Mixed Student Ideas about Mechanisms of Human Weight Loss

**Kamali N. Sripathi,[†]* Rosa A. Moscarella,[‡] Rachel Yoho,[§] Hye Sun You,[∥]
Mark Urban-Lurain,[†] John Merrill,[¶] and Kevin Haudek[#]**

[†]CREATE for STEM Institute, [¶]Department of Microbiology and Molecular Genetics, and
[#]Department of Biochemistry and Molecular Biology, Michigan State University, East Lansing,
MI 48824; [‡]Biology Department, University of Massachusetts Amherst, Amherst, MA 01003;
[§]Project Dragonfly and Department of Biology, Miami University, Oxford, OH 45056;
[∥]Department of Physical Sciences, Arkansas Tech University, Russellville, AR 72801

ABSTRACT

Recent calls for college biology education reform have identified "pathways and transformations of matter and energy" as a big idea in biology crucial for students to learn. Previous work has been conducted on how college students think about such matter-transforming processes; however, little research has investigated how students connect these ideas. Here, we probe student thinking about matter transformations in the familiar context of human weight loss. Our analysis of 1192 student constructed responses revealed three scientific (which we label "Normative") and five less scientific (which we label "Developing") ideas that students use to explain weight loss. Additionally, students combine these ideas in their responses, with an average number of 2.19 ± 1.07 ideas per response, and 74.4% of responses containing two or more ideas. These results highlight the extent to which students hold multiple (both correct and incorrect) ideas about complex biological processes. We described student responses as conforming to either Scientific, Mixed, or Developing descriptive models, which had an average of 1.9 ± 0.6, 3.1 ± 0.9, and 1.7 ± 0.8 ideas per response, respectively. Such heterogeneous student thinking is characteristic of difficulties in both conceptual change and early expertise development and will require careful instructional intervention for lasting learning gains.

INTRODUCTION

Biology education is undergoing a transformation across the entire range of K–16 education. The National Research Council (NRC, 2012) has underscored the need for students to become critical consumers of the scientific information that permeates their lives, regardless of whether they pursue a STEM (science, technology, engineering, and mathematics) career. The NRC thus developed guidelines for effective STEM K–12 education (NRC, 2012) with the goal of helping students become critical scientific thinkers. The American Association for the Advancement of Science (AAAS, 2011) issued a similar call specifically for undergraduate biology education, hereafter referred to as *Vision and Change*, citing the increasingly interdisciplinary nature of research in all biological fields and the need to incorporate scientific practices into biology curricula. Both reports define "disciplinary core ideas" (NRC, 2012) or "core concepts" (AAAS, 2011) that students should be able to understand and apply to new situations in order to be savvy scientific consumers. As such, these core ideas are not only essential to learning expectations for being scientific consumers, but can also be present as expectations of practitioner societies for developing the next generation of professionals in each society's respective area (Yoho *et al.*, 2018).

One of the *Vision and Change* core concepts (AAAS, 2011) is "pathways and transformations of matter and energy" (p. 13). This concept emphasizes the ubiquity of chemical and physical principles underlying complex biological systems from the microscopic, cellular, and molecular levels to the organismal and ecosystem levels. For

example, the cells of living organisms are held together by a multitude of intermolecular interactions between phospholipids in the membrane. Photosynthetic cells in plants use energy harvested from light radiation to fix carbon into organic molecules, which other organisms can use by consuming the plants. These organisms are then food sources themselves for higher-order consumers in the ecosystem. These are only a few examples of the relationships between chemical properties across scalar levels in biology. Students must comprehend and apply these principles to fully understand such systems in both familiar and novel contexts. Two central biochemical processes that determine the flow of matter and energy are cellular respiration and photosynthesis. As the pathway by which organisms of all kingdoms transform energy, cellular respiration is a cornerstone process in undergraduate introductory biology classrooms. Likewise, because glucose produced as a result of photosynthesis can be directly metabolized by enzymes of cellular respiration, this process also enjoys almost universal treatment in undergraduate courses ranging from introductory biology to upper-level biochemistry.

Students have significant difficulties understanding these biological pathways, however, likely due to these processes' highly interconnected nature. Examples of student confusion include difficulty between scalar processes, tracing matter across scales and phases, and lack of attention to atoms when comparing reactants and products. Many studies have separately reported students' tendencies to confuse cellular and physiological respiration (Bell, 1985; Anderson *et al.*, 1990; Driver *et al.*, 1994). Additionally, Hartley and colleagues (2011) found that students struggled to reason about biological processes across scales. The authors observed that students focused on explaining phenomena at a macroscopic level rather than describing underlying molecular processes. Hartley and colleagues further theorized that this may be because students are more comfortable reasoning at the macroscopic level or because students do not know that larger-scale phenomena can be explained by processes at smaller scales (Hartley *et al.*, 2011). Mohan and coworkers (2009) also identified "reasoning about systems and processes at multiple scales" (p. 678) as a key feature of scientific explanations with which students wrestle.

Other works have highlighted student problems in tracing matter and energy across complex biological systems. For example, studies have shown that students find it especially difficult to trace matter when gases are involved as either the products or reactants of processes (Anderson *et al.*, 1990; Wilson *et al.*, 2006; Jin *et al.*, 2013). Jin and colleagues (2013) found that students tended to believe that air and gases have no mass. The authors further described a learning progression focused on "carbon-transforming processes" (Jin *et al.*, 2013, p. 1663) and organized it according to five "progress variables" (p. 1667), each with four progressive levels of student growth in specific areas. Two progress variables focused on tracing matter are 1) *explaining materials* and 2) *explaining mass* (Jin *et al.*, 2013). The authors found significant student difficulties along both of these progress variables. For example, in the *explaining materials* progress variable, less advanced students often described vague processes converting one material into another. Regarding the *explaining mass* (or changes in weight/mass) progress variable, less advanced students were able to recognize that gain and loss of mass are associated with the subsequent gain and loss of materials; however, these students often conflated matter and energy. To address such student difficulties, Wilson and colleagues (2006) suggested that teaching students to trace matter can be an effective strategy for them to learn cellular respiration and photosynthesis. They assessed student ability to employ this strategy through essay prompts and interviews and designed multiple-choice items with distractors based on the revealed student difficulties. This work showed that students had ideas of varying correctness regarding tracing matter in familiar contexts such as human weight loss. When discussing weight loss mechanisms, students were able to correctly identify products, but did not explain, or explained incorrectly, the processes by which these molecules were produced. Additionally, some students incorrectly described fat being changed into other molecules or into energy, while others discussed the mass rejoining the atmosphere (Wilson *et al.*, 2006). These and other studies illustrate the many-sided nature of student thinking about matter-transforming processes.

Student thinking about matter-transforming processes is further confused when students are required to apply knowledge of these processes in a familiar context. One such context is human weight loss: fat molecules in human bodies must be broken down by beta-oxidation, whose by-products are then broken down by cellular respiration into CO_2 (84% of mass) and water (16% of mass; Meerman and Brown, 2014). CO_2 is exhaled through the lungs, while water is released by both breathing and physiological excretion. A recent study revealed that many health and fitness professionals do not understand this breakdown of fat (Meerman and Brown, 2014). The authors surveyed a total of 150 family doctors, dieticians, and personal trainers about where they think mass goes during weight loss. The majority believed the common, vague conception that "fat is converted to energy or heat" (Meerman and Brown, 2014, p. 1). Other prominent but erroneous ideas were that fat was solely defecated out or turned into muscle (Meerman and Brown, 2014; University of New South Wales, 2014). These results are particularly concerning in light of the fact that the public relies on professionals such as doctors and personal trainers to counsel them accurately about their health. This recent study, in addition to supporting the evidence of unclear understandings that students have about matter-transforming processes, reveals how those understandings persist throughout students' lives and careers.

However, student confusion in this field is not surprising considering the complicated nature of learning. The areas of cognitive psychology and learning theory have a long history of characterizing the complexity of how people learn. The constructivist paradigm (Cooper, 1993; Ertmer and Newby, 1993), for example, emphasizes that students learn new concepts by constructing meaningful relationships in their own minds. Work in the field of conceptual change explores how learning is accomplished when the learners' prior knowledge is "in conflict with" facts that must be acquired (Chi, 2008, p. 61). Both of these examples illustrate that learning is a complex process, well beyond the simplification of correct versus incorrect ideas. Rather than focusing on students' correct or incorrect ideas, educators should focus on how those ideas interact in the learners' minds, and what these interactions say about students' understandings about a given topic. In the context of human

weight loss and other carbon-transforming processes, much work has been conducted on the correct and incorrect ideas that students have, but little work has probed how students' ideas are connected in their minds.

Because of students' multifaceted understandings of complicated topics like human weight loss, it is crucial for instructors to accurately probe student understanding about these topics using appropriate types of assessment items. Multiple-choice assessments have historically been popular due to the ease of grading; however, more open-ended items such as constructed-response items have been shown to provide more nuanced pictures of student understanding. Previous work by Parker *et al.* (2012) compared the quality of student responses to multiple-choice items, multiple-true/false items, and essay prompts. The authors found that both multiple true/false and essay prompts revealed evidence that students hold both correct and incorrect ideas, a subtlety often missed by multiple-choice assessments. Additionally, Nehm and Schonfeld (2008) compared the ability of multiple-choice items, constructed-response items, and oral interviews to elicit students' correct and incorrect ideas about evolution. They found that both multiple-choice and constructed-response assessments effectively measured key concept occurrence and diversity; however, their results indicated that multiple-choice items may also have overemphasized students' knowledge of evolution concepts. Similarly, a recent study by Hubbard and coworkers (2017) found that, while multiple-true/false items may help students focus on specific ideas, constructed-response questions enable a more holistic view of student ideas about a topic.

The current study addresses the gaps in knowledge regarding how ideas about weight loss mechanisms are connected in students' minds to help instructors better understand student confusion about complex biological processes. The work presented here builds upon the extensive literature about student ideas of cellular processes. We probe student thinking on this topic using a constructed-response prompt adapted from Wilson *et al.* (2006) (hereafter the "weight loss item") to investigate the following research question: What molecular mechanisms and types of matter do students invoke to explain the process of weight loss?

Our results reveal that students have multiple ideas about weight loss. These ideas can be either scientific or more informal in nature, and contrasting ideas often coexist in students' responses and, by extension, in their minds. This heterogeneity is crucial for instructors to keep in mind when teaching matter-transforming processes and may be a key reason that students have trouble understanding these processes.

METHODS
Data Collection
We used the following constructed-response prompt to investigate student thinking about weight loss (the weight loss item): "Your friend lost 15 lbs on a diet. Where did the mass go?" This prompt was originally developed as part of a set of multiple-choice questions used to assess students' abilities to trace matter across complex biological processes in familiar contexts (Wilson *et al.*, 2006; Parker *et al.*, 2012). We administered the item at three large public research universities located in the Midwest and East Coast of the United States, all classified in the Carnegie definitions as either "higher" or "highest" for doctoral

university research activity (Carnegie Classification of Institutions of Higher Education, 2017). We collected a total of 2445 student responses from students in introductory biology classes for life science majors, because these courses often cover energy and matter transformations in cellular respiration. The prompt was administered online on each institution's course management software, as part of a homework assignment that was either assigned a small amount of credit or bonus points (usually <0.1% of total class points) only for completion. This assignment was prefaced with a request that the students not use outside resources to answer the question. This study was designated exempt by Michigan State University's Institutional Review Board (IRB x10-577 and STUDY00001648). Our IRB protocols enabled us to obtain anonymized data from these institutions, and as a result, we are unable to provide specific demographics for each class. However, because all data were obtained from courses for life sciences majors, we expect demographics of the student populations to echo typical trends in enrollment for such majors at these other institutions.

Rubric Generation and Refinement
We used a combination of a priori (based on trends described by Wilson *et al.*, 2006) and emergent coding to generate an analytic scoring rubric to capture prominent trends in student responses. The categories we adopted from Wilson and colleagues (2006) were (names listed are the finalized names of our rubric categories, described in Table 1) 1) mention of Correct Molecular Products; 2) named Molecular Mechanism; 3) vague description of Matter Converted to Energy; 4) Exhalation of mass; and 5) Excretion of mass through urine, feces, sweat, and/or tears. Although rubrics are typically used for grading student assignments, our group has capitalized on the value of these tools as a method of data analysis (e.g., Haudek *et al.*, 2012, 2015; Moscarella *et al.*, 2016). Employing an analytic rubric enabled us to capture multiple, distinct ideas present in the same student response in individual rubric categories (Yune *et al.*, 2018). Thus, it is possible for a single student response to be scored (i.e., classified) in single, multiple, and no rubric categories.

We modified our a priori rubric when we became aware that the rubric did not capture certain ideas present in our data. This required us to refine and expand the rubric to better capture the complexity of student thinking. The final result was an eight-category analytic rubric describing the specificity with which students traced matter when thinking about weight loss across molecular/cellular and organismal levels (Table 1). Responses that traced matter by identifying correct molecular products were scored in our Correct Molecular Products category, while responses that vaguely discussed carbon without reference to a specific molecular form were scored in our Carbon Alone category. Those responses that described one or more specific molecular processes were scored into our Molecular Mechanism category, while those discussing matter conversion more vaguely at the cellular level were scored in our General Metabolism category. Our Matter Converted to Energy category contained student responses that discussed vague conversions of mass to energy, often, but not always, at the cellular level. At the organismal level, responses correctly identifying exhalation as the exit route from the body were scored in Exhalation. Responses discussing the mass leaving the body as urine, feces,

TABLE 1. Finalized analytic scoring rubric for analyzing responses to the weight loss item

Analytic rubric category[a]	Criteria	Exemplar student response (key phrases underlined)[b]
Correct Molecular Products*	Responses that include correct molecular products in their explanation	"The mass in the body was loss [*sic*] through losing fat. The fat is converted into glucose, which then becomes energy. The energy then becomes CO2 [*sic*] and is left the body when breathing."
Carbon Alone*	Responses that do not indicate the molecular form (e.g., CO_2) of carbon as a product	"A friend who lost 15 lbs on a diet must have sweat[ed] and exhaled out enough cabon [*sic*] over the course of his diet to lose 15 lbs of fat."
Molecular Mechanism†	Responses that indicate correct processes (either by name or by description) through which mass is converted into other products	"As his fat stores were broken down (catabolism) to provide energy by cellular respiration, the molecules are broken down. The carbons of these organic molecules is [*sic*] converted to carbon dioxide and water, which are expelled from the body."
General Metabolism†	Responses that do not completely define or name a correct molecular mechanism (which would fall in the Molecular Mechanisms bin), but that do indicate some a certain degree of (either correct or incorrect) molecular understanding of molecular transformations or processes (either correct or incorrect)	"As the friend metabolized the fat, energy was withdrawn from the fat stores is cataloging the fasts. Simple molecules. These simple molecules were lost as co2 [*sic*], waste through respiration. A small portion of the fat stores were exited [*sic*] with other food waste as well."
Matter Converted to Energy	Responses that indicate incorrectly that mass is converted into energy or used up	"The 15 pounds of mass was converted into another source of energy or transferred. The 15 pounds may have also been lost as heat during the diet but in all cases it was not destroyed."
Exhalation	Responses that indicate mass has been released into the air	"The mass is breathed out as carbon dioxide in to [*sic*] the air."
Excretion	Responses that indicate mass leaves body as nongas waste	"The 15 pounds, through exercise, is lost as heat and water through sweating and exhalation."
How to Lose Weight	Responses that use "common knowledge" about dieting or exercise for weight loss	"It was used by the person to create energy. since they were using more than they were consuming, the matter used to create the energy was pulled from fat stores [ideally, although it could also have been pulled from muscle]. thus the weight went into creating energy [and probably waste too...]"

[a]An asterisk (*) indicates a pair of mutually exclusive categories. A dagger (†) indicates a second pair of mutually exclusive categories.
[b]All responses used here and henceforth are included verbatim as the students wrote them, including spelling and capitalization.

sweat, and/or tears were scored in the Excretion category. Finally, responses discussing conventional, nonscientific knowledge about weight loss (e.g., calories consumed less than calories spent, exercise as a means of weight loss) were scored in our How to Lose Weight category. Only 3% ($n = 39$) of responses were not classified into any category, because these responses lacked content, that is, they used unclear language (e.g., referring to "waste" without specifying what type) or restated scientific facts without clear explanation (see *Results* and Figure 2 later in this article for further details).

To make the rubric as easy as possible for scorers to understand and use, we established detailed scoring rules, some of which are summarized in Table 1. Because of our rules, the greatest number of ideas possible in a single response was six: we constrained the definition of the Carbon Alone and Correct Molecular Products categories and the Molecular Mechanism and General Metabolism categories to be two mutually exclusive pairs. We developed another rule for scoring the term "respiration," because it was sometimes unclear whether students meant physiological or cellular respiration, a well-documented lexical ambiguity (Bell, 1985; Anderson *et al.*, 1990; Driver *et al.*, 1994). We thus used context clues to decide whether the responses belonged in our Exhalation category or in our Molecular Mechanism category. If some mention was made of releasing products into the air, the response was categorized as an Exhalation response and not a Molecular Mechanism response. For instance,

the following student response suggested to the scorers that the student was referring to exhalation (this and all responses henceforth are reported verbatim, including spelling, capitalization, punctuation, and grammar): "The 15 pounds was turned into carbon that was stored in the body. Once stored it is released into the air. It goes into the air as carbon dioxide during respiration, as it leaves the body." Conversely, if the response included a sufficiently molecular interpretation of weight loss, we classified it under Molecular Mechanism and not Exhalation, as with the following response: "The 15 pounds, that the friend had originally gained by consuming carbon in food and incorporating it into their body, is released through the CO2 that we breathe out as a waste product of respiration. It is not converted to energy." Although these two responses appear essentially similar, the defining factor for our characterization was that the second response identified CO_2 "as a waste product of respiration," indicating to us that the second student was referring to the cellular process and not the physiological process.

Scorer Training and Rubric Reliability
For the training phase of rubric scoring, a total of six scorers (K.N.S., R.A.M., R.Y., H.-Y.S., J.M., and K.H.), all with PhDs in a biology-related discipline or science education, used an initial analytic rubric to reach consensus on the scores of 110 responses. For each response, each category was scored as being present or absent. The six scorers were then assigned to pairs

that were each assigned sets of 100 unscored responses. Individual scorers recorded their own scores for each rubric category for each response in an Excel spreadsheet. Each scorer pair then combined and compared their scores from their individual spreadsheets and met to discuss disagreements in score assignment that they had. Any disagreements that individual scorer-pairs could not resolve were sent to a third scorer who acted as a tiebreaker. The tiebreaker scoring was considered final. Disagreements that could not be resolved by the tiebreaker were brought to the entire group of six, with either the problematic scores or the rubric definition being modified. Scoring continued for 1100 responses until average Cohen's kappa values between scorers for all but one rubric category were 0.6 or better (the General Metabolism category had an average Cohen's kappa of 0.37, discussed later). We used kappa values of >0.6 as a benchmark, because a level of agreement of 0.6 or greater has been deemed "satisfactory" by Landis and Koch (1977). Responses on which all six scorers could not agree were removed, for a total data set of $N = 1192$ scored responses.

To improve the kappa value for the General Metabolism category, the group of six scorers analyzed human scores assigned to responses in this category using a suite of predictive machine-learning algorithms. The machine-learning ensemble predicted scores for the General Metabolism category using the human scores as a training set. Initial agreement between human scoring and machine scoring of the General Metabolism category was 0.622. To improve this kappa, the head scorer reviewed responses for which the machine-learning predictions disagreed with human scores: For responses for which the head scorer agreed with the human score, she retained the human score. For those responses for which the head scorer agreed with the machine prediction, she changed the human score to match the prediction. Responses whose scores the head scorer deemed ambiguous were discussed and resolved between the head scorer and three of the original scorers, and either the human scores or the category's rubric definition were modified as needed. After this process, the human-scored General Metabolism responses were again analyzed by the machine-learning ensemble, this time with a resulting kappa of 0.698 for the $N = 1192$ data set mentioned earlier. This method proved to be effective for improving the rubric definition of the General Metabolism category.

Larger-Grained Analysis of Student Responses

Once we had scored the data with the analytic rubric, we wanted to understand the relationships among the analytic rubric categories in student responses: Which categories tended to occur in the same responses more frequently? Which categories occurred together less frequently? We employed hierarchical clustering analysis of the rubric categories to initially characterize our student responses. We used the software SPSS v. 24 (IBM, 2016) to perform clustering analysis and correlation coefficient calculations. Because the measures are binary we used the average linkage between groups clustering method and the pattern difference dissimilarity measure (Choi et al., 2010).

In addition to using clustering analysis to characterize co-occurrences, we wanted to present important co-occurrences in a manner that could aid instructors in characterizing student responses to the weight loss item. We applied a modified version of the framework employed by the EvoGrader software (Moharreri et al., 2014). For the EvoGrader system, which automatically scores constructed responses to 83 evolution items (Nehm et al., 2012), Moharreri and colleagues (2014) developed scoring rubrics to characterize ideas in student responses as either Normative (or Scientific) or Nonnormative (or Naïve). The authors then classified responses as fitting Scientific, Mixed, or Naïve reasoning models based on whether the responses contained only Scientific ideas (a Scientific reasoning model), both Scientific and Naïve ideas (a Mixed model), or only Naïve ideas (a Naïve model).

Moharreri and colleagues' (2014) characterization scheme appealed to us because it allowed us to classify our rubric categories more subtly than a simple characterization of correct versus incorrect ideas. Additionally, the scheme provided a way for us to describe the connections between student ideas. Based on the work of Meerman and Brown (2014), we classified our Correct Molecular Products, Exhalation, and Molecular Mechanism rubric categories as Normative. Because each of the remaining categories may not be completely nonnormative (e.g., the water by-product of weight loss is predominantly lost through excretion), we labeled our remaining rubric categories as Developing ideas. Using these classifications, we were able to develop student *descriptive* models, in contrast to Moharreri and colleagues' *reasoning* models. We named our models "descriptive" to reflect the fact that students are describing the mechanism of weight loss rather than reasoning about it. To define our descriptive models, we used rules similar to those of Moharreri and colleagues (2014): 1) Scientific descriptive models contain only scientific ideas; 2) Mixed descriptive models contain at least one Scientific and at least one Developing idea; 3) Developing models (whose name was changed from "Naïve" models to reflect our modifications) contain only Developing ideas (Figure 1). In the *Results*, we present further investigations into these descriptive models as one example of a larger grain size that instructors can use to make sense of the varying ways students think about weight loss.

RESULTS

The research question driving our current work focuses on understanding the subtleties of student thinking about matter-transforming processes, so that educators can understand and use this complexity to support student-centered teaching practices. The complexity of student thinking that our analyses reveal is not trivial; thus, we have organized the following sections to methodically unpack these intricacies. We begin with a comparison of expert and student responses to the weight loss prompt to contrast the ideas in our analytic rubric that these two populations use to describe weight loss mechanisms. We then continue our analysis of student data for the simplest case of responses: "single category" responses. From here, we move on to investigating how two and more ideas can co-occur within student responses. We begin our co-occurrence analysis with a hierarchical clustering approach that reveals that Normative and Developing ideas largely cluster together (Figure 1). Next, we use a web diagram to dive deeper into two-category co-occurrences, because this type of analysis reveals both prominent and less prominent co-occurrences. We then analyze more complicated three- and four-category co-occurrences by tracing a single rubric category across different combinations. Because
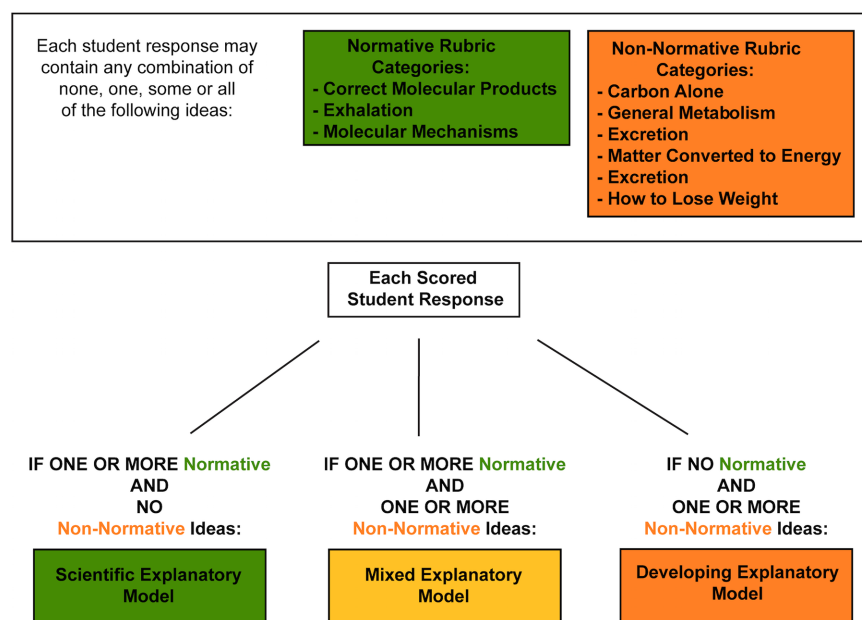
FIGURE 1. Diagram outlining classification of rubric categories and student responses into descriptive models. We classified our rubric categories (Table 1) as Normative or Developing ideas (based on Meerman and Brown, 2014) to avoid the more rigid classifications of correct vs. incorrect. We then fit each response to one of the following descriptive models: Scientific (only containing one or more Normative ideas), Mixed (containing both Normative and Developing ideas), or Developing (only containing one or more Developing ideas).

our analysis of five- and six-category responses revealed similar trends to our three- and four-category responses, this discussion can be found in the Supplemental Material. Taken together, the results presented here begin to elucidate the multiple, interrelated ideas that students have about metabolic processes in the human body.

Differences between Experts and Students in Tracing Matter in Human Weight Loss

To characterize how experts trace matter and energy in weight loss, we solicited responses from biology instructors. We requested information such as institution and typical courses in which they might teach cellular respiration and asked them to respond to the weight loss item. A total of 11 instructors teaching a variety of courses distributed over five public universities responded to our solicitation (Table 2). Introductory biology was strongly represented ($n = 5$ instructors), with additional

participants from physiology, general biology, and molecular and cellular biology.

We applied our eight-category analytic rubric (Table 1) to these responses. We found that high percentages of expert responses traced matter across the cellular/molecular and organismal scales. The highest-occurring categories were the Normative cellular/molecular Correct Molecular Products (100%) and Molecular Mechanism (81%) and the organismal Exhalation (90%) categories (Figure 2), as exemplified in the following instructor's response: "The fat was oxidized to form CO2 that was exhaled from the body."

Interestingly, all Developing categories, except for Carbon Alone, also appeared in responses from our instructor population (Figure 2). These categories describe less specific ways of tracing matter. Excretion (36%) and Matter Converted to Energy (27%) were the next highest occurring ideas in instructor responses. We saw that instructors used these Developing ideas to provide context for their tracing of matter. The following instructor prioritized the mass leaving the body as CO_2, but also added physiological routes of Excretion: "Through cellular respiration, fat molecules are converted to CO2 and H2O. CO2 is released from the body through the
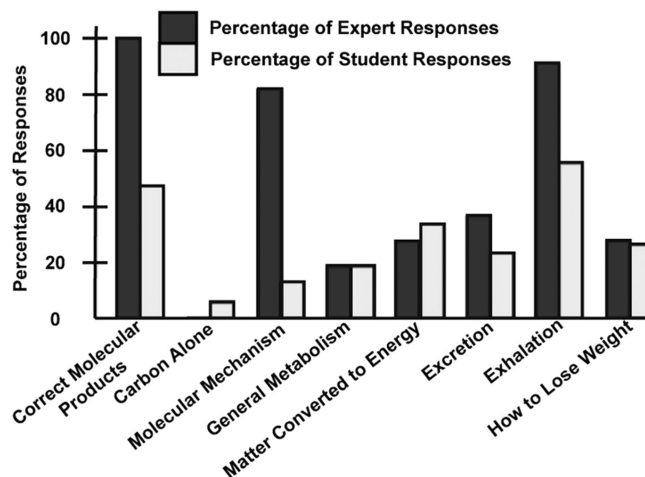

FIGURE 2. Normative and Developing rubric ideas in expert and student responses. We found that the Normative ideas of Correct Molecular Products and Exhalation were the highest-occurring categories in both expert and student responses. Developing ideas occurred in about the same percentages in both expert and student responses; however, we often found the important cellular-level context to be missing from student responses containing Developing ideas. Only one Developing category, Carbon Alone, occurred solely in student responses. We attribute this to student difficulty in thinking at the molecular level, a difficulty that experts typically do not have.

TABLE 2. Courses taught by instructors who provided expert responses

Course type	Number of instructors
Introductory biology	5
Physiology	2
General biology for majors	1
Molecular and cellular biology	1
Other (biological diversity, marine biology)	1
No course specified	1

lungs. H2O is released from the body via sweat and urine." Similarly, the instructor who provided the following example began with a vague conversion of matter (fat) to energy. However, this instructor continued to trace the matter through cellular processes in a more specific manner: "Fat is a source of energy for the body. You use fat as a fuel for cellular respiration and the carbon gets 'expelled' as carbon dioxide during exhalation." Instructor use of nonnormative (Developing) ideas in conjunction with Normative ideas further justifies our rationale in labeling the former as "Developing."

Analysis of instructor responses underscored the context-dependent nature of interpretation of our five Developing rubric categories. For example, in the following instructor response, the instructor begins with the less detailed Developing ideas of How to Lose Weight (occurring in 27% of instructor responses), and continues on to specify Normative routes of exit through Exhalation, Molecular Mechanism, and Correct Molecular Products: "My friend expended more energy than he/she consumed as food over a period of time. To make up the difference, some of the mass of my friend (fats, carbs, proteins) would be used as fuel for cell respiration. The products of cell respiration are about equal parts carbon dioxide and water. Because carbon dioxide has a much larger molar mass, most of the mass would be breathed out as carbon dioxide. The rest would be lost as water in some form (vapor, urine, sweating, etc.)." The first sentence by itself would be a nonnormative response. However, similar to the earlier examples, the response indicates a starting informal description of weight loss, followed by a more normative tracing of matter.

Our analytic rubric enabled us to track the multiple ideas we found in each instructor response. We found that the responses had an average of 3.82 ± 0.60 ideas. The fact that expert descriptions of matter-transforming processes contain such a diversity of ideas is striking. The nuances in our expert responses illustrate the reason for the extensive study of and teaching and learning interventions for this topic that have been put forth in an effort to enhance student understanding.

Our rubric categories highlighted different trends in how students trace matter and energy when thinking about weight loss compared with experts. In contrast to Correct Molecular Products being the highest-occurring Normative idea in expert responses, the highest-occurring Normative idea in student responses was Exhalation (55%, Figure 2), as in the following student response: "They breathed it out." Correct Molecular Products (47% of responses) was the second-highest normative idea in student responses: "The mass left as water and CO2." The most significant difference in Normative ideas between student and instructor responses was seen in the occurrence of the Molecular Mechanism category, which occurred in only 12% of student responses compared with 81% of expert responses. When Molecular Mechanism did occur in student responses, the responses were as specific as those of experts: "The fat was burned to run cellular respiration."

Although the percentages of student and expert responses in our Developing categories were comparable, student responses typically provided a less specific context for these ideas than did those of experts. Student use of the General Metabolism category was mostly analogous to that of experts (18% for both populations); the differences occurred in the context provided by the two groups. Student responses were often more vague with respect to the process or mechanism by which mass is transformed than were expert responses: "The fat was transformed into glucose and used by the body to make ATP and do work." The previous response is not complete, because there was no description or naming of the processes that convert fat to glucose. Other student responses in this category vaguely referred to the fat or mass being "metabolized" or used up in "metabolism" or "metabolic processes," often with no further explanation, such as the following response: "It was used up in the form of energy during metabolism." A comparable number of both student and expert responses also exhibited incorrect or vague ways of tracing matter. About a third (33%) of student responses, compared with 27% of expert responses, described vague matter-to-energy conversions in the Matter Converted to Energy category without specifying a mechanism for conversion. Such an idea is a common misconception among students (e.g., Wilson *et al.*, 2006), and appeared in our data set in various forms. Some responses specified a mass input (typically the fat) and energy output (typically just "energy"): "Our bodies convert molecules in our fat cells to energy that we can use, causing the fat cells to shrink." Unlike the expert responses presented earlier, this response references no specific mechanism for matter conversion. Furthermore, the vagueness in this language is an important point for instructors to consider as they choose language for their own instruction. Other responses in this category instead stated "heat" as a result of weight loss without further explanation: "the food/fat previously stored on the person's body doesn't go anywhere it is burned up & used to create energy (heat) etc." Student use of the Excretion category was similar to that of experts (23 vs. 36% respectively), but also lacked specific matter-transforming processes: "It is lost over time through excretion and sweat." We observed similar trends in student and expert contexts for the How to Lose Weight category, which students also used at a rate comparable to that of experts (26 vs. 27% respectively): "The 15 lbs were lost due to the fact that her caloric intake was less than the calories she used in a day for energy. To make up for this difference her body resorted to stored energy to burn to match the calories used." Other student responses in this category simply stated exercise or physical activity as the reason for losing weight: "The 15 pounds got used up during exercise, so she basically burned all of the calories that made up the 15 pounds." Both of these examples, however, lack a cellular explanation for weight loss, which was present in expert responses that used this category.

A notable exception to a similar student and instructor use was our smallest category in student responses: Carbon Alone. No experts used this category in their responses, but 5% of student responses did. For responses in this category, students traced matter by discussing the term "carbon" by itself, not in the context of other molecular compounds like CO_2. We chose to keep this category separate from Correct Molecular Products, because it was often unclear what students meant by their use of "carbon." In some cases, it was likely that the students were using "carbon" as a shorthand for "carbon dioxide": "A friend who lost 15 lbs on a diet must have sweat[ed] and exhaled out enough cabon [*sic*] over the course of his diet to lose 15 lbs of fat." In others, however, it is less clear what molecular form exactly the students meant in their responses: e.g., "The Carbon was released during cellular respiration." Such confusion is aligned with student difficulties with the particulate nature of

matter, as extensively documented in the chemistry education literature (e.g., Harrison and Treagust, 1996; Talanquer, 2009). These documented difficulties explain why we did not observe this category in our expert responses.

Similar to our instructor responses described earlier, it was quite common to find student responses that could be classified into more than one of our rubric categories. In fact, on average, student responses included slightly more than two ideas ($X = 2.19 \pm 1.07$), as identified by our scoring rubric. This average is less than that of instructor responses ($X = 3.82 \pm 0.60$), which may be attributed to experts' ability to "chunk" related pieces of information together due to their advanced proficiency. When we plotted total student responses versus the number of rubric categories (Figure 3), we found that a majority of student responses we analyzed (74.4%, $n = 887$) contained two or more ideas. As stated in the *Methods*, we also found that ~3% of responses ($n = 39$) did not contain any ideas described by our analytic rubric. Because we focused our rubric on commonly occurring ideas, we expected that there would be some aspects of student responses that occurred too infrequently (e.g., restatement of the law of conservation of mass) for us to document. Such infrequent ideas make up the 3% of "0 categories" responses in our data set.

Single-Category Student Responses Mostly Focus on Vague Matter-to-Energy Conversions

We began our analyses of student responses with the simplest case: those containing a single idea from our analytic rubric. From our analysis in Figure 3, 22% ($n = 265$ responses) could be categorized as "single-category" responses. The most common rubric category represented in this subset (Figure 4) was Matter Converted to Energy ($n = 115$). The majority of these responses discussed a vague mass-to-energy conversion without explicitly mentioning the organismal scalar level: "The 15 pounds of mass will be used as energy." Other responses were
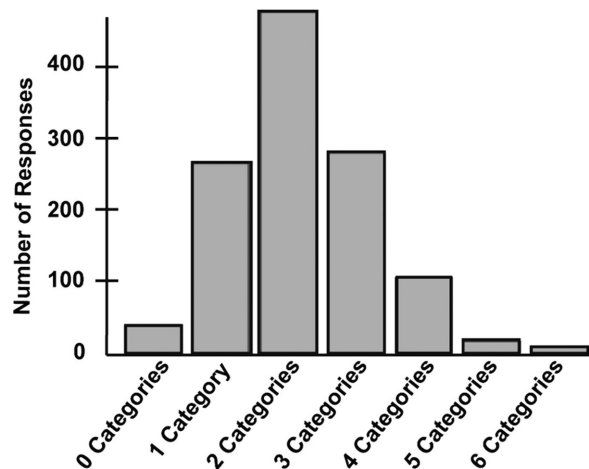
more descriptive of the body as a scalar level, but still vague about the molecular/cellular-level details of Matter Converted to Energy: "It was used up as energy in the body." Fewer responses focused exclusively on the cellular level: "Molecules in fat cells are converted into energy so the cells end up shrinking." A very small number of these responses traversed both the organismal and cellular levels: "THe [*sic*] body converts the fat in the fat cells into energy instead of fat. So the 15 pounds of mass is converted to energy. Fat cells never decrease but they can get smaller." The second most commonly occurring idea in these single-category responses was the organismal-level idea of Exhalation ($n = 54$; Figure 4). Similar to the single-category responses in Matter Converted to Energy, some of the Exhalation responses did not explicitly mention the organismal-level human body beyond discussing the process of exhalation, while other responses specifically referenced the body.

A small number of single-category responses were focused at the cellular level, containing the Normative ideas of either Correct Molecular Products or Molecular Mechanism. Of these responses, about half traced matter across the organismal and cellular levels: "The 15 pounds that my friend lost went into his/her growth, maintenance, waste, and cellular respiration, because matter can never be created or destroyed, only transferred." A few responses focused only on the cellular level: "CO_2."

Clustering and Correlational Analyses Reveal Complex Relationships between Normative and Developing Ideas in Student Responses

To further examine relationships between rubric categories in the student responses, we employed hierarchical clustering
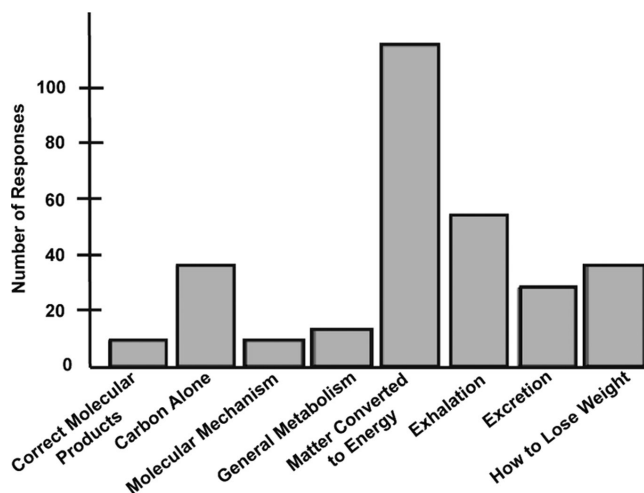

FIGURE 3. The majority of student response contain between one and three ideas. The graph depicts the total number of student responses vs. increasing total numbers of ideas (e.g., how many responses contain 0, 1, 2, 3, 4, 5, and 6 total ideas). We had 39 responses that did not contain any of the ideas from our analytic rubric. Based on the definitions of our analytic rubric categories (Table 1), the maximum number of ideas a response can contain is six.

FIGURE 4. The majority of single-category responses focus either on the organismal scale or no scale at all. The graph depicts the number of responses ($n = 265$) from our data set that contained a single idea from our analytic rubric. Of these, the majority of responses contained the Developing Matter Converted to Energy idea ($n = 115$), followed closely by responses containing the Normative Exhalation idea ($n = 54$). Most of these responses discussed weight loss at the organismal scale or no scale at all. A small portion of these single-category responses discussed ideas at the cellular level.
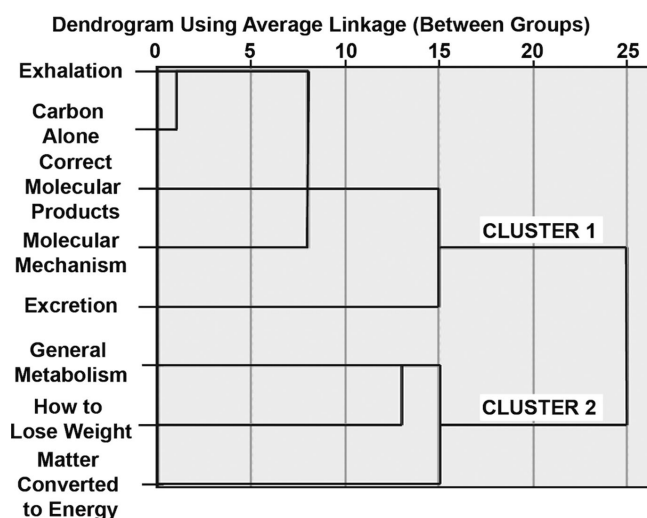
Dendrogram Using Average Linkage (Between Groups)

FIGURE 5. Normative and Developing rubric ideas cluster together. Our hierarchical clustering analysis shows that our Normative ideas of Correct Molecular Products, Exhalation, and Molecular Mechanism cluster together, while the Developing ideas of General Metabolism, How to Lose Weight, and Matter Converted to Energy cluster together. Exceptions are the Developing ideas of Carbon Alone (found in the normative cluster), and Excretion (found equidistant between the Normative and Developing clusters). These analyses, together with correlation analyses (see Supplemental Table S1), underscore the complicated relationships between our rubric categories in student answers.

analysis using rubric categories as the clustering variable. We found our rubric categories to cluster in two major groupings, as shown in Figure 5. The first cluster (cluster 1) contains the Normative ideas of Exhalation, Correct Molecular Products, Molecular Mechanism, and the Developing idea of Carbon Alone. The second cluster (cluster 2) contains the Developing ideas of General Metabolism, How to Lose Weight, and Matter Converted to Energy. Although technically placed in cluster 1, the Developing idea of Excretion is equidistant between both clusters. We were interested to see that the hierarchical clustering breakdown mostly follows our designation of Normative versus Developing ideas. The exception in cluster 1 is the idea of Carbon Alone, which, although indicative of student descriptions at a molecular level, is not a completely correct description of weight loss. The position of Excretion between the two clusters reflects the conflicting feedback from our expert users: Some gave us feedback that they believed Excretion should be part of normative ideas, while others agreed that this was a less normative idea. This result supports the action of labeling this idea as Developing. In some cases, Excretion is used with other Normative ideas, for example, when explaining what happens to water molecules produced during catabolism of fats, as in the following student response: "The mas [*sic*] was exhaled as co2 [*sic*] and excreted as water as urine, sweat, and even tears." In other cases, Excretion is used in a more naïve way, suggesting that physiological waste is the process that accounts for the significant portion of weight loss (Wilson *et al.*, 2006): "The mass was excreted out of her system or burned off during physical activity." These responses also emphasize the importance of context for the Developing ideas. In the former response, the

Developing Excretion ("excreted as water as urine, sweat") occurs with the Normative Exhalation ("exhaled") and Correct Molecular Products ("co2"), and is used in a largely normative way. In the latter response, Excretion ("excreted out of her system") appears to be used, together with the Developing idea of How to Lose Weight ("or burned off during physical activity"), as the main way the mass exits the body. The second response is a less normative and vague use of Excretion, without accounting for molecular processes and products.

Correlation analysis reveals varying degrees of correlation between most pairs of rubric categories (see Supplemental Table S1). After a Bonferroni correction for multiple significance tests, we discuss only correlations that are significant at < 0.001. The Normative Correct Molecular Products and Exhalation have the highest positive correlation coefficient ($r = 0.653$; two-tailed $p = 0.000$), commensurate with their categorization as Normative ideas. Both of these ideas are significantly (two-tailed $p = 0.000$) negatively correlated with the Developing idea of Matter Converted to Energy ($r = -0.392$ for Correct Molecular Products; $r = -0.476$ for Exhalation). The Normative idea of Molecular Mechanism*s* occurs relatively infrequently (12% of student responses), which results in weak correlations. This idea is most significantly correlated with the Normative idea of Correct Molecular Products ($r = 0.156$; $p = 0.000$). Molecular Mechanism is significantly negatively correlated with the ideas of General Metabolism ($r = -0.177$; $p = 0.000$), consistent with our rubric rules. It is also negatively correlated with the Developing idea of Matter Converted to Energy ($r = -0.098$; $p = 0.001$). Interestingly, Molecular Mechanism is also positively correlated with the Developing idea of How to Lose Weight ($r = 0.114$; $p = 0.000$). This may be represented by responses that discuss ideas of How to Lose Weight and explain these ideas using the underlying Molecular Mechanism.

Our Developing ideas also have complicated correlations with other rubric categories. For example, the Developing idea of Matter Converted to Energy has a small positive correlation with the Developing idea of How to Lose Weight ($r = 0.192$; $p = 0.000$). The Developing idea of Carbon Alone also exhibits complicated correlations. We labeled this idea as "Developing," because responses containing this idea give an incomplete description of where the mass goes. Consistent with our definition, the developing Carbon Alone is negatively correlated with the Normative idea of Correct Molecular Products ($r = -0.218$; $p = 0.000$). However, it is positively correlated with the Normative idea of Exhalation ($r = 0.156$; $p = 0.000$). Responses that contain this combination may be largely normative, because the idea of Exhalation may give context for discussions of Carbon Alone.

Two-Category Responses Trace Matter across Scales
To dive deeper into specific co-occurrences, we next analyzed how many of our responses contained two ideas from our analytic rubric using our web diagram (Figure 6). This visualization allowed us to quantify specific co-occurrences in a way that our hierarchical cluster analysis from the preceding section could not. Just over one-third (40%, $n = 477$) of responses in our data set were two-category responses. Figure 6 shows the relative co-occurrences of all possible pairs of rubric categories. We were surprised at the diversity of co-occurring pairs of rubric categories, and we began our analysis by focusing on how responses in the most prominent pairs of rubric categories
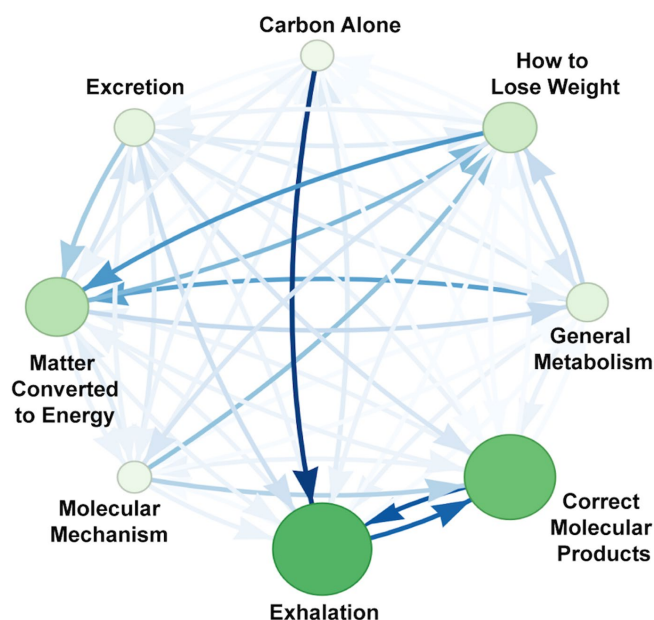
FIGURE 6. Categories can co-occur with each other in multiple combinations. This web diagram shows the co-occurrences of rubric categories for all "two-category" responses. Circles (nodes) represent rubric categories, while the arrows between nodes represent co-occurrences of category pairs. The size and color of each node indicate the number of responses in each category. The arrows point in the direction of connection; for example, the arrow between Carbon Alone and Exhalation indicates the percentage of responses containing ideas of Carbon Alone that also contain ideas of Exhalation (the reverse is not true). The color of the arrow represents the shared percentage, the larger the percentage, the darker the arrow. The largest co-occurrence was that of Exhalation and Correct Molecular Products (*n* = 215), followed by Matter Converted to Energy and How to Lose Weight (*n* = 64).

traced matter across scales. The combination of Normative Exhalation and Normative Correct Molecular Products was found in 215 of the total 477 responses. The majority of these traced matter across organismal and cellular levels: "Exhaled as CO2 [*sic*]." A small number of responses provided an unclear source for the cellular CO_2 product: "Most of the weight was released as co2 [*sic*] into the air." The second most common co-occurrence was that of the Developing ideas of Matter Converted to Energy and How to Lose Weight (*n* = 64; Figure 6). Of these, we found about half of responses focused on vague matter conversions at the organismal level, either through direct references to the body or to activities that the body performs (e.g., exercise): "The fat was converted into energy through exercise, which remained in his body and was likely consumed." Some responses did trace matter across organismal and cellular levels: "The fat is used to fuel the body. When you lose fat it is burned up to create energy. The mass is burned up and the fat cells shrink." Many responses discussed "fat cells" without discussing specifics about the processes of adipocytes, similar to the preceding example. This made it challenging for us to determine whether the students truly understood the function of "fat cells" or had incorrect conceptions.

We were also interested to see other less prevalent co-occurrences. For example, although 45% of two-category responses

contained the Normative Exhalation and Correct Molecular Products ideas, as described earlier, a small number (*n* = 28) of responses contained Normative Exhalation ideas together with the Developing idea of Carbon Alone. All of these responses traversed scales similarly, linking the cellular-level discussion of what appears as elemental carbon in descriptions of the organismal process of exhaling out excess weight. Some responses seemed to use it as a shorthand for CO_2: "The 15 pounds were exhaled as carbon." Others appeared to trace it correctly as building blocks of organic matter: "A large portion of one's mass is comprised of carbon. When glucose and fat stores are used up to facilitate activity, the body tissue releases carbon as a waste product. That carbon is carried to the lungs by the veins where it then can be exhaled into the air." These multiple contexts of Carbon Alone justifies its classification as a Developing idea.

The less frequent co-occurrences of our Developing Matter Converted to Energy category highlighted responses that traced matter across scales in largely nonnormative ways. Most responses in which Matter Converted to Energy co-occurred with How to Lose Weight (*n* = 64) focused on the organismal level of description: "He used his mass as energy. Because he was no longer creating as much (by eating better) his body used what was stored." Very few of these responses discussed cells (in our opinion, a minimum requirement of the cellular level) of any kind. The following response is another example of a description missing a cellular description: "Our bodies convert molecules in our fat cells to energy that we can use, causing the fat cells to shrink." Similarly, about half of the responses in which the Matter Converted to Energy idea co-occurred with General Metabolism (a total of *n* = 33) traced matter to the cellular level: "The weight that was lost was fat from her body which was converted to glucose as a source of energy for the body. The 15 pounds were turned from fat into ATP and released as heat in chemical reactions." These two co-occurrences of a common rubric category with different partners (Matter Converted to Energy + How to Lose Weight and Matter Converted to Energy + General Metabolism) further illustrated to us the context-dependent nature of our rubric categories.

In summary, we have described here the ways in which the most common two-category co-occurrences in our data set traversed biological scales. The majority of these responses traced matter across scales involving ideas about Exhalation and Correct Molecular Products. The second-highest co-occurrence consisted of less normative, developing descriptions discussing ideas about Matter Converted to Energy and How to Lose Weight, which were mostly confined to the organismal level. Both the Normative Exhalation and the Developing Matter Converted to Energy categories also co-occurred with other rubric categories, whose combined presence in the response determined whether the response was complete or incomplete. Overall, these findings begin to illustrate the extent of mixed ideas that students have when they think about organismal-level processes like weight loss.

Three- and Four-Category Responses Highlight Students' Mixed Thinking about Weight Loss
We next analyzed more elaborate answers that included three or four categories. These responses contain ideas comparable to the average number of ideas in expert responses (*X* = 3.81 ± 0.60 ideas). Analysis of these responses proved to be more

challenging than the analysis of two-category responses, due to the myriad possible combinations of three and four rubric categories. To illustrate trends for these responses, we here choose a single rubric category and trace it across exemplar combinations. For example, the highest-occurring rubric category for both three- and four-category responses was that of Exhalation. The majority of three-category responses that contained Normative ideas about Exhalation also contained the Developing idea of Excretion ($n = 78$). The following response contains these two ideas as well Normative Correct Molecular Products and is largely normative*:* "It was mostly breathed out in the form of CO2 [*sic*] while some was excreted as urine or sweat." Responses containing these three ideas typically traced matter in normative ways at the level of physiological systems and the entire body; however, this example missed a complete conversion of fat into CO_2 (i.e., Molecular Mechanism). Four-category responses containing Exhalation ideas highlighted other trends in student thinking. Similar to the normative three-category example, some four-category responses containing Exhalation were largely normative. The following response is generally correct, because it also contains the Normative ideas of Correct Molecular Products and Molecular Mechanism in addition to the Developing idea of How to Lose Weight: "When exercising cellular respiration removes carbon from glucose and other molecules that follow glycolysis and the CO2 [*sic*] combines with oxygen. His breathed out his mass (CO2) [*sic*]." Alternatively, some four-category responses containing Exhalation ideas were less specific, and thus less normative: "Fats are converted into glucose, the glucose is broken down into energy and co2 [*sic*] which get expelled by breathing." Despite the Normative ideas of Exhalation and Correct Molecular Products ("co2"), this response also describes the Developing, less specific ideas of General Metabolism ("Fats are converted into glucose") and Matter Converted to Energy ("glucose is broken down into energy").

Similarly, tracing the Developing category of Matter Converted to Energy across three- and four-category responses highlighted rich diversity in the ideas that students have about weight loss. The majority of three-category responses containing ideas of Matter Converted to Energy ($n = 75$) also contained ideas about How to Lose Weight. Most of these responses contained nonnormative, superficial ideas about tracing matter during weight loss, such as the following response that also contains the Developing Excretion idea: "My friend was on a diet and lost 15 pounds. because [*sic*] of his diet he was able to lose mass but the matter was not destroyed. It was converted into heat through exercise and sweat and was lost in waste (stool). The energy and matter was [*sic*] converted." A small number of these responses, however, contained the Normative idea of Molecular Mechanism: "Her intake of fats and sugars has decreased. Therefore, her stored fat is being broken down and used by the cell for glycolysis. The weight she lost was consumed by her cells and used to make energy." Although this response does show some nonnormative ideas (e.g., of Matter Converted to Energy: "The weight she lost was consumed by her cells and used to make energy"), the student correctly identifies the process of glycolysis as a means of mass transformation. A larger fraction of four-category rather than three-category responses containing Matter Converted to Energy were of a normative nature. In four-category responses, the Normative ideas of Correct Molecular Products ($n = 34$) and Exhalation ($n = 35$) occurred most commonly with Matter Converted to Energy. Thus, many four-category responses contain a large fraction of normative ideas, such as the following response that also describes Normative Molecular Mechanism: "The body uses it in cellular respiration it is given off in heat water and carbon dioxide when you breath out." Here, the Matter Converted to Energy category is expressed in the student's description of heat as a by-product of cellular respiration, which is not inherently an incorrect idea. The three Normative ideas in the response also provide a largely normative context for the Developing Matter Converted to Energy idea. Some four-category responses, however, only hinted at a correct cellular-level understanding, similar to three-category responses containing Matter Converted to Energy: "Fats are converted into glucose, the glucose is broken down into energy and co2 [*sic*] which get expelled by breathing." This contained ideas about Matter Converted to Energy ("the glucose is broken down into energy"), Correct Molecular Products ("co2"), and Exhalation ("get expelled by breathing"), together with ideas about General Metabolism ("Fats are converted into glucose"). Although the overall tracing of the mass is correct, the details are less clear. The student does not explain the processes by which fat is converted to glucose, for example, and the student's language also seems to indicate an immediate conversion of glucose into the nebulous concept of "energy." Both the context provided by the General Metabolism idea and the wording of the Matter Converted to Energy idea contribute to this response's vague and nonnormative character. Tracing the Matter Converted to Energy category, as an exemplar, across three- and four-category responses thus highlighted how important the other co-occurring categories are in determining whether a student response is normative or nonnormative.

In summary, we have found that responses containing three and four categories illustrate a rich heterogeneity in student thinking about weight loss. Five- and six-category responses showed similar heterogeneity, whose description can be found in the Supplemental Material. Our analyses indicate that combinations of ideas contribute to the overall Normative or Developing nature of student responses, which is a consideration that we suggest instructors keep in mind when teaching about this topic. In the next sections, we present one possible way to assess responses based on co-occurrence of ideas: our descriptive model framework (Figure 1).

Characterization of Student Ideas in Descriptive Models

Our results indicated that student tracing of matter in the context of weight loss is highly heterogeneous and rich in nature. To characterize this heterogeneity, we applied an adapted framework of student descriptive models to our data, as explained in *Methods* and Figure 1. Briefly, 1) if student responses contained one or more only Normative ideas, they fit Scientific descriptive models; 2) if student responses contained both Normative and Developing ideas, they fit Mixed descriptive models; and 3) if student responses contained one or more only Developing ideas, they fit Developing descriptive models. Following our definition of a Scientific model of student description, we found that 28% of the responses fell within this category (Figure 7). As for our Developing models of student description, 33% of the responses contained at least one Developing idea and no Scientific idea. Finally, 35% of the responses contained at least one Developing and one Scientific idea and were classified as
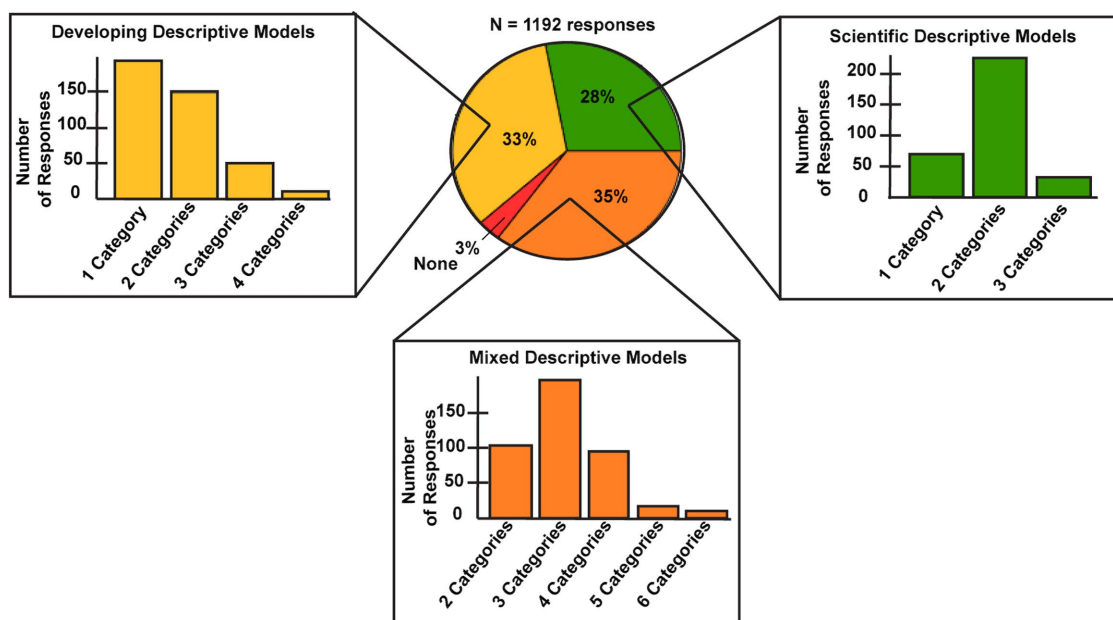
FIGURE 7. Ideas identified in each of the student descriptive models. The figure outlines the percentage of Developing, None, and Mixed and Scientific student descriptive models (adapted from Moharreri *et al.*, 2014) that we found in our data set (*N* = 1192). As the separate column graphs show, each descriptive model (except for the None model) is made up of a diversity of ideas. The average number of categories per response is 1.9 ± 0.6 for a Scientific model, while it is 1.7 ± 0.8 for the Developing model. The average number of responses for a Mixed model is slightly higher at 3.1 ± 0.9 (1.7 ± 0.6 Scientific ideas and 1.4 ± 0.7 Developing ideas).

Mixed student descriptive models. Responses that contained none of our rubric ideas were classified as "None" models (3% of responses). We would like to note that, because of the presence of at least one Scientific idea in our Mixed descriptive models, we place this model type as having intermediate sophistication between Scientific and Developing descriptive model types.

We were interested in gaining further insight into the diversity of ideas in each of our student descriptive models. Similar to the trends shown earlier, each of these student descriptive models (except for the None model) contained responses with a varying number of ideas (Figure 7). We found that Scientific descriptive models contained an average of 1.9 ± 0.6 ideas per response, with the majority of responses containing two ideas (*n* = 227). Developing models contained a similar average of 1.7 ± 0.8 ideas per response, with the largest number of responses containing a single idea (*n* = 193). Mixed models contained an average of 3.1 ± 0.9 ideas (with an average of 1.7 ± 0.6 Scientific ideas and an average of 1.4 ± 0.7 Developing ideas) per response. The majority of Mixed responses contained three ideas (*n* = 196), similar to the average of ideas in our instructor responses (*X* = 3.82 ± 0.6 ideas).

In summary, we have framed our results within the context of descriptive models based on the context provided by co-occurring ideas. We present this analysis with the intent to illustrate to instructors the complex connections between the ideas that students have in thinking about biological processes in the familiar context of weight loss.

Descriptive Models Contain Diverse Student Ideas across Scales

Given the multiple ideas that we found in each of our student descriptive models, we were interested in how our rubric cate-gories were distributed across the descriptive models (Figure 8). In terms of our Scientific ideas, we found that Correct Molecular Products and Exhalation ideas occurred almost equally in Scientific (Correct Molecular Products = 48%; Exhalation = 46%) and mixed (Correct Molecular Products = 52%; Exhalation = 54%) model responses. This finding indicates that these ideas were reasonably accessible to students but could easily be combined with less scientific ideas. Interestingly, Molecular Mechanism occurred more frequently in responses fitting a Mixed (62%) rather than a purely Scientific (38%) descriptive model. This may be because this category's cellular scale made it slightly more difficult for students to incorporate in their descriptions than the previous two Scientific ideas.

For our Developing ideas, we were curious to investigate the proportion of ideas that occurred in Developing versus Mixed descriptive models. The Mixed descriptive models were especially interesting, because these responses exhibit at least one Scientific idea in addition to the Developing idea under consideration. Carbon Alone occurred significantly more often in Mixed (92%) rather than Developing (8%) model responses, indicating that this idea is more often associated with Scientific-like explanations in our data set. Such a trend may be consistent with the literature that indicates that thinking at the cellular/molecular level is difficult for students: thus, when ideas at this level do occur, students are more likely to group them with other Scientific ideas. The next-highest Developing idea in Mixed models was that of Excretion (64%), similarly indicating that this idea occurred frequently with Scientific ideas. These results are also consistent with our hierarchical clustering results, in which Excretion was equidistant from our Normative and Developing clusters. Similar to the Normative ideas of Correct Molecular Products and Exhalation, responses containing
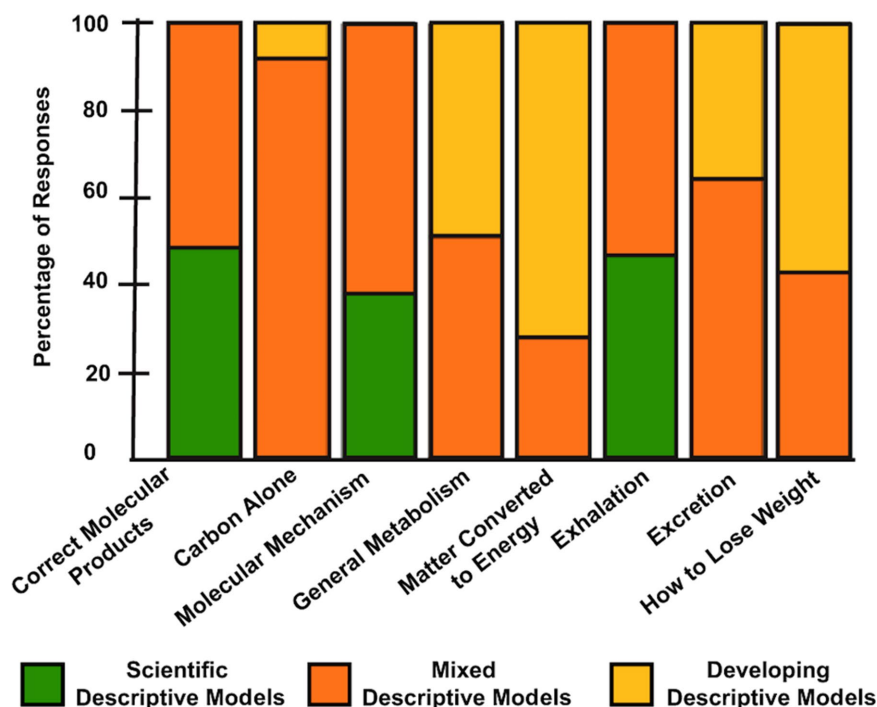
FIGURE 8. Student ideas distributed across different descriptive models. Each of our rubric categories occurred in two separate student descriptive model types. The Scientific ideas of Correct Molecular Products and Exhalation occur about equally in Scientific and Mixed descriptive models, while the third scientific idea of Molecular Mechanism occurs less frequently in Scientific models (38%) than in Mixed models (62%). Of the Developing ideas, that of Carbon Alone occurred most frequently in Mixed (92%) than in Developing (8%) models.

were interested to see that greater than 50% of responses classified as Scientific and Mixed descriptive models traversed scales (75 and 90%, respectively). About half of responses that fit Developing descriptive models (41%) exhibited the ability to traverse scales, which may indicate an additional dimension along which these students may need support in developing their understanding.

In summary, we found that about half of responses containing the Normative ideas of Correct Molecular Products and Exhalation aligned with Scientific descriptive models, perhaps indicating that students find these normative ideas easier to access when describing weight loss. Similarly, about half of responses containing each of the Developing ideas of General Metabolism, Excretion, and Carbon Alone occurred in Mixed as opposed to Developing descriptive models. This is in contrast to student responses discussing the other Developing ideas of Matter Converted to Energy and How to Lose Weight; less than 50% of responses containing each of these ideas fit a Mixed model of description. Additionally, our analyses revealed that responses fitting Scientific and Mixed models of description tended to move across scales more often than those aligned with a Developing model of description. Taken together, our results illustrate diverse, heterogeneous ways in which students combine ideas and traverse scales in their descriptions of weight loss.
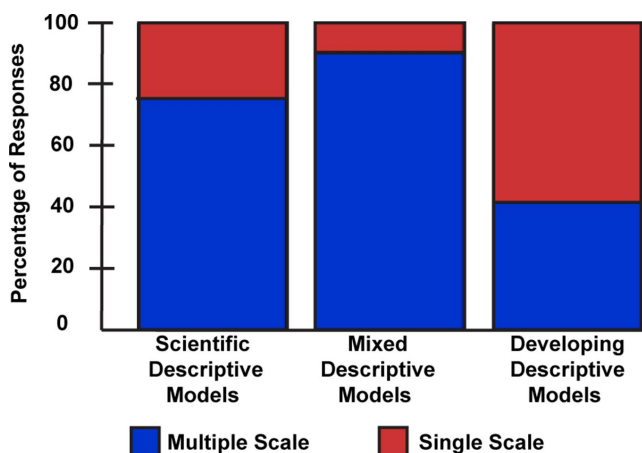
the Developing idea of General Metabolism occurred about equally in Mixed (51%) and Developing (49%) descriptive models. Matter Converted to Energy and How to Lose Weight occurred in responses that most frequently fit Developing descriptive models (Matter Converted to Energy = 72%; How to Lose Weight = 57%). The fact that these two Developing ideas occurred more frequently in less normative Developing models is consistent with the vague nature of these two ideas. Students' conversion of mass to energy by unspecified processes has been well documented (Wilson *et al.*, 2006), and comparable vague language is sometimes reinforced by instruction and/or learning materials such as textbooks. How to Lose Weight is a similarly vague idea, often supported by popular culture and organismal-level reasoning.

Furthermore, we investigated how students explicitly traversed scales as part of their explanations across the various reasoning models. We analyzed our student descriptive models for the distribution of responses at either a single scale or multiple scales (Figure 9). Responses were classified as incorporating only a single scale if all the ideas contained in that response were at either the cellular (i.e., Correct Molecular Products, Carbon Alone, Molecular Mechanism, General Metabolism, Matter Converted to Energy) or the organismal (Exhalation, Excretion, How to Lose Weight) level. Responses were classified as being at multiple scales if they contained at least one idea at the cellular level and a second idea at the organismal level. We

FIGURE 9. Distribution of scales across student descriptive models. We tracked the distribution of responses traversing levels (blue) or confined to a single scalar level (red) across different descriptive models. Note that Scientific and Mixed models show a greater number of responses that traverse scalar levels than do responses fitting a Developing descriptive model.

DISCUSSION

"Pathways and transformations of matter and energy" (AAAS, 2011, p. 13) have been identified as a key concept in biology that students must master to become scientifically literate citizens. Previous work investigated students' ability to trace matter as a potential learning strategy for complex biological processes (Wilson *et al.*, 2006). Here, we present our investigations into the subtleties of student tracing of matter in the context of human weight loss. Our results reveal three Normative and five Developing ideas across both the organismal and cellular scales that students use to think about this familiar process (Table 1).

In addition to observing these categories of tracing matter individually in student responses, our analyses also enabled us to observe how students combine these ideas when thinking about weight loss. We found that most students discussed two or more ideas when describing weight loss (in contrast to expert responses containing an average of about three ideas) and that co-occurrence of ideas provides important context regarding the normative or nonnormative nature of student understanding.

Student Ability to Traverse Scales in the Context of Weight Loss

An added concept that students need to master to gain expertise in biological processes is that of traversing biological scales. Hartley and coworkers' (2011) work with diagnostic question clusters revealed that students often used an organismal scale in their responses. The authors hypothesized that this may be due to the fact that students are more comfortable at this level. Our results support this hypothesis, because our organismal categories (Exhalation, Excretion, and How to Lose Weight) occurred in higher percentages in our data set compared with most of our cellular categories (with the exception of Correct Molecular Products; see Figure 2). Mohan and colleagues (2009) showed that students at lower levels of their learning progression on carbon cycling either did not reason beyond the organismal scale or reasoned to the organ, but not the cellular, scale. Our tracking of scales across student descriptive models exhibited slightly different trends. Of the three descriptive model types, responses fitting the Developing model showed the lowest percentage of traversing scales (41%), while well more than 50% of Scientific-model and Mixed-model responses traversed scales (75 and 90%, respectively). We were interested to see that more Mixed-model than Scientific-model responses traversed scales. This may indicate that students with a wider diversity of normative and nonnormative ideas are better able to reason across scales than students with only normative or only nonnormative ideas.

Students' Mixed Ideas in the Development of Expertise about Weight Loss Mechanisms

Our analysis of student responses to our weight loss item enabled us to characterize the diversity of ideas that students hold when discussing cellular processes. We were interested to find that student responses in our data set contained an average of about two ideas, and that most responses contained between one and three ideas. We thus applied a modified approach of Moharreri and colleagues (2014) to characterize models of student description of weight loss (see *Results*). Scientific and Developing descriptive models contain similar average numbers of ideas (1.9 ± 0.6 and 1.7 ± 0.8 respectively), while Mixed models had a slightly higher average number (3.1 ± 0.9) of ideas. These results emphasize that heterogeneity is characteristic of student thinking about weight loss, regardless of the normative or nonnormative nature of ideas.

The Developing and Mixed descriptive models contain Developing ideas that indicate lack of clarity regarding transformations of matter and energy. Developing ideas such as How to Lose Weight, Matter Converted to Energy, and General Metabolism are examples of vague descriptions that appear in both these models. The imprecision of these ideas can provide various insights from both the near-term perspective of student conceptual understanding and the longer-term perspective of development of expertise.

Regarding the near-term perspective of conceptual change, the Developing ideas in our rubric (Table 1) are consistent with persistent and incomplete conceptions that students have about matter and energy transformations. Extensive previous research has outlined similar conceptual difficulties that students have in describing matter-transforming processes. Wilson and colleagues (2006) described the variation that students have when thinking about mass in human weight loss, which ranged anywhere from correct identification of processes and products to incorrect or oversimplified conversions of matter. The trends in vague descriptions that Wilson and colleagues uncovered coincide with the significant percentages of our student responses that occur in the categories Excretion (23%) and General Metabolism (18%). The work of Hartley and colleagues (2011) and Jin and colleagues (2013) showed that students have difficulty reasoning at the level of atoms and molecules, which may further complicate how students understand carbon-transforming processes. In their progress variables of *explaining matter* and *explaining mass*, Jin and colleagues (2013) specify that student reasoning with atoms and molecules occurs at high level 3 (out of four levels). The difficulty students have reasoning at this scalar level coincides with the relatively low percentage of our student responses that fall into the Developing Carbon Alone category (5%), as well as the Normative category Molecular Mechanism (12%).

Student conceptions of energy appears to be even more complicated and varied. Anderson and colleagues (1990) described students' broad application of the term "energy." Their work showed that significant numbers of students believe that humans can gain energy from nonfood sources such as sunlight and exercise, similar to our Developing Matter Converted to Energy and How to Lose Weight categories. As Anderson and colleagues (1990) note, students may be unable to fully grasp the types, complexity and centrality of energy transformations in biology because of such informal, incomplete, and persistent misunderstandings. Although Hartley and colleagues (2011) developed their principle-based assessments to specifically address students' incomplete and informal understandings about matter and energy transformations, they found that 16% of students still used informal descriptions to explain such processes. They found informal reasoning to be persistent, and cited students' lifelong exposure to such colloquial descriptions (e.g., "'burning off' fat," Hartley *et al.*, 2011, p. 73) as the culprit. Furthermore, the authors found that students often used

energy as a "fudge factor" (Hartley *et al.*, 2011, p. 69) to avoid providing specifics about cellular processes.

These varied student conceptions about matter and energy pathways may have complicated long-term effects during the development of expertise. The presence of informal and fragmented ideas like those described in our Matter Converted to Energy and How to Lose Weight categories is consistent with the acclimation stage in Alexander's model of domain learning (Alexander, 2003). Alexander (2003) described this stage as the earliest in the development of expertise, when "acclimating learners' ability to discern the difference between accurate or inaccurate and relevant or tangential information is understandably hampered" (p. 11). Similarly, the trends we observed in our expert responses were consistent with Alexander's later competency and proficiency stages of expertise. Characteristic of the competency stage, our expert responses showed the quantitative difference from student responses by containing more ideas on average per response compared with student responses: 3.82 ± 0.6 and 2.19 ± 1.07 ideas, respectively. Likewise, our expert responses showed a "synergy among components … required for movement from competence to expertise" (Alexander, 2003, p. 12): We found that even when instructors included Developing ideas in their responses, they followed up with more specific, Normative ideas to support their description. The changes between the three stages of expertise are nontrivial and are likely to take time as students make the necessary connections between ideas.

Additionally, our Developing ideas are consistent with seminal work in expert–novice literature detailing novices' tendencies to focus on surface features of problems and representations. Chi and colleagues (1981) performed several studies to characterize the problem-sorting abilities of experts and novices in physics. Their work revealed that novices tended to focus on surface features (e.g., objects or terms directly mentioned in the problems) when grouping problems, whereas experts grouped problems based on their underlying concepts. The high occurrence of the How to Lose Weight category in our data set (26%) may be explained by students focusing on informal or superficial explanations of weight loss to which they are exposed in everyday life. Chi and colleagues (1981) also found that an "advanced novice" (p. 133) sorted problems with patterns distinct from both less experienced novices and an expert. Such diversity is captured in the variety of combinations of ideas in our Developing and Mixed descriptive models. This diversity is one reason that student understanding is slow to develop and is thus a key factor instructors must keep in mind when teaching complex biological processes.

Students' Heterogeneous Thinking: Instructional Implications

The work presented here shows that students in introductory biology courses across a wide range of institutions have a range of normative, scientific ideas and nonnormative ideas about these core concepts that have also been documented in other research. How should instructors address this heterogeneity? While detailed instructional suggestions are beyond the scope of this paper, the conceptual change literature provides some possible approaches. Research on conceptual change in science

education over decades has moved from simple ideas of student "misconceptions" being coherent and theory-like toward diSessa's view of "knowledge-in-pieces," in which students hold multiple, contradictory ideas concurrently (diSessa, 2006, 2008; Vosniadou *et al.*, 2008). Furthermore, this heterogeneity may make this topic more challenging for instructors to design instruction to address than concepts for which students have no prior knowledge or conceptions. Chi (2008) identifies three types of conceptual change learning. In the first, students have no prior knowledge, and learning consists of *adding* new knowledge. In the second, students have some correct prior knowledge, and learning consists of *gap filling*, providing additional details. For example, students may know that matter cannot be converted to energy, but may not understand the cellular processes by which the matter is converted to CO_2 and lost through exhalation. The third, and most difficult, is *changing* prior incorrect knowledge. This is further complicated by the grain size of the incorrect knowledge. If the grain size is at the level of single ideas, then instruction that refutes the ideas by showing how the incorrect idea is not compatible with the correct one can be successful. If the nonnormative ideas are a collection of ideas, then simple refutation is less successful, and instruction should seek to transform student mental models into a more normative model. This is the more difficult conceptual change to accomplish. To do so, instructors must provide students with multiple opportunities to create models and use those models to make and test predictions from their models. Simply telling students that their ideas are incorrect does not produce lasting conceptual change that persists beyond the next multiple-choice exam.

Here, we have presented evidence for the complex and heterogeneous thinking of college students about the core concept of energy and matter transformations. Owing to various ideas and connections that students can make when thinking about these matter-transforming processes, we encourage instructors to design assessments and instructional interventions to accurately assess the extent of students' expertise and the areas in which they need support. Assessments designed with student thinking in mind will enable instructors to meet each student where he or she is in his or her development in thinking about the matter- and energy-transforming processes. Constructed-response items such as the one we have described here are effective for such assessments. Our group has a long history of developing such items, along with computerized scoring models to provide instructors with rapid feedback about the types of ideas their students employ when answering these assessments. The weight loss item presented here is available for instructor use, along with its automated scoring model. The prompt, along with other items, scoring models, and instructional tools focusing on big ideas in biology can be found here: www.msu .edu/~aacr.

Limitations of Our Classification of Student Responses

Our analytic rubric and automated scoring model for our weight loss item provide rapid analysis of large sets of student responses. However, there are limitations to our approach. Application of our rubric categories is limited in that the categories depend heavily on the students' written words. Although we attempt to score precisely what the students have written and not what we *interpret* the responses to mean, student

language is not correspondingly precise. Sometimes our analysis may not accurately capture student meaning. This limitation affects the performance of the predictive model: Our ensemble of machine-learning algorithms are restricted by the upper limits of human interrater reliability. Therefore, the ensemble may sometimes mischaracterize student responses. Additionally, because the impartiality we attempt to impart to our scores and corresponding model may differ from learning goals that a given instructor has for a class, our scores are not meant to act as grades for student responses. Despite these limitations, however, our tools are still able to provide a broad overview of a collection of student responses. We encourage others to pursue the causes and further subtleties of student misunderstandings and effective interventions that correct students' Mixed and Developing descriptions about weight loss and other matter-transforming processes.

REFERENCES

Alexander, P. A. (2003). The development of expertise: The journey from acclimation to proficiency. *Educational Researcher*, *32*(8), 10–14.

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.

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

Carnegie Classification of Institutions of Higher Education. (2017). About Carnegie Classification. Retrieved October 17, 2018, from http://carnegieclassifications.iu.edu

Chi, M. T., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science*, *5*(2), 121–152.

Chi, M. T. H. (2008). Three types of conceptual change: Belief revision, mental model transformation, and categorical shift. In Vosniadou, S. (Ed.), *Handbook of research on conceptual change* (pp. 61–82). Hillsdale, NJ: Erlbaum.

Choi, S.-S., Cha, S.-H., & Tappert, C. C. (2010). A survey of binary similarity and distance measures. *Systemics, Cybernetics and Informatics*, *8*(1), 43–48.

Cooper, P. (1993). Paradigm shifts in designed instruction: From behaviorism to cognitivism to constructivism. *Educational Technology*, *33*(5), 12–19.

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, UK: Cambridge University Press.

diSessa, A. A. (2008). A bird's-eye view of the "pieces" vs. "coherence" controversy (from the "pieces" side of the fence). In Vosniadou, S. (Ed.), *International handbook of research on conceptual change* (pp. 35–60). New York: Routledge.

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

Ertmer, P. A., & Newby, T. J. (1993). Behaviorism, cognitivism, constructivism: Comparing critical features from an instructional design perspective. *Performance Improvement Quarterly*, *6*(4), 50–72.

Harrison, A. G., & Treagust, D. F. (1996). Secondary students' mental models of atoms and molecules: Implications for teaching chemistry. *Science Education*, *80*(5), 509–534.

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.

Haudek, K. C., Moscarella, R. A., Weston, M., Merrill, J., & Urban-Lurain, M. (2015, April 11–14). Construction of rubrics to evaluate content in students' scientific explanation using computerized text analysis. *Paper presented at National Association for Research in Science Teaching Conference (Chicago, IL)*.

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.

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.

IBM. (2016). *IBM SPSS Statistics* (*Version 24*). Armonk, NY.

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.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174.

Meerman, R., & Brown, A. J. (2014). When somebody loses weight, where does the fat go? *BMJ*, *349*, g7257.

Mohan, L., Chen, J., & Anderson, C. W. (2009). Developing a multi-year learning progression for carbon cycling in socio-ecological systems. *Journal of Research in Science Teaching*, *46*(6), 675–698.

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), 1–15.

Moscarella, R. A., Haudek, K. C., Knight, J. K., Mazur, A., Pelletreau, K. N., Prevost, L. B., … & Merrill, J. E. (2016, April 11–14). Automated analysis provides insights into students' challenges understanding the processes underlying the flow of genetic information. *Paper presented at National Association for Research in Science Teaching Conference (Baltimore, MD)*.

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

Nehm, R. H., Beggrow, E. P., Opfer, J. E., & Ha, M. (2012). Reasoning about natural selection: Diagnosing contextual competency using the ACORNS instrument. *American Biology Teacher*, *74*(2), 92–98.

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.

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.

Talanquer, V. (2009). On cognitive constraints and learning progressions: The case of "structure of matter." *International Journal of Science Education*, *31*(15), 2123–2136.

University of New South Wales. (2014, December 16). When you lose weight, where does the fat go? Most of the mass is breathed out as carbon dioxide, study shows. *ScienceDaily*, Retrieved March 21, 2019, from www.sciencedaily.com/releases/2014/12/141216212047.htm

Mixed Student Ideas about Weight Loss

Vosniadou, S., Vamvakoussi, X., & Skiopeliti, I. (2008). The framework theory approach to the problem of conceptual change. In Vosniadou, S. (Ed.), *International handbook of research on conceptual change* (pp. 3–34). New York: Routledge.

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. *Cell Biology Education*, *5*(4), 323–331.

Yoho, R., Urban-Lurain, M., Merrill, J., & Haudek, K. (2018). Structure and function relationships in the educational expectations of professional societies across the STEM disciplines. *Journal of College Science Teaching*, *47*(6), 24–31.

Yune, S. J., Lee, S. Y., Im, S. J., Kam, B. S., & Baek, S. Y. (2018). Holistic rubric vs. analytic rubric for measuring clinical performance levels in medical students. *BMC Medical Education*, *18*(1), 124.