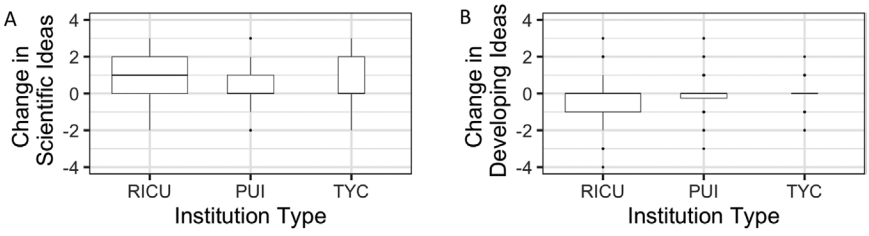

FIGURE 4. Students use different numbers of Scientific and Developing ideas after completing the tutorial. (A) Students from all three institution types add Scientific ideas in CRs after completion of the tutorial. (B) Students from all three institution types use similar numbers of Developing ideas pre- and post-tutorial. TYC, $n = 69$; PUI, $n = 212$; RICU, $n = 560$. Width of boxes is proportional to number of students in each group; box shows median and quartiles; whiskers are 1.5 times the interquartile range; outliers shown as dots.

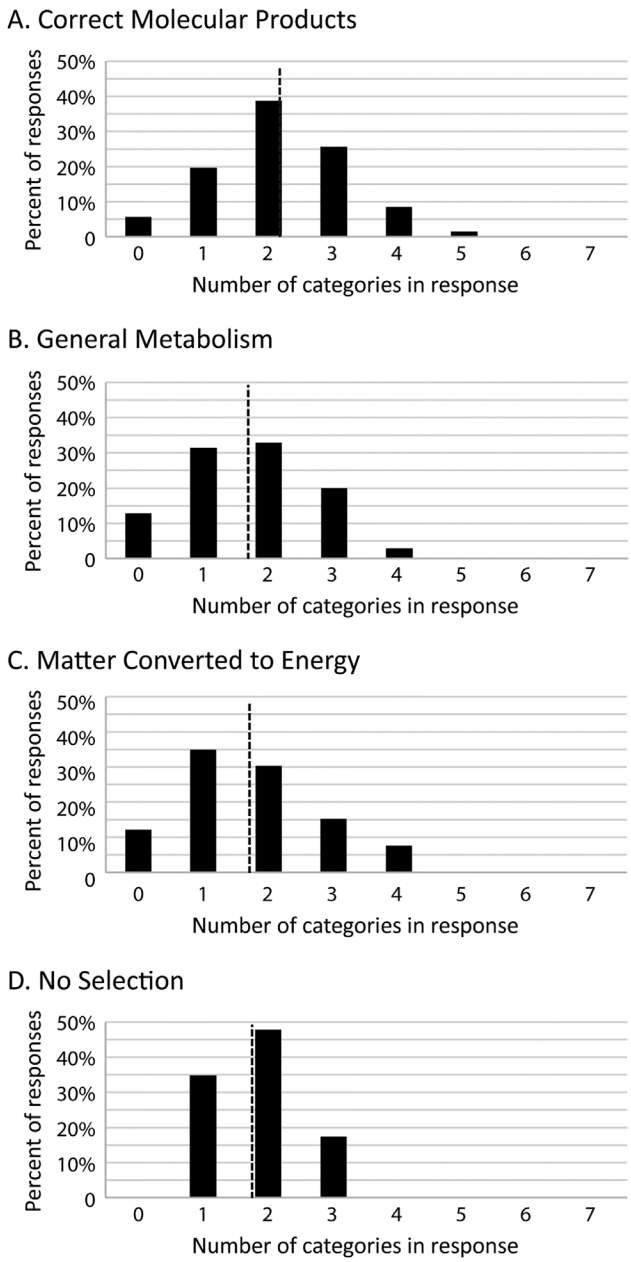

FIGURE 5. Students include multiple ideas in CR regardless of MC selection. Histograms of numbers of ideas in post-tutorial CRs based on MC selection. Students who selected the MC selection (A) Correct Molecular Products, *n* = 678; (B) General Metabolism, *n* = 70; (C) Matter Converted to Energy, *n* = 66; (D) no selection, *n* = 23. Dashed lines indicate the average number of ideas in CRs.

Energy misconception included Scientific ideas like Correct Molecular Products (35%) or Exhalation (29%) in their CRs. This mix of ideas is reflected in students' descriptive thinking models, where the most frequent descriptive model used by students who chose the correct MC option was Mixed, indicating inclusion of both Scientific and Developing ideas (Supplemental Table 5). Thus, while most students can identify the correct MC option, they exhibit mixed thinking about Cellular Respiration, which is most evident in their CRs.

DISCUSSION

This study examined undergraduate student thinking about cellular respiration after completing an interactive, computer-based tutorial. We found that student responses became more expert-like as evidenced by addition of Scientific ideas about cellular respiration. In two CR questions targeting the core concept of matter and energy, fewer students used a Developing idea related to a common misconception (Matter Converted to Energy) post-tutorial. Student CRs to a question on the core concept of enzyme structure and function, used as a control, changed very little after completion of the tutorial. This suggests student learning about cellular respiration was specifically due to concepts presented in the tutorial.

After Completion of the Tutorial, Student Thinking Becomes More Scientific

Studies of student learning in many STEM disciplines, including biology, show that students often hold both Scientific and Developing ideas as they learn new content, and that nonscientific ideas may persist for some time (e.g., Couch *et al.*, 2018; Hartley *et al.*, 2011; Moharreri *et al.*, 2014; Shtulman and Valcarcel, 2012; Vosniadou and Brewer, 1992). The assessments and automated categorization models based on the rubric developed by Sripathi *et al.* (2019) detected students' Mixed models of thinking, along with completely Scientific (and completely Developing) models.

We found that most student CRs for the Weight Loss question were more expert-like after the tutorial. More students included Scientific ideas (Correct Molecular Products and Exhalation) post-tutorial. These changes are consistent with expected learning outcomes targeted by the Cellular Respiration Explored tutorial. Few students (13%) included the Scientific idea of Molecular Mechanism post-tutorial. The tutorial includes the four processes of cellular respiration (glycolysis, pyruvate processing, the citric acid cycle, with a focus on the electron transport chain), which would all be categorized as Molecular Mechanism in a CR. Perhaps students completing the tutorial do not consider metabolic processes that might be categorized as a Molecular Mechanism to be important and do not include them in their responses.

Post-tutorial, fewer students used the Matter Converted to Energy misconception in CRs to the Energy from Glucose and Weight Loss questions. However, the wording of the MC question labels the Glucose Converted to Energy idea as a misconception. One interpretation of this reduction in frequency of this misconception in CRs is that students used the MC question wording as a hint. However, that does not explain why students also added Scientific ideas to their CRs for these two questions. Further, while the coding rubric for the Energy from Glucose question directly captures the Sugar Converted to Energy misconception, the coding rubric for the Weight Loss question captures a broader usage of the misconception. For example, this category also includes language about burning fat for energy or mass leaving the body as heat. We conclude that, while students might use the MC wording as a hint toward a correct response, this does not completely explain more Scientific responses.

Consistent with studies showing nonscientific ideas persist (e.g., Price *et al.*, 2016), not all changes in student responses were consistent with a more expert-like explanation. We found that students added the Developing idea of Excretion. We also

TABLE 4. Phi coefficient for ideas included in student CRs associated with MC selections

	Ideas represented by MC option[a]		
CR rubric category	**Correct Molecular Products[c]**	**General Metabolism**	**Matter Converted to Energy**
Correct Molecular Products[b]	0.270**	–0.159**	–0.177**
Exhalation[b]	0.172**	–0.084*	–0.139**
Molecular Mechanism[b]	0.064	–0.050	–0.044
General Metabolism	–0.050	0.049	0.026
Matter Converted to Energy	–0.231**	0.089*	0.196**
Excretion	0.118**	–0.055	–0.082*
How to Lose Weight	–0.007	–0.032	0.012

[a]MC selections: Correct Molecular Products, $n = 678$; General Metabolism, $n = 70$; Matter Converted to Energy, $n = 66$.
[b]Scientific ideas as defined by Sripathi *et al.* (2019).
[c]Correct MC option.
*$p < 0.05$.
**$p < 0.005$.

note that Sripathi *et al.* (2019) regarded the Excretion idea as highly context dependent and perhaps not representing a fully nonscientific idea.

Therefore, we examined the language of the formative assessments included throughout the tutorial. One of these items (called "LabLibs"; Meir *et al.*, 2019) states, "When your body uses glucose for energy during exercise, you weigh less afterward. Why?" The correct response, with student-selected options underlined, is: "You weigh less because glucose is transformed into carbon dioxide and water that is lost through exhaling and sweating." This response would be categorized by the automated categorization models used in this study as containing the Correct Products, Exhalation, and Excretion ideas. In this context, water lost through sweating (categorized as Excretion) is part of a Scientific description. This is not always the way students use the idea of Excretion in CRs (see Wilson *et al.*, 2006; Sripathi *et al.*, 2019). For this reason, we do not consider the addition of the idea of Excretion in post-tutorial responses strictly indicative of either Developing or Scientific thinking among students.

The Cellular Respiration Explored Tutorial Aids in All Students' Learning

Calls to include more diverse student populations in educational research studies include studying students at TYCs, which enroll a different demographic than 4-year institutions like RICUs and PUIs (Schinske *et al.*, 2017). Using data from multiple institution types, we found evidence that this tutorial is effective for increasing the number of Scientific ideas included in student responses from all three institution types. On average, students from all institutional types included more Scientific ideas after completion of the tutorial, while they maintained a similar number of Developing ideas. Thus, we suggest that the tutorial is beneficial for students from all institutional types. However, we are cautious in interpreting this as reflective of the entire population of students from any institutional type, as we do not have information about how instructors used the tutorial in class or the resources students used when completing the assessments. Despite this limitation, we argue that completion of the tutorial promotes an increase in use of Scientific ideas in an explanation about the cellular processes involved in weight loss for all students.

Student Ideas in CRs Are Not Limited by the Options of an MC Question

Because previous studies demonstrated that MC questions may not detect when students exhibit mixed thinking (e.g., Wilson *et al.*, 2006; Nehm and Schonfeld, 2008), we were interested in examining the relationship between student answers to related MC and CR questions. The post-tutorial assessment included an MC and a CR question about the same concept (cellular respiration) in the same context (weight loss). Students included ideas in their CRs that align with MC options other than the ones they selected, as well as additional ideas. For example, some students who selected the correct MC option included ideas from incorrect distractors (e.g., Matter Converted to Energy or How to Lose Weight) in their CRs. This is consistent with other studies comparing MC questions to other question types (Parker *et al.*, 2012; Couch *et al.*, 2018). We also found that about one-third of students who selected one of the MC distractors based on a common misconception included Scientific ideas in their CRs. This suggests that MC results alone may also underestimate students' ability or learning and label their response as "wrong" or "incorrect." Thus, MC responses may mask mixed or incomplete understanding.

This emphasizes a benefit of using CR questions to capture mixed student thinking. Some MC writing guidelines suggest the use of as many functional distractors as are feasible and that distractors include plausible ideas validated by other measures (e.g., Haladyna and Downing, 1989). This is challenging to do for complex processes like cellular respiration during weight loss, because a fully Scientific description includes molecular and physiological processes at multiple biological scales, as well as accurate identification of the molecules produced.

Implications for Educators

Instructors and instructional designers must consider which types of questions to include in assessments. Instructors may consider including and using a combination of highly constrained (e.g., MC) and limited-constraint questions (e.g., CR) in their assessments. Including CR questions can help illustrate the complex ways students think about weight loss and other biological concepts and improve instructors' capacity to diagnose student thinking and instructional efficacy. Similar to Sripathi *et al.* (2019) we found that student thinking about cellular respiration is often represented by Mixed descriptive

models, in which scientific and non-scientific ideas coexist in the learner.

This nuanced knowledge about student thinking may be beneficial when designing instructional practices or targeted feedback for students as part of formative assessment practices (Black and Wiliam, 2009). Providing students feedback about their performance along with guidance for improvement as part of formative assessment can lead to improved learning gains (Wiliam, 2011) and reduced achievement gaps (Freeman *et al.*, 2007; Pennebaker *et al.*, 2013). During self-assessments, students who use scores only from MC questions may overestimate their understanding, as they have little feedback to indicate whether they maintain a misconception or Developing idea. However, students who are aware that they also used Developing ideas as part of their explanations receive a more accurate assessment of their own understanding and can use that to improve and focus their study. For example, Lee *et al.* (2019) supplied individualized, automated feedback to students, which led students to revise and improve their written scientific arguments.

Other work supplements the idea that adding intermediate-constraint items to computer-based tutorials improves student learning. A study by Meir *et al.* (2019) found that intermediate-constraint questions improved evolutionary understanding, as measured by pre- to post-tutorial changes. Automated categorization models like those used in this study ease some time constraints on instructors by reducing grading time while still allowing capture of Mixed student thinking. We recommend using the Weight Loss question and other CRC-associated questions as formative assessments. With recent advances in computer scoring of CR questions, it is now feasible to give automated feedback to students (Linn *et al.*, 2014) and has been demonstrated in some specific cases (Nakamura *et al.*, 2016; Gerard *et al.*, 2019; Lee *et al.*, 2019). Thus, we suggest that the Weight Loss CR question and other similar questions with automated categorization models can be used to provide targeted feedback to students.

Students' descriptions about the mechanisms of cellular respiration became more Scientific as a result of completing the Cellular Respiration Explored tutorial. This shows that well-structured, targeted instructional aids can be used to assist student learning. The Cellular Respiration Explored tutorial includes formative assessments with targeted feedback to help reinforce Scientific ideas while addressing Developing ideas or misconceptions. This aligns with studies showing that assignment of formative assessments improves learning for students from multiple institutional types (Freeman *et al.*, 2011; Orr and Foster, 2013; Pape-Lindstrom *et al.*, 2018). Goff *et al.* (2018) demonstrated learning gains from supplementing lectures with a different online learning module on cellular respiration. Importantly, our results also show that a single instructional intervention (i.e., an online tutorial) is unable to guide all students to provide completely Scientific explanations. Development of expertise by students likely requires instructors to incorporate multiple varied instructional interventions over longer periods of time.

Instructors and instructional designers can use published resources (e.g., Bentley and Connaughton, 2017; Bergan-Roller *et al.*, 2017; Freeman *et al.*, 2017) or they may create their own interventions to help students in learning this and other core biological concepts. The CR questions and associated automated categorization models developed by our group are freely available (at beyondmultiplechoice.org) for instructors to include in their courses and use in similar ways to measure effects of instructional resources or interventions.

Limitations

While the current data set is geographically diverse, all institutional and student data were de-identified before analysis, and we cannot provide student demographic information. As noted earlier, we do not know the administrative conditions nor instructional sequencing of the tutorial in each course. We did not collect any data about how students interacted with the tutorial, such as view time or progression through the tutorial, for this study. Therefore, we cannot make claims about exactly how students interacted with the formative questions and the tutorial.

A possible limitation of the CRC tool is that it was developed using student responses from RICUs and may not accurately predict all ideas written by students from other populations. Currently, we are analyzing the performance of the CRC tool for responses by students from different institutional types.

Although students showed substantial increases in Scientific ideas, we do not know from these data how long those changes may persist. There is extensive literature on the difficulty of promoting long-term conceptual change (e.g., Tanner and Allen, 2005; diSessa, 2006; Chi, 2008; Duit *et al.*, 2008; Linn, 2008; Maskiewicz and Lineback, 2013), and instructors should ideally evaluate longer-term change by assessing key concepts after additional time has passed since the related instruction.

ACKNOWLEDGMENTS

The authors thank the Automated Analysis of Constructed Response (AACR) collaboration for helpful conversations while preparing this manuscript, especially Drs. Megan Shiroda, Alex Lyford, and Jennifer Kaplan. We also thank the SimBio content team for preparing a version of the Cellular Respiration Explored tutorial to be used in this study and thank participating instructors and students. We also thank the editor and two anonymous reviewers for their helpful comments. This material is based upon work supported by the National Science Foundation (DUE 1323162). Details about the questions used in this study and accompanying predictive models to score new student responses, can be found at beyondmultiplechoice.org.

REFERENCES

American Association for the Advancement of Science. (2011). *Vision and change in undergraduate biology education: A call to action*. Washington, DC.

Anderson, C. W., Sheldon, T. H., & Dubay, J. (1990). The effects of instruction on college nonmajors' conceptions of respiration and photosynthesis. *Journal of Research in Science Teaching*, *27*(8), 761–776. https://doi.org/10.1002/tea.3660270806

Bell, B. (1985). Students' ideas about plant nutrition: What are they? *Journal of Biological Education*, *19*(3), 213–218. https://doi.org/10.1080/00219266.1985.9654731

Bentley, M., & Connaughton, V. P. (2017). A simple way for students to visualize cellular respiration: Adapting the board game Mousetrap™ to model complexity. *CourseSource*, *4*. https://doi.org/10.24918/cs.2017.8

Bergan-Roller, H. E., Galt, N. J., Dauer, J. T., & Helikar, T. (2017). Discovering cellular respiration with computational modeling and simulations. *CourseSource*, *4*. https://doi.org/10.24918/cs.2017.10

Beyond Multiple Choice. (n.d.). *Automated Analysis of Constructed Response*. Retrieved February 11, 2020, from https://beyondmultiplechoice.org

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

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability*, *21*(1), 5–31. https://doi.org/10.1007/s11092-008-9068-5

Bogart, S. (n.d.). *SankeyMATIC*. Retrieved March 12, 2020, from http://sankeymatic.com

Brassil, C. E., & Couch, B. A. (2019). Multiple-true-false questions reveal more thoroughly the complexity of student thinking than multiple-choice questions: A Bayesian item response model comparison. *International Journal of STEM Education*, *6*(1), 16. https://doi.org/10.1186/s40594-019-0169-0

Carnegie Classification of Institutions of Higher Education. (2018). *About Carnegie Classification*. Retrieved November 20, 2020, from http://carnegieclassifications.iu.edu/

Chi, M. (2008). Three types of conceptual change: Belief revision, mental model transformation, and categorical shift. In Vosniadou, S. (Ed.), *International handbook of research on conceptual change* (pp. 61–82). New York: Routledge.

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

Coppock, A. (2020). Randomization inference for randomized experiments. *R package version 0.2.0*. https://CRAN.R-project.org/package=ri2

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

diSessa, A. A. (2006). A history of conceptual change research: Threads and fault lines. In Sawyer, K. (Ed.), *Cambridge handbook of the learning sciences* (pp. 265–281). Cambridge: Cambridge University Press.

Driver, R., Squires, A., Rushworth, P., & Wood-Robinson, V. (1994). *Making sense of secondary science: Research into children's ideas*. New York: Routledge.

Duit, R., Treagust, D. F., & Widodo, A. (2008). Teaching science for conceptual change: Theory and practice. In Vosniadou, S. (Ed.), *International handbook of research on conceptual change* (pp. 629–646). New York: Routledge.

Freeman, P. L., Maki, J. A., Thoemke, K. R., Lamm, M. H., & Coffman, C. R. (2017). Evaluating the quick fix: weight loss drugs and cellular respiration. *CourseSource*, *4*. https://doi.org/10.24918/cs.2017.17

Freeman, S., Haak, D., & Wenderoth, M. P. (2011). Increased course structure improves performance in introductory biology. *CBE—Life Sciences Education*, *10*(2), 175–186. https://doi.org/10.1187/cbe.10-08-0105

Freeman, S., O'Connor, E., Parks, J. W., Cunningham, M., Hurley, D., Haak, D., … & Wenderoth, M. P. (2007). Prescribed active learning increases performance in introductory biology. *CBE—Life Sciences Education,*, *6*(2), 132–139. https://doi.org/10.1187/cbe.06-09-0194

Gerard, L., Kidron, A., & Linn, M. C. (2019). Guiding collaborative revision of science explanations. *International Journal of Computer-Supported Collaborative Learning*, *14*(3), 291–324. https://doi.org/10.1007/s11412-019-09298-y

Goff, E. E., Reindl, K. M., Johnson, C., McClean, P., Offerdahl, E. G., Schroeder, N. L., & White, A. R. (2018). Investigation of a stand-alone online learning module for cellular respiration instruction. *Journal of Microbiology & Biology Education*, *19*(2). https://doi.org/10.1128/jmbe.v19i2.1460

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

Hake, R. R. (1998). Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. *American Journal of Physics*, *66*(1), 64–74. https://doi.org/10.1119/1.18809

Haladyna, T. M., & Downing, S. M. (1989). Validity of a taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education*, *2*(1), 51–78. https://doi.org/10.1207/s15324818ame0201_4

Hartley, L. M., Momsen, J., Maskiewicz, A., & D'Avanzo, C. (2012). Energy and matter: Differences in discourse in physical and biological sciences can be confusing for introductory biology students. *BioScience*, *62*(5), 488–496. https://doi.org/10.1525/bio.2012.62.5.10

Hartley, L. M., Wilke, B. J., Schramm, J. W., D'Avanzo, C., & Anderson, C. W. (2011). College students' understanding of the carbon cycle: Contrasting principle-based and informal reasoning. *BioScience*, *61*(1), 65–75. https://doi.org/10.1525/bio.2011.61.1.12

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., 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

Hussar, B., Zhang, J., Hein, S., Wang, K., Roberts, A., Cui, J., … & Dilig, R. (2020). *The condition of education 2020* (NCES 2020-144). Washington, DC: National Center for Education Statistics. Retrieved November 30, 2020, from https://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2020144

Jin, H., Zhan, L., & Anderson, C. W. (2013). Developing a fine-grained learning progression framework for carbon-transforming processes. *International Journal of Science Education*, *35*(10), 1663–1697. https://doi.org/10.1080/09500693.2013.782453

Kim, K. J., Meir, E., & Steinberg, E. (2014). *Cellular respiration explored*. Retrieved February 6, 2020, from simbio.com

Kim Yoon, Y. H., & Goetz, E. T. (1993). Strategic processing of test questions: The test marking responses of college students. *Learning and Individual Differences*, *5*(3), 211–218. https://doi.org/10.1016/1041-6080(93)90003-B

LaFleur, B. J., & Greevy, R. A. (2009). Introduction to permutation and resampling-based hypothesis tests. *Journal of Clinical Child & Adolescent Psychology*, *38*(2), 286–294. https://doi.org/10.1080/15374410902740411

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174. JSTOR. https://doi.org/10.2307/2529310

Lee, H., Pallant, A., Pryputniewicz, S., Lord, T., Mulholland, M., & Liu, O. L. (2019.). Automated text scoring and real-time adjustable feedback: Supporting revision of scientific arguments involving uncertainty. *Science Education*, *103*(3), 590–622. https://doi.org/10.1002/sce.21504

Linn, M. C. (2008). Teaching for conceptual change. In Vosniadou, S. (Ed.), *International handbook of research on conceptual change* (pp. 694–722). New York: Routledge.

Linn, M. C., Gerard, L., Ryoo, K., McElhaney, K., Liu, O. L., & Rafferty, A. N. (2014). Computer-guided inquiry to improve science learning. *Science,*, *344*(6180), 155–156. https://doi.org/10.1126/science.1245980

Maskiewicz, A. C., & Lineback, J. E. (2013). Misconceptions are "So yesterday!" *CBE—Life Sciences Education*, *12*(3), 352–356. https://doi.org/10.1187/cbe.13-01-0014

Meir, E., Wendel, D., Pope, D. S., Hsiao, L., Chen, D., & Kim, K. J. (2019). Are intermediate constraint question formats useful for evaluating student thinking and promoting learning in formative assessments? *Computers & Education*, *141*, 103606. https://doi.org/10.1016/j.compedu.2019.103606

Moharreri, K., Ha, M., & Nehm, R. H. (2014). EvoGrader: An online formative assessment tool for automatically evaluating written evolutionary explanations. *Evolution: Education and Outreach*, *7*(1), 15. https://doi.org/10.1186/s12052-014-0015-2

Nakamura, C. M., Murphy, S. K., Christel, M. G., Stevens, S. M., & Zollman, D. A. (2016). Automated analysis of short responses in an interactive synthetic tutoring system for introductory physics. *Physical Review Physics Education Research*, *12*(1), 010122. https://doi.org/10.1103/PhysRevPhysEducRes.12.010122

National Academies of Sciences, Engineering, and Medicine. (2016). *Barriers and opportunities for 2-year and 4-year STEM degrees: Systemic change to support students' diverse pathways*. Washington, DC: National Academies Press. https://doi.org/10.17226/21739

National Research Council. (2012). *A framework for K–12 science education: Practices, crosscutting concepts, and core ideas*. Washington, DC: National Academies Press. https://doi.org/10.17226/13165

National Research Council. (2014). *Developing assessments for the Next Generation Science Standards*. Washington, DC: National Academies Press. https://doi.org/10.17226/18409

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

Nehm, R. H., & Schonfeld, I. S. (2008). Measuring knowledge of natural selection: A comparison of the CINS, an open-response instrument, and an oral interview. *Journal of Research in Science Teaching*, *45*(10), 1131–1160. https://doi.org/10.1002/tea.20251

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

Opitz, S. T., Blankenstein, A., & Harms, U. (2017). Student conceptions about energy in biological contexts. *Journal of Biological Education*, *51*(4), 427–440. https://doi.org/10.1080/00219266.2016.1257504

Orr, R., & Foster, S. (2013). Increasing student success using online quizzing in introductory (majors) biology. *CBE—Life Sciences Education*, *12*(3), 509–514. https://doi.org/10.1187/cbe.12-10-0183

Pape-Lindstrom, P., Eddy, S., & Freeman, S. (2018). Reading quizzes improve exam scores for community college students. *CBE—Life Sciences Education*, *17*(2), ar21. https://doi.org/10.1187/cbe.17-08-0160

Parker, J. M., Anderson, C. W., Heidemann, M., Merrill, J., Merritt, B., Richmond, G., & Urban-Lurain, M. (2012). Exploring undergraduates' understanding of photosynthesis using diagnostic question clusters. *CBE—Life Sciences Education*, *11*(1), 47–57. https://doi.org/10.1187/cbe.11-07-0054

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

Pennebaker, J. W., Gosling, S. D., & Ferrell, J. D. (2013). Daily online testing in large classes: Boosting college performance while reducing achievement gaps. *PLoS ONE*, *8*(11), e79774. https://doi.org/10.1371/journal.pone.0079774

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

Price, R. M., Pope, D. S., Abraham, J. K., Maruca, S., & Meir, E. (2016). Observing populations and testing predictions about genetic drift in a computer simulation improves college students' conceptual understanding. *Evolution: Education and Outreach*, *9*(1), 8. https://doi.org/10.1186/s12052-016-0059-6

R Core Team. (2020). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/

Roediger, H. L., & Marsh, E. J. (2005). The positive and negative consequences of multiple-choice testing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(5), 1155–1159. http://dx.doi.org.proxy1.cl.msu.edu/10.1037/0278-7393.31.5.1155

RStudio Team. (2020). *RStudio: Integrated Development for R. RStudio*. Boston, MA: PBC. http://www.rstudio.com/

Scalise, K., & Gifford, B. (2006). Computer-based assessment in e-learning: A framework for constructing "intermediate constraint" questions and tasks for technology platforms. *The Journal of Technology, Learning and Assessment*, *4*(6). Retrieved February 14, 2020, from https://ejournals.bc.edu/index.php/jtla/article/view/1653

Schinske, J. N., Balke, V. L., Bangera, M. G., Bonney, K. M., Brownell, S. E., Carter, R. S., … & Corwin, L. A. (2017). Broadening participation in biology education research: Engaging community college students and faculty. *CBE—Life Sciences Education*, *16*(2), mr1. https://doi.org/10.1187/cbe.16-10-0289

Shtulman, A., & Valcarcel, J. (2012). Scientific knowledge suppresses but does not supplant earlier intuitions. *Cognition*, *124*(2), 209–215. https://doi.org/10.1016/j.cognition.2012.04.005

Sripathi, K. N., Moscarella, R. A., Yoho, R., You, H. S., Urban-Lurain, M., Merrill, J., & Haudek, K. (2019). Mixed student ideas about mechanisms of human weight loss. *CBE—Life Sciences Education*, *18*(3), ar37. https://doi.org/10.1187/cbe.18-11-0227

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

Tanner, K., & Allen, D. (2005). Approaches to biology teaching and learning: Understanding the wrong answers—teaching toward conceptual change. *Cell Biology Education*, *4*(2), 112–117. https://doi.org/10.1187/cbe.05-02-0068

Urban-Lurain, M., Moscarella, R. A., Haudek, K. C., Giese, E., Sibley, D. F., & Merrill, J. E. (2009). Beyond multiple choice exams: Using computerized lexical analysis to understand students' conceptual reasoning in STEM disciplines. In *2009 39th IEEE Frontiers in Education Conference* (pp. 1–6). https://doi.org/10.1109/FIE.2009.5350596

Vosniadou, S. (2012). Reframing the classical approach to conceptual change: Preconceptions, misconceptions and synthetic models. In Fraser, B. J., Tobin, K., & McRobbie, C. J. (Eds.), *Second international handbook of science education* (pp. 119–130). Dordrecht, Netherlands: Springer. https://doi.org/10.1007/978-1-4020-9041-7_10

Vosniadou, S., & Brewer, W. F. (1992). Mental models of the earth: A study of conceptual change in childhood. *Cognitive Psychology*, *24*(4), 535–585. https://doi.org/10.1016/0010-0285(92)90018-W

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

Wiliam, D. (2011). What is assessment for learning? *Studies in Educational Evaluation*, *37*(1), 3–14. https://doi.org/10.1016/j.stueduc.2011.03.001

Wilson, C. D., Anderson, C. W., Heidemann, M., Merrill, J. E., Merritt, B. W., Richmond, G., … & Parker, J. M. (2006). Assessing students' ability to trace matter in dynamic systems in cell biology. *CBE—Life Sciences Education*, *5*(4), 323–331. https://doi.org/10.1187/cbe.06-02-0142

Zhu, M., Liu, O. L., & Lee, H.-S. (2020). The effect of automated feedback on revision behavior and learning gains in formative assessment of scientific argument writing. *Computers & Education*, *143*, 103668. https://doi.org/10.1016/j.compedu.2019.103668