Adv Physiol Educ 47: 762–775, 2023.
First published August 24, 2023; doi:10.1152/advan.00156.2022

american physiological society

ADVANCES IN PHYSIOLOGY EDUCATION

EDUCATION RESEARCH

Physiology Core Concepts

Covariational reasoning and item context affect language in undergraduate mass balance written explanations

Megan Shiroda,[1] Jennifer H. Doherty,[2,3] Emily E. Scott,[4] and Kevin C. Haudek[1,5]

[1]*CREATE for STEM Institute, Michigan State University, East Lansing, Michigan, United States;* [2]*Department of Physiology, Michigan State University, East Lansing, Michigan, United States;* [3]*Lyman Briggs College, Michigan State University, East Lansing, Michigan, United States;* [4]*Department of Biology, University of Washington, Seattle, Washington, United States; and* [5]*Department of Biochemistry and Molecular Biology, Michigan State University, East Lansing, Michigan, United States*

Abstract

Mass balance (MB) reasoning offers a rich topic for examination of students' scientific thinking and skills, as it requires students to account for multiple inputs and outputs within a system and apply covariational reasoning. Using previously validated constructed response prompts for MB, we examined 1,920 student-constructed responses (CRs) aligned to an emerging learning progression to determine how student language changes from low (1) to high (4) covariational reasoning levels. As students' abilities and thinking change with Context, we used the same general prompt in six physiological contexts. We asked how Level and Context affect student language and what language is conserved across Contexts at higher reasoning Levels. Using diversity methods, we found student language becomes more similar as covariational reasoning level increases. Using text analysis, we found context-dependent words at each Level; however, the type of context words changed. Specifically, at Level 1, students used context words that are tangential to MB reasoning, while Level 4 responses used words that specify inputs and outputs for the given Item Context. Further, at Level 4, students shared 30% of language across the six contexts and leveraged context-independent words including *rate*, *equal*, and some form of *slower/lower/smaller*. Together, these data demonstrate that Context affects undergraduate MB language at all covariational reasoning levels, but that the language becomes more specific and similar as Level increases. These findings encourage instructors to foster context-independent, comparative, and summative language during instruction to functionally build MB and covariational reasoning skills across contexts.

NEW & NOTEWORTHY This article builds on the work of Scott et al. (Scott EE, Cerchiara J, McFarland JL, Wenderoth MP, Doherty JH. *J Res Sci Teach* 1: 37, 2023) and Shiroda et al. (Shiroda M, Fleming MP, Haudek KC. *Front Educ* 8: 989836, 2023) to quantitatively examine student language in written explanations of mass balance across six contexts using constructed response assessments. These results present an evaluation of student mass balance language and provide researchers and practitioners with tools to assist students in constructing scientific mass balance reasoning explanations.

constructed response; context; covariational reasoning; mass balance; student language

INTRODUCTION

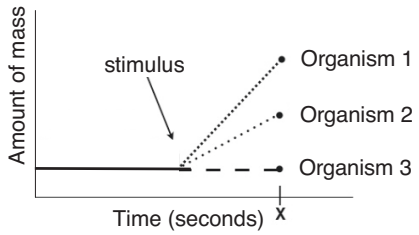
In the last decade, science, technology, engineering, and mathematics (STEM) education has sought to emphasize broad scientific concepts and skills that students can leverage to understand phenomena across multiple contexts. For example, the National Research Council put forth a framework on cross-cutting concepts to be taught in K12 science including the concept "cause and effect," which requires students to be able to "predict and explain events in new contexts" (1). Similarly, in undergraduate biology education, *Vision and Change* called for instruction to be focused on a set of core competencies, including mathematical reasoning, and core concepts, such as transformations of matter and energy, that can be applied across contexts (2). Within physiology, Michael and McFarland (3) described 15 core concepts that are important for students to master to succeed within the discipline. These core concepts are useful to students as they learn to reason as scientists explain novel phenomena using cause and effect to predict outcomes. One such core concept is mass balance (MB; sometimes called matter accumulation or pool and flux reasoning). MB, based on the law of conservation of mass, is a reasoning strategy that can be used to predict whether the amount of material in a compartment will stay the same, increase, or decrease by tracking inputs to and

Correspondence: M. Shiroda (shirodam@msu.edu).
Submitted 29 August 2022 / Revised 27 July 2023 / Accepted 22 August 2023

outputs from the compartment (4). The amount of material in a compartment can have important consequences on organismal function. For example, the amount of neurotransmitter in a synapse determines the relative magnitude of the signal on the postsynaptic neuron and is determined by the rate of release from the presynaptic neuron (input) and the rate of removal from the synapse (output). MB is broadly applicable across STEM and can be applied to understand disparate phenomena such as carbon accumulation in the atmosphere and the buildup of free calcium in the sarcoplasm and can even be applied to track population dynamics by attending to count instead of mass. The majority of studies examining MB reasoning have been within the context of climate change (5, 6); however, recent work has extended MB reasoning to physiology contexts in animals and plants (7). Due to its broad utility, MB is considered a conceptual scaffold that students can learn to apply across contexts to make sense of novel systems (4, 8).

In spite of the usefulness of MB, student difficulty with it has been documented across STEM (5–7, 9). This difficulty is thought to be in part due to the mathematical thinking required to accurately determine how mass will change (5–7). Scott et al. (7) have proposed that to successfully predict the changes in mass undergraduates must identify the relevant fluxes and apply covariational reasoning. Covariational reasoning is a skill that allows a person to conceive two variables changing simultaneously (10, 11). For example, within the context of glucose accumulation in an oak leaf, an important input of glucose in the leaf is photosynthesis while a critical output is cellular respiration, which breaks down glucose. To accurately describe the amount of glucose in a leaf, one must account for increasing or decreasing rates of either photosynthesis or cellular respiration. As part of their analysis, Scott et al. (7) described an emerging four-level learning progression that describes how students develop covariational reasoning in the context of MB reasoning. This learning progression was developed using student thinking about problems situated in six different Item Contexts within physiology across plants and animals. One such generic assessment prompt and the six different contexts are provided in Fig. 1. Each of the items describes a situation in which three organisms undergo an identical change or stimulus that results in an influx of matter to a given space and has an associated graph depicting a change in total mass. In addition to naming a specific organism, each Item Context names specific matter, inputs, and outputs that are relevant to the organism and context. In Level 1 of the emerging learning progression, students relate directional changes in nonflux variables to the changing amount, i.e., students discuss surface features that do not directly relate to MB. In Level 2 explanations, students relate magnitude changes in a single flux with the changes in amounts. In Level 3 explanations, students relate a single flux rate of change variable correctly or relate a net flux rate of change incorrectly with changes in amounts. Finally, in Level 4 explanations, students are integrating magnitude changes of both fluxes in the system to accurately explain how the amounts of mass change in the system. This learning progression also forms the basis for a coding rubric to assess the Level of student reasoning in a written constructed response (CR).

CR is one of many ways to assess student thinking. This assessment method allows students to use their own words to explain complex topics, thereby providing students with an opportunity to engage in the practice of crafting an explanation (12, 13). CRs provide meaningful insight into student thinking for both instructors and researchers, as responses produced by students can reveal performance differences, complex thinking, and unexpected language (14, 15). Previously, student CRs have been used to increase understanding of diverse STEM topics including tracking matter and mass across scales (16) and mechanistic reasoning (17, 18). As previous work has demonstrated that language used in these responses is reflective of student thinking revealed during interviews (7, 15), CRs are a convenient and appropriate source of student language for assessing covariational reasoning and MB explanations.



In **context**, input increases mass while output reduces mass.

a) Explain how organism 1 has more mass compared to organism 2 and 3 at time x given that all organisms have the same rate of input.
b) Explain how organism 3 has no change in mass at time x while organisms 1 and 2 both show increased object given that all three organisms had the same rate of input.

Cat: In the brain, serotonin enters the synapse when it's released from a neuron and is removed from the synapse by a cotransporter back into the neuron.	**Rat:** Blood enters the aorta (large blood vessel) from the heart and exits by flowing to the rest of the body.
Hawk: In a muscle cell, Ca2+ enters the cytoplasm through a channel in the sarcoplasmic reticulum (SR) and is removed back into the SR from the cytoplasm by a pump.	**Pea:** Auxin, a hormone involved in plant growth, is transported into a cell through auxin/H+ co-transporters (AUX) and out of a cell via carrier proteins called PINs that are located on the cell membrane.
Oak: In the leaves of an oak tree, photosynthesis makes glucose while cellular respiration breaks down glucose.	**Human:** Oxygen enters the lungs through inhalation and leaves the lungs by diffusing into blood vessels.

Figure 1. Generic prompt and individual contexts. Students were provided with 1 of the 6 physiological contexts for the prompt below. A generalized graph and question structure are provided on the left, while Item Context is given on the right with the name of the item bolded.

Written student language in CRs can be analyzed and compared through text analysis. Typically these methods determine words, phrases, or themes that are associated with a given text or set of CRs and have been used previously to understand student thinking (19–21). Recently, diversity methods traditionally used in ecology have been applied to short CRs to quantify the language diversity within a CR corpus (22). Briefly, an individual CR is treated as a sample, while each word within the CR is treated as an individual in the sample, with repeated words being individuals of the same species. Ecologists use measures, including Bray-Curtis dissimilarity and species turnover, to quantify the differences between samples in the corpus by comparing the species (words, in this study) present in each sample (CR). This allows researchers to quantitatively examine differences in text (e.g., words and phrases) between individual CRs or categorical groups of CRs in the corpus. We also apply a data reduction technique called ordination, which projects complex data into a two-dimensional "map" that places similar samples close to one another and dissimilar samples further away. When applied to language, these measures can be used to visualize the language diversity of a corpus or compare groups of CRs based on categorical data (22).

To better understand student progression in MB reasoning, we can use text diversity methods to compare and contrast the language students use in CR explanations at different covariational reasoning levels and within different physiological Contexts of the assessment items. Previous research has found that Item Context can greatly affect student ideas and language (20, 23, 24). Specifically, novice students are more likely to incorporate surface features of the context in comparison to experts (25, 26). However, similar work examining changes to students' explanations when reasoning about MB using covariational reasoning in physiological contexts has not been reported. We believe examining CRs at increasing reasoning Levels and across multiple Contexts will reveal how students build mass balance explanations and demonstrate covariational reasoning and mass balance skills. As the ideas students express change at each level of the learning progression, we expect to observe differences among the different covariational reasoning Levels but expect that many Context words will be maintained across all Levels. Additionally, we expect some shared language will exist in CRs across Contexts, because mass balance reasoning represents a core concept that can be applied to many phenomena in STEM. From examining the effects of both Context and Level, we expect to determine shared language at higher Levels of covariational reasoning that are Context independent, which may guide practitioners in key language that students use to build understanding and explanations that can be applied to novel situations.

To this end, we examined student language in a large corpus of CRs that was previously collected using a set of six prompts, each representing a different physiological context (see Fig. 1). These responses were each previously assigned to one of the four Levels of covariational reasoning by coders using a single rubric that reflects a covariational reasoning learning progression (7). Using this data set, we sought to answer the following research questions (RQs):

1) How do Item Contexts and the Level of covariational reasoning affect undergraduate language in mass balance explanations?
2) What shared language do undergraduates use to demonstrate covariational reasoning in spite of different Item Contexts?

MATERIALS AND METHODS

Data Collection

CRs were collected as previously described (7). Briefly, a large set of CRs ($n = 4{,}470$) was collected from undergraduates across multiple institutions, instructors, and classes via an online survey. These CRs are in response to items describing phenomena that can be explained with MB reasoning in six physiological contexts (described below). Experts had previously assigned these CRs to a Level of covariational reasoning using a coding rubric aligned to an emerging learning progression for MB reasoning (described below). From this large data set, we randomly selected CRs to create a data set that allowed us to examine responses by two variables of interest: Level of covariational reasoning and Item Context. We prioritized balancing the Item and Level of Covariational Reasoning to allow for easier comparison of the diversity metrics and statistical tests. We randomly selected 80 CRs from each of six Item Contexts in each of the four Levels of covariational reasoning resulting in a data set of 1,920 CRs. These responses come from undergraduates in science majors (biology, physics, life sciences, and STEM) and nonmajor disciplines at eight institutions, including three research-intensive colleges and universities, two master's colleges and universities, and three community colleges. Due to the large number of responses collected from students at the research-intensive institutions, these responses are heavily represented in the data set (95%).

Items and Rubric Description

In this work, we examine student language in response to prompts that were previously developed to examine students' mass balance reasoning (7). To explain the phenomena, students should use covariational reasoning in their explanations. As noted above, the assessment items describe a situation in which three organisms of a species undergo an identical change that results in an influx of matter to a given space and has an accompanying graph (Fig. 1). In addition to naming a specific organism, each Item Context names specific matter, inputs, and outputs that are relevant to the organism and context. In *part A* of each item, students are prompted to explain how *organism 1* had a larger pool of mass than the other two organisms. In *part B*, students are asked how *organism 3* showed no change in the total mass despite the increased influx. For all analyses in this study, *parts A* and *B* were combined into a single response for each student. This basic item structure was used in six different physiological contexts (denoted as Item Context). Each Item Context is named in this work based on the organism that is described in the item (Cat, Hawk, Oak, Rat, Human, and Pea). All six items can be found in the Supplemental Materials of Scott et al. (7) or online at beyondmultiplechoice.org. As part of their analysis of student CRs, Scott et

al. (7) described an emerging learning progression and associated coding rubric that assigns each response to one of four Levels of covariational reasoning. Individual coding rubrics were created for each context to specify pertinent inputs and outputs, but the levels of reasoning in each rubric aligned across context, allowing comparison of the Levels across Item Contexts. Each response in this data set therefore has an associated Context, based on the version of the prompt and an assigned Level, based on the associated coding rubric to categorize the reasoning used.

Bray-Curtis Dissimilarity Calculations

For a more detailed explanation of application of both ecological diversity metrics and ordinations to text data, please see Shiroda et al. (22). All diversity metrics were calculated in version 7.08 of PC-ORD (27) using word matrices created in WordStat (v.8.0.23, 2004–2018, Provalis Research). For these matrices, no words were excluded, but stemming (described below) with English (snowball) was applied. Bray-Curtis dissimilarity (or Sorensen dissimilarity) is a measure of percent dissimilarity and is calculated using the formula $1-\frac{2W}{A+B}$, in which W is the sum of shared abundances between two responses and A and B are the sums of abundances in individual responses (28). Bray-Curtis dissimilarities are calculated using only the responses of interest and, more importantly, only the words that are present within that subset of responses. This means that the values presented in RQ1 and 2 are calculated with different matrices in terms of the number of words (columns) and responses (rows). As Bray-Curtis distance measures are calculated using all columns, using different matrices results in different values even though the responses themselves have not changed.

Ordinations

Dimension reduction techniques can be used to project the data matrix into a two- or three-dimensional plot. Typically with ordination, each axis represents multiple facets with each axis being a vector that explains the highest percentage of the data or columns in the data matrix used. In this work, CRs (represented as data points) are placed on the axes based on the presence or absence of words and their frequencies. Two responses that are lexically very similar to each other will be found close to one another in a biplot, while two responses that are lexically dissimilar will be further away from each other. Such plots allow us to visually compare responses (or groups of responses) to one another.

In this work, we selected detrended correspondence analysis (DCA) as our ordination technique. DCA is unique from other ordination techniques in that the first vector (x-axis) is directly related to species turnover, which is a useful metric of diversity. Species turnover is equivalent to half changes, with 100 DCA units on the x-axis representing one half change. One "half change" is when 50% of the words in two responses are shared and the remaining 50% are not. As the number of half changes increases, responses share fewer and fewer words; after four half changes, responses essentially do not share any words. The x-axis can be used to quantify language differences between two responses or between groups of responses using the centroids. To calculate the half changes between two groups, we subtracted the x-coordinates of the centroids of two groups. It is important to note, that the origin (0) does not represent a certain amount of diversity; therefore, responses close to the origin are not more or less diverse than those elsewhere. In contrast to the x-axis, the y-axis is more difficult to interpret for two reasons. As explained above, in ordination, each axis typically represents more than one dimension of the data and therefore does not have a defined label. In addition to this, DCAs use a detrending method that stretches and condenses the y-axis in a nonstandard way to correct an arching effect in the data (29); therefore, the raw values of the y-axis should only be interpreted relative to each other, not as values. For example, the y-axis can be used to compare points to each other (i.e., *point A* is further from *point B* than *point C* is) but not to calculate distances between points (i.e., these 2 points are each 10 units away from this point).

Data matrices for the ordinations were created in WordStat (v.8.0.23, 2004–2018, Provalis Research) after the use of stemming (English snowball) and removal of a custom word exclusion list (containing articles, conjunctions, and prepositions) to reduce the number of uninformative, but frequent words, including articles (a, an, the), conjunctions (as, and, but, like, or) and prepositions (about, above, across, after, against, at, before, behind, below, beneath, beside, concerning, considering, despite, down, during, except, from, minus, near, over, past, per, plus, than, through, to, toward, towards, under, upon, vs., via, with, within, without). We specifically confirmed that none of these words were important to student explanations by examining the context of the words in student responses. We also excluded any words not appearing in at least three responses as ordinations require at least three instances to detect a pattern (29). Spelling errors were not corrected in this work as stemming removed the majority of these errors.

We performed the DCAs in version 7.08 of PC-ORD (27). Depending on the data set, the raw data can be transformed for it to be used with certain methods; however, we did not need any transformations to perform this analysis. The calculations needed to perform ordination techniques are performed within PC-ORD but require settings to be selected. First, ordinations are calculated using a seed number which can be randomly selected or entered. Each seed number results in similar patterns, but with slightly different numbers; therefore, we selected the seed number 999 to ensure the ordination can be replicated. For DCA, we elected to down-weight rare words due to the large size of the data set. This focuses the ordination on overarching patterns in the data. Scores were calculated for words using weighted averaging. We examined the significance of each axis using 999 randomizations. The percent inertia (or variance explained) for each axis is provided in the outputs of the PC-ORD file and included in our results. We compiled categorical data (Context and Level) associated with the CRs into a separate secondary matrix to to be overlaid on the ordination plots. Applying this test means that 1,000 total plots are created, revealing if the one presented best represents the data. One plot (Cat) required one response (*response 51*) to be removed from the ordination as it was an extreme outlier that made interpretation of the plot impossible (27). Centroids shown in

the plots are calculated by averaging the *x*- or *y*-coordinates of the group of interest in PC-ORD.

Text Analysis of CRs

To understand the differing language within categorical groupings of CRs (i.e., by covariational reasoning level or Item Context), we determined words that are significantly associated with a given category of CRs using WordStat (v.8.0.23, 2004–2018, Provalis Research). This analysis compares the total count of a given word in each category (covariational reasoning Level or Item Context) and returns words that are associated with a given category. We performed text analysis after stemming (English snowball) and removing the default Exclusion list in WordStat. Stemming removes the end of words to mitigate the effect of tenses, singular/plural, and common spelling errors. Words that have undergone stemming are noted in the text as the stemmed root with a dash (e.g., *releas-* could include *release*, *released*, *releases*, *releasing*). We examined any words that returned a $P \leq 0.05$ from a chi-square test.

Statistical Analysis

Statistical calculations were performed using an online calculator (https://www.statskingdom.com/kruskal-wallis-calculator.html). For the Kruskal-Wallis test, we tested a significance level of 0.05, included outliers, and used an effect size of 0.3. We used Dunn's multiple comparison method with a Bonferroni correction. Permutational multivariate analysis of variance (PERMANOVA) is a type of statistical *F* test that compares the differences in the mean within-group distances among all the tested groups (30). PERMANOVAs were performed in PC-ORD with 4,999 randomizations and an assigned seed number of 999.

RESULTS

First, we provide some examples of student responses to an item across covariational reasoning levels. We will continue to use the example of glucose accumulation in the leaf of an oak tree from the introduction (i.e., *Oak* item from Fig. 1). The full prompt reads:

> 'In the leaves of an oak tree, photosynthesis makes glucose while cellular respiration breaks down glucose. The graph shows the amount of glucose in three oak tree leaves over time. A botanist is studying how the amount of glucose in oak tree leaves is impacted by different light conditions. Initially under medium light, there is the same amount of glucose in three oak leaves. Then, the botanist turns on a high intensity grow light, causing the rate of photosynthesis in each oak leaf to increase to 13 units/second (i.e., "high light"). After some time (time point X), the amount of glucose is different for each oak leaf despite having the same rate of photosynthesis in each oak leaf. a) Explain how oakleaf 1 has more glucose compared to oak leaves 2 and 3 given that all oak leaves have the same rate of photosynthesis. b) Explain how oakleaf 3 has no change in glucose while oak leaves 1 and 2 both show increased glucose in their leaves given that all three oak leaves had the same rate of photosynthesis.'

An example Level 1 student response reads, "Leaf 1 could have a faster NADPH production rate or a faster citric cycle rate. The oak leaf may have a mutation where it has a lower amount of electron acceptors. Thus, during medium light, it is already at max photosynthesis rate." This student focuses on ways the input could be affected without considering the outputs of the system. Reasoning at Level 2, another student responds, "Oakleaf 1 has more glucose at the end because there must be less usage or need for glucose in the plant versus the other plants. There must be a lesser need. Oakleaf *[sic]* 3 must only need a certain amount of light to produce glucose and it reached its cap." While this response attends to an output (glucose usage), the student does not properly integrate the input and output when explaining Oakleaf 3 maintaining the same level of glucose. Reasoning at Level 3, another student writes, "Oak 1 has more glucose compared to oak 2 and 3 because it has a faster rate of cellular respiration where glucose is being broken down faster. oak leaf 3 has no change in glucose because its rate of photosynthesis is the same as its rate of cellular respiration." This student, while attempting to balance both the input and output, makes a mistake in that a higher rate of cellular respiration would result in more glucose when it would result in less. Finally, a Level 4 response explains, "Rate in is the same for all three (photosynthesis) and rate out is lowest for leaf 1 (cellular respiration). Rate in and rate out are the same causing the same glucose levels. Cellular respiration is the same as photosynthesis rates." This student correctly integrates the input and output in glucose to account for different amounts of glucose in each leaf.

We began by examining the length of responses in the corpus, as response length can provide general insight into changes in student explanations. Overall, the CRs range from 4 to 259 words in length. We then compared response length based on which Item Context students responded to and which covariational reasoning level was assigned to the response (Table 1). We did find significant statistical differences in length based on the Item Context (Kruskal-Wallis-test with Bonferroni correction; $P = 0.003$); however, the effect size is small ($\eta^2 = 0.027$), indicating the magnitude of

Table 1. *Diversity metrics of CR data set*

	Length (Words)	Bray-Curtis Dissimilarity	Centroid (*x*-Axis)	Average Distance to Centroid (*x*-Axis)
Corpus	51.7	79.2	121.7	52.0
Item				
Cat	44.1	73.8	65.9	19.3
Hawk	54.9	72.0	111.0	11.9
Human	54.8	72.3	207.2	18.6
Oak	55.1	71.6	74.4	17.6
Pea	46.9	72.9	83.3	13.3
Rat	54.6	69.7	189.0	17.4
Level				
1	39.8	84.1	123.7	50.9
2	51.0	80.4	122.8	52.8
3	54.6	75.5	121.1	51.0
4	61.5	70.4	119.5	53.1

CR, constructed response.

the differences is small. Hawk, Human, Oak, and Rat have similar average response lengths of 54.9, 54.8, 55.1, and 54.6 words, respectively. None of these pairs are significantly different from each other ($P > 0.05$). Responses to the items in the Pea and Cat contexts are significantly shorter ($P < 0.0005$), with average lengths of 44.1 and 46.9, respectively. There is no significant difference between the length of Cat and Pea responses.

We also observed a statistically significant difference in response length based on the Level of covariational reasoning exhibited (Kruskal-Wallis-test with Bonferroni correction; $P = 0.0083$). On average, Level 1 responses are shortest with a mean of 39.8 words per response, followed sequentially by Levels 2 (51.0 words), 3 (54.6 words), and 4 (61.5 words). We compared each possible level to other levels individually and found that each pairing of the groups is significant ($P < 0.05$). In comparison to differences in response length based on Context, there is a higher effect size ($\eta^2 = 0.086$) of covariational reasoning level based on response length. While we found here that higher Level answers are longer on average, it is important to note that the length of a response does not always indicate its level of reasoning. Indeed, two of the longest responses in the corpus (257 and 259 words) are Level 1 and 2 responses, respectively. In comparison, a Level 4 response within the Oak context contains only 25 words.

RQ1: How Do Item Context and the Level of Covariational Reasoning Affect Student Language in Mass Balance Explanations?

Using diversity metrics such as Bray-Curtis dissimilarity, we can measure the amount of similarity among the responses within the corpus or a subset of the corpus (i.e., a group based on Level or Context). The Bray-Curtis dissimilarity has a value of 0 when two responses are exactly the same and 100 when no words are shared between responses. We calculated this measure for each pairing of responses in the entire corpus and for pairing responses within the different categorical groupings. The corpus has a dissimilarity of 79.2, meaning CRs in the corpus share ~20% of words. Each of the Item Contexts subgroups has a lower Bray-Curtis value in comparison to the corpus, indicating that the responses within a given Item Context are more similar to each other than to the corpus overall. All the Bray-Curtis values for each of the Item Contexts are relatively similar to each other, ranging from 69.7 to 73.8. This means that no single Item Context results in more restrictive or varied language usage in comparison to another Context. In comparison, we observe a greater difference in Bray-Curtis dissimilarity measures based on covariational reasoning Levels. First, Level 1 has a higher Bray-Curtis dissimilarity value (84.1) than the corpus, demonstrating that responses coded at Level 1 share fewer words with each other on average than the corpus does overall. The Bray-Curtis value for the Level 2 subgroup is similar to the corpus at 80.4. In contrast, the Bray-Curtis measures for Level 3 and Level 4 are lower than the value of the corpus at 75.5 and 70.4, respectively. The value for Level 4 means that despite responses belonging to six different Item Contexts, responses share ~30% of their language. This is interesting because it is similar to those for the individual Item Contexts subgroups, indicating responses with the highest level of covariational reasoning share as much language as responses within an individual Item Context despite being from six different Contexts.

To further compare the CRs based on language, we performed detrended correspondence analysis (DCA) on the entire corpus. To interpret these plots, points that are closer together are more similar and those that are further apart are more dissimilar. Similar to other ordination techniques, the y-axis has no defined units as the axis collapses many dimensions of the data. In contrast, the x-axis is defined with 100 DCA units on the x-axis representing a one half change. One "half change" is when 50% of the words in two responses are shared and the remaining 50% are not. As the number of half changes increases, responses share fewer and fewer words; after 4 half changes, responses essentially do not share any words.

The plot explains 7.4% of the variation in the data with the first two axes. While this value seems low, it is within a normal range for a large dataset (1,920 rows and 937 columns; Ref. 31), and each axis is significant ($P = 0.001$; 999 randomizations). By overlaying the different categorical variables on the responses, we can observe how subgroups relate to each other (Fig. 2). Hulls (lines connecting different points) are drawn to define the outer perimeter of categorical groupings. Centroids (large filled circles) are points defined by the average of the x- and y-coordinates of the categorial group. CRs appear to tightly cluster based on the Item Context (Fig. 2*A*) but not based on the covariational reasoning Level (Fig. 2*B*). Using the x-coordinates of the group centroids, we find greater spread between subgroups based on Item Contexts (range: 65.9–207.2) than Level (range: 119.5–123.7). Since the detrending process used on the y-axis stretches and collapses the axis in a non-linear way, the y-axis cannot be used to calculate area. We can, however, calculate the distance of each response with a given subgroup from the centroid using the x-coordinates to approximate the splay of the data. This allows us to approximate how different responses are in word usage from the "average" response within a subgroup. For Item Context, we observe average distances to centroid for each Context (range: 11.9–19.3), with the Hawk context having the tightest cluster based on x-coordinates (Table 1). This small range of values indicates that each subgroup of Context exhibits a similar splay on the x-axis to other Item Contexts. These values are also much lower than the corpus overall (52.0), indicating these groupings are indeed more similar to each other. In contrast to Item Context, there are much higher average distances based on Level (range: 50.9–53.1).

These groupings of data can also be statistically compared. PERMANOVA tests determine if categorical groupings are distinct by comparing the differences in the mean within-group distances in multiple dimensions (30). We conducted PERMANOVA tests and found there is significant variation among the subgroups in both the Item Contexts ($p = 0.002$) and the Levels of covariational reasoning ($P = 0.002$). While the Level groups are visually indistinguishable in the DCA plot, this biplot represents only two dimensions of the data, while PERMANOVAs utilize the entire matrix used to create the DCA into account. This indicates that groupings of responses by Level are in fact different based on other

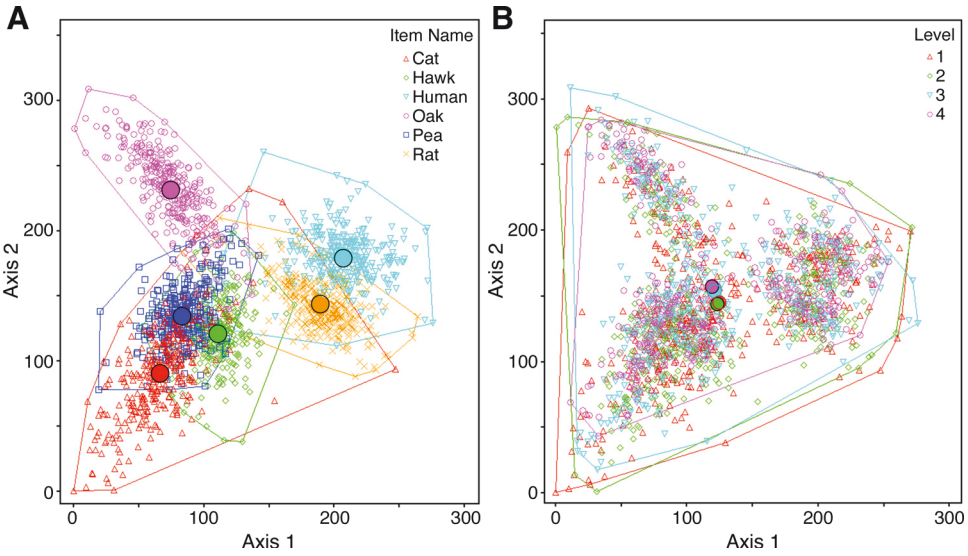


Figure 2. Detrended correspondence analysis (DCA). DCA was performed on a single matrix of 1.920 responses without any data transformation. Individual constructed responses (CRs) are represented by data points. The categorical data of Item Context (*A*) or covariational reasoning Level (*B*) are overlaid using shape and color. Centroids of a given grouping variable are represented by large, filled circles.

dimensions in the data set. These results from PERMANOVA tests concur with the diversity metrics and text analysis that indicate differences in subgroupings based on both Item Context and Level. In combination, these results indicate Context better explains the variance in language in the CRs, but that covariational reasoning Level still affects student language.

To gain a better understanding of what language in the student CRs is responsible for the differences between the groupings, we also examined which words are associated with the different groupings using text analysis software. Overall, the groupings based on the six Item Contexts had 226 significantly predictive words ($P \leq 0.05$), while fewer words were predictive of the groupings of the covariational Level (156, $P \leq 0.05$; Table 2). Most words closely associated with a given Item Context are expected based on the question stem (i.e., the organism, substance, and location) or the specific scenario described in the question setup. For example, in the Oak context, student responses are more likely to include *oak*, *leaf*, *photosynthesis*, *cellular respiration*, and *light* (see Table 2), which represent the organism, input, output, and stimulus in this Item Context and are also included in the prompt itself. Most other predictive words for Item Context groupings, while not in the prompt, are relevant to the specific context or phenomenon presented in the item. Continuing the Oak context example, other significant words for this grouping are associated with the system, such as *tree* or *sunlight*, or specific processes of photosynthesis and respiration, including *Calvin Cycle*, *intensity*, *chlorophyll*, *chloroplast*, *photosystem*, *photon*, *electrons*, and *energy*. However, there are some words that are significantly increased in the Oak context that do not appear to directly relate to the context, such as *differ*, *because*, *factor*, *depend*, and *happen*. In general, we found similar patterns for each Item Context.

For covariational reasoning Level, most words associated with a given Level align well with the coding rubric. For example, most of the words in Level 1 are context-specific words (or stems of words) without directly relating to the relevant fluxes or covariational reasoning (see Table 2), such as *physic*, *bacteria*, *exercis-*, *intens-*, *light*, *photosystem*, *absorb*, *person*, *shape*, or *size*. Increasing to Levels 2 and 3, the words remain context-heavy but are more closely associated with the identification of relevant fluxes in the item or how to change fluxes. For example, the words *channel*, *remov-*, *co-transport*, *produc-*, *uptake*, *reuptake*, *gradient*, *degrade*, *diameter*, *resist-*, and *open* all relate to how the accumulation of a specific material could be affected in one or more of the contexts. Finally, at Level 4, there are fewer context words and more words associated with covariational reasoning. Specifically, there is an increase in comparative language, such as *equal*, *smaller*, *lowest*, *balance*, *greater*, *compar-*, *than*, *slower*, *fewer*, *less*, and *lower*, and other words that explain inputs and outputs, including *rate*, *out*, *in*, *result*, *enter*, *exit*, *decreas-*, *inceras-*, and *net*. In addition, if a context-dependent word was more common in Level 4 responses, it was likely to be relevant to the MB problem. For example, the compartment (*cytoplasm*, *vessel*) and the organism (*pea*) were more frequent in Level 4 responses than the lower Levels. We think this is because Level 4 responses compare the three organisms in the prompt and therefore use the same word multiple times.

RQ2: What Shared Language Do Undergraduates Use to Demonstrate Covariational Reasoning in Spite of Different Item Contexts?

In RQ1, we observed differences in student language based on the Level of covariational reasoning Bray-Curtis dissimilarity, PERMANOVA, and predictive words associated with each Level. However, the Bray-Curtis dissimilarities and PERMANOVA values also indicate that Level 4 responses are more similar to each other as a group than the corpus overall, but these similarities are not apparent in the ordination plots, in which we observed almost no separation of the groups of responses based on Level. We believe this may be because words relevant to the Item Context are obscuring the effect of the Levels. To better understand how student language changes at each Level of covariational reasoning, we examined the effect of the Level of covariational reasoning within each Item Context individually. We then looked for patterns and similarities across the Item Contexts to

Table 2. *Words that are predictive of Item Contexts or Level*

Item	Total	Predictive Words
Cat	32	cat, cotransport-, enzym-, fit*, highest, immedi, inhibit*, level*, molecul-*, ms, neuron, neurotransmitt-, part*, present, probabl-, proper, react, reaction*, receptor, releas-, respond, reuptak-, sensit-*, serotonin, shape*, signal, slower, stimuli, synaps-, synapt-, unit, uptak-
Hawk	46	action, activ-, atp, bind, block, bring, ca, calcium, channel, close, contract, cytoplasm, decreas-, determin-, equilibrium, fact, forc, function, genet-, gradient, graph, hawk, insid-, ion, larg-, many, mass, move, movement, muscl-, number, open, outsid-, potenti-, relat-, remov-, reticulum, return, ryr, sarcoplasm, sensit*, SR, stimul-, stimulus, very, work
Human	43	abl-, addit-, air, alveoli, bloodstream, breath-, capac-, capillary, concentr-, condit-, diffus-, dure-, effici-, enter, exchang-, exercis-, exhal-, experi-, fit*, hemoglobin, higher, hold, inhal-, intak-, lower, lung, metabol-, oxygen, pace, person, quick, rapid, requir-, row, rower, satur, shape*, state, steady, tissu-, util, vessel, why
Oak	57	absorb, any, area, avail-, becaus-, break, broken, calvin, cellular, chlorophyl, chloroplast, convert, CR, creat, cycl, depend, differ, distanc-, electron, energy, expos-, factor, glucos-, grow, happen, high, increas-, intens-, leaf, leav, level*, light, limit, make, max, medium, molecul-*, oak, oakleaf, occur, only, part*, photon, photosynthesi-, photosystem, process, produc-, product, rate, respir-, someth-, store, sunlight, surfac-, tree, usag-, water
Pea	24	after, aux, auxin, bacteria, cell, degrad-, fewer, flux, impact, infect, inhibit*, maintain, membran-, mutat-, pea, pin, plant, protein, reaction*, resist, respons-, rhizobia, root, transport
Level 1	45	absorb, action, adapt, affect, any, bacteria, bind, capac, condit, contract, dure, exercis-, expos-, factor, fit, genet-, hold, intens-, light, limit, max, mayb-*, only, person, photon, photosystem, physic, potenti-, reach, react, reaction, receiv-, receptor, respond, row, sensit-, shape, signal, size, stimul-, stimuli, sunlight, system, util-, water
Level 2	51	air, already, anoth-, area, aux, avail-, begin, bigger, block, breath, calvin, caus-, channel, chloroplast, close, convert, cycl-, diamet-, exhal-, experi-, faster, gradient, graph, happen, heart, impact, infect, inhal-, issu-, larger, mayb-*, mutat-, normal, open, perhap-, possibl-, prevent, produc-, product, requir-, resist, run, ryr, signific, someth-, start, stimulus, stronger, surfac-, uptak-, volum-
Level 3	19	after, calcium, capillary, cotransport, differ, effici-, excess, extrem-, left, mass, minut-, pump, quick, remov-, reuptak-, stay, synaps-, synapt-, transport
Level 4	42	accumul-, amount, aorta, balanc-, blood, break, build, buildup, cell, cellular, chang-, compar-, cytoplasm, decreas-, determin-, diffus-, enter, equal, exit, fewer, greater, increas-, leav-, lower, lowest, match, ms, net, overal-, pea, photosynthesi-, pin, rate, relat-, remain, respir-, result, slower, smaller, smallest, unit, vessel

WordStat was used to determine significantly predictive words of each subcategory. These analyses were performed separately for item context and covariational reasoning level. *Words that were predictive of more than one subcategory within either context or covariational reasoning level. Words are stemmed and represent more than one tense or the singular and plural form of a word. Students also used abbreviations such as millisecond (ms), cellular respiration (CR), and sarcoplasmic reticulum (SR).

reveal the shared language students use to successfully construct MB explanations.

As in RQ1, we examined language diversity using Bray-Curtis dissimilarity (Table 2), but here we are examining responses within a single Item Context. We found that the Level of covariational reasoning shows a similar effect across all six Item Contexts. As the covariational reasoning Level increases, the Bray-Curtis dissimilarity decreases with each consecutive Level, meaning language expressed by students in their responses becomes more similar at higher levels of covariational reasoning. The values observed within each context at a specific Level are much lower in comparison to those seen by covariational reasoning Levels in the overall data set (compared to Table 1), indicating there is even greater similarity in student language at higher covariational reasoning within a given Item Context. This indicates that Level has its own additional effect on language diversity beyond that which is observed when separating responses by Item Context. Indeed, as noted above (Table 1), across the corpus, Level 4 responses share ~30% similarity to each other, while within a given Item Context, Level 4 responses exhibit more similarity to each other and share on average 40% of the language (Table 3).

The effect of covariational reasoning on student language within an Item Context can also be observed in the DCA plots (Fig. 3). Here, we observe a progression of student language in CRs according to covariational reasoning Level, in that each Level appears to be a subset of the previous Level. Indeed, for each Item Context, the group centroids of each Level generally follow a pattern of having sequential x-axis centroids based on Level. For example, in the Oak Context, the centroids are 196.4, 173.8, 118.1, and 95.5 for Levels 1, 2, 3, and 4, respectively. As the positioning of responses towards or away from the origin is not important, the apparent flip in

Table 3. *Bray Curtis dissimilarity and ordination calculations within Contexts*

Item	Corpus	Level 1	Level 2	Level 3	Level 4
		Bray-Curtis dissimilarity			
Cat	73.7	79.2	75.0	68.8	62.1
Hawk	72.0	78.7	72.2	68.2	59.9
Human	72.3	76.0	73.7	69.5	61.5
Oak	71.6	77.8	73.5	65.1	59.9
Pea	72.9	80.2	71.3	69.2	62.3
Rat	69.7	79.1	69.9	63.7	57.5
		Centroid (x-axis)			
Cat	181.2	231.0	206.6	150.5	137.3
Hawk	137.5	173.5	170.9	111.9	93.7
Human	151.5	117.5	115.3	174.0	199.2
Oak	146.0	196.4	173.8	118.1	95.5
Pea	169.3	198.6	204.7	146.4	127.5
Rat	119.4	157.9	145.0	94.7	80.1
		Average distance from centroid (x-axis)			
Cat	32.7	33.7	32.6	38.0	26.3
Hawk	24.5	25.2	25.9	28.0	19.0
Human	32.2	26.2	34.5	42.5	25.4
Oak	29.8	29.3	32.0	31.8	26.2
Pea	31.0	37.4	29.1	30.2	27.3
Rat	25.7	31.1	22.9	29.4	19.6

Detrended correspondence analysis data for the cat context is only representative of 79 responses instead of 80, as an outlier had to be removed from the analysis.

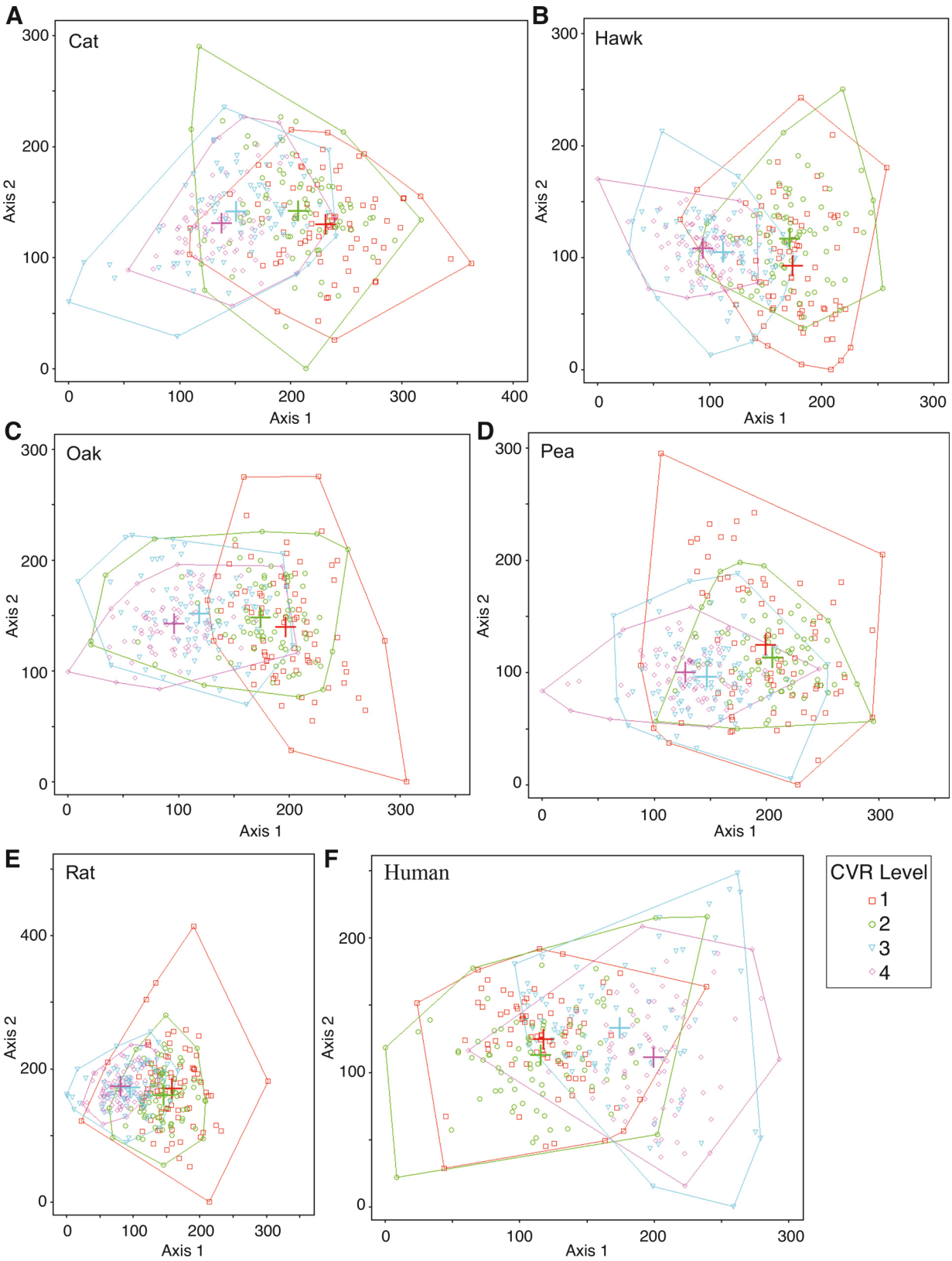


Figure 3. Detrended correspondence analysis (DCA) of individual Item Contexts. DCA was performed without any data transformation. *A–F*: each graph is a separate ordination that represents 80 responses, except Cat, which has only 79 as *response 51* was removed as an outlier; Item Context is overlaid: Cat (*A*), Hawk (*B*), Human (*C*), Oak (*D*), Pea (*E*), or Rat (*F*). Covariational reasoning Level (CVR) is overlaid. Centroids of a given grouping variable are represented by large plus signs.

the direction of the pattern observed in the Human context is not relevant, only their relative position to each other. Further, we observe the general pattern that student language in CRs converges as covariational reasoning Level increases. We can calculate the average distance of the responses with a Level grouping from its centroid on the *x*-axis (Table 3) for each Item Context. These distance averages show Level 4 (average: 24.0; range 19.0–27.3) is the most condensed grouping on the *x*-axis in comparison to the other Levels, which have an average distance of 30.5, 29.5, and 33.3 sequentially. The distance values for responses in Level 4 are markedly lower than for other Levels within each Context, with the exception of the Human context. Within this context, we observed a unique pattern that Level 1 (26.2) and Level 4 (25.4) both have lower average distances between the responses and the centroid in comparison to Level 2 (34.5) and Level 3 (42.5). The lower splay in Human Level 1 CRs compared to the other Contexts is not reflected in the Bray-Curtis Dissimilarity, suggesting it may be an artifact of the ordination procedure comparing all responses during the data reduction process, not just those within a single Context. This could mean that there is less overlap between Levels 1 and 2, for example, than the other contexts resulting in a tighter grouping of Level 1 responses. Overall, these ordinations reveal greater separation based on Level that we could observe in the ordinations from RQ1. This separation is supported by PERMANOVAs ($P = 0.0002$). In combination, the Bray-Curtis dissimilarities and ordinations indicate that within each context, student language converges in a similar way from lower covariational reasoning Levels to higher Levels.

Following our examination of diversity metrics and ordination, we also performed text analysis by covariational reasoning Level within each of the six Item Contexts. Within each context, there is a varied number of total words that are predictive of a given Level for each context (Table 4; range: 62–93). Hawk and Oak are contexts that have the highest number of words that are predictive for a certain Level within the Context (93 words each). Pea is the context that shows the fewest predictive words for Level (62), while Human (72), Rat (78), and Cat (79) are near the middle of the range. The difference in the number of predictive words across the Contexts is not surprising as each context requires a different language to describe the phenomena that are occurring. As we observed in the larger data set, most of the predictive words for Level 4 are heavily associated with reasoning about MB (see Table 4), including *lower*, *rate*, *net*, *less*, *change*, and *than*. For example, in the Human context, a Level 4 response reads, "[Human] 1 has a lower rate of oxygen diffusion out of the lungs than the other [humans.] [Human] 3′s lungs diffuse oxygen into blood vessels at the same rate it enters the lungs[.]" This response is similar in expressed reasoning to a Level 4 response from the Pea context, with only the organism and substance changing: "There is a smaller rate out of auxin in plant 1. there is an equal rate in of auxin as there is a rate out of auxin." Indeed, many of these words were found to be predictive of reasoning associated with Level 4 were found in all six contexts individually, including *rate*, *equal*, and some form of *slower/lower/fewer/smaller*. *Accumul-* and *net* were also found as predictive words for Level 4 in five of the six contexts. In contrast, for the other covariational reasoning levels, there are no predictive words that are present in all six contexts, though for Level 3, some words appear in three (*transport*, *mass*, *balanc-*, *effici-*, *product*, *mutat-*, *receiv-*) or four (*transport*) contexts. As was observed with the larger data set, the lower two levels of covariational reasoning mostly contain words that are related to the Item Context but not MB reasoning. For example, in the Oak context, predictive words for Level 1 include *photosystem*, *intens-*, *water*, *sunlight*, *Calvin*, *chloroplast*, and *rubisco* and only play an indirect role in MB reasoning, in comparison to *cellular respiration* and *photosynthesis* (Oak context, Level 4) which are the immediate inputs and outputs of the compound that students are tracking in this item. Such differences can be observed in the example responses provided at the beginning of RESULTS.

DISCUSSION

We sought to understand how language in student explanations about phenomena that invoke MB reasoning changes with increased level of reasoning and across physiological contexts. Overall, we found that as students' MB explanations increased in level of covariational reasoning, their language focused on specific, relevant physiological structures and they described material flows with comparative context-independent words. Teaching students to use physiological core concepts as a reasoning strategy has been suggested as a critical component to helping students effectively transfer their understanding from one physiological context to another (8, 32, 33). These data provide empirical support for this recommendation, suggesting that using a physiological core concept-based approach to instruction can help students leverage language that focuses on the key elements and mechanisms involved in MB phenomena across contexts.

Using the entire corpus, we found that Item Context influences student MB language in CRs more heavily than the Level of covariational reasoning expressed. This is supported primarily by results from the ordination and text analysis. In the ordinations, individual responses are grouped more distinctly based on Item Context than by their coded covariational reasoning Level, which heavily overlap on the biplot (Fig. 2). This suggests that responses grouped by Item Context are very similar to each other, yet distinct from responses in other Item groupings. From text analysis, we find that there are many more total words that are predictive of Item Context than are predictive of Level, suggesting that language in CRs differ more by Item Context than across reasoning Levels. In contrast, diversity measures reveal that responses at Level 4 share a similar amount of language with each other, regardless of Item Context, as all responses share with each other within a single context. It is especially impressive that students at Level 4 share ~30% of the language across six different contexts. In combination with results from the text analysis, these findings support that student language in explanations becomes more specific to appropriate MB language with increased covariational reasoning. Examination of predictive words for reasoning levels suggests this may be because students at lower levels use language consistent with a variety of Item Context features not important for reasoning about MB, while students at higher

Table 4. *Words that are predictive of Item Contexts or covariational reasoning Level*

Level	Total	Predictive Words
		Cat (n = 79)
1	20	action, affect, age, communic, differ, genet, level, mayb, potenti, react, reaction, receiv, respond, respons, signal, space, stimul, stimuli, stimulus, type
2	13	avail, bind, block, break, channel, higher, mutat, open, product, receptor, resist, rid, similar
3	20	concentr, cotransport, effici, excess, extrem, function, immedi, insid, low, measur, neuron, period, protein, remain, stay, synaps, synapt, time, transport, work
4	26	accumul, activ, amount, balanc, build, cat, chang, compar, constant, despit, enter, entranc, equal, explain, fewer, greater, leav, lower, lowest, match, ms, net, rate, releas, remov, slower
		Hawk (n =93)
1	18	bind, action*, alreadi, ap, area, befor, contract, excit, fire, fulli, hold, motor, neuron, reach, receiv, size, stimul, stimuli
2	32	absorb, action*, begin, block, broken, channel, close, creat, depolar, dhp, exchang, explain, flux, gradient, graph, henc, longer, membran, mutat, open, outsid, perhap, period, provid, releas, requir, resist, respons, rye, ryr, stimulus, stronger
3	17	after, avail, balanc, becaus, calcium, effici, extrem, function, long, mass, posit, possibl, pump*, push, quick, remov
4	28	accumul, amount, ca, cancel, cell, chang, cytoplasm, determin, enter, equal, exit, fewer, hawk, ion, leav, lower, match, ms, net, observ, output, overal, pump*, rate, remain, result, slower, unit
		Human (n = 72)
1	19	athlet, becaus, capac, dure, endur, exercis, exert, forc, heart, higher, hold, peopl, person, residu, row, similar, strenuous, therefor, work
2	22	air, airway, alreadi, anoth, area, atmospher, bigger, breath, deeper, deepli, exhal, factor, fewer, harder, hyperventil, larger, normal, part, shallow, surfac, tidal, ventil
3	9	capillari, cellular, intak, problem, proport, resist, stay, suppli, transport
4	22	accumul, balanc, blood, build, diffus, enter, equal, fast, leav, lower, lowest, lung, match, min, net, pulmonari, rate, rower, slower, smaller, time, vessel
		Oak (n = 93)
1	27	absorb, advantag, anymor, block, closer, condit, diffus, factor, flow, gradient, intak, intens, light, max, number, photon, photosystem, pigment, posit, receiv, resist, rubisco, sourc, sunlight, system, water, wavelength
2	23	area, avail, calvin, chloroplast, convert, cycl, effici, function, generat, matter, mayb, nadph, normal, nutrient, photosynthes, prevent, produc, product, reactant, reduct, rubp, surfac, work
3	18	balanc, consum, creat, depend, differ, glucos, highest, increas, level, mass, move, phloem, repair, respons, storag, thing, transport, turn
4	25	break, breakdown, build, cellular, compens, concentr, equal, exact, insid, larg, leaf, leav, lower, match, net, oak, photosynthesi, rate, respir, result, slower, smaller, smallest, unit, usag-
		Pea (n = 62)
1	11	bacteria, exchang, expos, follow, gradient, immedi, light, limit, receptor, stop, sunlight
2	28	activ, ani, aux, auxin, begin, bind, cascad, case, close, cotransport, excess, factor, faster, grow, hormon, infect, make, mutat, plant, possibl, prevent, produc, product, signal, therefor, transport, uptak, wherea
3	7	amount, cell*, differ, leav, pump, rid, total
4	17	accumul, balanc, cell*, decreas, equal, fewer, low, lower, match, move, pea, pin, pins, protein, rate, result, smaller
		Rat (n = 78)
1	22	activ, adapt, affect, capac, cell, endur, exercis, fit, flexibl, healthier, hr, node, oxygen, physic, restrict, shape, show, size, space, tissu, transport, work
2	27	alreadi, beat, cardiac, caus, consist, contract, diamet, dilat, doe, dure, effect, faster, heart, issu, larger, mayb, min, muscl, onc, pace, perhap, pump, realli, requir, run, stronger, weaker
3	12	amount*, balanc, becaus, compar, defect, downstream, effici, exceed, excess, healthi, mass, prevent
4	18	accumul, amount*, aorta, blood, chang, enter, equal, exit, highest, leav, left, lower, match, net, rat, rate, result, slower

WordStat was used to determine significantly predictive words of each subcategory. These analyses were performed separately for each item context. *Words that were predictive of more than one covariational reasoning level within the context. Words are stemmed and represent more than one tense and can also represent both the singular and plural form of a word.

Levels of reasoning use language focused on specific context features, such as relevant fluxes, that better address the question.

These conclusions are further supported by examining Item Contexts individually. Here, we observe the same patterns of language change in responses within each context. We found higher Bray-Curtis dissimilarities at the lower levels of covariational reasoning, indicating that even within an individual Item Context, students at these lower levels do not use similar language in their responses. As we balanced the data set to equally represent each reasoning Level within each of the Item Contexts, we cannot examine if a single Item Context presented students with more difficulty engaging in covariational reasoning than others. Ordination plots of the six different contexts also show grouping features similar to the Bray-Curtis dissimilarity, with tightest groupings in the biplots for responses at Level 4 than groupings for the other covariational reasoning Levels. We did observe Level 1 within the Human context had a similar average distance to the centroid to that of Level 4. On deeper examination of the Level 1 Human responses, we found students often wrote about exercising, making their language similar (though inaccurate). For example, two student responses read, "Rowers 3 oxygen could not have changed because they are already adapted to that amount of exercise and therefore do not need to intake *[sic]* more oxygen." and "I think that rower 3 was physically able to withstand the exercise." This demonstrates an important point that similarity in language between responses does not always mean the language is accurate or productive within the explanation. While we found here and in other work (22) that language generally becomes less diverse as students increase in scientific thinking, this is not necessarily a rule. Another important aspect of interpreting these results is understanding the differences

in how these diversity measures are calculated. Namely, ordinations use and account for all language in the entire corpus and across all groupings. In contrast, Bray-Curtis Dissimilarities are calculated with just the subset of the data of interest (e.g., Human Level 1 responses). The differing results suggest that there is only increased similarity with Level 1 if they are examined with the entire data set, while the similarity in Level 4 is apparent whether examined alone or within the whole data set.

Examining the student language across all Item Contexts more closely shows that students are removing some of the item surface features from their responses but do not remove all context-dependent words. Instead, the language becomes more specific to the context that is needed to respond to the question. MB reasoning in these systems requires students to use the same skill, reasoning with inputs and outputs, in spite of the context, but answering the question requires context language specific to the fluxes. For example, to explain how an oak leaf has more, less, or the same amount of glucose as another leaf, a student must understand that the input is photosynthesis, and the output is cellular respiration. These context words would not be relevant in the Pea Context since the pertinent input is controlled by the AUX cotransporters and output by the PIN proteins. However, in addition to these specific context words, at Level 4, students are also using language that is productive in building a MB explanation within any context, including *rate*, *out*, *in*, *result*, *enter*, *exit*, and *net*, resulting in Level 4 responses sharing 30% of language all contexts. Given that responses within a single Context at Level 4 only share 40% of language, we find this to be an impressive amount of similarity over six different Item Contexts. Overall, this language analysis supports that when successfully applying covariational reasoning to MB phenomena, students use both relevant context-dependent and context-independent comparative language, while students who are less successful, use mostly context-dependent language that reflects a variety of surface features of the item or system.

Conclusions

Broadly, student language usage in textual CRs is understudied. Here, we have shown the utility of diversity measures and text analysis as part of investigating changes to explanations to better understand how students develop proficiency in reasoning about key disciplinary ideas. While this work focuses on covariational reasoning in MB, similar investigations would be useful for other core concepts or other science disciplinary ideas to gain a better understanding of how students construct explanations and the impact of different variables on the student, classroom, and/or assessment. We have previously used these methods to examine the effect of instruction on student language, which could be useful in determining the effect of classroom interventions (22).

This study examined how student explanations using MB reasoning in physiology changed over six different Item Contexts and across covariational reasoning levels. Physiology core concepts have been proposed as a tool to help students transfer their reasoning across many diverse contexts (7, 8, 32, 33). Our findings of how students use similar language to explain phenomena related to MB across contexts provide empirical support for this suggestion. Across six widely different physiological contexts, we found that students who engage in high-level MB reasoning use fewer, more productive context-dependent words, while also increasing the amount of context-independent, comparative language. These results encourage instructors to model context-independent language in addition to ensuring students recognize the important context-dependent inputs and outputs. Namely, words such as *rate*, *out*, *in*, *result*, *enter*, *exit*, and *net* are not technical or context specific but are equally important to higher levels of mass balance reasoning explanations. This shared language provides educators with

Table 5. *Example responses from oak leaf with instructional recommendations*

Level	Level Description	Example Response (Oak Leaf)	Instructional Recommendation
4	Integrate magnitude changes of both fluxes to accurately explain how the amounts of mass change in the system	The rate of cellular respiration is different for the leaves and for leaf 1, it is slower. The rate of photosynthesis is equal to rate of cellular respiration.	Student explanation is using language reflective of the correct context-specific inputs and outputs and uses context-independent comparative language.
3	Relate a single flux rate of change variable correctly or relate a net flux rate of change incorrectly with changes in amounts	Oak leaf 1 may have transported or used less glucose, therefore glucose built up in that leaf giving it more glucose than the other 2 leaves. The rate out must be higher than the rate in, so the phloem might be more efficient and glucose moved out at a faster rate than the other leaves.	The student has identified the correct mass and is using context-independent comparative language. They have not identified the correct inputs and outputs. Encourage students to identify the specific physiological processes.
2	Relate magnitude changes in a single flux with the changes in amounts	Leaf one's chloroplast could be best fit for the type of light that is being shined on(better absorption), it could have more chloroplast then the others, might just produce more products needed. It could have not enough NADPH for the Calvin cycle to continue or the Calvin cycle in general could be messed up.	The student is focused on surface features of the system. Encourage students to identify the mass, inputs, and outputs.
1	Relate directional changes in non-flux variables to the changing amount	Each leaf absorbed variable intensities of pigment, or blue light. It didn't absorb any of the blue light.	The student is focused on surface features of the system. Encourage students to identify the mass, inputs, and outputs.

Words that were found to be increased in predictive analysis are underlined in the example responses. Instructional recommendations are included to better focus student language.

building blocks for improving student explanations and understanding of MB problems, regardless of physiological context. We recommend that educators practice this language with students regardless of context or what course is being taught. The language we observed in higher level explanations is indicative of students who can reason productively and would be more likely to succeed in another context. Therefore, we recommend educators use simpler contexts to help students acquire mass balance language and reasoning strategies that can help students reason in more complex contexts.

To assist with this recommendation, we include example responses for each of the covariational reasoning Levels described in the paper, using the Oak Leaf item as an example (Table 5). Within the responses, we underline the language that was found to be specific to that Level of covariational reasoning and provide a recommendation for how an instructor could encourage students to improve based on the language used in the response. The language in the response with the highest mass balance reasoning level and our recommendations reflect the Mass Balance Reasoning tool presented by Scott et al. (7). This MB Reasoning tool is designed to provide educators and students with a structure for how to approach mass balance reasoning problems that can be applied to any context. In this approach, students first identify the mass and compartment and determine the processes that affect the mass. These steps require the context-specific language such as *glucose*, *leaf*, *cellular respiration*, and *photosynthesis*. Next, students determine the relative sizes of the rates and identify the net rate. These steps require the context-independent language, including *equal*, *slower*, and *rate*. Just as we scaffold and support students in using appropriate physiological vocabulary words through consistency and repetition, our results emphasize that we should also prioritize students learning context-independent language, which is key to understanding physiology.

DATA AVAILABILITY

Data will be made available upon reasonable request.

GRANTS

This work was funded by National Science Foundation Grants DUE 1660643 (to M. Shiroda and K. C. Haudek) and 1661263 (to J. H. Doherty and E. E. Scott).

DISCLOSURES

J. Doherty is an editor of *Advances in Physiology Education* and was not involved and did not have access to information regarding the peer-review process or final disposition of this article. An alternate editor oversaw the peer-review and decision-making process for this article.

No conflicts of interest, financial or otherwise, are declared by the authors.

AUTHOR CONTRIBUTIONS

M.S. conceived and designed research; M.S. analyzed data; M.S., J.H.D., E.E.S, and K.C.H. interpreted results of experiments; M.S. prepared figures; M.S. drafted manuscript; M.S., J.H.D., E.E.S., and K.C.H. edited and revised manuscript; M.S., J.H.D., E.E.S., and K.C.H. approved final version of manuscript.

REFERENCES

1. **National Research Council.** *A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas*. Washington, DC: National Academies Press, 2012. doi:10.17226/13165.
2. **American Association for the Advancement of Science.** Vision and change in undergraduate biology education: a call to action. http://visionandchange.org/finalreport [2011].
3. **Michael J**, **McFarland J.** The core principles ("big ideas") of physiology: results of faculty surveys. *Adv Physiol Educ* 35: 336–341, 2011. doi:10.1152/advan.00004.2011.
4. **Michael J**, **Modell H.** Validating the core concept of "mass balance". *Adv Physiol Educ* 45: 276–280, 2021. doi:10.1152/advan.00235.2020.
5. **Covitt BA**, **Parker JM**, **Kohn C**, **Lee M**, **Lin Q**, **Anderson CW.** Understanding and responding to challenges students face when engaging in carbon cycle pool-and-flux reasoning. *J Environ Educ* 52: 98–117, 2021. doi:10.1080/00958964.2020.1847882.
6. **Sterman JD**, **Sweeney LB.** Understanding public complacency about climate change: adults' mental models of climate change violate conservation of matter. *Climatic Change* 80: 213–238, 2007. doi:10.1007/s10584-006-9107-5.
7. **Scott EE**, **Cerchiara J**, **McFarland JL**, **Wenderoth MP**, **Doherty JH.** How students reason about matter flows and accumulations in complex biological phenomena: an emerging learning progression for mass balance. *J Res Sci Teach* 1: 37, 2023. doi:10.1002/tea.21791.
8. **Michael J.** Use of core concepts of physiology can facilitate student transfer of learning. *Adv Physiol Educ* 46: 438–442, 2022. doi:10.1152/advan.00005.2022.
9. **Cronin MA**, **Gonzalez C**, **Sterman JD.** Why don't well-educated adults understand accumulation? a challenge to researchers, educators, and citizens. *Org Behav Human Decision Process* 108: 116–130, 2009. doi:10.1016/j.obhdp.2008.03.003.
10. **Carlson M**, **Jacobs S**, **Coe E**, **Larsen S**, **Hsu E.** Applying covariational reasoning while modeling dynamic events: a framework and a study. *J Res Math Educ* 33: 352–378, 2002. doi:10.2307/4149958.
11. **Thompson P**, **Carlson M.** Variation, covariation, and functions: foundational ways of thinking mathematically. In: *Compendium for Research in Mathematics Education*, edited by Cai J. Reston, VA: National Council of Teachers of Mathematics, 2017, p. 421–456.
12. **NGSS Lead States.** *Next Generation Science Standards: for States, by States*. Washington, DC: The National Academies Press, 2013. https://www.nextgenscience.org/.
13. **Krajcik JS.** Commentary—applying machine learning in science assessment: opportunity and challenges. *J Sci Educ Technol* 30: 313–318, 2021. doi:10.1007/s10956-021-09902-7.
14. **Birenbaum M**, **Tatsuoka KK**, **Gutvirtz Y.** Effects of response format on diagnostic assessment of scholastic achievement. *Appl Psychol Measure* 16: 353–363, 1992. doi:10.1177/014662169201600406.
15. **Nehm RH**, **Schonfeld IS.** Measuring knowledge of natural selection: A comparison of the CINS, an open-response instrument, and an oral interview. *J Res Sci Teach* 45: 1131–1160, 2008. doi:10.1002/tea.20251.
16. **Sripathi KN**, **Moscarella RA**, **Yoho R**, **You HS**, **Urban-Lurain M**, **Merrill J**, **Haudek K.** Mixed student ideas about mechanisms of human weight loss. *CBE Life Sci Educ* 18: ar37, 2019 [Erratum in *CBE Life Sci Educ* 19: co3, 2020]. doi:10.1187/cbe.18-11-0227.
17. **Noyes K**, **McKay RL**, **Neumann M**, **Haudek KC**, **Cooper MM.** Developing computer resources to automate analysis of students' explanations of London dispersion forces. *J Chem Educ* 97: 3923–3936, 2020. doi:10.1021/acs.jchemed.0c00445.
18. **Uhl JD**, **Shiroda M**, **Haudek KC.** Developing assessments to elicit and characterize undergraduate mechanistic explanations about information flow in biology. *J Biol Educ* 1–20, 2022. doi:10.1080/00219266.2022.2041460.
19. **Haudek KC**, **Prevost LB**, **Moscarella RA**, **Merrill J**, **Urban-Lurain M.** What are they thinking? Automated analysis of student writing about acid-base chemistry in introductory biology. *CBE Life Sci Educ* 11: 283–293, 2012. doi:10.1187/cbe.11-08-0084.

20. **Prevost LB**, **Smith MK**, **Knight JK.** Using student writing and lexical analysis to reveal student thinking about the role of stop codons in the central dogma. *CBE Life Sci Educ* 15: ar56, 2016. doi:10.1187/cbe.15-12-0267.
21. **Weston M**, **Haudek KC**, **Prevost L**, **Urban-Lurain M**, **Merrill J.** Examining the impact of question surface features on students' answers to constructed-response questions on photosynthesis. *CBE Life Sci Educ* 14: ar19, 2015. doi:10.1187/cbe.14-07-0110.
22. **Shiroda M**, **Fleming MP**, **Haudek KC.** Ecological diversity methods improve quantitative examination of student language in shore constructed responses in STEM. *Front Educ* 8: 989836, 2023. doi:10.3389/feduc.2023.989836.
23. **De Lima J.** *Contextual Influences on Undergraduate Biology Students' Reasoning and Representations of Evolutionary Concepts*. East Lansing, MI: Michigan State University, 2021.
24. **Federer MR**, **Nehm RH**, **Pearl D.** Examining gender differences in written assessment tasks in biology: a case study of evolutionary explanations. *CBE Life Sci Educ* 15: ar2, 2016. doi:10.1187/cbe.14-01-0018.
25. **Nehm RH**, **Ridgway J.** What do experts and novices "see" in evolutionary problems? *Evol Educ Outreach* 4: 666–679, 2011. doi:10.1007/s12052-011-0369-7.
26. **Smith JI**, **Combs ED**, **Nagami PH**, **Alto VM**, **Goh HG**, **Gourdet MAA**, **Hough CM**, **Nickell AE**, **Peer AG**, **Coley JD**, **Tanner KD.** Development of the biology card sorting task to measure conceptual expertise in biology. *CBE Life Sci Educ* 12: 628–644, 2013. doi:10.1187/cbe.13-05-0096.
27. **McCune B**, **Mefford MJ.** PC-ORD. Multivariate Analysis of Ecological Data. Version 7.08. Gleneden Beach, OR: MjM Software Design, 2018.
28. **Bray JR**, **Curtis JT.** An ordination of upland forest communities of southern Wisconsin. *Ecological Monographs* 27: 325–349, 1957. doi:10.2307/1942268.
29. **Peck JE.** Multivariate Analysis for Community Ecologists: Step-by-Step Using PC-ORD. Gleneden Beach, OR: MjM Software Design, 2010.
30. **Anderson MJ.** Permutational multivariate analysis of variance (PERMANOVA). In: *Wiley StatsRef: Statistics Reference Online*. Hoboken, NJ: Wiley, 2017. doi:10.1002/9781118445112.stat07841.
31. **Goodrich JK**, **Di Rienzi SC**, **Poole AC**, **Koren O**, **Walters WA**, **Caporaso JG**, **Knight R**, **Ley RE.** Conducting a microbiome study. *Cell* 158: 250–262, 2014. doi:10.1016/j.cell.2014.06.037.
32. **Modell HI.** How to help students understand physiology? Emphasize general models. *Adv Physiol Educ* 23: 101–107, 2000. doi:10.1152/advances.2000.23.1.S101.
33. **Doherty JH**, **Cerchiara JA**, **Scott EE**, **Jescovitch LN**, **McFarland J**, **Haudek KC**, **Wenderoth MP.** Oaks to arteries: the physiology core concept of flow down gradients supports transfer of student reasoning. *Adv Physiol Educ* 47: 282–295, 2023. doi:10.1152/advan.00155.2022.