Investigating the Influence of Assessment Question Framing on Undergraduate Biology Student Preference and Affect

Jeremy L. Hsu,* Noelle Clark, Kate Hill, and Melissa Rowland-Goldsmith
Schmid College of Science and Technology, Chapman University, Orange, CA 92866

ABSTRACT

Nearly all undergraduate biology courses rely on quizzes and exams. Despite their prevalence, very little work has been done to explore how the framing of assessment questions may influence student performance and affect. Here, we conduct a quasi-random experimental study where students in different sections of the same course were given isomorphic questions that varied in their framing of experimental scenarios. One section was provided a description using the self-referential term "you", placing the student in the experiment; another section received the same scenario that used classmate names; while a third section's scenario integrated counterstereotypical scientist names. Our results demonstrate that there was no difference in performance throughout the semester between the sections, nor were there differences in students' self-reported stress and identity. However, students in all three sections indicated that they most preferred the self-referential framing, providing a variety of reasons that suggest that these variants may influence how well a student reads and processes the question. In addition, our results also indicate that the framing of these scenarios can also have a large impact on some students' affect and attitude toward the question. We conclude by discussing implications for the biology education research community and biology instructors.

INTRODUCTION

Undergraduate biology courses tend to rely heavily on exams and quizzes as means of assessment. Such assessments usually play a large role in determining students' performance in the course and therefore may have a significant impact in how a student perceives success in the field and how likely a student will be to persist within biology (Wright *et al.*, 2016). Given the ubiquity of such assessments and their large influence, there is a robust body of literature that has examined assessment questions in biology and other science, technology, engineering, and mathematics (STEM) fields. Such work has focused on characterizing and investigating the impact of question format (e.g., multiple choice vs. free response, etc.) and cognitive level of questions in biology classes on student affect (e.g., emotions, attitudes, etc.) and performance. This work has identified that most introductory biology classes tend to focus on lower-level cognitive skills, with the cognitive level of questions impacting student learning (Momsen *et al.*, 2010, 2013; Williams *et al.*, 2011), that instructors' approaches to creating assessment questions vary substantially (Wright *et al.*, 2018), that question format (e.g., multiple choice vs. free response) can lead to differences in affect and cognitive strategies (O'Neil Jr. and Brown, 1998), and that certain question formats and cognitive levels may cause different demographic groups to perform differently (Wright *et al.*, 2016).

Despite the prevalence of such assessments and their importance in undergraduate biology, very little work has been done to examine how the wording of different assessment questions may influence student affect and performance in undergraduate biology courses. This paucity of work is particularly striking given that past work in other STEM fields across both K–12 and higher education has revealed that relatively

Luanna Prevost *Monitoring Editor*
Submitted Dec 15, 2022; Revised Aug 15, 2023; Accepted Aug 24, 2023
CBE Life Sci Educ December 1, 2023 22:ar45
DOI:10.1187/cbe.22-12-0249
*Address correspondence to: Jeremy L. Hsu (hsu@chapman.edu).

minor changes in otherwise isomorphic questions can have a large impact on student performance and affect. For instance, negatively worded multiple choice questions tend to reduce student performance and can lead to more confusion on both assessments and surveys (Johnson *et al.*, 2004; Roszkowski and Soven, 2010; Sonderen *et al.*, 2013; Chiavaroli, 2019). Similarly, personalizing questions in mathematics, physics, and engineering courses (i.e., grounding scenarios in students' academic and extracurricular interests) can lead to increases in motivation and learning (Davis-Dorsey *et al.*, 1991; Ku and Sullivan, 2001; Akinsola and Awofala, 2009; Awofala, 2014; Bernacki and Walkington, 2014; D'Agata, 2015; Melsky, 2021). In addition, there is evidence that personalized messages during multimedia science lessons can also lead to increases in problem-solving abilities (Moreno and Mayer, 2000). While this work on personalization has been done predominantly at the primary and secondary levels and there have been conflicting results on its impact, such work suggests that it is worth examining how differences in the wording and framing of biology assessments may be contributing to different impacts on student affect and learning (López and Sullivan, 1992; Bates and Wiest, 2004; Cakir and Simsek, 2010; Cakir *et al.*, 2016).

We thus conducted a quasi-random experimental study to address two research questions:

1. How does the different framing (i.e., the use of authentic scientist names, classmate names, or first-person usage in experimental scenarios) of otherwise isomorphic assessment questions impact student performance and affect in an undergraduate biology course?
2. What framing do students prefer in authentic, constructed-response assessment questions in an undergraduate biology course?

Our research was done in the context of scenario-based, constructed-response assessment questions that ask students to consider a real-world, authentic scenario or scientific study and to answer questions that predict what would happen if they ran the experiment or varied an aspect of the experiment. Such questions can be characterized as authentic assessments, which are described as questions where students are challenged to think critically in an open-ended task that mimics or involves thinking through real-world applications (Koh, 2017; Wiggins, 2019). Such scenario-based questions often involve higher-order cognitive skills and thus require significant cognitive load – or the amount of mental processing when reading and thinking about a question – for students responding to such questions (Villarroel *et al.*, 2018). We chose to study such questions for several reasons. First, the use of problems that rely on authentic scenarios allows us to vary both the framing of the scenarios presented as well as the questions asked, enabling us to explore how these potentially more noticeable differences influence student affect. Second, the use of such scenario-based questions allows for a larger set of possible variations in question wording as compared with lower-order cognitive questions. For instance, there are more limited ways to vary the question framing of a recall-level question that asks a student to define a term, as compared with a higher-order scenario-based question that presents more complex information. In addition, our work focuses on this type of question given the calls for instructors to include more higher-order cognitive questions in undergraduate biology classes, leading to an urgent need to better comprehend the impact of different pedagogical choices when writing higher-order assessment questions (Momsen *et al.*, 2010). Finally, this type of question was chosen for practical reasons as well, given the existing structure of undergraduate biology courses at our institution that use such scenario-based questions, enabling us to conduct a quasi-random study where we could compare between groups of students enrolled in different sections.

Conceptual framework

Past work that has varied the wording and framing of assessment questions has been conducted primarily in mathematics courses and have varied the length of questions, the level of specificity provided (word hypernymy), consistency of sentences, and problem topic (Walkington *et al.*, 2019). There has also been past work that has examined the use of different genders and pronouns in a question (Walsh *et al.*, 1999; Walkington *et al.*, 2019). However, the most common study design consists of testing the impact of personalization, e.g., modifying problems to fit students' interests (López and Sullivan, 1992; Ku and Sullivan, 2000; Bates and Wiest, 2004; Cakir *et al.*, 2016). Such work is of more limited interest to undergraduate biology courses, though, given that personalization to students' interests is less likely in such contexts because assessment questions are likely more constrained with class sizes usually larger than in K–12 schools. Instead of testing personalization of questions or varying the structure of the question, we focus on investigating the impact of varying the framing of who is conducting the experiment in the scenario, with the goal of investigating whether such differences lead to any changes in how students visualize, conceptualize, and relate to the given scenario.

We situate our work in the theoretical framework of discourse comprehension, given the necessity for students answering an assessment question to read and process the given situation/experimental setup and question. Under the theory (sometimes referred to as the construction-integration framework) proposed by Van Dijk and Kintsch (1983), students must build both a textbase and situation model when reading a new scenario. The textbase represents a more basic understanding of the language used in the question and contains only a minimal level of inferences needed to allow the student to make meaning of what the situation is describing, while the situation model represents a more complex mental representation and model of the given situation and experiment (Van Dijk and Kintsch, 1983; Kintsch, 1986; Graesser and Zwaan, 1995; Gunel *at al.*, 2009). Under this framework, changing how a question is worded can impact both the students' textbase (how they process the information while reading) and situation model (their ability to process and generate a mental framework and representation of the scenario).

Affective constructs

We examine different affective constructs that we hypothesized could be influenced by assessment question framing and potentially impact students' textbase or situation model. We were first interested in examining the impact on stress given that students often experience stress and anxiety when taking assessments, negatively impacting performance (Jamieson *et al.*, 2016; Harris *et al.*, 2019; Hsu and Goldsmith, 2021). Past work

TABLE 1. Variants of constructed-response assessment questions

Variant name	Description	Example (Question scenario derived from Chen *et al.*, 2018)
Authentic	The use of actual scientist names conducting a real study, drawn from diverse examples of scientists; this version also included the year of the study to convey the authenticity of the example used	In a 2017 study, Dr. Meiya Chen examined KRAS expression in different cells. Suppose Dr. Chen and her colleagues do an experiment to compare KRAS expression in two cells (cell A and cell B). They run out both KRAS mRNA (gel A) and protein (gel B). Given this information, predict how the Ct values for KRAS will compare between the two cells, if Dr. Chen ran a quantitative-reverse transcription PCR using primers complementary to KRAS. Explain your reasoning.
Self-referential	The use of the second person "you" to provide self-referential framing to the student reading the scenario, and situating the student reading the scenario as conducting the experiment	Suppose you examine KRAS expression in different cells. Suppose you do an experiment to compare KRAS expression in two cells (cell A and cell B). You run out both KRAS mRNA (gel A) and protein (gel B). Given this information, predict how the Ct values for KRAS will compare between the two cells, if you ran a quantitative-reverse transcription PCR using primers complementary to KRAS. Explain your reasoning.
Classmate referential	The use of a classmate's name as the person conducting the experiment	In a research study, Veronica examined KRAS expression in different cells. Suppose Veronica and her colleagues do an experiment to compare KRAS expression in two cells (cell A and cell B). Veronica runs out both KRAS mRNA (gel A) and protein (gel B). Given this information, predict how the Ct values for KRAS will compare between the two cells if Veronica ran a quantitative-reverse transcription PCR using primers complementary to KRAS. Explain Veronica's reasoning.

has also identified that changes in assessment wording can influence stress during an assessment (Riley, 2001) and that students' stress during tests can likewise influence their reading comprehension and memory retrieval (Cassady, 2004; Rai *et al.*, 2011). While there has not been any direct work linking changes in students' stress with students' textbase or situation models, we speculate that differences in question variant may influence students' test anxiety, in turn potentially impacting how students build their textbase and situation model as they read and process the experimental scenario in the problem.

We were similarly interested in determining the impact of assessment question framing on different aspects of students' STEM identity and how they perceived the question. While there have been variations in how science identity has been defined, the construct is generally recognized to measure how well someone feels like they fit in within the science community and if they think of themselves as a scientist (Singer *et al.*, 2020). Others have also argued that science identity is a type of social identity and is inherently linked to feeling a sense of belonging within the science community and feeling part of the ingroup (Kim *et al.*, 2018). Given that showing students a diverse set of scientists can improve STEM identity and sense of belonging, we speculate that variation in scenario-based assessment questions may influence how well a student feels like they are a scientist or a part of the science community (Sharkawy, 2012; Schinske *et al.*, 2016; Yonas *et al.*, 2020). In particular, we hypothesize that how a scenario-based question is presented may impact how well students feel like the examples used reflect their identities (a measure we call identity reflection) and how well they relate to the people performing the experiments in the scenario (relationship to people performing experiments), both of which can influence students' interest in the experiments and scenarios. Since students' levels of interest can shape how well students read and comprehend text (Aprilia *et al.*, 2020), we speculate that changes in these affective constructs may impact students' textbase and situation model. We also explore how well students can visualize the experiments being described in the scenario (visualization of experiments), i.e., their ability to build a cogent situation model from the description of the experiments.

Taken together, this work indicates that variation in how assessment questions are framed can potentially play a large role in influencing student science identities and related constructs. Similarly, these changes may lead to differences in students' textbase or situation model when reading and processing these assessment questions, thus also potentially impacting student affect and performance on the assessments.

Question framing variants

We identify three main variants of scenario-based questions, based on both theoretical and practical perspectives. We describe these three variants below and refer to them as authentic, self-referential, or classmate referential (Table 1).

The first variant, or framing, presents an authentic research study and includes both the lead scientist's name as well as the year the study was conducted to further establish the authenticity of the example (Table 1). We chose authentic framing for several reasons. First, the use of real scientists' names allowed us to choose names that were diverse in terms of both the ethnicity and gender that they presented. Past work has revealed that presenting students with a diverse set of scientists and including counterstereotypical descriptions of these scientists can lead to positive affect, including higher STEM identity and sense of belonging and lower stereotype threat (Sharkawy, 2012; Schinske *et al.*, 2016; Yonas *et al.*, 2020). Similarly, many students may have little knowledge of actual scientists, contributing to students stereotyping who conducts science, suggesting that students may benefit from seeing a range of diverse examples (Schinske *et al.*, 2015). There have also been repeated calls in professional development literature for instructors to include a diverse set of scientist names in a class in order to highlight achievements from many different groups and promote multicultural education in STEM (The National Association for Multicultural Education, n.d.; Reflections on Improving Diversity and Inclusion in Science Teaching | Diverse Educators, 2021). While these past studies and calls have primarily

relied on using in-class examples and longer activities that introduce more background on the scientists, we are not aware of any work that has examined the impact of diverse scientist names in assessments and whether this authentic framing can have similarly positive impacts on student affect. Finally, we note that while there is variation in how authentic assessments are defined, some past work has characterized such assessments as those that "create an atmosphere that is more life-like for stronger engagement and connection… [with] real world applications to perform tasks" (Steppler, 2020), suggesting that the use of real scientist names may be associated with increasing the authenticity of a task.

The second variant we use presented the scenario and problem using the second person "you" to situate the student reading the problem as the one conducting the work (Table 1). This self-referential framing is based upon work in the learning sciences that has identified that students who think about themselves while processing and encoding new information tend to retain and process that information better (Craik and Tulving, 1975; Conway and Dewhurst, 1995; Symons and Johnson, 1997; Rogers *et al.* 1999; Mayer *et al.*, 2004; Turk *et al.*, 2015). While there have been multiple hypotheses proposed to explain this self-reference effect for memory, the two most commonly accepted hypotheses cite that self-referencing either triggers internal elaboration of other items associated with the task to the student, or better allows the student to organize the information provided in relation to the student and to each other (Klein and Kihlstrom, 1986; Klein and Loftus, 1988; Symons and Johnson, 1997; Klein, 2012; Turk *et al.*, 2015). Given this self-reference effect, there have been studies that have applied such self-referential framing in the context of assessment questions, primarily in the context of mathematics education for K–12 students. These studies have found divergent results. For example, in a study involving elementary school children solving mathematics problems orally, the use of the word "you" led to fewer requests to repeat the question, faster problem-solving, and higher accuracy than when compared with isomorphic problems using randomized first names, with evidence suggesting that the self-referential framing changed how students cognitively processed the information in the question (d'Ailly *et al.*, 1997). Similarly, when university students were tasked with completing a linear ordering task designed to simulate a mathematical problem, use of the self-referential "you" again led to benefits in problem-solving ability for students (D'Ailly *et al.*, 1995). University students' ability to retain information also increased after seeing a narrated animation that used self-referential framing, suggesting that the use of "you" in the multimedia presentation led to higher student interest in the problem and a clearer textbase and situation model (Mayer *et al.*, 2004). However, other studies have found different results. For example, a study involving middle school and high school students completing an electronic algebra program with built-in assessments varied the use of pronouns in their assessments, with no benefits reported for self-referential framing (Walkington *et al.*, 2015). A follow-up study that integrated the use of an online homework platform in mathematics classes for middle-school and high-school students found limited impact of individual language features on student problem-solving, though the results indicated that self-referential framing could potentially have a larger influence in students who had less familiarity with the types of problems presented (Walkington *et al.*, 2019). Finally, a study involving fourth graders solving mathematical problems also found that self-referential framing had no influence on students' performance (de Koning and van der Schoot, 2019). Taken together, these contrasting results suggest that there is a need to further investigate the impact of self-referential framing on student performance and affect, particularly in the context of undergraduate biology education, where no previous work has been done.

The third variant was classmate referential, when the scenario and problem referred to a classmate conducting the experiment (Table 1). In this variant, names of students enrolled in the class were randomly selected for each problem. This framing follows the model of past studies that have personalized problems by including friends' names. These studies – which have only been done in the context of mathematical problems for elementary- and middle-school students – have provided contradictory results: some studies indicate that using friends' names contributed to lower student stress, greater interest, and higher performance, arguing that using friends' names lowers cognitive load and increases intrinsic motivation (i.e., motivation driven by seeing inherent value or satisfaction, rather than for an external reward) to the problem (Hart, 1996; Riley, 2001), while others find an impact only in certain scenarios or for certain types of mathematical problems (Davis-Dorsey *et al.*, 1991; López and Sullivan, 1992). Yet other studies find no differences in performance (Ku and Sullivan, 2000; D'Agata, 2015). Despite these variable results, this classmate referential model in assessment questions has been suggested for K–12 educators as a way to potentially increase student interest and motivation across books and websites geared for instructors (Krawec and Warger, 2015; Findley, 2016; 3 Tips for Creating Math Word Problems That Boost Critical Thinking, 2021; 10 Fun and Original Ways to Teach Math Word Problems, n.d.), and a research team was recently awarded a prize in an educational technology competition to develop an electronic platform for K–12 mathematics education that includes personalization of common first names in the students' school (Carnegie Learning Only K–12 Finalist for Dept. of Education $1Million XPRIZE, 2022). Given this past work in mathematics education, we included this variant as a comparison to better investigate the impact of assessment framing on student affect, because this variant provides a version that is not a self-referential framing but also does not rely on authentic scientist names, thus potentially allowing us to draw more inferences about the impact of using self-referential framing or incorporating authentic scientist names.

MATERIALS AND METHODS

Study context and demographics

The study was conducted at a private, comprehensive university in southern California with a R2 designation under the Carnegie Classification (McCormick and Zhao, 2005). At the institution, classes are often separated into multiple sections of the same class, with each section taught by one instructor. Our study relied on a quasi-random design that involved three sections of the same introduction to molecular genetics course taught by two of the authors (J.H. and M.R.G.) in spring 2022. Quasi-random studies involving parallel sections of the same course have been used in past biology education research

TABLE 2. Summary of enrollment for the three sections, as well as the assessment-framing variant each class received throughout the semester

Section	Enrollment	Instructor	Assessment framing
1	54	Instructor 1 (J.H.)	Authentic
2	55	Instructor 1 (J.H.)	Self-referential
3	53	Instructor 2 (M.R.G.)	Classmate referential

studies, and allow for the comparison of different treatments across the sections (Pape-Lindstrom *et al.*, 2018; Harris *et al.*, 2019). This course was chosen for several reasons. First, the two instructors collaborate on the course design and structure, sharing activities, slides, and assessments, thus controlling for most potential impacts from course factors. Second, one of the instructors (J.H.) taught two of the three sections, thus further controlling for instructor impact on student affect and allowing for more robust comparisons. Third, the course prepares students for authentic, scenario-based questions through in-class activities and formative assessments and relies almost entirely on scenario-based constructed response questions for assessments that involve higher-order cognitive skills.

The three sections had roughly equal enrollment, ranging from 53 to 55 students (Table 2). This course is predominantly taken by first-year students from allied health fields (health science and applied human physiology majors) as well as students enrolled in an accelerated pre-pharmacy program during spring semesters. While there was variation in the student demographics, we note that approximately two-thirds of each section consisted of either health sciences or prepharmacy students, with Asians representing the plurality or majority of students in each of the sections (Supplemental Table S1). Most of the students identified as female, with approximately one-fifth of all students identifying as first-generation students and 10% as transfer students (Supplemental Table S1). These statistics largely align with university demographics.

Study design and instrument development

The study relied on a quasi-random experimental design, where each of the three sections was randomly assigned to a different variant of how assessment questions were framed for the whole semester (Table 2; Figure 1). Students were not made aware of this difference in framing until the end of semester survey, after all assessments in the study had been completed. The course consisted of three quizzes (taking 30–40 min each, with the same amount of time given to each section) and two exams (taking 50 min), which were all conducted in class and consisted of between three to four questions, each with multiple subparts. Quiz and exam questions were isomorphic between sections, with only the framing of the question varying between sections. Each assessment was collaboratively written by the two instructors, who reviewed each assessment before deployment to verify that the questions were scenario-based, constructed-response questions that required higher-order cognitive skills. Examples and practice problems used in class largely varied between the authentic- and self-referential variants. For the third section that used the classmate-referential framing, randomized student names were used for each question. We ensured that student names were not repeated in more than one problem throughout the semester.

We gathered several sources of data to investigate the impact of these different assessment-framing variants on student affect and examine student preferences. We first compared student performance across the sections in each of the quizzes and exams to check whether there was any differential performance that may impact student affect and preferences. In addition, we deployed 1) a baseline survey given on the first day of the semester, 2) embedded surveys during quizzes and exams, and 3) a survey at the end of the semester (Figure 1). This third survey asked for students' preference between the variants, and we also compared student performance by preference to explore whether students with different preferences have differences in their textbase and situation model formation that are impacting their performance.

The study was reviewed and approved by the Chapman Institutional Review Board.

Baseline survey

First, we surveyed students on the first day of class to gather demographic data. In this survey, we also asked students two questions related to science identity: the first asked about their sense of belonging and the second about how much they felt like their identities were reflected in science classes (which we will refer to as identity reflection). Students responded to two five-point Likert-scale statements in this baseline survey:

1. I feel that I belong to the science community.
2. I see aspects of my identity reflected in my science courses.

The first statement is derived from a STEM sense of belonging instrument (Good *et al.*, 2012), which has been used in an

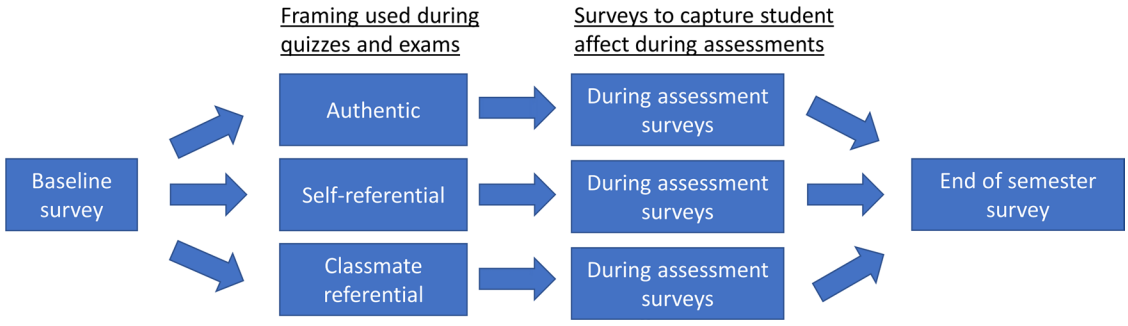

FIGURE 1. Schematic of experimental design and data collected

TABLE 3. Affective constructs measured during assessments.

Construct	Measure	Likert-scale item	Source or description about question development and other notes
Student stress	Calmness	I feel calm.	Drawn from published psychosocial scale for use in STEM contexts (Findley-Van Nostrand and Pollenz, 2017)
	Nervousness	I feel nervous.	
Science identity	Identity – science person	The examples used in this exam made me see myself as a science person.	Only included after the first quiz; derived from single-item identity instrument for use on STEM identity (Dou *et al.*, 2019)
	Identity reflection	I see aspects of my identity reflected on the examples used in this assessment.	Developed de novo through iterative process
	Sense of belonging	I see myself as part of the science community.	Derived from Sense of Belonging Scale (Hurtado and Carter, 1997; Trujillo and Tanner, 2014); only included after the first quiz
	Relationship to people performing experiments	I could relate to the people performing the experiments in this exam.	Developed de novo through iterative process; only included after the first quiz
	Visualization of experiments	I could visualize the experiments being described in this assessment.	Developed de novo through iterative process
	Interest	I am interested in the experiments described in this assessment.	Structured after Likert-scale measures used to measure interest in other STEM education studies (Blair and Frezza, 2020; Nawawi *et al.*, 2021)

abbreviated form in several other studies, including as a four-item scale with extremely high reliability between statements (Cronbach's alpha = 0.96; Lytle and Shin, 2020; Moudgalya *et al.*, 2021). Given this high reliability and past work that has indicated that some single-item instruments may be sufficient to characterize affect (McDonald *et al.*, 2019), we made the decision to only include one question asking about sense of belonging to parallel the survey structure used in the second part of the study, which was constrained by time logistics.

The second statement on identity reflection was developed de novo through an iterative process. While there are validated instruments examining STEM identity (Trujillo and Tanner, 2014), we chose to not use any of these previously published questions. Rather than directly characterizing a student's STEM identity, we were interested in exploring how specific classroom experiences, including assessment questions, would impact students' connections to the course. Thus, three of the four authors (J.L.H., K.H., and M.R.G.) independently brainstormed possible Likert-scale questions and iteratively discussed until identifying a statement that was clear to each of the authors. Second, to check response process validity, we conducted a post-hoc cognitive interview with an undergraduate member of the research team (N.C.) who had not been involved with instrument development as well as three undergraduates not part of the research team.

During assessment surveys

Following these baseline surveys, we also embedded several questions into the course's five assessments (three quizzes and two mid-semester exams) to capture student affect during assessments (Table 3). Students were asked a series of five-point Likert-scale questions at the end of the quiz or exam, including a directed-response control question (e.g., "Please select disagree if you are reading this.") We discarded any responses that did not meet the directed-response control question. We were interested in capturing the influence of the assessment framing questions on student affect during the exam, a period that can involve higher stress in students (Jamieson *et al.*, 2016; Harris *et al.*, 2019; Hsu and Goldsmith, 2021). We therefore relied on embedded-survey questions on the assessments rather than deploying a survey after the exam, when student affect may be influenced by other factors, such as talking to classmates and reflecting on the exam (Santoro and Bunte, 2023). However, embedding in assessment questions during the assessments constrained what questions we could ask. Given the timed nature of quizzes and exams conducted in class, we could only rely on Likert-scale questions that were relatively fast for students to respond to and did not include any free-response questions. We also deliberately limited the number of Likert-scale questions to ensure that students were able to have adequate time to complete the assessment as well as the survey questions. Each Likert-scale question was either derived from published and validated instruments or developed de novo following the iterative process described for the baseline survey. A post-hoc cognitive interview was again conducted with an undergraduate researcher (N.C.) not involved in instrument design as well as three undergraduates not part of the research team to verify the clarity of the questions. Three of the measures were not included in the first quiz and were only included starting with the first exam.

Factor analyses

We conducted an exploratory factor analysis (EFA), using Jeffrey's Amazing Statistics Program, to determine which of the different measures would load together on the same factor

(JASP Team, 2022). EFA is a commonly used technique within education research to determine whether several measures on a survey are correlated and should be collapsed into a single measure for analysis if they are measuring the same variable (Beck and Blumer, 2016; Knekta *et al.*, 2019). Here, we ran an EFA using the data from the first exam, which was the first assessment to include all measures. This EFA was also repeated with the combined data across all assessments except the first quiz, which did not include all questions. For each EFA, we used minimum-residual estimation and oblique-factor rotations, which has become the standard for factor analyses in education research given that factors are likely to correlate with each other (Leandre *et al.*, 2012; Knekta *et al.*, 2019). In addition, we only retained factors which had an eigenvalue greater than one, a cutoff that has been suggested for use within education research given that factors with an eigenvalue lower than one likely do not contribute substantially to explaining the given variance (Beavers *et al.*, 2019; Knekta *et al.*, 2019).

End of semester survey

Finally, to determine student preferences of the variants, we conducted an end of semester survey the last week of the semester, after the last mid-semester quiz and exam had been completed. The survey (Supplemental Materials) was conducted before the final examination, which was not included in this study for logistical reasons. The end of semester survey was the first time that students were made aware of the different possible variants. Students were asked to read the three isomorphic versions of the question listed in Table 1; because all three sections of the course had a student with the same first name, we used this name in the example for the classmate referential version. We refer to this shared student name as "Veronica" in Table 1, though this is a pseudonym.

After reading the three versions, students were asked which version they most and least preferred and were also prompted to explain their reasoning. Responses to the two open-ended questions that asked students to explain their reasoning were read by two of the authors (J.L.H. and N.C.), who independently came up with codebooks following the principles of grounded theory (Locke, 2002). The two researchers then discussed and came up with a consensus codebook. Next, the two researchers independently coded 30 randomly selected responses (representing approximately a quarter of the total responses), which included 10 responses from each of the three classes, to compare interrater reliability. Cohen's kappa was calculated using ReCal2 (Freelon, 2013) and was 0.71 and 0.77 for the two free-response questions, respectively. Given that these values of kappa indicate substantial agreement (Landis and Koch, 1977), one coder (N.C.) independently coded the remaining responses.

Statistical analysis

We first compared student performance across sections by 1) comparing individual scores in each of the five assessments, as well as 2) aggregated scores from all five assessments, using one-way ANOVAs. We also compared performance of students based on the variant they preferred. Given the unequal numbers of students who preferred each of the three variants (see results section below) and the difference in variance between the groups, we used a non-parametric Kruskal-Wallis test to compare performance.

Next, we compared the average Likert-scale response at the beginning of semester survey between sections (one-way ANOVA). For analyzing the responses to the Likert-scale questions regarding student affect during the assessments, we relied on the factor analyses, collapsing interrelated Likert-scale items into one construct by averaging the Likert-scale scores for each of the measures part of the construct. Reverse-coded items were flipped before being included in the analysis. This method has been used in previous biology education research studies and allows a robust statistical comparison of constructs (Beck and Blumer, 2016). Student affect was then compared between sections, as well as between students who indicated that they preferred different variants, in each of the assessments (two-way ANOVAs with post-hoc Bonferroni corrections). All comparisons were performed using R.

RESULTS

No difference in student performance across sections and by variant preference

We first compared student performance across the three sections (Supplemental Table S2). We compared performance in each of the five assessments, as well as with the aggregated scores across the assessments, and found no differences by section in any of these comparisons (one-way ANOVA; all p values > 0.05). In addition, we compared performance of students based on the variant they preferred, which was identified in the end of semester survey. We identified that there was no difference in how students performed based on the variant that they preferred for all five assessments, as well as the combined scores across all the assessments (Kruskal-Wallis tests; $p > 0.05$). We were unable to compare student performance by section and preference, given the relatively few students who indicated that they preferred certain versions in each section (see section below). Given the lack of evidence that student performance on these assessments varied either by section or by student preference of variants, we did not include student performance as a factor in any of our further analyses.

Factor analyses show two main constructs

Our EFA revealed that there were two main factors in our data from the Likert-scale questions done during the assessments (Table 4). First, the measures for calmness and nervousness loaded together across each of the assessments, with negative loading scores for nervousness. These results are consistent with the data from the published psychosocial scale these measures were drawn from and also support reverse coding the nervousness measure when aggregating these measures into the stress construct (Findley-Van Nostrand and Pollenz, 2017). Second, all the other affective measures loaded together with each other, suggesting that the responses to these questions are correlated and supporting our hypothesis that each of the six measures relate to some aspect of students' identity. Quiz 1 did not include three of these measures; however, the other three measures still loaded together. These results were consistent when we conducted factor analyses for each of the other assessments and when we ran the data combined across all assessments except quiz 1. Together, these two constructs accounted for 54.3% of the variance (39.8% from identity, and 14.6% from stress).

TABLE 4. EFA for combined data across all assessments (exams one and two, and quizzes two and three), except quiz one. Quiz one was excluded from this analysis because it did not include all Likert-scale items used in the later assessments.

		Factor	
Construct	**Measure**	**1**	**2**
Student Stress	Calmness	0.829	
	Nervousness	–0.676	
Science Identity	Identity-science person		0.774
	Identity reflection		0.597
	Sense of belonging		0.708
	Relationship to people performing experiments		0.786
	Visualization of experiments		0.721
	Interest		0.747

No difference in student affect by section or preference

We identified that there were no differences among the sections in the initial baseline survey conducted either for sense of belonging or identity reflection (Supplemental Table S3; one-way ANOVA, $p > 0.05$). Given this, we did not include data from the start of the semester as a factor in any of our other analyses.

Next, we compared the two measures of stress and identity during the assessments across the three sections (Supplemental Tables 4 and 5), each of which had been provided a different variant during the assessments. We found no difference in either construct between the sections in any of the assessments (two-way ANOVA with post-hoc Bonferroni correction).

In addition, we also compared stress and identity across the groups of students who preferred the different variants (Supplemental Tables 6 and 7). We found no differences between students who preferred different versions across any of the five assessments for both stress and identity (two-way ANOVA with post-hoc Bonferroni correction).

Most students prefer the self-referential framing, though preference varies depending on which version students have been using throughout the semester

We find that the majority (62.1%) of students indicated that they preferred the self-referential framing, followed by 19.4% preferring authentic. Only 11.7% preferred the classmate referential variant, with the remainder (6.8%) indicating that they had no preference. These trends held across each of the sections (Figure 2). However, there were differences in the strength of this preference by section, and there appears to be an increase in preference for the variant that students were using all semester as compared with the other sections. For instance, students in section 2, who had the self-referential framing in their assessments throughout the semester, showed the highest percentage of students who preferred that version across the three sections. Similarly, students in section 1 (who had the authentic variant) had the highest percentage of students preferring this version as compared with the other two sections, and students in section 3 (who had the classmate-referential variant) had the highest percentage of students preferring this version compared with the other two sections.

Students indicate that authentic and classmate referential are their least preferred versions

The plurality of students (40.8%) indicated that they least preferred the authentic variant, followed by 31.1% of students indicating that they least preferred the classmate-referential variant. Only 14.6% of students indicated that the self-referential variant was their least preferred version, with 13.6% indicating no preference. There were again differences by section (Figure 3). For instance, a slightly greater number of students in section 1 (who had the authentic variant) indicated that they least preferred the classmate referential (32.4%) as compared

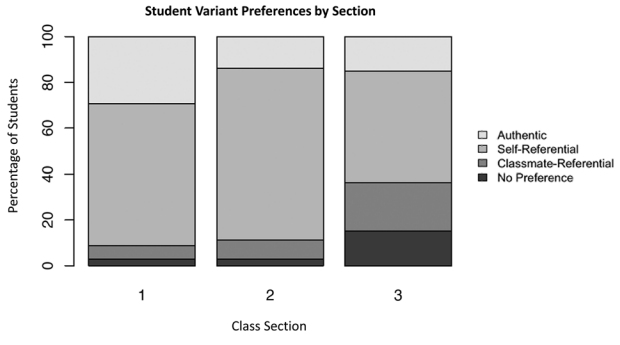

FIGURE 2. Student preferences by section. Section 1 was given the authentic framing in their assessments throughout the term; section 2 was given self-referential framing; and section 3 was given classmate-referential framing.

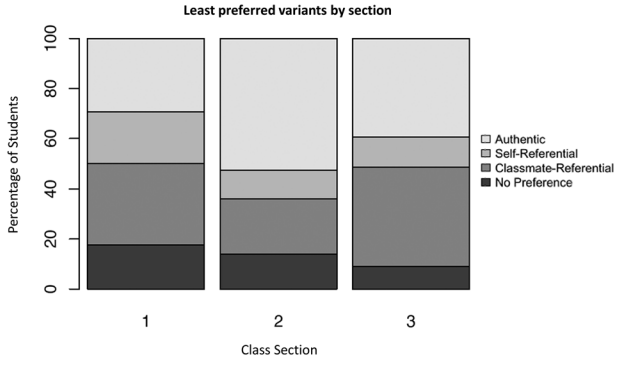

FIGURE 3. Students' least-preferred variants by section. Section 1 was given the authentic framing in their assessments throughout the term; section 2 was given self-referential framing; and section 3 was given classmate-referential framing.

with authentic (29.4%) variants. In contrast, students were evenly divided between these two variants when identifying which version they least preferred in section 3 (who had the classmate-referential version), while the majority of students (52.8%) in section 2 (who had the self-referential variant) stated that they least preferred the authentic variant.

Students provide a variety of different reasons to explain their preferences

We identified multiple emergent themes when students were queried as to why they preferred a given variant (Table 5), and similarly when they were asked to explain which variant they least preferred (Table 6). For instance, students often cited preferring a version because they indicated it was shorter, had less information, or was easier to read (or conversely, indicated that they did not like a version because of its length or challenging nature to read). "The question is worded in a concise yet easy to understand manner," one student wrote when explaining why they preferred the self-referential version. Similarly, others viewed the authentic- or classmate-referential versions as being more conducive to read, writing that these versions were "easy to read without unnecessary parts to solve the question." However, some students report preferring a version with more complexity. "I would choose [authentic version] because although it is longer, it provides a more thorough explanation of what the question is asking and better explains the circumstances," one student wrote, indicating that they perceived the additional details in this version as helpful for facilitating their understanding of the problem.

Students also frequently reported that they preferred certain versions because they could visualize the experiments better (or conversely, did not like other versions if it prevented them from visualizing the scenarios). "I feel like I could really picture the experiment in my head," one student wrote when explaining why they preferred the classmate-referential version. Similarly, another student cited how they could not visualize the experiment for the self-referential version, commenting that "this prompt is non-specific, which doesn't allow me to get a good visualization of the question." Other students reported that they felt more connected or related to a given version. "It uses second person which makes me feel more invested in the question," one student wrote when describing their preference for the self-referential version.

However, the frequency of these explanations differed by preference, and there were also some themes unique to a given version (Supplemental Tables 8 and 9). For instance, some students reported preferring the authentic version because it felt more applicable to real-world experiments and scenarios (the authentic code). "It is nice to know that there have been other scientists who have studied/study the concepts we have learned throughout the course," one student commented, indicating that the wording reinforced their view of themselves as a scientist (i.e., increasing STEM identity). This student continues, "relating topics such as KRAS and its expression is something we have learned in class and seeing it in the context of what appears to have been real experiments is nice to see." Another student described not liking the self-referential or classmate-referential version for similar reasons: "When I see that it's just a scenario-based question with no people mentioned in it then I think that this experiment can't actually be performed in real life and I might not care as much about it." Other students, in contrast, did not prefer the authentic version, citing how seeing a real-world example made them feel less confident. "By using a study it puts me in the mindset that the problem will be difficult to understand," one student commented.

DISCUSSION

Our work provides the first study examining the impact of different assessment framing variants on students' performance and affect in undergraduate biology and provides one of the first studies to do so for any STEM course at the collegiate level. In addition, our results generate new insight into the variants that students prefer and how students perceive these different versions. These results extend existing literature on different assessment question variants, and in some cases conflict with data from K–12 contexts in other STEM disciplines, generating new questions into how the wording and framing of different questions can impact student affect.

Impact on performance

First, our study found that there was no difference in average quiz or exam score between the sections, and similarly no difference in performance based on students' preferences for the different versions. This result conflicts with some of the previous findings, mostly in the context of K–12 mathematics classes, that have found that self-referential framing or classmate-referential framing leads to positive impacts on student performance (D'Ailly *et al.*, 1995; Hart, 1996; d'Ailly *et al.*, 1997; Riley, 2001; Mayer *et al.*, 2004). We note, however, that such studies have been conducted in very different contexts than ours: for example, one of the studies measured the amount of time elementary school students took to respond to mathematics problems orally as well as the students' rate of accuracy (d'Ailly *et al.*, 1997), while another study focusing on university students similarly measured time of response to a linear ordering task (D'Ailly *et al.*, 1995). The only other study we are aware of that tested self-referential framing in university students relied on a multimedia video on human respiration and used different framing variants for the narration before assessing student transfer of knowledge (Mayer *et al.*, 2004). We are not aware of any previous work that has studied assessment question variation in biology classes, or with authentic scenario-based questions. We speculate that the drastically different contexts of these studies may explain our conflicting results. For instance, it is possible that the self-referential variant on our assessments enables some students to form a textbase and situation model more easily than in the other variants, given that some students cited how this version appeared to have the lowest complexity and lead to the lowest cognitive load required to read and process the question. This may result in students being able to process and respond to these questions faster, in line with the results from mathematics education finding that self-referential framing leads to improved speed of answering mathematical problems (D'Ailly *et al.*, 1995; d'Ailly *et al.*, 1997). However, given that most students had sufficient time to respond to each of the questions in the assessments and had time to review and check their responses at the end, it is possible that any differences in time processing and solving these problems would not equate to a difference in performance. In addition, we note that

TABLE 5. List of emergent themes when students were queried which variant they preferred and why. Only codes at 5% or more frequency in at least one of the sections are included.

Code name	Code description	Example code	Percent of responses across all students
Easy to read	Stated that a given version was the most clear and straightforward, or least confusing and complicated phrasing or wording of the question	"I feel as if this question was the easiest to understand and straight to the point. The other questions had other topics that were not essential to the question and would be distracting."	25.0%
Visualization	Stated that they preferred a given version because they could more easily see themselves conducting an experiment or the process of conducting an experiment	"I liked how the scenario put me in the question because I was able to picture me doing the experiment and it made me feel more like a science person"	23.3%
Less complexity is good	Cited that they did not like a version because it provided too much information or details, or alternatively liked a version because it contained less information	"I would prefer [this version] on an exam because the question is not diluted with unnecessary information such as names of individuals or experimental goals. Instead, it gets straight to the point and I know exactly what I need to answer."	17.7%
Fewer words	Explicitly mentioned having less amount of reading for a given version	"It's shorter"	11.8%
Identity	Stated that they felt more connected or related to a given version, or increased STEM identity	"I like that it's supposing 'I' do this experiment. I can't relate to it as much when it's telling me about a random Doctor who performed the experiment."	7.1%
Distraction due to names	Describes how seeing names detracted from thinking through the problem and why this led to the self-referential version being their favorite	"It is a little distracting to read questions with someone's name. It is also easier to imagine the experiment when I am put into the scenario."	4.7%
No preferences or major differences	Stated that they did not see any major changes between the versions	"To be honest I can't really seem to tell the major difference between all of these studies so I don't have a strong preference."	4.1%
More complexity is good	Cited that they preferred a version because it provided more information or did not prefer a version because it lacked information	"I feel like [classmate-referential variant] is more in depth, therefore makes more sense to me"	3.5%
Authentic	Stated that a given version felt more applicable to real-life experiments and science	"Providing background information about research that has been done in relation to the questions make the questions seem more applicable to the real world as these experiments were completed in real life."	3.0%
Similarity to past exams	Describes that the question felt more comfortable because they had seen the same framing in past examples and assessments in the class	"This version is the most similar to practice and past-exam problems so it is the easiest for me to understand"	3.0%
Stereotype threat	States that reading the word "you" led to negative emotions or connotations for that version	"Furthermore, "you" seems too personal and in a way scares me."	1.8%

multiple other studies from K–12 mathematics classes, most using self-referential framing in electronic problem-solving platforms for students, found that self-referential framing did not improve student performance (Walkington *et al.*, 2015, 2019; de Koning and van der Schoot, 2019), and that some studies found that certain variants only improved performance

TABLE 6. List of emergent themes when students were queried which variant they least preferred and why. Only codes at 5% or more frequency in at least one of the sections are included.

Code name	Code description	Example code	Percent of responses across all students
More complexity is bad	Cited that they disliked a version because it provided too much information or preferred a version because it had less information	"I just feel like it's a lot of extra information that is distracting me"	23.6%
Length	Stated that a version was too long or wordy	"I would least prefer [classmate referential] because it seems very wordy compared with the two other options even though they ask the same question. While reading [this version] I caught myself having to reread it to understand it while in the other two versions I perfectly understood what the question was asking. It seems like there's unnecessary words and a repetition of phrases that makes it confusing"	18.8%
Less authentic	Mentioned that the wording makes the question feel less scientific	"[Self-referential version] is the least I would prefer because it's just giving me the scenario. It's not telling me who's testing it and for what purpose. When I see that it's just a scenario-based question with no people mentioned in it then I think that this experiment can't actually be performed in real life and I might not care as much about it."	11.0%
No preferences or major changes	Stated that they did not see any major changes between the versions	"I have no preference for the choice that I would least prefer when taking a quiz of exam because both [authentic version] and [classmate referential] version are formatted the same."	9.4%
Distraction due to names	Describes how seeing names detracted from thinking through the problem	"The names and details add an extra layer of confusion because it's a word you repeatedly have to keep reading and can get somewhat stress inducing. I'd rather get straight to the point. I focused more on 'Veronica' than the problem or case at hand."	7.9%
Intimidation of real study or scientist	Described that the use of authentic science or scientists led to barriers processing information or changes in affect toward the question	"Using a professional (Dr.) and sayings it's a study may be more intimidating when reading a question on a test."	6.3%
Less connected (identity)	Stated that they felt less related to the experiment	"[It's] not something that I would relate to personally. It feels like a watered down version of [authentic] version."	5.5%
Hard to visualize	Stated that they did not like a given version because they could not easily see the process of conducting an experiment	"Not having it be done by an actually biologist or by yourself makes it harder to visualize"	4.7%
Similarity to class and past exams	Describes how the question felt more uncomfortable because they had seen the same framing in class on exams or examples.	"Because it sounds like what we already do often."	1.6%
Stereotype threat	States that reading the word "you" led to negative emotions or connotations for that word	"When 'you' is put in a question I tend to feel overwhelmed."	1.6%
Less complexity is bad	Cited that they did not like a version because it did not provide enough information or details, or alternatively liked a version because it had sufficient information	"Though more straight to the point, the question [for self-referential] does not provide context in the way that [authentic version] does. It is nice knowing that there were others who have done the experiments about KRAS, but additionally, the context gives great meaning and helps understand what process is being done."	0.8%

in specific scenarios or for a given type of mathematical problem (Davis-Dorsey *et al.*, 1991; López and Sullivan, 1992). These results again suggest that the type of assessment question, the context where it is given, and the demographics of the students responding to the question all likely play roles in influencing how students respond to the different variants and whether there is a difference in performance.

Impact on affect

We similarly found no difference in either student stress or student identity between students taking the different assessment question variants or between students who preferred each of the versions. This again contradicts several past studies, done exclusively in primary and secondary mathematics education, that have found that different variants of problems have impacted student stress, interest, and motivation (Davis-Dorsey *et al.*, 1991; López & Sullivan, 1992; Hart, 1996; Riley, 2001). Like with performance, we speculate that the context of the assessments may influence how much the question variants may influence different affective constructs. For instance, college students have different stressors than students in K–12 and likely rely on different coping strategies as well (Zeidner, 1996; Hicks and Heastie, 2008). Similarly, college students, by virtue of being more advanced than students in primary and secondary education, have had significantly more experiences with STEM disciplines than the elementary students included in some of these other studies. These experiences may cause college students to have different motivations, values, and ability beliefs than elementary school students, which could mediate the influence of the different assessment framing variants (Robnett and Leaper, 2013; LaForce *et al.*, 2017; Starr, 2018).

In addition, our exploratory work was constrained to a very limited number of Likert-scale questions, given the time constraints of students answering questions during the assessment. We relied on questions answered during the assessment, given our goal of capturing affect during the act of completing such assessments in biology classes. However, most validated instruments for measuring different aspects of student affect, including science identity, attitudes toward science, and sense of belonging, rely on 20 or more Likert-scale items (Lovelace and Brickman, 2013; Trujillo and Tanner, 2014; Chen and Wei, 2022). It is possible that the limited number of Likert-scale questions we used, combined with the relatively small class sizes at our institution, limited the ability of our study to detect small but meaningful changes in student affect between students taking the different variants (Al-Subaihi, 2003). Despite this, we note that our results from the end of semester survey indicate that the different question variants may be influencing affective factors for many students, and likewise may have impact on some students' ability to build textbase and situation models (see section below).

Student preferences reveal that the different variants may have potentially different affective impacts and influence students' ability to read and process a scenario

Despite no measurable affective impact during assessments, students demonstrated strong preferences for different variants when presented with all three versions and provided a range of different reasons to justify which version they most and least preferred. Most students preferred the self-referential version and identified either the authentic or classmate-referential variants as their least preferred versions. We find that students cite a range of different reasons for their preferences, suggesting that there are differences in students' cognitive load and affect when reading and processing different variants of the question. These data, in turn, suggest that these differences likely impact how students build both their textbase and situation models of the scenarios and corresponding questions. We discuss the different reasons students provide for why they prefer (or do not prefer) each of the different variants below.

Self-referential

First, most students identified that they preferred the self-referential version, a preference that was observed in all three sections. The most frequently cited reason, stated by over half of the students who chose this variant, was that the self-referential version was the easiest of the three versions to read. These responses indicate that these students likely had an easier time constructing a textbase with this variant. While these students did not provide any reasons for why this version was easier to read, 27 % and 10% stated that they preferred the self-referential variant because it had fewer words and less complexity, respectively (these responses were coded separately from those who indicated a version was easier to read but did not provide a reason why). These reasons provide additional insight into why many students may find the self-referential variant to be easier to read, and we draw upon cognitive-load theory to further situate our results. Cognitive load theory has been used as a way to augment the textbase-situation model framework in interpreting results from assessment question wording in the past (Walkington *et al.*, 2019). This theory states that there are limits to how much each student can mentally engage with each problem, with a limited capacity of working memory (Plass *et al.*, 2010; Sweller, 2011; Walkington *et al.*, 2019). Too much information in a problem, including any information that is perceived as extraneous, may overwhelm the working memory, interfere with the ability to build a textbase, and lead to reduced cognitive abilities (Sweller, 2011). Here, our results indicate that the self-referential version may trigger the lowest cognitive load in students. For instance, students noted that this version had the fewest words, and similarly cited how the authentic and classmate-referential versions had the most words, given that the latter versions include an additional name and (for the authentic version) the year of the study. These details – while integral parts of the classmate-referential and authentic versions – thus are likely perceived by these students as extraneous details that the self-referential version avoids, thus reducing complexity and cognitive load. These results align with past work that has found that reducing unnecessary sentences and details in problems can improve problem-solving abilities (Sweller, 1988; Tarmizi and Sweller, 1988).

In addition, a fourth of students who indicated that they preferred the self-referential variant stated that their preference was because this version made it easier for them to see themselves conducting the experiment itself. For instance, one student wrote that this version "was straight to the point and wasn't wordy in any way. I liked how the scenario put me in the question because I was able to picture me doing the experiment and it made me feel more like a science person." This reasoning is consistent with past work that has found that students, when

presented with a self-referential variant, may adopt a self-anchoring strategy where they center a given scenario around themselves, leading to faster processing of the cognitive information (D'Ailly *et al.*, 1995; Scheller and Sui, 2022). We hypothesize that a similar effect may be occurring here, where students are able to form a situation model more easily when centering themselves as the ones performing the experiment.

However, despite nearly two-thirds of students identifying this variant as the one they most preferred, 15% of students identified this variant as their least-preferred version. Only three reasons were provided by more than 10% of this group when prompted why this version was their least favorite. First, over a fourth of students in this group stated that this variant made it harder for them to visualize the experiment, followed by an eighth of students each indicating that the framing led to negative emotions or connotations, or that the version did not feel like authentic, real science. For instance, one student wrote that this variant was their least preferred version because "it lacks any personal description of the experiment which is what helps me visualize the experiment best", directly conveying how this version challenges their ability to form a situation model. Similarly, another student wrote "it's not telling me who's testing it and for what purpose…I might not care as much about it." This response highlights how some students may have lower situational interest - defined as a "temporary interest that arises spontaneously due to environmental factors" (Schraw *et al.*, 2001) – in this variant due to the lack of details about who is performing the experiment and instead relying upon the self-referential "you."

In sum, these reasons indicate that affective impacts may vary drastically within students. For instance, while most students had an easier time forming a situation model with this variant, others disagreed, suggesting that they may not be adopting a self-anchoring strategy when reading and processing this variant. Similarly, other students indicate that reading this variant prompted negative feelings. We speculate that the self-referential wording may be triggering stereotype threat, the implicit danger of confirming a negative stereotype about a given identity or group (Steele and Aronson, 1995). While we were not able to link demographic questions to individual responses on the end of semester survey, it is possible that some of these students may hold identities that they perceive as having negative stereotypes about abilities in biology. Past work has identified that many factors, such as a given survey question, the asking of demographic information, and the identities of the instructor, can all trigger stereotype threat (Lauer *et al.*, 2013). Thus, it is possible that the self-referential framing of "you" sparked introspection and these feelings of risk, causing these students to not prefer this version.

Authentic variant

The second-most preferred version, selected by a fifth of students, was the authentic variant. Students gave a variety of reasons for preferring this version. Nearly a third of students in this group cited how they felt that this version was the easiest to read, though this was lower than the percent of students who identified that the self-referential version was the easiest to read. Similarly, a fourth of students stated that they felt they could see themselves conducting this experiment, appreciated the real-life examples and how applicable the concepts were to authentic experiments, and liked the additional details that the authentic framing conveyed that the other versions did not have. For example, one student wrote that "I just personally like the idea of using real-life examples from actual scientific studies done. It makes me feel like the work I am doing is more integrated outside the classroom."

These results are consistent with past work that has found that students have an easier time constructing a situation model, if they are interested in the given topic (Walkington *et al.*, 2015). For instance, the authentic scenario may have sparked a greater situational interest in these students who likely recognized that some of the biological principles they were learning in class were applicable to authentic scientific studies. In addition, the counterstereotypical scientist names used may also play a role in sparking situational interest. There are several possible reasons. First, past work has identified that many adolescent students find STEM role models who deviate from stereotypes as more interesting, indicating that the use of counterstereotypical names may have sparked greater-situational interest (Steinke *et al.*, 2022). Second, students are likely drawing inferences about the gender and race or ethnicity of the scientist from the given name, which may shape students' interest (Kozlowski *et al.*, 2022). For example, one student stated "Doctors are cool! And I love seeing female scientists," indicating that the student had ascribed a gender to the scientist based on the name and pronouns used, and hints at an increased interest and excitement because of these potentially shared identities. This idea connects to past work that has identified that students in STEM connect more with mentors and role models who share a similar set of identities and/or experiences as the student (Buck *et al.*, 2008; Atkins *et al.*, 2020). Thus, it is possible that students who identify with the same gender, race, or ethnicity as those of the scientist (as perceived by the student) in the problem may see an increase in their situational interest (and thus ability to build a situational model) because of these shared identities.

However, this variant was identified as the least-preferred version by the plurality of students (40.8%). Students gave several reasons. First, nearly half of students cited the increased complexity of the authentic version, which was the only variant to include the year of the study, with others citing the increased length of the problem due to the addition of the year and title of the scientist. "The [year] 2017 study is not necessary," one student wrote. "I prefer if the question was straight to the point and did not have extra information that I do not need to answer the question." These responses reveal that these students may have a more challenging time reading and formulating a textbase for this scenario, with the additional details and length making the scenario harder to understand for these students and increasing their cognitive load.

Interestingly, a fifth of students identified that this version was their least-preferred version because they had negative emotions from seeing an authentic scientist's name. "I don't know who this doctor is and it makes it sound like they are doing something hard so [the problem] subconsciously makes me feel like it is going to be a hard question even if it isn't," one student wrote. Similarly, another student conveyed how reading about "Dr. Chen['s] research makes me think they spent a long time on it studying it, so how would we be able to know the answer." These responses reveal that the self-efficacy of

these students, or their confidence in their ability to succeed, is likely being negatively impacted by seeing an authentic scientist name and title. Physiological and affective states have been identified in Bandura's framework of self-efficacy as one of the four main factors that influence students' self-efficacy (Bandura *et al.*, 1999; Trujillo and Tanner, 2014), and these responses reveal that seeing a scientists' name, along with their designation as a "Dr.", may be triggering an affective response where the students become more intimidated by the scenario and question. While we are not aware of any previous work that has examined the impact of changes in self-efficacy on students' ability to build a textbase and situation model, this decrease in self-efficacy may have negative impacts on these students' performance as well as interest and motivation in STEM (Trujillo and Tanner, 2014; Ballen *et al.*, 2017).

Classmate-referential variant

The classmate-referential version was chosen by the fewest number of students as their preferred version, with fewer than 12% selecting this variant. Interestingly, a plurality of students in this group cited how this version helped connect them to and relate back to the scenario. "Seeing Veronica sounds like a friend's name or peer," one student wrote, indicating that while they did not directly recognize Veronica as a classmate, they still ascribed the name to a fellow undergraduate student. Others directly recognized Veronica as a classmate. For example, one student wrote that they preferred this version "because while I am taking the quiz/exam, I can visualize my classmates doing the experiments and have more confidence in my abilities. Furthermore, if my classmates are making genetic modifications, there is no reason that I can too one day. I also feel like having a classmate's name makes it easier to read/follow because I know who they are and do not get confused." This student's response highlights several themes: first, the student indicates that they have an easier time forming a textbase and situation model with this version, suggesting that using classmates' names is likely beneficial for some students. Second, the student describes how reading about a classmate conducting experiments increases their self-efficacy. This response corresponds with one of Bandura's identified sources of self-efficacy, where vicarious experiences – observing peers complete tasks successfully – can increase self-efficacy (Bandura *et al.*, 1999; Usher and Pajares, 2008). It is intriguing that this vicarious experience does not come from a direct observation, but instead from the student reading about a fictional case of a classmate conducting an experiment. More work is needed in the future to investigate how widespread potential changes in self-efficacy are due to students interpreting an assessment-question scenario as a vicarious experience.

Despite this, many students identified this version as their least preferred version, with half stating that they viewed the length of this version as a barrier to them forming a textbase. One student commented that this version "has the most words compared with the other versions. It confuses me and makes it harder to hone on what needs to be identified." Several students commented upon how the repeated use of the classmate's name was distracting and again inhibited their formation of a textbase and situation model: "The names and details add an extra layer of confusion because it's a word you repeatedly have to keep reading and can get somewhat stress inducing," one student commented. "I'd rather get straight to the point. I focused more on 'Veronica' than the problem or case at hand." Finally, a third of students in this group indicated that they disliked this version because of its lack of authenticity. "Reading something about a scientist doing a study like in [the authentic version] is more interesting to me and grabs my attention more than reading about some random student/person like in [this version]," one student commented. "So I would least prefer [the classmate-referential version]."

We speculate that some of this variance in how students responded to the classmate-referential version may depend on if the student is familiar with Veronica, the student named in the example. Given our class sizes of nearly 60 students in each section, it is plausible that many students may not recognize the name as that of a classmate, and those students may have a different affective response to this version than those who are friends with Veronica or know of her as a classmate. For instance, situational interest can be sparked by thinking about people that a student knows, and it is possible that those students who know Veronica may have a more favorable view of this version than those who do not know Veronica (Walkington, *et al.*, 2014; Walkington *et al.*, 2015). This difference in situational interest may thus in turn impact the students' ability to build their textbase and situation model (Walkington *et al.*, 2015). Similarly, it is possible that students may not want to be perceived as judging the thinking of one of their classmates.

Student preferences may be influenced by familiarity with a given version

Our data also suggests that student preferences may be influenced by what version they were given during the in-class assessments. For example, while each of the three sections preferred the self-referential version, the section that was given that version all semester showed the highest percentage of students, indicating that they preferred this version (Figure 2). Similarly, the section that was given the authentic variant showed the highest percentage of students preferring that version, and the section that was given the classmate-referential had the highest percentage of students preferring that version across the three sections (Figure 2). Several students commented upon this similarity in the survey. For example, one student wrote "this version is the most similar to practice and past exam problems so it is the easiest for me to understand." These results suggest that some of the student preferences may be driven by familiarity with a given version, with some students potentially gravitating toward the framing that they had seen on assessments in the class throughout the semester. Future work that surveys students on their preferences before the start of a term and then again after the term is needed to investigate the influence of this level of familiarity on student preferences.

Limitations and future directions

Our work has several limitations. First, we recognize that our study relies on students enrolled in one class (albeit across multiple sections) at one institution, and specific institutional attributes may limit the generalizability of the results. Similarly, our measures of affect were constrained to a limited number of Likert-scale items, due to the limitations of having students

answering such questions during assessments, and we were also unable to examine the impact of the variants on students' speed of answering each question. Future work that expands such work across a broader set of classes and instructors and utilizes interviews or other measures of potential affective change will provide more insight into the impact of differential assessment framing variants. We also acknowledge that our study is limited in the context of higher-order, scenario-based constructed response questions, and that many assessments in undergraduate biology classes may be lower-order questions and/or multiple-choice questions (Momsen *et al.*, 2010; Williams *et al.*, 2011; Momsen *et al.*, 2013). Future work is needed to investigate whether our results are applicable in the context of other types of assessment questions and across different demographics of students. Finally, our work is limited by relying on students conveying the impact of the different versions on their affect when describing their preferences in a survey; we were not able to directly measure what factors shape these preferences and why some students prefer a given version and others prefer a different version. We also did not directly ask about students' perceptions of each of these versions beyond asking them their preferences and reasoning for why they preferred (or did not prefer) a version. We were also not able to examine whether demographic factors may help drive some of these differences. Future work that relies on think-aloud interviews may be helpful to investigate the impact of these different assessment-question variants and the factors that shape these differential impacts.

Despite these limitations, our exploratory study is the first we are aware of to investigate the framing of assessment questions in the context of undergraduate biology. We did not find evidence that the different variants impacted either performance or affect based on the during-assessment measures. However, our end-of-semester survey reveals that such differences in framing are likely impacting students' discourse comprehension, including some students' ability to build a textbase and situation model, and that differences in framing do cause changes in how some students relate to the experiment. In sum, our work provides the first evidence that small changes of wording in biology assessment questions may influence students' affect and is the first work we are aware of that informs instructors of students' preferences and perceptions of these different versions. Future work is needed to continue studying how the framing of different examples used in class, in formative assessments, and in summative assessments may impact different students' affect and performance.

Implications for instructors

Our work provides several implications for instructors:

- **Reduce cognitive load in assessment questions.** Our results demonstrate that some students have challenges building their textbase and situation model if they see information they perceive as extraneous, such as scientists' names and the year of such studies, and that these students may prefer the self-referential model because of the reduced cognitive load. Instructors should be cognizant that any extra details not pertinent to the question may increase cognitive load and thus influence students' ability to form a situation model.
- **Consider student preferences when writing assessment questions.** Our results demonstrate that although there was no difference in performance, students largely prefer the self-referential version. Many of these students viewed the authentic- and classmate-referential variants negatively. However, there was no consensus, with a smaller number of students preferring either the authentic- or classmate-referential versions. Instructors may thus wish to survey their students on their preference of assessment-framing variant and use the results to guide how they present their assessments. For instance, instructors may wish to vary their assessment-framing variant based on the preferences in their class. Instructors of smaller classes may even consider tailoring the variant used in each individual assessment based on that student's preference, given the potentially positive impacts on student affect.
- **Maintain consistency with how assessment questions are framed and align in-class examples.** Our work showed that students in class exhibited a greater preference towards the version they had seen in assessments all semester, with multiple students commenting upon the familiarity of this variant when they were given the choice of versions. Instructors should thus consider aligning the examples used in class, practice problems, and other formative assessments in the same manner as those used in quizzes and exams. In addition, we encourage instructors to use the same framing throughout each of the summative assessments in the course, given the potential benefits of maintaining consistency with how these questions are framed.

ACKNOWLEDGMENTS

We thank Desiree Forsythe, Zach Thammavongsy, and Briana Craig for helpful feedback and comments on this manuscript.

REFERENCES

3 Tips for Creating Math Word Problems That Boost Critical Thinking. (2021). Retrieved September 26, 2022, from www.edutopia.org/article/3-tips-creating-math-word-problems-boost-critical-thinking

10 Fun and Original Ways to Teach Math Word Problems. (n.d.). Retrieved September 26, 2022, from https://boredteachers.com/post/teach-math-word-problems

Akinsola, M. K., & Awofala, A. O. A. (2009). Effect of personalization of instruction on students' achievement and self-efficacy in mathematics word problems. *International Journal of Mathematical Education in Science and Technology*, *40*(3), 389–404. doi:10.1080/00207390802643169

Al-Subaihi, A. A. (2003). Sample size determination. Influencing factors and calculation strategies for survey research. *Neurosciences Journal*, *8*(2), 79–86.

Aprilia, F., Lustyantie, N., & Rafli, Z. (2020). The Effect of Reading Interest and Achievement Motivation on Students' Discourse Analysis Competence. *Journal of Education and E-Learning Research*, *7*(4), 368–372.

Atkins, K., Dougan, B. M., Dromgold-Sermen, M. S., Potter, H., Sathy, V., & Panter, A. T. (2020). "Looking at Myself in the Future": How mentoring shapes scientific identity for STEM students from underrepresented groups. *International Journal of STEM Education*, *7*(1), 42. doi: 10.1186/s40594-020-00242-3

Awofala, A. O. A. (2014). Examining Personalisation of Instruction, Attitudes toward and Achievement in Mathematics Word Problems among Nigerian Senior Secondary School Students. *International Journal of Education in Mathematics, Science and Technology*, *2*(4), 273–288.

Ballen, C. J., Wieman, C., Salehi, S., Searle, J. B., & Zamudio, K. R. (2017). Enhancing Diversity in Undergraduate Science: Self-Efficacy Drives Performance Gains with Active Learning. *CBE—Life Sciences Education*, *16*(4), ar56. doi: 10.1187/cbe.16-12-0344

Bandura, A., Freeman, W. H., & Lightsey, R. (1999). Self-Efficacy: The Exercise of Control. *Journal of Cognitive Psychotherapy*, *13*(2), 158–166. doi: 10.1891/0889-8391.13.2.158

Bates, E. T., & Wiest, L. R. (2004). Impact of Personalization of Mathematical Word Problems on Student Performance. *The Mathematics Educator*, *14*(2), 17–26. https://openjournals.libs.uga.edu/tme/article/view/1876

Beavers, A., Lounsbury, J., Richards, J., Huck, S., Skolits, G., & Esquivel, S. (2019). Practical Considerations for Using Exploratory Factor Analysis in Educational Research. *Practical Assessment, Research, and Evaluation*, *18*(1) doi: https://doi.org/10.7275/qv2q-rk76

Beck, C. W., & Blumer, L. S. (2016). Alternative Realities: Faculty and Student Perceptions of Instructional Practices in Laboratory Courses. *CBE—Life Sciences Education*, *15*(4), ar52. doi: 10.1187/cbe.16-03-0139

Bernacki, M., & Walkington, C. (2014). The Impact of a Personalization Intervention for Mathematics on Learning and Non-Cognitive Factors. In Stamper, J., Pardos, Z., Mavrikis, M.,& McLaren, B. M. (Eds.), *Proceedings of the 7th International Conference on Educational Data Mining, London, UK*.

Blair, M., & Frezza, S. (2020). Assessing interest and confidence as components of student motivation in informal STEM learning. *2020 IEEE Frontiers in Education Conference (FIE)*, 1–5. doi: 10.1109/FIE44824.2020.9273939

Buck, G. A., Clark, V. L. P., Leslie-Pelecky, D., Lu, Y., & Cerda-Lizarraga, P. (2008). Examining the cognitive processes used by adolescent girls and women scientists in identifying science role models: A feminist approach. *Science Education*, *92*(4), 688–707. doi: 10.1002/sce.20257

Cakir, O., Cakmak, S., & Yilmaz, F. (2016). The Effect of Personalization in Physics Teaching. *New Trends and Issues Proceedings on Humanities and Social Sciences*, *6*, 146–156. https://un-pub.eu/ojs/index.php/pntsbs/article/view/1165

Cakir, O., & Simsek, N. (2010). A comparative analysis of the effects of computer and paper-based personalization on student achievement. *Computers & Education*, *55*(4), 1524–1531. doi: 10.1016/j.compedu.2010.06.018

Carnegie Learning Only K-12 Finalist for Dept. of Education $1Million XPRIZE. (2022). Retrieved September 26, 2022, from www.yahoo.com/now/carnegie-learning-only-k-12-183000976.html

Cassady, J. C. (2004). The impact of cognitive test anxiety on text comprehension and recall in the absence of external evaluative pressure. *Applied Cognitive Psychology*, *18*(3), 311–325. doi: 10.1002/acp.968

Chen, M., Lin, M., & Wang, X. (2018). Overexpression of miR-19a inhibits colorectal cancer angiogenesis by suppressing KRAS expression. *Oncology Reports*, *39*(2), 619–626. doi: 10.3892/or.2017.6110

Chen, S., & Wei, B. (2022). Development and Validation of an Instrument to Measure High School Students' Science Identity in Science Learning. *Research in Science Education*, *52*(1), 111–126. doi: 10.1007/s11165-020-09932-y

Chiavaroli, N. (2019). Negatively-Worded Multiple Choice Questions: An Avoidable Threat to Validity. *Practical Assessment, Research, and Evaluation*, *22*(1) https://doi.org/10.7275/ca7y-mm27

Conway, M. A., & Dewhurst, S. A. (1995). The self and recollective experience. *Applied Cognitive Psychology*, *9*(1), 1–19. doi: 10.1002/acp.2350090102

Craik, F. I. M., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General*, *104*(3), 268–294. doi: 10.1037/0096-3445.104.3.268

d'Ailly, H. H., Simpson, J., & MacKinnon, G. E. (1997). Where should" you" go in a math compare problem? *Journal of Educational Psychology*, *89*(3), 562.

D'Agata, B. (2015). The Influence of Personalized Mathematical Word Problems on Second Graders' Performance, Attitudes Toward Word Problems, and Difficulty Ratings. *Student Research Submissions*. https://scholar.umw.edu/student_research/118

D'Ailly, H. H., Murray, H. G., & Corkill, A. (1995). Cognitive Effects of Self-Referencing. *Contemporary Educational Psychology*, *20*(1), 88–113. doi: 10.1006/ceps.1995.1005

Davis-Dorsey, J., Ross, S. M., & Morrison, G. R. (1991). The role of rewording and context personalization in the solving of mathematical word problems. *Journal of Educational Psychology*, *83*(1), 61.

de Koning, B. B., & van der Schoot, M. (2019). Can "you" make a difference? Investigating whether perspective-taking improves performance on inconsistent mathematical word problems. *Applied Cognitive Psychology*, *33*(5), 911–917. doi: 10.1002/acp.3555

Dou, R., Hazari, Z., Dabney, K., Sonnert, G., & Sadler, P. (2019). Early informal STEM experiences and STEM identity: The importance of talking science. *Science Education*, *103*(3), 623–637. doi: 10.1002/sce.21499

Findley, J. (2016, August 21). 8 Ways to Help Students Be Successful with Word Problems in Upper Elementary. Retrieved September 26, 2022, from https://jenniferfindley.com/help-students-word-problems-upper-elementary/

Findley-Van Nostrand, D., & Pollenz, R. S. (2017). Evaluating Psychosocial Mechanisms Underlying STEM Persistence in Undergraduates: Evidence of Impact from a Six-Day Pre–College Engagement STEM Academy Program. *CBE—Life Sciences Education*, *16*(2), ar36. doi: 10.1187/cbe.16-10-0294

Freelon, D. (2013). ReCal OIR: Ordinal, interval, and ratio intercoder reliability as a web service. *International Journal of Internet Science*, *8*(1)

Good, C., Rattan, A., & Dweck, C. S. (2012). Why do women opt out? Sense of belonging and women's representation in mathematics. *Journal of Personality and Social Psychology*, *102*(4), 700–717. doi: 10.1037/a0026659

Graesser, A. C., & Zwaan, R. A. (1995). Inference generation and the construction of situation models. In Weaver III, C. A., Mannes, S., & Fletcher, C. R. (Eds.), *Discourse Comprehension: Essays in Honor of Walter Kintsch* (pp. 117–139). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Gunel, M., Hand, B., & McDermott, M. A. (2009). Writing for different audiences: Effects on high-school students' conceptual understanding of biology. *Learning and Instruction*, *19*(4), 354–367. doi: 10.1016/j.learninstruc.2008.07.001

Harris, R. B., Grunspan, D. Z., Pelch, M. A., Fernandes, G., Ramirez, G., & Freeman, S. (2019). Can Test Anxiety Interventions Alleviate a Gender Gap in an Undergraduate STEM Course? *CBE—Life Sciences Education*, *18*(3), ar35. doi: 10.1187/cbe.18-05-0083

Hart, J. M. (1996). The effect of personalized word problems. *Teaching Children Mathematics*, *2*(8), 504–506.

Hicks, T., & Heastie, S. (2008). High School to College Transition: A Profile of the Stressors, Physical and Psychological Health Issues That Affect the First-Year On-Campus College Student. *Journal of Cultural Diversity*, *15*(3), 143–147. https://digitalcommons.uncfsu.edu/soe_faculty_wp/14

Hsu, J. L., & Goldsmith, G. R. (2021). Instructor Strategies to Alleviate Stress and Anxiety among College and University STEM Students. *CBE—Life Sciences Education*, *20*(1), es1. doi: 10.1187/cbe.20-08-0189

Hurtado, S., & Carter, D. F. (1997). Effects of College Transition and Perceptions of the Campus Racial Climate on Latino College Students' Sense of Belonging. *Sociology of Education*, *70*(4), 324–345. doi: 10.2307/2673270

Jamieson, J. P., Peters, B. J., Greenwood, E. J., & Altose, A. J. (2016). Reappraising Stress Arousal Improves Performance and Reduces Evaluation Anxiety in Classroom Exam Situations. *Social Psychological and Personality Science*, *7*(6), 579–587. doi: 10.1177/1948550616644656

JASP Team. (2022). *JASP (Version 0.16.4)*. https://jasp-stats.org

Johnson, J. M., Bristow, D. N., & Schneider, K. C. (2004). Did You Not Understand The Question Or Not? An Investigation Of Negatively Worded Questions In Survey Research. *Journal of Applied Business Research (JABR)*, *20*(1), 75–86. doi: 10.19030/jabr.v20i1.2197

Kim, A. Y., Sinatra, G. M., & Seyranian, V. (2018). Developing a STEM Identity Among Young Women: A Social Identity Perspective. *Review of Educational Research*, *88*(4), 589–625. doi: 10.3102/0034654318779957

Kintsch, W. (1986). Learning from text. *Cognition and Instruction*, *3*(2), 87–108.

Klein, S. B. (2012). Self, Memory, and the Self-Reference Effect: An Examination of Conceptual and Methodological Issues. *Personality and Social Psychology Review*, *16*(3), 283–300. doi: 10.1177/1088868311434214

Klein, S. B., & Kihlstrom, J. F. (1986). Elaboration, organization, and the self-reference effect in memory. *Journal of Experimental Psychology: General*, *115*(1), 26–38. doi: 10.1037/0096-3445.115.1.26

Klein, S. B., & Loftus, J. (1988). The nature of self-referent encoding: The contributions of elaborative and organizational processes. *Journal of Personality and Social Psychology*, *55*(1), 5–11. doi: 10.1037/0022-3514.55.1.5

Knekta, E., Runyon, C., & Eddy, S. (2019). One Size Doesn't Fit All: Using Factor Analysis to Gather Validity Evidence When Using Surveys in Your Research. *CBE—Life Sciences Education*, *18*(1), rm1. doi: 10.1187/cbe.18-04-0064

Koh, K. H. (2017, February 27). *Authentic Assessment*. doi: 10.1093/acrefore/9780190264093.013.22

Kozlowski, D., Murray, D. S., Bell, A., Hulsey, W., Larivière, V., Monroe-White, T., & Sugimoto, C. R. (2022). Avoiding bias when inferring race using name-based approaches. *PLOS ONE*, *17*(3), e0264270. doi: 10.1371/journal.pone.0264270

Krawec, J., & Warger, C. (2015). *Solve It! Teaching Mathematical Problem Solving in Inclusive Classrooms—Grades 5-6*, 1st ed., Reston, VA: Exceptional Innovations.

Ku, H.-Y., & Sullivan, H. J. (2000). Personalization of mathematics word problems in Taiwan. *Educational Technology Research and Development*, *48*(3), 49–60. doi: 10.1007/BF02319857

Ku, H.-Y., & Sullivan, H. J. (2001). Effects of Personalized Instruction on Mathematics Word Problems in Taiwan. In: *Annual proceedings of Selected Research and Development [and] Practice Papers presented at the National Convention of the Association for Educational Communications and Technology from November 8-12, 2001 at Atlanta, GA* Vol 1–2, (pp. 85–94). https://eric.ed.gov/?id=ED470135

LaForce, M., Noble, E., & Blackwell, C. (2017). Problem-Based Learning (PBL) and Student Interest in STEM Careers: The Roles of Motivation and Ability Beliefs. *Education Sciences*, *7*(4), 92. doi: 10.3390/educsci7040092

Landis, J. R., & Koch, G. G. (1977). An Application of Hierarchical Kappa-type Statistics in the Assessment of Majority Agreement among Multiple Observers. *Biometrics*, *33*(2), 363–374. doi: 10.2307/2529786

Lauer, S., Momsen, J., Offerdahl, E., Kryjevskaia, M., Christensen, W., & Montplaisir, L. (2013). Stereotyped: Investigating Gender in Introductory Science Courses. *CBE—Life Sciences Education*, *12*(1), 30–38. doi: 10.1187/cbe.12-08-0133

Leandre, R., Fabrigar, L. R., & Wegener, D. T. (2012). *Exploratory factor analysis*. Oxford, UK: Oxford University Press.

Locke, K. (2002). The grounded theory approach to qualitative research. In *The Jossey-Bass Business & Management Series*. *Measuring and analyzing behavior in organizations: Advances in measurement and data analysis*, San Francisco, CA, US, pp. 17–43.

López, C. L., & Sullivan, H. J. (1992). Effect of personalization of instructional context on the achievement and attitudes of hispanic students. *Educational Technology Research and Development*, *40*(4), 5–14. doi: 10.1007/BF02296895

Lovelace, M., & Brickman, P. (2013). Best Practices for Measuring Students' Attitudes toward Learning Science. *CBE—Life Sciences Education*, *12*(4), 606–617. doi: 10.1187/cbe.12-11-0197

Lytle, A., & Shin, J. E. (2020). Incremental Beliefs, STEM Efficacy and STEM Interest Among First-Year Undergraduate Students. *Journal of Science Education and Technology*, *29*(2), 272–281. doi: 10.1007/s10956-020-09813-z

Mayer, R. E., Fennell, S., Farmer, L., & Campbell, J. (2004). A personalization effect in multimedia learning: Students learn better when words are in conversational style rather than formal style. *Journal of Educational Psychology*, *96*(2), 389.

McCormick, A. C., & Zhao, C.-M. (2005). Rethinking and reframing the Carnegie classification.*Change: The Magazine of Higher Learning*, *37*(5), 51–57.

McDonald, M. M., Zeigler-Hill, V., Vrabel, J. K., & Escobar, M. (2019). A Single-Item Measure for Assessing STEM Identity. *Frontiers in Education*, *4*, www.frontiersin.org/articles/10.3389/feduc.2019.00078

Melsky, K. (2021). *Effect of Personalized Problems in Undergraduate Thermal Fluid Transport Courses*. PhD Thesis, Tufts University.

Momsen, J. L., Long, T. M., Wyse, S. A., & Ebert-May, D. (2010). Just the Facts? Introductory Undergraduate Biology Courses Focus on Low-Level Cognitive Skills. *CBE—Life Sciences Education*, *9*(4), 435–440. doi: 10.1187/cbe.10-01-0001

Momsen, J., Offerdahl, E., Kryjevskaia, M., Montplaisir, L., Anderson, E., & Grosz, N. (2013). Using Assessments to Investigate and Compare the Nature of Learning in Undergraduate Science Courses. *CBE—Life Sciences Education*, *12*(2), 239–249. doi: 10.1187/cbe.12-08-0130

Moreno, R., & Mayer, R. E. (2000). Engaging students in active learning: The case for personalized multimedia messages. *Journal of Educational Psychology*, *92*(4), 724.

Moudgalya, S. K., Mayfield, C., Yadav, A., Hu, H. H., & Kussmaul, C. (2021). Measuring Students' Sense of Belonging in Introductory CS Courses. *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education*, (pp. 445–451). New York, NY: Association for Computing Machinery. https://dl.acm.org/doi/abs/10.1145/3408877.3432425

Nawawi, N. M., Sout, N. M., Hassan, K. B., Samah, N. N. A., Kamaruddin, H. H., Khalid, R. M., & Azman, H. H. (2021). The perception of pre-university students on STEM. *Journal of Physics: Conference Series*, *1882*(1), 012155. doi: 10.1088/1742-6596/1882/1/012155

O'Neil, Jr., H. F., & Brown, R. S. (1998). Differential Effects of Question Formats in Math Assessment on Metacognition and Affect. *Applied Measurement in Education*, *11*(4), 331–351. doi: 10.1207/s15324818ame1104_3

Pape-Lindstrom, P., Eddy, S., & Freeman, S. (2018). Reading Quizzes Improve Exam Scores for Community College Students. *CBE—Life Sciences Education*, *17*(2), ar21. doi: 10.1187/cbe.17-08-0160

Plass, J. L., Moreno, R., & Brünken, R. (Eds.) (2010). *Cognitive load theory* (pp. 253–272). Cambridge University Press. https://doi.org/10.1017/CBO9780511844744

Rai, M. K., Loschky, L. C., Harris, R. J., Peck, N. R., & Cook, L. G. (2011). Effects of Stress and Working Memory Capacity on Foreign Language Readers' Inferential Processing During Comprehension. *Language Learning*, *61*(1), 187–218. doi: 10.1111/j.1467-9922.2010.00592.x

Reflections on Improving Diversity and Inclusion in Science Teaching | Diverse Educators. (2021, July 16). Retrieved August 30, 2022, from www.diverseeducators.co.uk/reflections-on-improving-diversity-and-inclusion-in-science-teaching/

Riley, M. M. (2001). *The effects of problem revision of mathematics word problems on performance and anxiety for fourth-grade students.* Dissertation for PhD in Education. University of South Carolina, United States – South Carolina. www.proquest.com/docview/250185330/abstract/59812E5E8D25481BPQ/1

Robnett, R. D., & Leaper, C. (2013). Friendship Groups, Personal Motivation, and Gender in Relation to High School Students' STEM Career Interest. *Journal of Research on Adolescence*, *23*(4), 652–664. doi: 10.1111/jora.12013

Rogers, T. B., Kuiper, N. A., & Kirker, W. S. (1999). *Self-reference and the encoding of personal information*, New York, NY, US: Psychology Press, pp. 149.

Roszkowski, M. J., & Soven, M. (2010). Shifting gears: Consequences of including two negatively worded items in the middle of a positively worded questionnaire. *Assessment & Evaluation in Higher Education*, *35*(1), 113–130. doi: 10.1080/02602930802618344

Santoro, L. R., & Bunte, J. B. (2023). What Did You Get? Peers, Information, and Student Exam Performance. *Research in Higher Education*, *64*(3), 423–450. doi: 10.1007/s11162-022-09711-w

Scheller, M., & Sui, J. (2022). The power of the self: Anchoring information processing across contexts. *Journal of Experimental Psychology: Human Perception and Performance*, *48*(9), 1001–1021.

Schinske, J., Cardenas, M., & Kaliangara, J. (2015). Uncovering Scientist Stereotypes and Their Relationships with Student Race and Student Success in a Diverse, Community College Setting. *CBE—Life Sciences Education*, *14*(3), ar35. doi: 10.1187/cbe.14-12-0231

Schinske, J. N., Perkins, H., Snyder, A., & Wyer, M. (2016). Scientist Spotlight Homework Assignments Shift Students' Stereotypes of Scientists and Enhance Science Identity in a Diverse Introductory Science Class. *CBE—Life Sciences Education*, *15*(3), ar47. doi: 10.1187/cbe.16-01-0002

Schraw, G., Flowerday, T., & Lehman, S. (2001). Increasing Situational Interest in the Classroom. *Educational Psychology Review*, *13*(3), 211–224. doi: 10.1023/A:1016619705184

Sharkawy, A. (2012). Exploring the potential of using stories about diverse scientists and reflective activities to enrich primary students' images of scientists and scientific work. *Cultural Studies of Science Education*, *7*(2), 307–340. doi: 10.1007/s11422-012-9386-2

Singer, A., Montgomery, G., & Schmoll, S. (2020). How to foster the formation of STEM identity: Studying diversity in an authentic learning environment. *International Journal of STEM Education*, *7*(1), 57. doi: 10.1186/s40594-020-00254-z

Sonderen, E. van, Sanderman, R., & Coyne, J. C. (2013). Ineffectiveness of Reverse Wording of Questionnaire Items: Let's Learn from Cows in the Rain. *PLOS ONE*, *8*(7), e68967. doi: 10.1371/journal.pone.0068967

Starr, C. R. (2018). "I'm Not a Science Nerd!": STEM Stereotypes, Identity, and Motivation Among Undergraduate Women. *Psychology of Women Quarterly*, *42*(4), 489–503. doi: 10.1177/0361684318793848

Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology*, *69*, 797–811. doi: 10.1037/0022-3514.69.5.797

Steinke, J., Applegate, B., Penny, J. R., & Merlino, S. (2022). Effects of Diverse STEM Role Model Videos in Promoting Adolescents' Identification. *International Journal of Science and Mathematics Education*, *20*(2), 255–276. doi: 10.1007/s10763-021-10168-z

Steppler, K. C. (2020). *Evaluating the Use of Writing Prompts & Graphic Organizers in Middle School Mathematics: Action Research to Improve Mathematical Achievement and Students' Attitudes with Authentic Assessments*. Ed.D., University of South Carolina. University of South Carolina, United States – South Carolina. www.proquest.com/docview/2503956414/abstract/728C8985E5B34CEDPQ/1

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, *12*(2), 257–285. doi: 10.1016/0364-0213(88)90023-7

Sweller, J. (2011). CHAPTER TWO - Cognitive Load Theory. In Mestre, J. P., Ross, B. H. (Eds.), *Psychology of Learning and Motivation* Vol. 55, (pp. 37–76). Academic Press. doi: 10.1016/B978-0-12-387691-1.00002-8

Symons, C. S., & Johnson, B. T. (1997). The self-reference effect in memory: A meta-analysis. *Psychological Bulletin*, *121*(3), 371–394. doi: 10.1037/0033-2909.121.3.371

Tarmizi, R. A., & Sweller, J. (1988). Guidance during mathematical problem solving. *Journal of Educational Psychology*, *80*, 424–436. doi: 10.1037/0022-0663.80.4.424

The National Association for Multicultural Education. (n.d.). *I teach science. Can I be a multicultural educator?* Retrieved August 30, 2022, from www.nameorg.org/learn/i_teach_science_can_i_be_a_mu.php

Trujillo, G., & Tanner, K. D. (2014). Considering the Role of Affect in Learning: Monitoring Students' Self-Efficacy, Sense of Belonging, and Science Identity. *CBE—Life Sciences Education*, *13*(1), 6–15. doi: 10.1187/cbe.13-12-0241

Turk, D. J., Gillespie-Smith, K., Krigolson, O. E., Havard, C., Conway, M. A., & Cunningham, S. J. (2015). Selfish learning: The impact of self-referential encoding on children's literacy attainment. *Learning and Instruction*, *40*, 54–60. doi: 10.1016/j.learninstruc.2015.08.001

Usher, E. L., & Pajares, F. (2008). Sources of Self-Efficacy in School: Critical Review of the Literature and Future Directions. *Review of Educational Research*, *78*(4), 751–796. doi: 10.3102/0034654308321456

Van Dijk, T. A., & Kintsch, W. (1983). *Strategies of discourse comprehension*. New York, NY: Academic Press.

Villarroel, V., Bloxham, S., Bruna, D., Bruna, C., & Herrera-Seda, C. (2018). Authentic assessment: Creating a blueprint for course design. *Assessment & Evaluation in Higher Education*, *43*(5), 840–854. doi: 10.1080/02602938.2017.1412396

Walkington, C., & Bernacki, M. L. (2014). Motivating students by "personalizing" learning around individual interests: A consideration of theory, design, and implementation issues. In Karabenick, S., & Urdan, T. (Eds.), *Advances in motivation and achievement*, *18* (pp. 139–176). Bingley, England: Emerald.

Walkington, C., Clinton, V., Ritter, S. N., & Nathan, M. J. (2015). How readability and topic incidence relate to performance on mathematics story problems in computer-based curricula. *Journal of Educational Psychology*, *107*(4), 1051.

Walkington, C., Clinton, V., & Sparks, A. (2019). The effect of language modification of mathematics story problems on problem-solving in online homework. *Instructional Science*, *47*(5), 499–529. doi: 10.1007/s11251-019-09481-6

Walsh, M., Hickey, C., & Duffy, J. (1999). Influence of Item Content and Stereotype Situation on Gender Differences in Mathematical Problem Solving. *Sex Roles*, *41*(3), 219–240. doi: 10.1023/A:1018854212358

Wiggins, G. (2019). The Case for Authentic Assessment. *Practical Assessment, Research, and Evaluation*, *2*(2) doi: https://doi.org/10.7275/ffb1-mm19

Williams, A. E., Aguilar-Roca, N. M., Tsai, M., Wong, M., Beaupré, M. M., & O'Dowd, D. K. (2011). Assessment of Learning Gains Associated with Independent Exam Analysis in Introductory Biology. *CBE—Life Sciences Education*, *10*(4), 346–356. doi: 10.1187/cbe.11-03-0025

Wright, C. D., Eddy, S. L., Wenderoth, M. P., Abshire, E., Blankenbiller, M., & Brownell, S. E. (2016). Cognitive Difficulty and Format of Exams Predicts Gender and Socioeconomic Gaps in Exam Performance of Students in Introductory Biology Courses. *CBE—Life Sciences Education*, *15*(2), ar23. doi: 10.1187/cbe.15-12-0246

Wright, C. D., Huang, A. L., Cooper, K. M., & Brownell, S. E. (2018). Exploring Differences in Decisions about Exams among Instructors of the Same Introductory Biology Course. *International Journal for the Scholarship of Teaching and Learning*, *12*(2). https://eric.ed.gov/?id=EJ1186071

Yonas, A., Sleeth, M., & Cotner, S. (2020). In a "Scientist Spotlight" Intervention, Diverse Student Identities Matter. *Journal of Microbiology & Biology Education*, *21*(1), 25. doi: 10.1128/jmbe.v21i1.2013

Zeidner, M. (1996). How do high school and college students cope with test situations? *British Journal of Educational Psychology*, *66*(1), 115–128. doi: 10.1111/j.2044-8279.1996.tb01181.x