INVESTIGATING THE IMPACT OF HIGHLIGHTING COUNTERSTEREOTYPICAL SCIENTISTS IN BIOLOGY TESTS

Subject/Problem

Tests remain prevalent in both secondary and post-secondary biology courses, and past work has found that small changes in how questions are worded can have a large influence on students' performance and attitudes (López and Sullivan, 1992; Bates and Wiest, 2004; Johnson et al. 2004; Cakir and Simsek, 2010). Despite their prevalence, there has been relatively limited work examining how the framing of scenario-based test questions shape students' attitudes. Our past work identified that most students preferred assessment questions framed in a self-referential manner (in which the student is situated as the scientist conducting the experiment) as compared to isomorphic questions framed around a scientist conducting the experiment (authors, 2023). Many students attributed this preference to a lower perceived cognitive load and a greater affirmation of their STEM identity when reading the self-referential version, despite a plethora of other papers finding that highlighting counterstereotypical scientists in other contexts (e.g., in class and in homework assignments) can lead to increases in students' STEM identity, self-efficacy, and interest in STEM (e.g., Yonas et al. 2020, Costello et al. 2025).

However, our past study had several key limitations. First, students' preferences were measured retroactively in a survey at the end of the semester, and students were never presented with both framings in the same test. Similarly, the course did not include any explicit narratives or spotlights on counterstereotypical scientists outside of the test, potentially limiting the impact of seeing counterstereotypical scientists on tests.

Here, we build upon our original work and address the following research questions: 1) How does exposure to counterstereotypical scientist spotlights in formative assignments influence students' preference for self-reference versus scientist-framed test questions? 2) How does this exposure affect their attitudes (e.g., motivation, STEM identity) when viewing test questions?

Design/Procedure

Institutional and Course Context: We conducted this study at a public, liberal arts college in upstate New York, in the context of a 200-level ecology course taken by mostly biology majors. The course had 49 students enrolled in two sections and included personal stories of counterstereotypical scientists embedded throughout the curriculum, including in assignments where students reflect on the relatability of these scientists and their perceptions of who can be a scientist. Tests in the course are designed around scenarios written about a recently published primary research article, with the role of the author highlighted. For each of the class's five tests, students were provided with two versions of the test scenario (but not the test questions) one week in advance of the test, written entirely with self-referential framing or scientist framing, and then allowed students to choose which version they desired for the test. Other than the self-referential or scientist framing, questions were isomorphic. We recorded students' choice of versions for each test. Data collection was approved by the university's IRB (#202324063), and we obtained informed consent from participants.

Survey Design and Deployment: We deployed a survey after each of the five tests; students completed each survey the day of the test. Survey questions were derived from our past work (authors, 2023) and included open-ended questions asking students to explain why they

chose their version of the test and closed-ended questions measuring the impact of the version on students' STEM identity and other affective constructs.

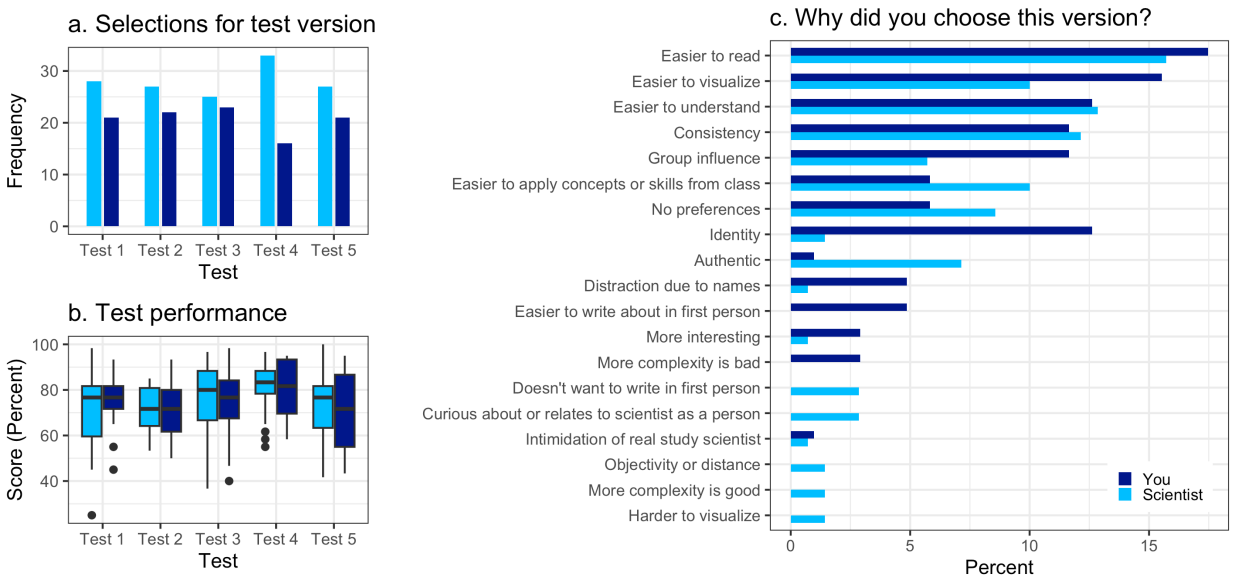
Interviews: We rely on a mixed-method approach with interviews following the survey to generate further insight into the impact of the different framings on students' attitudes and preferences. We developed an interview protocol through an iterative process, with several researchers independently brainstorming questions aligned with our research questions. Interested students consented to be contacted for interviews through the surveys. To minimize power differential, a trained undergraduate student conducted all interviews, which lasted up to one hour each. All interviews were audio recorded and transcribed.

Data Analysis: Closed-ended survey questions, as well as student preferences on each test, were analyzed with Fisher's or binomial exact tests. Two-sample tests (Wilcoxon rank sum and Student's *t*) were used to compare scores for each test. We used thematic analysis (Clark and Braun, 2017) for open-ended questions on the survey. In brief, multiple researchers independently reviewed a subset of survey responses, generated codebooks, before discussing to come up with a consensus codebook. Two researchers then independently coded 40 survey responses, representing 16.5% of all responses. Cohen's kappa was 0.86. Given this high interrater reliability, one researcher coded all remaining survey responses. Interview analysis followed a similar process, with two researchers independently generating codebooks before discussing to form a consensus codebook. Interrater reliability was again high (kappa = 0.84), allowing one researcher to code the remaining interviews. A two-proportions *z*-test or Fisher's exact test was used to compare the proportion of codes appearing in responses between versions. We used R (R Core Team 2025) for all statistical analyses.

Analyses and Findings

Our current study results contrast strongly with the previous study, which found a preference for self-referential framing (authors, 2023). Instead, for four of the five tests, students selected the two versions equally (all $p > 0.05$), though students showed a preference for the scientist-framed version on the fourth test ($p = 0.021$; Figure 1a). Notably, for each test, scores were the same across versions (all $p > 0.05$; Figure 1b), suggesting that performance was not tied to a preference for a specific test framing. Furthermore, post-test survey responses to Likert items revealed that student attitudes – STEM identity, stress – were similar between versions (except for survey item 2 on the second test: students who chose the scientist-framed version reported that they related more to the researchers than those who chose the student version; $p = 0.044$).

Figure 1: Test Version Selections, Student Performance, and Reasons for Choices. (a) Frequency each version was chosen for each test. (b) Scores for each test, based on version. Boxes represent the interquartile range (25th-75th percentiles), with the horizontal line denoting median. Whiskers extend to the most extreme data points within 1.5 times the IQR; individual dots represent statistical outliers. (c) Proportion of responses per code, for each version.



Through our qualitative analyses, we found that the most frequent reasons given for selecting a specific version were also similar between versions: the version they chose was easier to read, easier to visualize, easier to understand, and consistent with prior experiences (Figure 1c). A few differences arose for reasons that arose less commonly. For example, about 8% of responses identified that the version they chose felt more authentic, with those who chose the scientist-framed version reporting this reason more frequently than those who chose the self-referential framing (odds ratio = 7.32, $p = 0.03$). On the other hand, about 14% of responses mentioned the relatability of the version. Within those responses, students who chose the self-referential version were over eight times more likely to report that they could relate to it (identity; odds ratio = 8.77, $p = 0.0009$). For instance, one student wrote:

> It is easier to visualize what is going on when the test is worded in terms of me conducting the experiment. I also am still deciding on what I want to do with my biology degree, so reading about myself in STEM careers and situations is interesting. The questions are then easier to follow along since I already pictured myself carrying out the experiment. (Self-referential; Easier to visualize, Identity)

However, while relatability or validation of STEM identity appeared to be more commonly mentioned in the open-ended survey questions asking students to explain their preference, student responses to the Likert item "I see aspects of my identity reflected in the examples used in this test" did not differ between those that chose the self-referential vs. scientist-framed formats (per test, all $p > 0.05$). Also, there were no differences between students who chose each format for the Likert item "I could relate to the people performing the

experiments in this test" except for the instance mentioned above, in which those who chose the scientist-framed version related more strongly (Test 2 only). The frequency of codes may reflect their relative importance for the students in the study, with cognitive demand taking precedence over other reasons such as authenticity and identity. Indeed, follow-up interviews in the subsequent semester corroborated the post-test surveys, with students expressing most prominently that their choice of version was primarily driven by perceived ease of understanding and visualizing the experiment.

The follow-up interviews also reflected how preferences were a reaction to the high-stakes nature of a summative assessment. For instance, familiar formats appear to have been chosen to alleviate test-anxiety. For example, one student stated:

> I'm an anxious test taker, so, it's just because of prior testing and schooling… I'm used to this type of test, so I feel like it was… it made me more comfortable, because I knew I could read through it and get through the reading and get done what I needed to get done, rather than this 'yourself' version. (Scientist-framed)

Students also expressed a preference for the version they thought would optimize the time available to complete the test. When considering the version not chosen, one interviewed student stated "It probably would have taken me longer to read through the scenario, and then I probably would have run out of time for the questions." (Self-referential)

The evidence for the impact of prior exposure to counterstereotypical scientists on student preferences for question framing was inconclusive. Only two of the nine interviewees mentioned a connection between the exposure to scientists in formative assessments and the subsequent appearance of other, contemporary scientists on tests; thus, we are unable to directly attribute an influence of prior exposure to counterstereotypical scientists in formative assessments to a shift away from a preference for self-referential test question framing. Additional research is needed to ascertain if the positive impacts of featuring counterstereotypical scientists in curricula (e.g., Yonas et al. 2020, Costello et al. 2025) extend to the test-taking context.

For this population of students, the scientist-framed format was mentioned as feeling more familiar than the self-referential format. Students may feel stronger in their ability to succeed (i.e., increased self-efficacy) with a format they are familiar with, much like how prior exposure to content in prerequisite courses facilitates student learning (Shaffer et al. 2016). We suggest that practitioners use a mix of formats in order to give students greater familiarity with both types of tests, and explain the reasons why both formats should be practiced. Providing students with a choice of framing may reduce student test anxiety.

Contribution and General Interest

In summary, we found students were equally split in their preferences for self-referential and scientist-framed question formats and cited the same primary reason of reduced cognitive load, in contrast to prior work that showed a preference for self-referential framing. Our findings suggest that preferences depend on the context in which self-referential and scientist-framed questions are implemented. Our work also emphasizes that for high-stakes summative assessments, test question framing can have an impact on student test anxiety when the framing used is perceived as unfamiliar. Furthermore, our study is of broad interest to the NABT membership, from practitioners to education researchers. Tests are an element across biology courses at all levels, and our findings have implications for composing assessment questions. Our work also shows that the impact of question framing is complex and needs further study, especially for the context-dependence of student preferences.

References

Authors (2023). Investigating the influence of assessment question framing on undergraduate biology student preference and affect. CBE-Life Sciences Education 22(4).

Bates, E. T., & Wiest, L. R. (2004). Impact of personalization of mathematical word problems on student performance. The Mathematics Educator, 14(2), 17–26.

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

Clarke, V., & Braun, V. (2017). Thematic analysis. The journal of positive psychology, 12(3), 297-298.

Costello, R. A., Ewell, S. N., Adams, P. E., Aranda, M. L., Curry, A., De Jesus, M. M., ... & Ovid, D. (2025). Highlighting counterstereotypical scientists in undergraduate life science courses. CBE—Life Sciences Education, 24(2), es1.

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

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

R Core Team (2025). _R: A Language and Environment for Statistical Computing_. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Shaffer, J. F., Dang, J. V., Lee, A. K., Dacanay, S. J., Alam, U., Wong, H. Y., ... & Sato, B. K. (2016). A familiar (ity) problem: Assessing the impact of prerequisites and content familiarity on student learning. PloS one, 11(1), e0148051.

Yonas, A., Sleeth, M., & Cotner, S. (2020). In a “Scientist Spotlight” intervention, diverse student identities matter. Journal of microbiology & biology education, 21(1), 10-1128. doi: 10.1128/jmbe.v21i1.2013.