Summer 6-14-2023

Evaluating Assessment Score Validity and Characterizing Undergraduate Biology Exam Content

Crystal Uminski
University of Nebraska–Lincoln, crystal.uminski@huskers.unl.edu

EVALUATING ASSESSMENT SCORE VALIDITY AND

CHARACTERIZING UNDERGRADUATE BIOLOGY EXAM CONTENT

by

Crystal Uminski

A DISSERTATION

Presented to the Faculty of

The Graduate College at the University of Nebraska

In Partial Fulfillment of Requirements

For the Degree of Doctor of Philosophy

Major: Biological Sciences

Under the Supervision of Professor Brian A. Couch

Lincoln, Nebraska

June, 2023

EVALUATING ASSESSMENT SCORE VALIDITY AND

CHARACTERIZING UNDERGRADUATE BIOLOGY EXAM CONTENT

Crystal Uminski, Ph.D.

University of Nebraska, 2023

Advisor: Brian A. Couch

The landscape of undergraduate biology education has been shaped by decades of reform efforts calling for instruction to integrate core concepts and scientific skills as a means of helping students become proficient in the discipline. Assessments can be used to make inferences about how these reform efforts have translated into changes in department curriculum and course practices. Such changes can be measured using student scores on researcher-developed programmatic and concept assessments. Scores on these assessments are often assumed to be accurate representations of student biology content knowledge, but my work indicates that the validity of these interpretations may be threatened when students complete the assessments in low-stakes contexts that are more likely to elicit low test-taking effort. Score validity is also threatened in high-stakes out-of-class contexts in which students may be incentivized to leverage external resources to increase their score. My findings suggest that departments and instructors using programmatic and concept assessments to evaluate the progress of their curriculum and courses in meeting the goals of reform effort should carefully interpret scores in light of the conditions in which students completed the assessment. The impacts of reform efforts may also be detected in the types of skills and content that are assessed on course exams. I studied the skills and content of lower-division undergraduate biology exams in the context of a three-dimensional framework consisting of scientific practices,

interdisciplinary crosscutting concepts, and disciplinary core ideas. I found that very few exam items were three-dimensional, primarily due to the low number of items assessing scientific practices. Although there were few three-dimensional items, those items were more likely to use a constructed-response format and assess higher-order cognitive skills compared to items not aligned with all three dimensions. To achieve the goals of reform efforts in undergraduate biology education, my research indicates instructors may need time, resources, and training for writing and grading three-dimensional assessments. Altogether, this dissertation sheds critical insight into the process and content of evaluating student learning, thereby refining our understanding of the impact of education reforms.

DEDICATION

This dissertation is dedicated to my parents, Denise and Alan Uminski, who sparked my interest in science. Their support and encouragement have reignited that spark countless times across my life.

AUTHOR'S ACKNOWLEDGEMENTS

Writing a dissertation has given me a lot of time to reflect upon the knowledge and skills that I have learned and the many people who have played an important role in that learning process. I am incredibly thankful for my teachers, mentors, friends, and family who have shaped the way I think about the world, and I am happy to have the opportunity to share this work with the people who helped me along the way.

This dissertation is a testament to the unwavering support from my advisor, Dr. Brian Couch. Brian challenged me to think big and aim high, and his confidence in me inspired me to achieve more than I thought was possible. Brian invested thousands of hours in supporting me throughout graduate school, and his time and dedication set me up for success as a discipline-based education researcher. I look forward to continuing to share my successes with him in the future.

I thank my committee members, Drs. Chad Brassil, Jenny Dauer, Anna Hiatt, and Alena Moon for their contributions to this dissertation and for being role models on campus. Thank you to Chad for illustrating how data can drive important discussions about equity in undergraduate education, Jenny for highlighting the necessity of theory in education research, Anna for being an advocate for useful research that extends its reach beyond the ivory tower, and Alena for championing scientific practices and making me feel like part of the discipline-based education research (DBER) community on campus.

I have tremendous gratitude for Dr. Keenan Noyes. I feel incredibly fortunate to have gotten to know Keenan as a friend and as a scholar over the last year in the Couch Lab. Keenan has demonstrated unparalleled kindness and has set the bar for what I aspire to be as a researcher and as a mentor. I am thankful for the many insightful and inspiring conversations about three-dimensional learning, and I appreciate his generosity in sharing his time, knowledge, experiences, and cat photos. DBER is a better field with Keenan in it.

Enormous thanks to Dr. Sarah Kennedy, who has been a close friend and confidant for over a decade. I've benefited from her wisdom (and good taste in media) more than she can imagine. Since our time working together at the Etownian, Sarah has encouraged me to pursue opportunities to learn and grow in a wide range of content areas and disciplines, and it has been exciting to see where that knowledge and those skills have become useful later in life. I'm grateful to Sarah for reminding me to pursue all my endeavors with Big Suit energy. My life is richer because of our conversations.

I extend a heartfelt thanks to the friends I've made in graduate school who have been sources of inspiration and encouragement. Bailey McNichol has been a powerhouse of productivity that I can only aspire to replicate, and her involvement within the Lincoln community is commendable. Brandi Pessman frequently reminded me the importance of celebrating little accomplishments and taught me to appreciate spiders. I think of her fondly when I see an eight-legged creature scuttle past. Dr. Laura Segura Hernández's genuine love for natural history is contagious. It was hard not to smile when hearing her talk about pseudoscorpions. Ashley Foltz helped me to remember that there's more to graduate school than just graduate school. Stephanie Berg made me feel welcomed and valued in DBER, and I am in awe of her efforts as an advocate and her knack for bringing a group together. Taylor Rosso's enthusiasm for outreach is inspiring to both current and future scientists. Nicole Fiore helped me to keep things in perspective and prioritize what is really important. I have full confidence that these amazing women will continue to inspire the people around them, will break boundaries in science, and will change the world for the better.

I owe thanks to Dr. Mike Herman, whose leadership in the School of Biological Sciences helped to make my work in DBER feel very welcome within this discipline. Mike's leadership extended beyond campus, and I thank him for the many times he supported science communication by attending the Nerd Nite events I organized in Lincoln.

I am grateful to Drs. A. Kelly Lane and Ariel Marcy, whose mentorship in the Couch Lab and advice set me on the right path early in my graduate career.

I acknowledge the other members of the Couch Lab that I overlapped with during my time in Lincoln: Allison Upchurch, Dana Kirkwood-Watts, and Mojtaba Khajeloo.

I thank Jeff Ackley, who helped frame my thinking about assessments and shaped the path of my career in science assessments.

I also thank Dr. David Bowne, who fostered my early interest in biology education research at Elizabethtown College and has continued to inspire me to be a better biology educator.

I want to recognize my grandparents, Monica and Edward Kalinski, who have helped celebrate every academic milestone in my career. There's nothing quite like the warm feeling of having your grandparents say they are proud of you.

Throughout graduate school, Michelle and Kim kept me up to date with photos from home and pictures of the pets. Even though I was a thousand miles away, thank you for making me feel a bit closer to New Jersey.

Words can't quite adequately express the gratitude that I have for my parents, Denise and Alan Uminski, who have provided me with a lifetime of support and encouragement for my scientific interests and my academic pursuits. This support has ranged from, but is certainly not limited to, science summer camps and after school programs, camping at Round Valley, trips to the Rutgers Geology Museum, summer college courses, excursions to find fossils in Big Brook, and allowing me to stay up late during middle school to watch the newest episodes of MythBusters. My mom instilled in me the value of education and my dad inspired me with his wealth of information about the natural world which he was able to communicate with ease and enthusiasm. My parents set the foundation of my interest in science and provided me with the tools and resources to build upon it over a lifetime. This dissertation is for them.

As I wrap up this degree, it is easy to look back and see how many fantastic teachers whose effort and dedication to science education have molded me into the scientist I am today. From kindergarten through college, I am grateful for the teachers who inspired curiosity, encouraged inquiry, and made science fun, and I'd like to thank and acknowledge Ms. Kerry, Ms. Bosco, Ms. Columbus, Mrs. Cole, Ms. DeSesso, Mrs. Sherman, Ms. Murray, Mrs. Dowe, Mrs. Kenney, Mr. Kowalski, Mr. Moncelsi, Mr. Pirog, Dr. Bowne, Dr. Wohl, Dr. Bridge, Professor Martin, and Dr. Murray.

Lastly, I am thankful for my cats, Delilah and Travolta, who were my research companions during graduate school and whose antics made me smile even in the most stressful parts of this degree. They can't read this acknowledgement, but they know how much I appreciate them because I tell them so every day.

TABLE OF CONTENTS

LIST OF MULTIMEDIA OBJECTS

INTRODUCTION

The landscape of undergraduate biology education has been shaped by decades of reform efforts calling for instruction to integrate core concepts and scientific process skills as a means to gain disciplinary proficiency (American Association for the Advancement of Science [AAAS], 1989, 1993, 2011; National Academies of Sciences, Engineering, and Medicine [NASEM], 2016b, 2021, 2022; National Research Council [NRC], 1996, 2003b, 2012a). These calls for reform have identified core concepts within biology (NRC, 1996, 2012a; AAAS, 2011) and have focused on different aspects the scientific process, including inquiry (NRC, 2000), competencies (AAAS, 2011), and scientific practices (NRC, 2012a). While the terminology and focus of these calls for reform may differ slightly, a central theme across the calls is the common goal of getting students in science courses to be actively involved in deeply understanding and doing science. This goal has largely been informed by the anticipated demands and needs of the future workforce (NASEM, 2016b; National Center on Education and the Economy, 2008; NRC, 2007; Olson & Riordan, 2012). Students entering both science and non-science careers need to be prepared for a data-driven world where information—and misinformation—is increasingly accessible. Thus, these calls and reform efforts emphasize that science education needs to move away from rote memorization of facts and towards providing students knowledge and skills they can use to critically analyze and evaluate the vast amount of information they will encounter in many facets of their lives.

Many of the calls to incorporate scientific skills into science education stem from the K-12 education system where they were enacted as national-level standards (NRC, 1996; NGSS Lead States, 2013). Currently, 44 of the of U.S. States have adopted a common vision for K-12 science education by using the Next Generation Science Standards or its adaptations (NASEM, 2021; NGSS Lead States, 2013), and the adoption of these standards into schools and classrooms was facilitated by accountability policies and federal intervention programs (Hardy & Campbell, 2020). Standardized science assessments designed to measure students' conceptual understanding and application of scientific skills provide data about the progress of these K-12 reform efforts at both state and nationwide levels (e.g., California Department of Education, 2023; Maryland State Board of Education, 2022; U.S. Department of Education et al., 2019). Our understanding of the progress of these reform efforts at the undergraduate level is less clear as there are few analogous policies, programs, and assessments within undergraduate science (NASEM, 2016a); thus, it is difficult to determine how these calls have permeated into college courses and this is a challenge that warrants additional research.

To monitor instructional transformation in undergraduate biology, we can rely on the information provided by assessments. Assessments, broadly defined in the context of biology education, are tools for collecting information that can be used to make inferences about student understanding of biology concepts. These inferences are often made under the assumption that assessment scores accurately reflect student knowledge, so it is important to consider test-taking behaviors to determine whether scores represent valid depictions of student understanding. Valid interpretations of assessment scores are crucial for accurately determining the impact of reform efforts. In addition to providing information about student knowledge, assessments also provide insight into what instructors value in their courses as the content and skills that are assessed reflect instructors' prioritized learning outcomes (NRC, 2003a; Scouller, 1998). As such, assessments can frame the picture of how the goals of reform efforts have been translated into practice in undergraduate biology classrooms. Accordingly, in this dissertation I investigate biology assessments with an eye towards score validity and to characterize the content that is being assessed on undergraduate biology course exams.

I focus my research on three types of assessments that are commonly used in undergraduate biology: programmatic assessments, concept assessments, and summative assessments in the form of tests or exams. These three types of assessments provide snapshots of undergraduate biology ranging from the wide scope of an entire department to the detailed portrait of a single course. Programmatic assessments, such as the suite of Biology Measuring Achievement and Progression in Science (Bio-MAPS) diagnostic assessments (Couch et al., 2019; Couch, Wood, et al., 2015; Semsar et al., 2019; Summers et al., 2018), are tools to measure student understanding of foundational core concepts across a degree program. Given that programmatic assessments contain content that spans a four-year biology degree, these assessments are often administered under low stakes conditions where students are given participation credit and are not graded based on the correctness of their responses. Concept assessments, such as the Introductory Molecular and Cell Biology Concept Assessment (Shi et al., 2010), are similar to programmatic assessments in that they measure student conceptual knowledge, but concept assessments are more often used in individual courses or units rather than across an entire department. As concept assessments may be more closely aligned with course learning objectives, biology instructors may use a broader range of administration conditions in terms of how they assign credit. Compared to programmatic and concept assessments, which are intentionally designed to assess concepts from frameworks such as *Vision and Change* (AAAS, 2011; Branchaw et al., 2020), there is much more variability in the content assessed on course exams. Undergraduate instructors have a high degree of autonomy when designing their exams (Couch et al., 2023), and the design of these exams can signal the prioritized learning outcomes in instructors' courses (Wiggins & McTighe, 2005). Thus, the content and skills included on course exams can be used as a way to determine course curriculum alignment to reform efforts and their associated frameworks.

The following sections of this introduction briefly introduce the rationale, research questions, and main findings of the four studies included in this dissertation. Chapters 1 and 2 focus on programmatic and concept assessments, respectively. These chapters provide evidence of score validity for programmatic and concept assessments. Specifically, I examined how administration conditions can affect student engagement on assessments in ways that shape score validity interpretations. These chapters highlight that although biology content knowledge is what is being tested, biology content knowledge is not always what is being measured. I provide recommendations to instructors and departments on how to administer assessment instruments and how to appropriately interpret assessment scores in light of student test-taking behaviors. Chapters 3 and 4 of this dissertation characterize exams from undergraduate biology courses. These chapters investigate what content and skills are being assessed in biology courses using the lens of a three-dimensional framework. These chapters illustrate how instructors may be better supported in aligning their assessments with the goals of national calls and reform efforts in science education. Altogether, this dissertation provides a broad-scoping answer to the question: What are we assessing in undergraduate biology?

Chapter 1: GenBio-MAPS as a Case Study to Understand and Address the Effects of Test-Taking Motivation in Low-Stakes Program Assessments

Vision and Change (AAAS, 2011) represents a landmark report calling for the reform of undergraduate biology education. The advent of this report created the need for tools that biology departments can use to self-assess their progress in meeting curricular reform goals (Branchaw et al., 2020; Smith et al., 2019). Thus, discipline-based education researchers created a suite of programmatic assessments designed to measure biology students' understanding of *Vision and Change* core concepts across a major. General Biology – Measuring Achievement and Progression in Science (GenBio-MAPS) is one such programmatic assessment (Couch et al., 2019).

GenBio-MAPS is intended to be administered as a low-stakes assessment (i.e., students receive participation credit for submitting the assessment). While low-stakes assessments have benefits in that they provide flexible testing locations and might minimize testing anxiety, the low stakes also have the potential to elicit low test-taking effort from students in ways that threaten test score validity (Wise & DeMars, 2005; Wise & Kong, 2005). Low test-taking effort is often reflected in short test completion times, rapid selection of responses to test items, or self-reports of low effort, and these low-effort behaviors may yield scores that underestimate student understanding. Such underestimations of student understanding may misinform department-level decisions about teaching and curriculum that can have consequences for student learning outcomes. Previous research on test-taking effort on low-stakes assessments had only been conducted on general education assessments (Cole et al., 2008; Hoyt, 2001; Sundre &

Wise, 2003; Swerdzewski et al., 2011; Thelk et al., 2009), but I anticipated that a biology-specific assessment may yield different test-taking behaviors from biology students. This study addressed five research questions to explore test-taking motivation in a disciplinary context:

1) How are students engaging with the GenBio-MAPS instrument?
2) Does self-reported effort align with observed test-taking behaviors?
3) How do different aspects of test-taking effort relate to GenBio-MAPS score?
4) To what extent do students demonstrate test-taking persistence?
5) How might departments filter student responses to reduce the influence of low test-taking effort?

I found that most students were using effortful behavior when completing GenBio-MAPS, but there was a small proportion of students who exhibited evidence of low test-taking effort in their short test-completion time, rapid selection of responses, and/or self-reports of low test-taking effort. Students with these low effort behaviors tended to have lower GenBio-MAPS scores, which are likely unrepresentative of their actual biology content knowledge. I identified a set of criteria and cutoffs to filter out the scores of students with low test-taking effort and proposed a motivation filtering protocol to yield datasets that better represents student understanding of biology core concepts.

Chapter 2: How Administration Stakes and Settings Affect Student Behavior and Performance on a Biology Concept Assessment

Concept assessments in biology are validated assessment instruments developed by discipline-based education researchers that instructors can deploy to diagnose student understanding of foundational biological concepts (Knight, 2010). As instructors often use student scores on concept assessments to inform their instructional choices, it is important that the scores provide a valid portrayal of student understanding. Scores may not be valid if students exhibit low test-taking effort (Wise & DeMars, 2005) or if students consult external resources when completing the concept assessment (Munoz & Mackay, 2019). Certain concept assessment administration conditions may make these test-taking behaviors more likely to occur, but there had not yet been an empirical comparison across the range of administration stakes and settings. In this study, I analyzed data from lower-stakes testing conditions (i.e., participation credit) and higher-stakes conditions (i.e., grading based on correctness of responses) for in-class and out-of-class settings. I used concept assessment score, completion time, and the correlation of concept assessment scores with previous course exams as indicators of underlying test-taking behaviors. The research question for this study was:

1. How do administration stakes and settings affect student test-taking behavior and performance and influence interpretation of student scores on a biology concept assessment?

Student performance on a biology concept assessment was similar across lower-stakes in-class, lower-stakes out-of-class, and higher-stakes in-class settings, suggesting a degree of equivalence between these administration conditions. Students spent more time, had higher scores, and had the lowest correlation with the previous test performance when they completed concept assessments in higher-stakes out-of-class conditions. This finding suggests that instructors should carefully interpret the scores from higher-stakes out-of-class conditions as the scores may be more of a reflection of accessing external resources and may not accurately reflect student understanding of biology concepts.

Chapter 3: Testing Scientific Practices: A Nationwide Analysis of Undergraduate Biology Exams

National calls have emphasized that incorporating scientific practices into undergraduate science education is key for addressing the needs of increasingly interdisciplinary science fields and to solve emerging global challenges (NASEM, 2021, 2022; NRC, 2007). The importance of the scientific practices is underscored by their inclusion as one of the dimensions in a three-dimensional framework for science education (NRC, 2012a). Yet, despite the importance of the scientific practices, previous work suggests that most undergraduate biology students are likely not encountering these practices in their course assessments, which mainly test memorized facts aligned to the lower-order cognitive skills on Bloom's Taxonomy (Momsen et al., 2010, 2013). To better understand the current state of scientific practices in undergraduate biology, I conducted a nationwide study of lower-division biology courses and analyzed how each instructor's exam questions aligned to the three-dimensional framework, with specific attention towards scientific practices. This research cast light on what content is being assessed in undergraduate biology courses and how instructors incorporate scientific practices into their assessments with regards to higher-order cognitive skills. The research questions for this study were:

1. To what extent do exams align to the three-dimensional framework with particular reference to the scientific practices?
2. What is the relationship between an exam's alignment to the three-dimensional framework and to Bloom's Taxonomy of cognitive skills?

Overall, I found that very few exams in a nationwide sample of undergraduate biology courses aligned to the three-dimensional framework, which was largely driven by a very small number of items meeting the criteria for scientific practices. The exams that incorporated a greater number of scientific practices tended to assess higher-order cognitive skills on Bloom's Taxonomy.

Chapter 4: Identifying Factors Associated with Instructor Implementation of Three-Dimensional Assessments in Undergraduate Biology Courses

The three-dimensional framework for science education suggests that students develop deep understanding of science when their learning integrates scientific practices with foundational disciplinary core ideas and interdisciplinary crosscutting concepts (NRC, 2012a). In Chapter 3, I found that the large majority of undergraduate biology exams did not assess the scientific practices dimension of this framework, and as such, were not three-dimensionally aligned. Previous work at a single institution had similar results and found that many assessments in introductory undergraduate science courses do not align to the three-dimensional framework, particularly when the courses were taught prior to reform efforts at the institution (Matz et al., 2018). Given the low use of three-dimensional assessments, Matz and colleagues (2018) raised a question about what factors in undergraduate education might be barriers to three-dimensional assessment. My work in this chapter builds off my previous findings in Chapter 3 and sought to answer the question posed by Matz et al. (2018). Drawing upon the conceptual model of coherence as a lens for this study, I used a generalized linear mixed model to identify factors across the levels of the undergraduate education system that may be helping or hindering biology instructors in using three-dimensional assessments. This work aimed to address the overarching research question:

1. What constraints and challenges are undergraduate biology instructors facing in implementing three-dimensional assessments in their courses and where may they need additional support?

My work here suggested that instructors may face constraints and challenges associated with the time needed to develop and grade three-dimensional assessments, as three-dimensional items were more likely to use a constructed-response format. I also identified that existing professional development opportunities and training may not have necessarily yielded measurable benefits to three-dimensional alignment, and this may be an area where instructors could use additional support. My work suggests that institutions and departments can support their instructors by providing the time, resources, and appropriate training needed to implement three-dimensional assessments in undergraduate biology courses.

In summary, assessments play a key role in shaping the future of undergraduate biology education, as the context and content of assessments signals the prioritized learning outcomes in courses and in departments. Across these four chapters, I aimed to provide actionable recommendations that instructors and departments can use to carefully consider how and what they are assessing in undergraduate biology. These chapters illuminate paths for using assessment tools to make data-driven decisions about curriculum and instruction and incorporating scientific practices as a means of aligning the content of assessments with national calls.

REFERENCES FOR THE INTRODUCTION

American Association for the Advancement of Science. (1989). Science for All Americans: A Project 2061 Report on Literacy Goals in Science, Mathematics, and Technology. American Association for the Advancement of Science.

American Association for the Advancement of Science. (1993). Benchmarks for Science Literacy. Oxford University Press.

American Association for the Advancement of Science. (2011). Vision and Change in Undergraduate Biology Education: A Call to Action. AAAS. https://live-visionandchange.pantheonsite.io/wp-content/uploads/2011/03/Revised-Vision-and-Change-Final-Report.pdf

Branchaw, J. L., Pape-Lindstrom, P. A., Tanner, K. D., Bissonnette, S. A., Cary, T. L., Couch, B. A., Crowe, A. J., Knight, J. K., Semsar, K., Smith, J. I., Smith, M. K., Summers, M. M., Wienhold, C. J., Wright, C. D., & Brownell, S. E. (2020). Resources for teaching and assessing the Vision and Change biology core concepts. CBE—Life Sciences Education, 19(2), es1. https://doi.org/10.1187/cbe.19-11-0243

California Department of Education. (2023). 2021–22 California Science Test Results at a Glance—CAASPP Reporting (CA Dept of Education). California Assessment of Student Performance and Progress: California Science Test (CAST) Test Results at a Glance. https://caaspp-elpac.ets.org/caaspp/DashViewReportCAST?ps=true&lstTestYear=2022&lstTestType=X&lstGroup=1&lstSubGroup=1&lstSchoolType=A&lstGrade=13&lstCounty=00&lstDistrict=00000&lstSchool=0000000

Cole, J. S., Bergin, D. A., & Whittaker, T. A. (2008). Predicting student achievement for low stakes tests with effort and task value. Contemporary Educational Psychology, 33(4), 609–624. https://doi.org/10.1016/j.cedpsych.2007.10.002

Couch, B. A., Prevost, L. B., Stains, M., Whitt, B., Marcy, A. E., Apkarian, N., Dancy, M. H., Henderson, C., Johnson, E., Raker, J. R., Yik, B. J., Earl, B., Shadle, S. E., Skvoretz, J., & Ziker, J. P. (2023). Examining whether and how instructional coordination occurs within introductory undergraduate STEM courses. Frontiers in Education, 8. https://www.frontiersin.org/articles/10.3389/feduc.2023.1156781

Couch, B. A., Wood, W. B., & Knight, J. K. (2015). The Molecular Biology Capstone Assessment: A concept assessment for upper-division molecular biology students. CBE—Life Sciences Education, 14(1), ar10. https://doi.org/10.1187/cbe.14-04-0071

Couch, B. A., Wright, C. D., Freeman, S., Knight, J. K., Semsar, K., Smith, M. K., Summers, M. M., Zheng, Y., Crowe, A. J., & Brownell, S. E. (2019). GenBio-MAPS: A programmatic assessment to measure student understanding of Vision and Change core concepts across general biology programs. CBE—Life Sciences Education, 18(1), ar1. https://doi.org/10.1187/cbe.18-07-0117

Hardy, I., & Campbell, T. (2020). Developing and supporting the Next Generation Science Standards: The role of policy entrepreneurs. Science Education, 104(3), 479–499. https://doi.org/10.1002/sce.21566

Hoyt, J. E. (2001). Performance funding in higher education: The effects of student motivation on the use of outcomes tests to measure institutional effectiveness. Research in Higher Education, 42(1), 71–85. https://doi.org/10.1023/A:1018716627932

Knight, J. (2010). Biology concept assessment tools: Design and use. Microbiology Australia, 31(1), 5–8. https://doi.org/10.1071/ma10005

Maryland State Board of Education. (2022). Standard Setting Progress, State Assessment Overview & Update for Spring 2022 Results. Maryland State Department of Education. https://go.boarddocs.com/md/msde/Board.nsf/files/CHKK5B4B80F3/$file/StandardSettingProgressOverviewStateAssessmentsTimelineUpdateSpring2022ResultsV2.pdf

Matz, R. L., Fata-Hartley, C. L., Posey, L. A., Laverty, J. T., Underwood, S. M., Carmel, J. H., Herrington, D. G., Stowe, R. L., Caballero, M. D., Ebert-May, D., & Cooper, M. M. (2018). Evaluating the extent of a large-scale transformation in gateway science courses. Science Advances, 4(10), eaau0554. https://doi.org/10.1126/sciadv.aau0554

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. https://doi.org/10.1187/cbe.10-01-0001

Momsen, J. L., 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. https://doi.org/10.1187/cbe.12-08-0130

Munoz, A., & Mackay, J. (2019). An online testing design choice typology towards cheating threat minimisation. Journal of University Teaching & Learning Practice, 16(3). https://doi.org/10.53761/1.16.3.5

National Academies of Sciences, Engineering, and Medicine. (2016a). Barriers and Opportunities for 2-Year and 4-Year STEM Degrees: Systemic Change to Support Students' Diverse Pathways (S. Malcom & M. Feder, Eds.; p. 21739). National Academies Press. https://doi.org/10.17226/21739

National Academies of Sciences, Engineering, and Medicine. (2016b). Developing a National STEM Workforce Strategy: A Workshop Summary. The National Academies Press. https://doi.org/10.17226/21900

National Academies of Sciences, Engineering, and Medicine. (2021). Call to Action for Science Education: Building Opportunity for the Future. National Academies Press. https://doi.org/10.17226/26152

National Academies of Sciences, Engineering, and Medicine. (2022). Imagining the Future of Undergraduate STEM Education: Proceedings of a Virtual Symposium (K. Brenner, A. Beatty, & J. Alper, Eds.). National Academies Press. https://doi.org/10.17226/26314

National Center on Education and the Economy. (2008). Tough Choices or Tough Times: The Report of the New Commission on the Skills of the American Workforce (Revised and Expanded edition). Jossey-Bass.

National Research Council. (1996). National Science Education Standards. National Academies Press. https://doi.org/10.17226/4962

National Research Council. (2000). Inquiry and the National Science Education Standards: A Guide for Teaching and Learning (S. Olson & S. Loucks-Horsley, Eds.). National Academies Press. https://doi.org/10.17226/9596

National Research Council. (2003a). Assessment in Support of Instruction and Learning: Bridging the Gap Between Large-Scale and Classroom Assessment - Workshop Report. National Academies Press. https://doi.org/10.17226/10802

National Research Council. (2003b). BIO2010: Transforming Undergraduate Education for Future Research Biologists. National Academies Press. https://doi.org/10.17226/10497

National Research Council. (2007). Rising Above the Gathering Storm: Energizing and Employing America for a Brighter Economic Future. National Academies Press. https://doi.org/10.17226/11463

National Research Council. (2012). A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas. National Academies Press. https://doi.org/10.17226/13165

NGSS Lead States. (2013). Next Generation Science Standards: For States, By States. The National Academies Press.

Olson, S., & Riordan, D. G. (2012). Engage to Excel: Producing One Million Additional College Graduates with Degrees in Science, Technology, Engineering, and Mathematics. Report to the President. In Executive Office of the President. Executive Office of the President. https://eric.ed.gov/?id=ED541511

Scouller, K. (1998). The influence of assessment method on students’ learning approaches: Multiple choice question examination versus assignment essay. Higher Education, 35(4), 453–472. https://doi.org/10.1023/A:1003196224280

Semsar, K., Brownell, S., Couch, B. A., Crowe, A. J., Smith, M. K., Summers, M. M., Wright, C. D., & Knight, J. K. (2019). Phys-MAPS: A programmatic physiology assessment for introductory and advanced undergraduates. Advances in Physiology Education, 43(1), 15–27. https://doi.org/10.1152/advan.00128.2018

Shi, J., Wood, W. B., Martin, J. M., Guild, N. A., Vicens, Q., & Knight, J. K. (2010). A diagnostic assessment for introductory molecular and cell biology. CBE — Life Sciences Education, 9(4), 453–461. https://doi.org/10.1187/cbe.10-04-0055

Smith, M. K., Brownell, S. E., Crowe, A. J., Holmes, N. G., Knight, J. K., Semsar, K., Summers, M. M., Walsh, C., Wright, C. D., & Couch, B. A. (2019). Tools for change: Measuring student conceptual understanding across undergraduate biology programs using Bio-MAPS assessments. Journal of Microbiology & Biology Education, 20(2). https://doi.org/10.1128/jmbe.v20i2.1787

Summers, M. M., Couch, B. A., Knight, J. K., Brownell, S. E., Crowe, A. J., Semsar, K., Wright, C. D., & Smith, M. K. (2018). EcoEvo-MAPS: An ecology and evolution assessment for introductory through advanced undergraduates. CBE—Life Sciences Education, 17(2), ar18. https://doi.org/10.1187/cbe.17-02-0037

Sundre, D. L., & Wise, S. L. (2003). Motivation filtering: An exploration of the impact of low examinee motivation on the psychometric quality of tests. National Council on Measurement in Education, Chicago.

Swerdzewski, P. J., Harmes, J. C., & Finney, S. J. (2011). Two approaches for identifying low-motivated students in a low-stakes assessment context. Applied Measurement in Education, 24(2), 162–188. https://doi.org/10.1080/08957347.2011.555217

Thelk, A. D., Sundre, D. L., Horst, S. J., & Finney, S. J. (2009). Motivation matters: Using the Student Opinion Scale to make valid inferences about student performance. The Journal of General Education, 58(3), 129–151. JSTOR.

U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, & National Assessment of Educational Progress. (2019). NAEP Report Card: 2019 NAEP Science Assessment. U.S. Department of Education. https://www.nationsreportcard.gov/science/supporting_files/2019_infographic_science.pdf

Wiggins, G. P., & McTighe, J. (2005). Understanding by Design. Association for Supervision and Curriculum Development.

Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. Educational Assessment, 10(1), 1–17. https://doi.org/10.1207/s15326977ea1001_1

Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. Applied Measurement in Education, 18(2), 163–183. https://doi.org/10.1207/s15324818ame1802_2

CHAPTER 1: GENBIO-MAPS AS A CASE STUDY TO UNDERSTAND AND ADDRESS THE EFFECTS OF TEST-TAKING MOTIVATION IN LOW-STAKES PROGRAM ASSESSMENTS[1]

ABSTRACT

The General Biology–Measuring Achievement and Progression in Science (GenBio-MAPS) assessment measures student understanding of the *Vision and Change* core concepts at the beginning, middle, and end of undergraduate biology degree programs. Assessment coordinators typically administer this instrument as a low-stakes assignment for which students receive participation credit. While these conditions can elicit high participation rates, it remains unclear how to best measure and account for potential variation in the amount of effort students give to the assessment. To better understand student test-taking motivation, we analyzed GenBio-MAPS data from more than 8000 students at 20 institutions. While the majority of students give acceptable effort, some students exhibited behaviors associated with low motivation, such as low self-reported effort, short test completion time, and high levels of rapid-selection behavior on test questions. Standard least-squares regression models revealed that students' self-reported effort predicts their observable time-based behaviors and that these motivation indices predict students' GenBio-MAPS scores. Furthermore, we observed that test-taking behaviors and performance change as students progress through the assessment. We provide recommendations for identifying and filtering out data from

[1] This research was first published with minor formatting differences as Uminski C., & Couch B. A. (2021). CBE—Life Sciences Education, 20(2), ar20. https://www.lifescied.org/doi/10.1187/cbe.20-10-0243

students with low test-taking motivation so that the filtered data set better represents student understanding.

INTRODUCTION

Biology departments use program assessments to measure students' understanding of biology topics as they progress through an undergraduate degree program. General Biology–Measuring Achievement and Progression in Science (GenBio-MAPS) is one such assessment that focuses on student understanding of the *Vision and Change* core concepts (American Association for the Advancement of Science [AAAS], 2011; Couch et al., 2019). GenBio-MAPS is part of the suite of Bio-MAPS program assessments that are designed to measure conceptual understanding of biology topics at key time points in a degree program (Smith et al., 2019). Specifically, GenBio-MAPS is administered at the beginning of the first introductory course, after completion of introductory courses, and in advanced courses before graduation. Biology departments can use the data gathered from GenBio-MAPS across these time points to monitor student learning gains, identify areas of curricular proficiency or deficiency, measure the impact of curricular changes, and understand student performance based on demographic characteristics (Couch et al., 2019). Biology departments may also use GenBio-MAPS data to satisfy departmental requirements for institutional reporting and accreditation.

GenBio-MAPS is administered to undergraduate students outside class time as an online survey. The online out-of-class format does not take time from class instruction and allows the instrument to be administered and scored consistently and efficiently across different courses and institutions. While the online out-of-class administration may be convenient for test administrators, this format necessitates low-stakes testing conditions in which students are not graded based on test performance. If GenBio-MAPS had higher stakes, there might be greater incentive for students to access external resources, and maintaining test security to prevent academic dishonesty in the out-of-class context would be difficult for departments to achieve. Under low-stakes testing conditions, prior research on a similar instrument (Couch et al., 2015) found that student performance in the out-of-class context does not differ significantly from an in-class administration, suggesting that students engage with the assignment to roughly the same degree as they would for an in-class activity (Couch and Knight, 2015).

While this finding provides some indication regarding student effort, departments using data from low-stakes administrations of GenBio-MAPS should still consider the potential effects of test-taking motivation on assessment scores. Researchers have noted that, without academic consequences for test performance, students may be less inclined to give their best effort on low-stakes assessments (Wise and DeMars, 2005). Students with low test-taking effort may exhibit behaviors such as guessing, omitting items, and rapid selection of responses (Wise and Kong, 2005). These behaviors present a concern for departments, because they can introduce construct-irrelevant variance to assessment scores (Swerdzewski et al., 2011; American Educational Research Association et al., 2014). Construct-irrelevant variance refers to the extent to which test scores are affected by processes outside the target the test is intending to measure. When construct-irrelevant variance occurs due to low test-taking effort, students' scores may not represent their conceptual understanding but instead reflect their low motivation for the task (Wise and DeMars, 2010).

Researchers studying low-stakes assessments have developed methods of "motivation filtering" to address the construct-irrelevant variance associated with low test-taking motivation (Sundre and Wise, 2003; Wise and DeMars, 2005). Motivation filtering relies on the assumption that motivation is associated with test performance but not associated with ability (Wise et al., 2006b). When these assumptions are met, motivation filtering methods can be applied to identify the test responses from students exhibiting low motivation and remove these scores from the data set. The motivation filtering process is expected to decrease construct-irrelevant variance due to low motivation and improve the validity of the inferences that can be drawn from test scores (Wise and DeMars, 2005, 2010). Although Wise and colleagues (Wise and DeMars, 2005, 2010; Wise and Kong, 2005; Wise et al., 2006b) have been proponents of the use of motivation filtering, this practice is not widely reported in the literature on low-stakes assessments and has not been studied in the context of a biology program assessment.

Test-taking motivation can influence test performance, so it is important to understand how students are engaging with diagnostic assessments under low-stakes conditions. Given its use in undergraduate biology programs, we use GenBio-MAPS as a case study to compare different metrics for test-taking motivation, including student self-reported survey perceptions and time-based behaviors. This research will help to reveal the relationship between self-reported and behavioral measures of motivation and their effect on test performance. Understanding these relationships will inform how data from GenBio-MAPS and similar discipline-based low-stakes assessments can be filtered to account for the influence of low test-taking motivation.

Theoretical Framework

The literature on motivation is vast, and the term "motivation" can have different meanings depending on context. For this research, "motivation" is defined as "the process whereby goal-directed activity is instigated and sustained" (Schunk et al., 2008, p. 4), and

Table 1.1: Behavioral indicators associated with test-taking motivation

Index of Motivation	Behavioral indicator of high test-taking motivation	Behavioral indicator of low test-taking motivation
Choice of tasks[a]	• Voluntary completion of test instrument under low-stakes conditions	• Student does not choose to complete test
Effort	• High self-reported effort • Student takes an adequate amount of time to read and contemplate each test question before responding (e.g., solution behavior) • Adequate test completion time	• Low self-reported effort • Student responds in less than the amount of time needed to read and contemplate the test questions (e.g., rapid-selection behavior) • Short test completion time
Persistence	• Consistent use of solution behavior throughout the test • Consistent amount of time spent on each test question as the test progresses	• Increase in rapid-selection behaviors as the test progresses • Decrease in the amount time spent on each test question as the test progresses
Achievement	• High score on test that reflects student ability	• Low score in relation to student ability

[a]Choice of tasks was not considered in this study, since we did not have any information from the students who chose not to complete the survey.

we refer to motivation specifically in the context of low-stakes testing. In this work, we studied motivation by examining students' test-taking behaviors related to the intended goal of students performing to the best of their abilities on GenBio-MAPS. Motivation can be inferred when student behavior aligns with the four indexes of motivation: choice of tasks, effort, persistence, and achievement (Lepper et al., 1973; Zimmerman and Ringle, 1981; Salomon, 1984; Pintrich and Schrauben, 1992; Schunk, 1995). Specific test-taking behaviors align with each index of motivation (Table 1.1). Choice of tasks would be evidenced by students initiating the assessment, but we will not study this here, as we have no information from students who chose not to complete GenBio-MAPS. In the current study, we will focus on test-taking effort (inferred by the three behavioral indicators of self-reported effort, solution behavior, and test completion time), persistence behavior (determined by the amount of time spent on each question as the test progresses), and achievement (measured by GenBio-MAPS score). Each of these indexes of motivation will be discussed in more detail in the following paragraphs.

Effort can be measured through self-reported means, often using Likert-type survey instruments. In our study, we used the Student Opinion Scale (SOS; Sundre and Moore, 2002) to collect self-reported data on student test-taking effort. This instrument is easily administered following an assessment and previous research has shown that the SOS collects reliable data on undergraduate test-taking motivation in a variety of low-stakes contexts (Wise and Kong, 2005; Sundre, 2007; Thelk et al., 2009). While the SOS reveals aspects of student test-taking effort, there are noted limitations in the use and interpretation of this instrument. One such limitation is that self-reported data rely on the assumption that students accurately gauge and report their levels of motivation (Wise, 2006; Swerdzewski et al., 2011), and students' self-reported motivation may not correspond to their behaviors for several reasons. Students may consciously alter and increase their self-reported motivation if they feel pressure to give socially acceptable answers (Fisher and Katz, 2000). Attribution bias may unconsciously influence self-reported motivation, because students who believe that they did not do well on a test may ascribe their poor test performance to a lack of effort over a lack of ability (Schunk et al., 2008; Duckworth et al., 2011). Other limitations present themselves in the methods in which the SOS instrument is administered to examinees. Collecting self-reported data at the end of an assessment does not allow for a more nuanced understanding of changes that occur as the test progresses (Wise and Kong, 2005). As a result of these limitations, we cannot rely on self-reported data alone to gauge the various dimensions of students' test-taking effort.

Effort can also be inferred based on timing data from students as they progress through a test, and these data are readily collected by computer-based testing platforms. The amount of time spent per question can be processed to determine the proportion of questions on which students exceed a minimal threshold time (i.e., solution behavior) or to quantify the amount of time students spend on the entire test (i.e., test completion time). We refer to solution behavior and test completion time as observable test-taking behaviors. Even though solution behavior and test completion time are strongly correlated, the two measures are distinct and provide different insights into student effort (Wise and Kong, 2005). Solution behavior provides information about whether students exceed the minimum time deemed necessary to read and process each test question. Traditionally, the literature has equated solution behavior with the active seeking of the correct response to a question by reading carefully and fully considering the options (Schnipke and Scrams, 1997; Wise and Kong, 2005; Kong et al., 2007; Setzer et al., 2013). However, there are limitations in this interpretation, and we note that response times can be classified as solution behavior, even if the student is disengaged or distracted by unrelated activities (Lee and Jia, 2014). Thus, solution behavior is necessary for, but not necessarily indicative of, test-taking effort (Kong et al., 2007). Conversely, rapid-selection behavior refers to student responses that were submitted in a time shorter than necessary to read and process the question stem and options (Wise and Kong, 2005). The degree to which students use solution behavior is associated with test completion time: students who use more solution behavior are also expected to spend a longer time on an assessment. While solution behavior can be used to indicate the presence of effort when completing an assessment, test completion time provides a window into how much effort was expended, with longer test completion times generally associated with higher effort (Wise and Kong, 2005).

Persistence behaviors provide another perspective on student motivation. In the context of test-taking motivation, persistence involves sustained effort throughout the duration of the test. This can be detected using both self-reported and time-based data. The effort subscale of the SOS instrument addresses persistence in items 2 and 10 ("I engaged in good effort throughout this test"; "While taking this test, I was able to persist to completion of the task"; (Sundre and Moore, 2002; Sundre, 2007). Persistence can also be identified by analyzing question-by-question changes in the use of solution behavior across an assessment. This approach was used in previous research and indicated that solution behaviors tend to decrease (i.e., rapid-selection behaviors tend to increase) as students move through a test (Wise, 2006; Wise et al., 2009). These changes in effort as the test progresses signal low persistence and thus low test-taking motivation. In addition to changes in solution behavior, changes in the amount of time spent on each question can also reflect test-taking persistence.

We use GenBio-MAPS score as a measure of achievement. Achievement is an indirect index of motivation and is affected by the other three indices. The students who choose a specific task, put effort into the task, and consistently engage with the task over the appropriate time span are expected to achieve at higher levels (Pintrich and Schrauben, 1992; Schunk, 1995). In the context of low-stakes assessments, highly motivated students are more likely to achieve higher test scores than unmotivated students (Wise and DeMars, 2005). As a result, the scores of students with high test-taking motivation may be more likely to reflect their true abilities, while the scores of students with low test-taking motivation may underestimate what the students are capable of achieving.

Research Questions

Previous research on test-taking motivation has largely been conducted using low-stakes general education assessments (Schiel, 1996; Hoyt, 2001; Sundre and Wise, 2003; Wise and Kong, 2005; Wise et al., 2006b; Cole et al., 2008; Thelk et al., 2009; Wise and DeMars, 2010; Swerdzewski et al., 2011). GenBio-MAPS is a discipline-specific biology assessment that was administered to students enrolled in biology courses, and there remains a need to explore test-taking motivation in this disciplinary context. Thus, we will pursue several research questions related to student motivation when completing GenBio-MAPS: 1) How are students engaging with the GenBio-MAPS instrument? 2) Does self-reported effort align with observed test-taking behaviors? 3) How do different aspects of test-taking effort relate to GenBio-MAPS score? 4) To what extent do students demonstrate test-taking persistence? 5) How might departments filter student responses to reduce the influence of low-test taking effort? Answering these questions will help biology departments better interpret data from GenBio-MAPS and make informed decisions about their degree programs. This work will also provide guidance for addressing the effects of low test-taking motivation on diagnostic assessments more broadly, including for similar types of instruments and within other science, technology, engineering, and math (STEM) disciplines.

METHODS

GenBio-MAPS Administration

GenBio-MAPS consists of 39 question stems with four to five true-false (T/F) statements each for a total of 175 accompanying T/F statements that assess Vision and

Change core concepts (AAAS, 2011). Each student was administered a random subset of 15 question stems and their associated T/F statements. The order of the question stems and T/F statements within each question stem were randomized for each student. Full details regarding the development and administration of the GenBio-MAPS instrument can be found in Couch et al. (2019).

Our analyses used the final data set from the instrument development process (Couch et al., 2019). These cross-sectional data were collected during the 2016 calendar year from students in 152 biology courses at 20 institutions (Supplemental Table 1.1). Each student responded at only a single time point and thus is only represented once in this data set. Students completed GenBio-MAPS as part of normal course or program requirements and received course credit or extra credit for completing the instrument. Credit was determined by course instructors, and there was no additional benefit to students based on correctness of responses or the decision to release their responses for research purposes.

GenBio-MAPS was administered using the Qualtrics survey platform (Qualtrics, 2019). On the first page of the survey, students were introduced to the assessment, asked to answer the questions to the best of their abilities in one sitting, and urged to refrain from using outside resources (e.g., peers, websites). GenBio-MAPS was designed to take approximately 30 minutes to complete, but there was no time limit on the assessment. The Qualtrics platform unobtrusively collected data about the amount of time students spent on each multiple–true-false (MTF) question, which corresponds to one survey page.

The SOS (Sundre and Moore, 2002) was administered in the survey after students completed the GenBio-MAPS assessment. The SOS contains two subscales designed to measure the perceived importance of doing well on the test and the amount of effort the student expended on the test. Each subscale contains five questions. Both subscales were administered, but only data from the effort subscale were used for this research, because students were not expected to place a high degree of personal importance on the test. The SOS items use a Likert-type response system, where 1 = strongly disagree, 2 = disagree, 3 = neutral, 4 = agree, and 5 = strongly agree. The two items on the effort subscale that have negative stems (e.g., "I did not give this test my full attention while completing it") were reverse coded before scores were calculated (Sundre, 2007). We calculated the average score that students reported on the SOS, using a range from 1 to 5. Higher scores on the SOS represent a greater amount of effort on GenBio-MAPS.

Data Processing, Participation Rates, and Student Demographics

We applied initial and minimal data processing to remove responses that were incomplete, duplicated, or unusable. Note that, although we used the same data set as Couch et al. (2019), we targeted a broader range of students in our study and accordingly used less-restrictive data-processing procedures. We first removed submissions from individuals who did not reach the end of the survey, reported being under 18 years of age, did not consent to release survey data, or had already submitted complete survey data in the same course. We also excluded data from individuals who had responded to fewer than 60 T/F statements, a cutoff selected because it represents the minimum number of statements that students could encounter in an administration of 15 GenBio-MAPS question stems. Our final data set contained 8185 responses (Table 1.2). Roughly 3% of students who remained in the data set did not complete the SOS instrument; these students were only excluded from analyses that involved SOS scores. Response times for individual questions that exceeded 15 minutes represented 1% of the response times recorded, and the data for those pages were replaced with the average page time of 1.5 minutes (Supplemental Table 1.2).

Table 1.2: Student self-reported demographics

Student characteristic	n[a]	%
Course time point		
Beginning of introductory series	3935	48
End of introductory series	3118	38
Advanced	1132	14
Gender		
Female	5223	65
Male	2829	34
Non-binary[b]	55	<1
Race/ethnicity[c]		
Non-underserved	6209	79
Underserved	1700	21
Highest level of parental education		
Completed bachelor's degree	5006	63
Did not complete bachelor's degree	2967	37
Language		
English spoken at home growing up	6966	86
English not spoken at home growing up	1140	14
Major		
Declared or intent to declare a major in biology	5830	72
Non-biology major	2235	28

[a] Numbers do not add to full sample size because some students left the given item blank.
[b] Due to low numbers, responses in this group were excluded from analyses.
[c] Underserved racial/ethnic groups included students who self-identified as African American/Black, Filipino, Hispanic/Latinx, Native American/Alaska Native, Native Hawaiian, and Pacific Islander. This grouping is not intended to obscure the unique histories and identities of any group.

Identifying Solution Behavior and Persistence Behaviors

We set response time thresholds based on the number of characters in the text of each GenBio-MAPS MTF question, including spaces. The standardized directions in each question and text within figures, graphs, or tables were excluded from the character count. We calculated thresholds based on a rate of 100 characters per second (Supplemental Table 1.3), which approximates threshold rates used in prior studies (Wise and Kong, 2005; Kong et al., 2007). Response times above the threshold (i.e., solution behavior) were assigned a value of 1, and response times below the threshold (i.e., rapid-selection behavior) were assigned a value of 0. We used the methods established by Wise and Kong (2005) and calculated the sum of the values for solution behavior, then divided by the number of questions on the assessment. The resulting value represented the proportion of test questions for which the student used solution behavior. Consistent with previous studies (Wise and Kong, 2005; Kong et al., 2007), we did not consider the readability of the text (e.g., Flesch reading ease or Flesch-Kincaid level [Flesch, 1948; Kincaid et al., 1975]) when setting the response time thresholds. We determined persistence behaviors by examining changes to the proportion of students using solution behavior and the length of response times for each page in the survey.

Statistical Analyses

For certain analyses, we identified arbitrary effort cutoffs based on the judgment that students below these cutoffs could be reasonably considered to be giving insufficient effort, a criterion that provides the basis for the filtering or removal of students from the data set. For the SOS effort subscale, we selected 2.5 as the cutoff, as students below this mark fall in the range of disagreeing or strongly disagreeing with effort statements. We used a cutoff of 0.6 for solution behavior, and students below this mark were engaging in solution behavior on fewer than 60% of the questions (i.e., students were using rapid-selection behavior on at least 40% of questions). Finally, based on prior estimates of how long it takes to read quickly through the assessment, we used 10 minutes as a cutoff for test completion time. We use these cutoffs to distinguish between what we hereafter refer to as “motivated” and “unmotivated” students.

We calculated overall score as the proportion of T/F statements answered correctly. Each T/F statement response was scored as 1 = correct or 0 = incorrect, and overall score was calculated by summing the number of correct T/F statements for each student and dividing by the total number of statements. We used JMP (SAS Institute Inc.,

2019) to calculate Cronbach's alpha to determine the estimated reliability of the items on the SOS instrument and to estimate standard least squares linear regression models to understand how different variables explained student effort, persistence, and overall score. Predictor variables were tested based on whether they had previously shown significant effects in Couch et al. (2019) or were hypothesized to explain variance in the outcome variable. We included self-reported demographic variables as fixed effects and institution as a random effect in our models predicting effort and overall score. Reference groups were selected based on the group having the larger sample size. We included student and question as random effects in our models for test-taking persistence. A correlation matrix for variables is provided in Supplemental Table 1.4. Given the correlations between predictor variables, we applied a backward stepwise model-selection procedure to address potential issues with multicollinearity (Akaike, 1973). Starting with the highest p-values, nonsignificant variables were individually tested for retention in the model and were only retained if the new model had an Akaike information criterion (AIC) value more than two units greater than the prior model.

Institutional Review Board Approval

This research was approved by the University of Nebraska–Lincoln (protocol 14618).

RESULTS

How Are Students Engaging with the GenBio-MAPS Instrument?

We examined student engagement with GenBio-MAPS based on self-reported effort, solution behavior, and test completion time (Figure 1.1). The estimated reliability of the SOS effort subscale (using Cronbach's alpha) was 0.81. Most students (86%) reported a score on the effort subscale greater than or equal to 2.5. The mean score on the

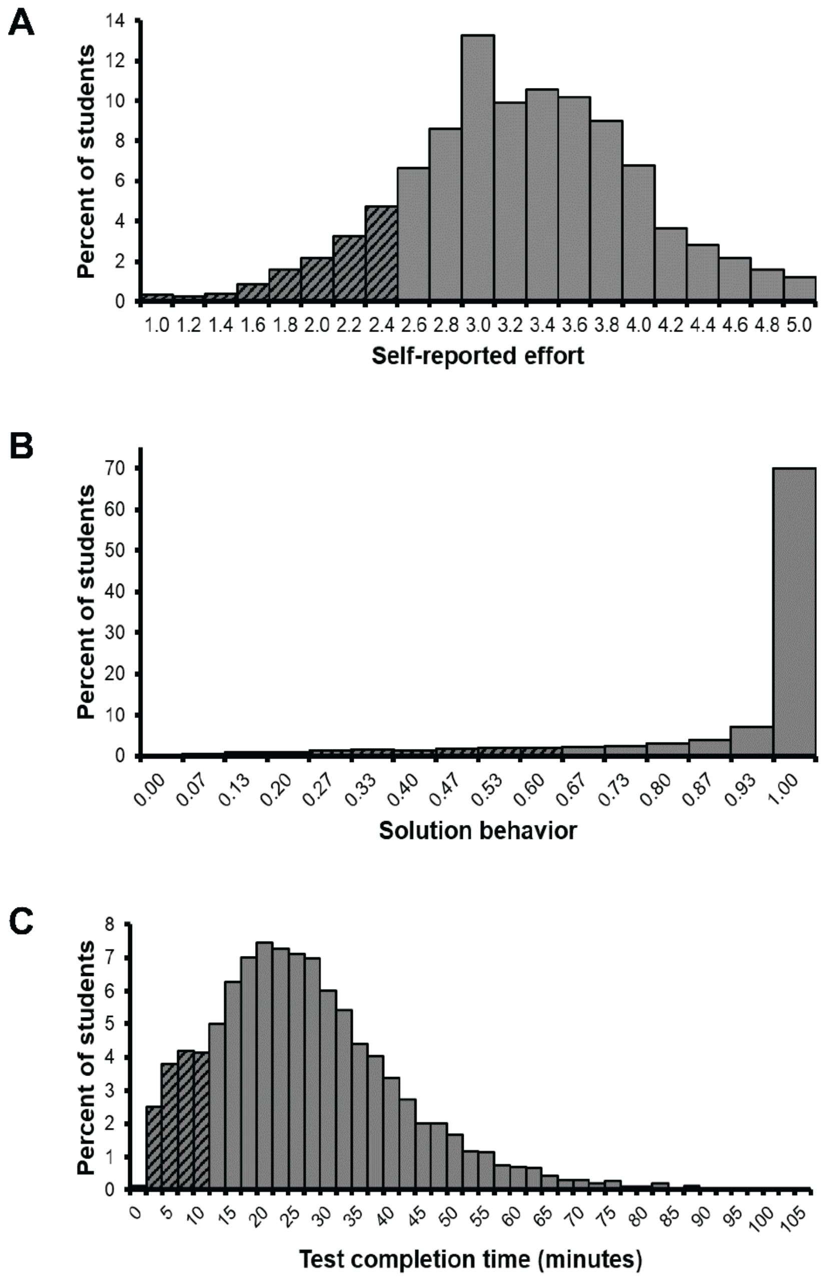


Figure 1.1: Distribution of (A) self-reported effort, (B) solution behavior, and (C) test completion time. The striped portion of each distribution represents the students considered to be demonstrating unmotivated test-taking behavior. (A) Self-reported effort was determined using the average of students' responses to the effort subscale of the SOS instrument. Higher average scores reflect student perception of using a greater amount of effort on GenBio-MAPS. (B) Solution behavior represents the proportion of questions for which a student did not use rapid-selection behavior. (C) The intended test completion time for GenBio-MAPS was 30 minutes.

effort subscale was 3.26, with an SD of 0.72. Most students (90%) used solution behavior on greater than 60% of GenBio-MAPS questions, and 64% of students used solution behavior on every question. Approximately 90% of students had test completion times longer than 10 minutes. The mean test completion time was 27.78 minutes with an SD of 15.11.

We found that the different measures of effort generally correlated with each other (Supplemental Table 1.4). To understand differences in student motivation classifications, we analyzed how commonly students received the same classification of either “motivated” or “unmotivated” across measures. There was a 72% agreement between self-reported effort and solution behavior. Self-reported effort and test completion time agreed 69% of the time. The two time-based indicators of effort, solution behavior and test completion time, had the largest agreement at 93%. Agreement across all three indicators of effort was 66%. Thus, while there is correspondence across these three indicators of test-taking effort, they each capture slightly different subsets of student behaviors.

Most of the demographic variables that we included in our models significantly predicted scores on the SOS effort subscale (Supplemental Table 1.5); however, the effect size for each variable was small and the adjusted R^2 for our model was low (0.033). Our results suggest that student demographic characteristics had negligible effects on self-reported effort, which provides further evidence that the SOS effort subscale consistently measures test-taking effort across diverse student populations.

Does Self-Reported Effort Align with Observed Time-Based Behaviors?

We examined the degree to which students’ self-reported effort predicted their observed time-based behaviors, using separate models to predict the effects of student demographics and self-reported effort on solution behavior and test completion time (Supplemental Table 1.6). We found that most demographic variables had significant ($p < 0.05$) but weak effects on solution behavior and test completion time. These findings suggest that variation in observed time-based behavior cannot be largely attributed to differences in student demographic characteristics.

Our models indicated that students at different time points in degree programs behaved differently when completing GenBio-MAPS. Compared with the beginning of the introductory series (first time point), students at the end of the introductory series (second time point) had lower solution behavior and shorter test completion times. These students at the end of the introductory series (second time point) also had lower time-based effort than students in advanced courses (third time point). The models further indicated that students with a higher score on the SOS effort subscale spend more time on GenBio-MAPS and used more solution behavior. Overall, student demographics and self-reported effort explained a relatively small amount of the variation in observed time-based behaviors (solution behavior: adjusted $R^2 = 0.145$; test completion time: adjusted $R^2 = 0.091$).

How Do Different Aspects of Test-Taking Effort Relate to GenBio-MAPS Score?

We hypothesized that self-reported effort and observed time-based behaviors affect student performance on GenBio-MAPS. Given the correlations between the three indicators of effort, we used regression models to separately test for the effects of self-reported effort, solution behaviors, and test completion time (Supplemental Table 1.7). In each model, each demographic variable significantly ($p < 0.0001$) predicted score, as we have found previously (Couch et al., 2019). We found that self-reported effort, solution behavior, and test completion time had positive effects on score, indicating that students who reported higher effort, used more solution behavior, or spent longer amounts of time on the test were likely to achieve higher scores. When considered separately, the model containing solution behavior explained more of the variance in score (adjusted R^2 = 0.418) compared with self-reported effort (adjusted R^2 = 0.343) or test completion time (adjusted R^2 = 0.350). When we added all three of these variables into one regression model to look at the combined effects of test-taking effort on score (Table 1.3), their effect sizes decreased, but the adjusted R^2 of the model increased to 0.452.

Table 1.3: Standard least squares linear regression model[a] on the effects of student demographic characteristics and test-taking effort on GenBio-MAPS score

Parameter[b]	Estimate	SE	*df*	*t*	*p*
Intercept	0.369	0.012	113.9	31.79	<0.0001
Gender: male (ref: female)	0.015	0.001	7519	13.96	<0.0001
Race/ethnicity: underserved (ref: non-underserved)	-0.012	0.001	7536	-8.80	<0.0001
Parental education: did not complete bachelors' degree (ref: parent completed bachelor's degree)	-0.012	0.001	7536	-10.74	<0.0001
Language: English not spoken at home (ref: English spoken at home)	-0.013	0.002	7531	-8.37	<0.0001
Major: not majoring in biology (ref: majoring in biology)	-0.006	0.001	7534	-5.06	<0.0001
Time point [2-1]: end of introductory series (ref: beginning of introductory series)	0.059	0.003	7429	23.14	<0.0001
Time point [3-2]: advanced series (ref: end of introductory series)	0.050	0.004	7536	14.06	<0.0001
Self-reported effort	0.024	0.002	7522	10.94	<0.0001
Time point [2-1]*self-reported effort	-0.001	0.003	7522	-0.45	0.6555
Time point [3-2]*self-reported effort	0.022	0.005	7519	4.53	<0.0001
Solution behavior	0.127	0.009	7529	13.42	<0.0001
Time point [2-1]*solution behavior	0.063	0.013	7526	4.79	<0.0001
Time point [3-2]*solution behavior	0.067	0.023	7518	2.97	0.0030
Test completion time	0.001	0.000	7533	6.41	<0.0001
Time point [2-1]*Test completion time	0.000	0.000	7526	2.65	0.0081
Time point [3-2]*Test completion time	-0.000	0.000	7519	-1.37	0.1694

[a] Score ~ institution + gender + race/ethnicity + parental education + language + major + time point + self-reported effort + time point*self-reported effort + solution behavior + time point*solution behavior + test completion time + time point*test completion time

[b] Estimates for nominal variables indicate the effect based on being a member of the focal group in comparison to the reference (ref) group.

Our models indicated that time point in a degree program largely affects GenBio-MAPS performance. As expected, students at later time points in a degree program were predicted to have higher GenBio-MAPS scores than students at earlier points in a degree program. We also examined the interactions between test-taking effort and time point in a degree program. These interactions allow us to determine how effort affects scores at each time point (Figure 1.2). For self-reported effort, advanced students show a disproportionate benefit as they report increasing effort. For solution behavior, as students reach later time points, their engagement in solution behavior increasingly results in higher scores. Both of these results are consistent with the idea that effort has a greater impact on the performance of students at later time points. For test completion time, students at the end of the introductory series see a disproportionate benefit from taking more time than students at the beginning of the introductory series, but advanced students do not see any further benefit from taking more time to complete the test.

To What Extent Do Students Demonstrate Test-Taking Persistence?

Students used the SOS instrument to report their test-taking effort after completing GenBio-MAPS, but this single data point was not sufficient to capture subtle changes in test-taking effort that may have occurred as the test progressed. Our results indicate that persistence behaviors generally decreased over the course of the test (Figure 1.3). When comparing the first and last questions on the test, the proportion of students using solution behavior decreased from 0.99 to 0.83, the average number of minutes per question decreased from 2.1 minutes to 1.3 minutes, and the proportion of students answering correctly decreased from 0.67 to 0.62. Regression models, which account for the difficulty of each randomly displayed question, confirm that the display order of questions had a significant ($p < 0.0001$) negative effect on solution behavior, the amount of time spent on the question, and the score that students achieved on the question (Supplemental Table 1.8).

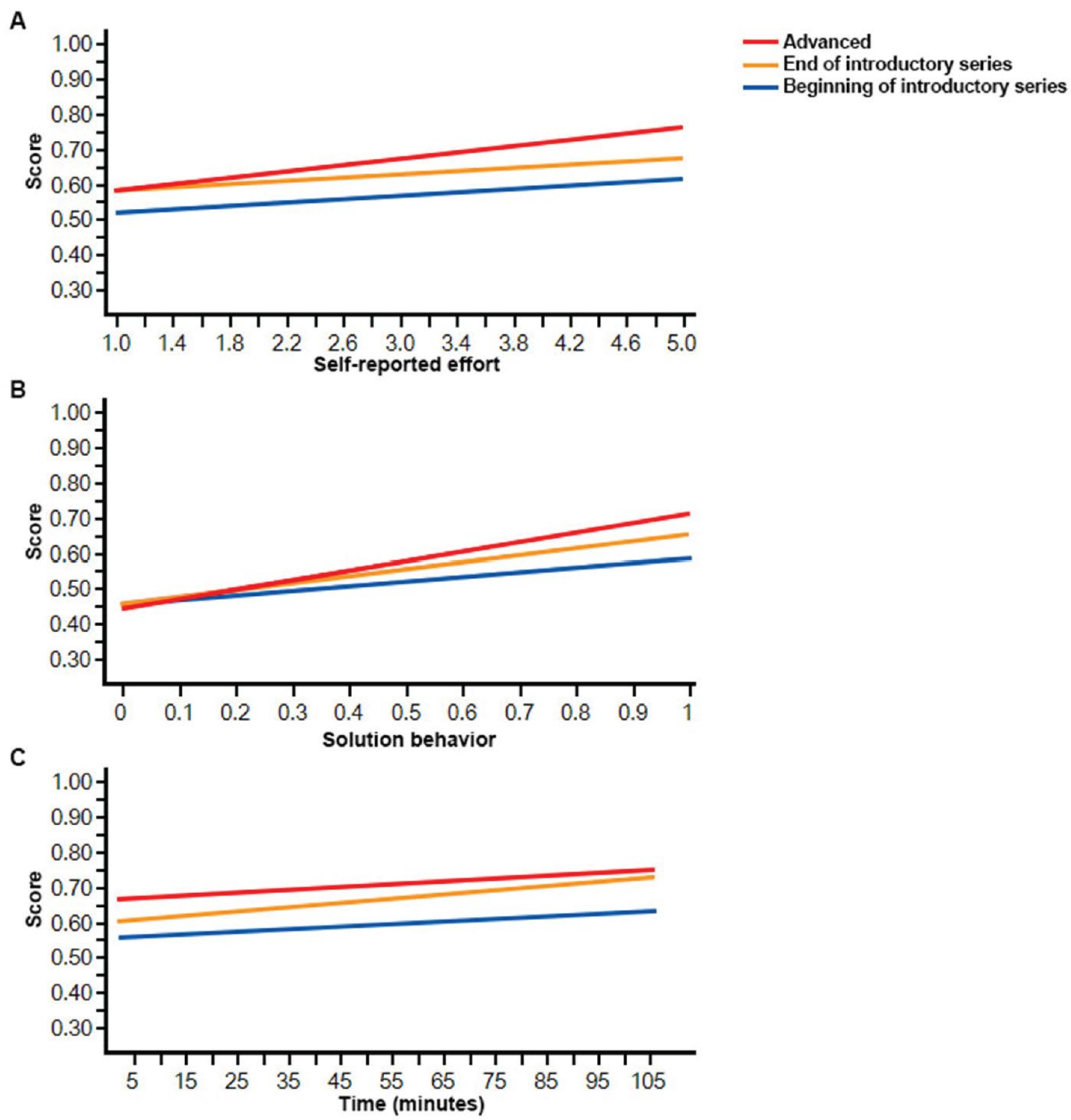


Figure 1.2: Modeled interaction effects between (A) self-reported effort, (B) solution behavior, and (C) test completion time and time point in a degree program on GenBio-MAPS score. Lines represent students enrolled in courses at the beginning of the introductory course series (blue), end of the introductory course series (orange), and end of advanced courses (red).

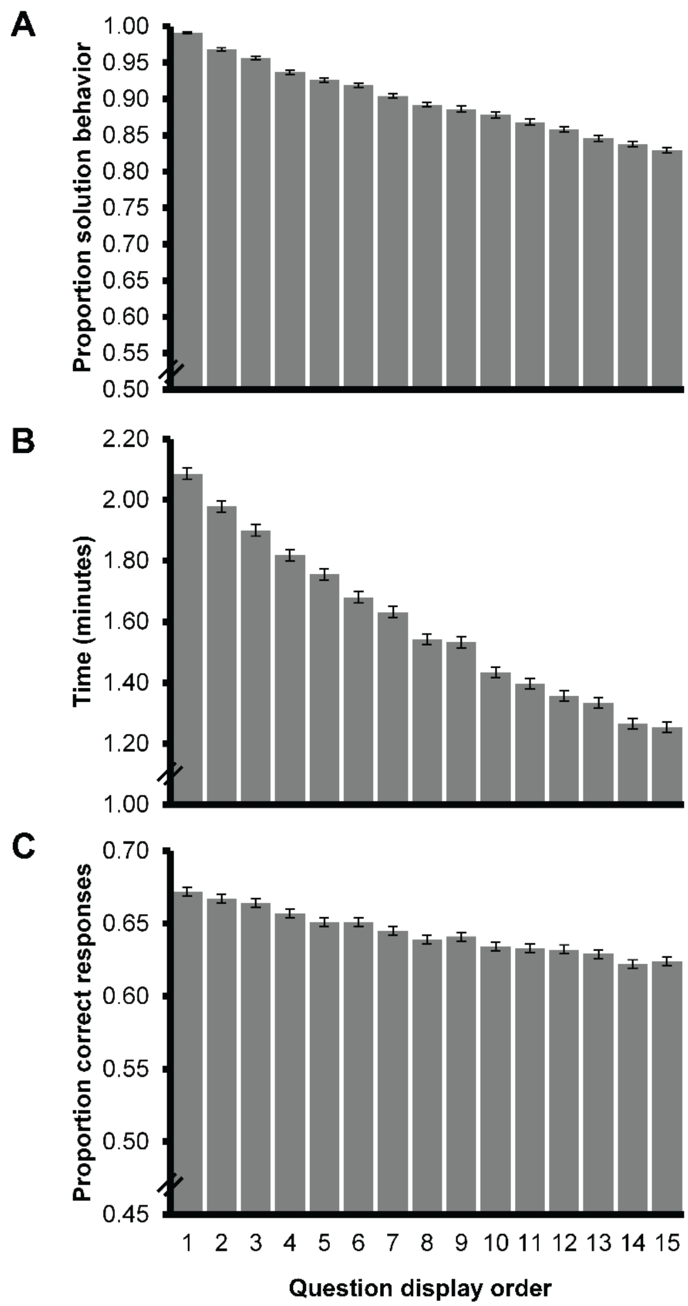


Figure 1.3: Effect of question display order on student test-taking behaviors and performance. Bars represent (A) the proportion of students using solution behavior, (B) the average minutes spent by each student, and (C) the proportion of correct responses for questions shown in each position on the test. Each student received a random subset of 15 GenBio-MAPS questions displayed in a random order, so differences between student behavior or performance on each question cannot be attributed to question characteristics. The y-axis for each graph was truncated for emphasis. Error bars represent standard errors.

How Might Departments Filter Student Responses to Reduce the Influence of Low Test-Taking Effort?

Two criteria should be considered before using motivation filtering techniques: test motivation and test score should be significantly correlated, and there should be a very low correlation between test motivation and student ability (Wise et al., 2006b). Our results satisfy the first criterion, because our three indicators of test-taking motivation (self-reported effort, solution behaviors, and test completion time) had significant effects on student scores. Our data also satisfy the second criterion. Students' self-reported grade point averages (GPAs) were correlated with the three effort indicators (self-reported effort: $r = 0.0673$; solution behavior: $r = 0.1109$; time: $r = 0.0434$), but these correlations are below the recommended threshold (Ferguson, 2009). Meeting this criterion is important to ensure the filtering process does not simply remove students with lower academic ability.

Given that data should not be removed without sufficient cause, we established the criterion that data should only be filtered when there is a compelling indication that a student expended very little effort. Thus, we explored how various filters affect the data set before making recommendations about which filtering strategy is appropriate. First, we analyzed the score distributions of students excluded by each of the filters (Figure 1.4). We found that students who self-reported low effort on the SOS (<2.5) could still achieve reasonably high scores (i.e., 60–90% correct), suggesting that some high-performing students may not perceive or report themselves to be giving high effort. Conversely, students with low solution behavior (<0.6) or time (<10 minutes) mostly scored below 60% correct, indicating that these filters capture far fewer students with high scores. This pattern also remained when using a dual filter that removed students if they had either low solution behavior or low test completion time. The test scores of students who were removed by this dual filter mirrored but did not completely align with a binomial distribution arising from random responses (Supplemental Figure 1.1).

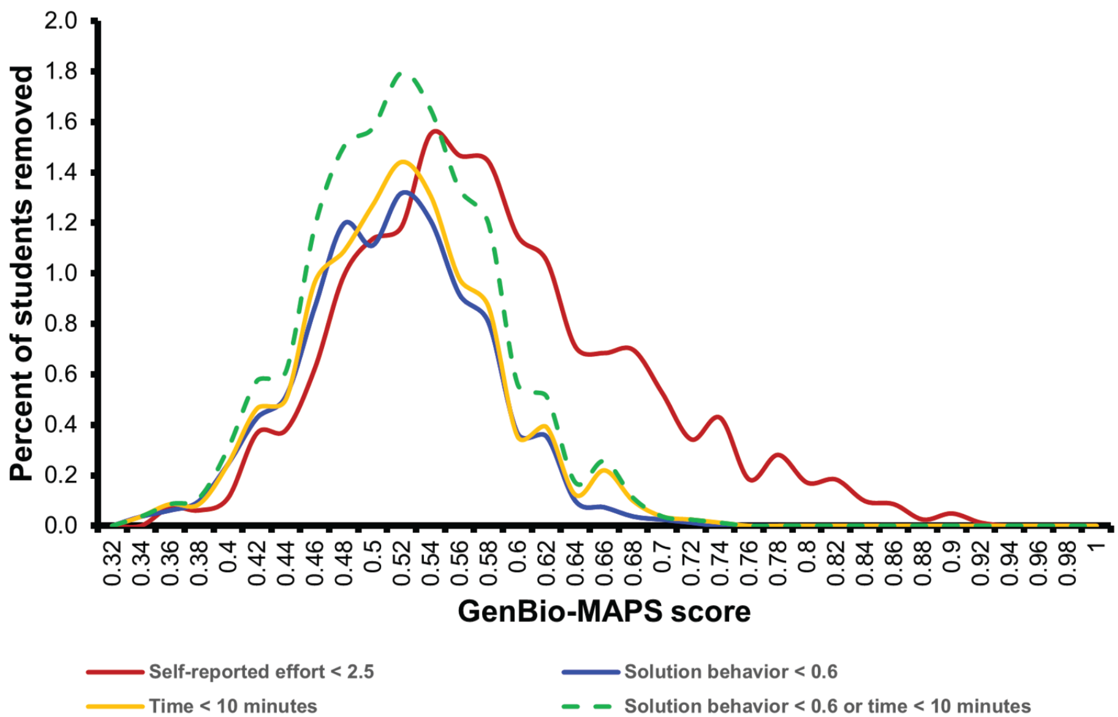


Figure 1.4: Distribution of student responses removed by each motivation filter. Lines represent the percentage of students who were removed by filters for self-reported effort (red), solution behavior (blue), and test completion time (yellow). The dashed green line represents the number of students removed by our recommended motivation filter, which removes students based on either low solution behavior or low test completion time.

We next examined test metrics for the students remaining after application of each filter (Table 1.4). The filter based on self-reported effort was the most restrictive filter (excluding 16% of the data set) but resulted in the smallest change on the mean test score for the remaining sample. The separate filters based on solution behavior or test completion time performed similarly, which can be attributed to the high agreement between the filters. However, these filters were not synonymous, as the dual filter removed a higher percentage of the sample and resulted in a slightly higher mean test score.

Table 1.4: Comparison of filtered scores across methods[a] of motivation filtering

	All students	**Self-reported effort ≥ 2.5**	**Solution behavior ≥ 0.6**	**Time ≥ 10 min**	**Solution behavior ≥ 0.6 and Time ≥ 10**
N	8185	6871	7385	7318	7068
Percent of sample excluded	0	16	10	11	14
Mean GenBio-MAPS score	0.639	0.649	0.653	0.653	0.658
SD	0.12	0.12	0.11	0.11	0.11
Standardized mean test score change[b]	0.00	0.08	0.10	0.11	0.15
Mean GPA[c]	4.23	4.23	4.25	4.25	4.26

[a] Filters listed represent the population that is included in the sample.
[b] Standardized mean score change = $(\text{Mean}_{\text{filtered}} - \text{Mean}_{\text{original}})/\text{SD}_{\text{original}}$
[c] GPA was self-reported on a scale where 5 = A- to A+ (3.70 - 4.00); 4 = B- to B+ (2.70 - 3.69); 3 = C- to C+ (1.70 - 2.69); 2 = D- to D+ (0.70 - 1.69); 1 = E or F (0.00 - 0.69)

Our analysis included the average self-reported GPA for each filtered subset of data. We used GPA as an indicator of bias, because GPA does not have a strong magnitude of correlation with the measures of test-taking effort. There was no statistical difference between the mean GPA in the unfiltered sample and data filtered using self-reported effort. There was a slight increase in the mean GPA for the remaining filters. These increases were statistically significant ($p < 0.05$); however, the statistical significance of the small changes in GPA may be attributed to the large sample size (7913 students reported their GPAs for analysis). We conclude that the overall distribution of student academic ability in the filtered samples is comparable to that of the unfiltered set.

DISCUSSION

GenBio-MAPS is a biology program assessment that is administered as an online survey outside class time under low-stakes conditions (i.e., participation credit for completion). This administration format has many practical advantages but introduces potential caveats to score interpretation. Under these conditions, student test-taking motivation cannot be assumed, and low test-taking motivation threatens valid score interpretation. Our research sought to characterize students' effort on GenBio-MAPS, understand how different effort metrics relate to performance, and outline appropriate ways to reduce the effects of low test-taking effort. Ultimately, these insights are intended to help test administrators process and interpret their data from low-stakes assessments in a way that accurately captures student understanding.

Most Students Used Motivated Behavior on GenBio-MAPS

While one of the goals of our work was to identify and remove scores from students with low test-taking effort, we want to emphasize that this group of students was only a small percentage of our data set. We found that the majority of students (>86%) reported and used motivated behavior when completing GenBio-MAPS and that there was a high degree of consistency across the self-reported and time-based effort measures (Figure 1.1). Student use of solution behavior on GenBio-MAPS was comparable to student behavior in other low-stake contexts (Wise et al., 2006a, 2009; Wise and DeMars, 2010); however, we observed a slightly higher percentage of students reporting motivated behavior on GenBio-MAPS compared with low-stakes general education tests (Schiel, 1996; Hoyt, 2001; Swerdzewski et al., 2011). The expectancy-value theory of achievement motivation (Eccles et al., 1983; Wigfield and Eccles, 2000) may provide an explanation for this result. This theory states that motivation to perform well on a task is influenced by expectancy for success on the task and the perception that the task is important or interesting. In our context, the task (GenBio-MAPS) is a discipline-specific test that was administered only to students enrolled in biology courses. Thus, the students may have had a greater expectancy to do well on a biology test and may have had greater interest in its biology content, which could have led them to report greater effort compared with a general education test outside the discipline. This interpretation also agrees with our finding that biology majors tended to have higher effort metrics than nonmajors (Supplemental Tables 1.5 and 1.6).

The Amount of Time Students Spend on Each Question Decreases across the Test

Although most students engaged in effortful behavior, we noticed a significant effect of question order on student behaviors. We found that test-taking persistence tended to decrease as students moved through the test (Figure 1.3; Supplemental Table 1.8). There was a decreasing proportion of solution behavior with increasing question position, which is a trend that has been documented in other low-stakes assessment contexts (Wise, 2006; Wise et al., 2009). The amount of time spent on a question as well as the percentage of correct responses also decreased as students moved through the test. The decrease in time spent on questions may be partially attributed to a growing familiarity with the test format. Each GenBio-MAPS question contains the same line of text providing instructions on how to respond to T/F statements, which students may have ignored later in the test. The decrease in solution behavior and decrease in time spent per question are closely related, because students who do not use solution behavior have inherently short question-response times. Changes in solution behavior and time spent per question both contribute to the decrease in the proportion of correct answers at the overall test level, but our results suggest that solution behavior has a greater influence on GenBio-MAPS score than time (Table 1.3; Supplemental Table 1.7).

While these patterns in persistence may seem discouraging, we note that even at the end of the test where we observed the least-persistent behaviors, we saw that the majority of students (83%) used solution behavior and that the average question time (1.25 minutes) represented a reasonable amount of time for answering GenBio-MAPS questions. Using motivation filtering on the data set will help to remove some of the effects of low test-taking persistence but may not capture the extent of low-effort responses that occur at the end of the test. Thus, we support the continued practice of randomizing the question order during GenBio-MAPS administrations, which distributes the effect of low-effort behaviors that occur toward the end of the test across the question pool.

Effortful Behavior Predicts Higher GenBio-MAPS Scores

Our research adds to the body of literature that demonstrates a positive relationship between test-taking motivation and student performance on low-stakes tests. Historically, most of the work on test-taking motivation has been completed in the context of general education assessments (Schiel, 1996; Hoyt, 2001; Sundre and Wise, 2003; Wise and Kong, 2005; Wise et al., 2006b; Cole et al., 2008; Thelk et al., 2009; Wise and DeMars, 2010; Swerdzewski et al., 2011). However, work from the broader suite of Bio-MAPS assessments has provided more recent evidence of a positive relationship between motivation and test score occurs in the context of discipline-specific tests. Higher scores on the effort subscale of the SOS instrument were predictive of higher scores on EcoEvo-MAPS (Summers et al., 2018) and Phys-MAPS (Semsar et al., 2019). Our work on GenBio-MAPS corroborates this finding about the effects of self-reported effort on biology program assessment scores, while also providing insights into the relationship between time-based behaviors and score on a discipline-specific assessment.

Our models predicted that students who reported and used effortful behavior were likely to have higher scores (Table 1.3; Supplemental Table 1.7). This important result is consistent with motivation theory (Pintrich and Schrauben, 1992; Schunk, 1995) and aligns with previous findings in the literature on low-stakes assessments (Wolf and Smith, 1995; Schiel, 1996; Wise and DeMars, 2005; Cole et al., 2008; Thelk et al., 2009). Our work bolsters existing theory and matches findings from other low-stakes contexts, but we also contributed a new perspective to the field by examining how test-taking motivation is affected by students' time point in a degree program. We found that test-taking effort has a greater effect on student performance at later time points (Figure 1.2). Our findings suggest that, when students in upper-level courses have low test-taking effort, there is likely to be a more pronounced discrepancy between their actual understanding of biology and the level of biology understanding that their low GenBio-MAPS score implies. This underestimation of students' skills and abilities threatens valid interpretation of GenBio-MAPS scores and provides support for the practice of motivation filtering to remove the scores of students with low test-taking effort.

Motivation Filtering Should Be Used to Remove Data from Low-Effort Students

Our findings support the conclusions drawn by Wise and DeMars (2005), which suggest that test scores from students with low test-taking motivation may be underestimating students' knowledge, skill, and abilities. For this reason, we encourage departments administering GenBio-MAPS to collect data on students' test-taking effort and use these data to inform their interpretation of test scores. We suggest that departments apply motivation filtering to reduce the negative influence of low test-taking effort on GenBio-MAPS scores.

While all the motivation filters addressed the effects of low test-taking effort, the filters did not address these effects equally, and they produced subtle differences in resulting scores (Table 1.4). Given that it is generally not ideal to remove responses from data sets, we sought to identify a filtering strategy that only eliminated data from students who clearly gave an insufficient effort. Based on our findings, we recommend using a dual filter that removes students who had either low solution behavior or short test completion time. While these individual filters largely overlap (93%), using the dual filter helps identify students who may have met one criterion, but who still gave an unsatisfactory effort. For example, a student may have spent just barely more than the threshold time on each question, or a student may have spent less than the threshold time on most questions and a considerable time on a few questions. This filter captures a range of low-effort behaviors that likely introduce construct-irrelevant variance, but it does not remove an excessive number of students from the data set.

Although the data from the SOS instrument are convenient to collect, we do not recommend using the data from the SOS effort subscale as a motivation filter. Compared with the time-based filters, we observed that the SOS filter captured a greater number of responses from students who achieved high scores (Figure 1.4), which also explains why there was a smaller effect on mean score with this filter. Steedle (2014) observed a similar trend in that many examinees who reported low effort using the SOS instrument actually performed well on the Collegiate Learning Assessment. Steedle proposed several explanations for this result and suggested that it may be attributed to students not accurately providing self-reported data, intentionally selecting inaccurate responses, or making errors when interpreting SOS item wording. Our recommended motivation filter avoids these potential problems with self-reported data and relies only on objective time-based behaviors. After applying the dual filter, departments may still incorporate SOS or time-based variables in their statistical models, although this option may not be viable at institutions with small student numbers.

Previous studies have called attention to the need for additional research on motivation filtering (Sundre and Wise, 2003; Wise and DeMars, 2005, 2010; Wise and Kong, 2005; Wise et al., 2006b). Only a small number of studies have been conducted since these calls to action were issued in the early 2000s (Swerdzewski et al., 2011; Waskiewicz, 2011; Steedle, 2014). The scant number of publications on motivation filtering is alarming, considering that Wise and DeMars (2010) suggested that "measurement practitioners routinely apply motivation filtering whenever the data from low-stakes tests are used to support program decisions" (p. 27). Our research with GenBio-MAPS contributes to the limited literature in the field by providing evidence that motivation filtering is an effective and generalizable technique that can be used to better inform decisions made about biology degree programs.

Recommendations for GenBio-MAPS Administration

Wise (2006) emphasized that, in addition to developing methods to identify and manage data from low-effort students, adopting test administration strategies that promote effort for low-stakes tests is important. While this was not the focus of the current research, we suggest that departments communicate and emphasize the importance and usefulness of GenBio-MAPS data. Students who perceive the importance or usefulness of an assessment are more likely to put forth more effort (Cole et al., 2008), and framing assessments as important tools to collect data for the student's institution has been an effective method to increase test-taking motivation in other low-stakes contexts (Huffman et al., 2011; Liu et al., 2015). We strongly recommend that instructors assign some amount of participation credit for completing the instrument, as we have found repeatedly that instructors who fail to provide this incentive obtain very low participation rates. We do not recommend that departments assign grades based on answer correctness as a way to increase student test-taking effort. Although previous studies (Wolf and Smith, 1995; Napoli and Raymond, 2004) have indicated that students who were told that test performance would count toward a course grade reported higher test-taking motivation and performed better on college-level standardized tests, these studies had the benefit of administering their graded versions under secure conditions. Most departments lack the resources to proctor program-level tests, and assigning grades to students taking the test outside a proctored environment would likely encourage students to seek external resources. Departments that can administer under secure conditions (e.g., in-person or video proctoring) face the possibility that students being graded may still attempt to obtain test materials before the assessment. Furthermore, previous work on a science literacy assessment established that assigning a small amount of performance-based course credit (i.e., part of a quiz grade) to increase the stakes of the test did not significantly affect students' self-reported effort or performance (Segarra et al., 2018). Assigning course grades for GenBio-MAPS may also result in other unintended consequences, such as increased test anxiety, which can threaten the interpretation of test scores (Cassady and Johnson, 2002).

CONCLUSIONS

Our work demonstrates that test-taking motivation represents an important consideration in the interpretation of scores from discipline-specific low-stakes assessments. While our study examined test-taking motivation for a biology program assessment, our results are likely generalizable to investigations of test-taking motivation in other contexts and STEM disciplines where assessment instruments are administered in low-stakes settings. Our results are also relevant to low-stakes administrations of other diagnostic tests or activities that share characteristics with GenBio-MAPS (e.g., pre–post concept inventories). We encourage test administrators to collect and report measures of effort (e.g., self-reported effort, solution behavior, test completion time) and to apply motivation filtering to address the negative effects of the low test-taking effort that can occur during low-stakes administration conditions. Our motivation filtering procedure can be adapted for other instruments, adjusting the thresholds for detecting low motivation accordingly based on the number or content of items. Taking these steps to identify and remove low-effort responses may provide departments with a more accurate representation of student understanding of assessed concepts, which can better inform decisions made using assessment data.

Accessing Instruments

GenBio-MAPS is published in its entirety in Couch et al. (2019) and can also be accessed through the online portal (http://cperl.lassp.cornell.edu/bio-maps). The SOS (Sundre and Moore, 2002), as well as an administration manual for the instrument, can be accessed at www.jmu.edu/assessment.

Acknowledgements

We thank Sara Brownell, Alison Crowe, Scott Freeman, Jennifer Knight, Katharine Semsar, Michelle Smith, Mindi Summers, Christian Wright, and Yi Zheng for their work on GenBio-MAPS development. We thank the site coordinators who helped coordinate GenBio-MAPS administration, the instructors who administered the survey in their courses, and the students who participated in the research. This material is based upon work supported by the National Science Foundation (NSF) Graduate Research Fellowship (DGE-1610400), Transforming Undergraduate Education in STEM (DUE-1322364), and Improving Undergraduate STEM Education (DUE-1610621) programs. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

REFERENCES FOR CHAPTER 1

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In Petrov, B. N.Csaki, F. (Eds.), Proceedings of the 2nd International Symposium on Information Theory (pp. 267–281). Budapest: Akademiai Kiado.

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

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). Standards for educational and psychological testing. Washington, DC.

Cassady, J. C., & Johnson, R. E. (2002). Cognitive test anxiety and academic performance. Contemporary Educational Psychology, 27(2), 270–295. https://doi.org/10.1006/ceps.2001.1094

Cole, J. S., Bergin, D. A., & Whittaker, T. A. (2008). Predicting student achievement for low stakes tests with effort and task value. Contemporary Educational Psychology, 33(4), 609–624. https://doi.org/10.1016/j.cedpsych.2007.10.002

Couch, B. A., & Knight, J. K. (2015). A comparison of two low-stakes methods for administering a program-level biology concept assessment. Journal of Microbiology & Biology Education, 16(2), 178–185.

Couch, B. A., Wood, W. B., & Knight, J. K. (2015). The Molecular Biology Capstone Assessment: A concept assessment for upper-division molecular biology students. CBE—Life Sciences Education, 14(1), ar10. https://doi.org/10.1187/cbe.14-04-0071

Couch, B. A., Wright, C. D., Freeman, S., Knight, J. K., Semsar, K., Smith, M. K., ... & Brownell, S. E. (2019). GenBio-MAPS: A programmatic assessment to measure student understanding of Vision and Change core concepts across general biology programs. CBE—Life Sciences Education, 18(1), ar1. https://doi.org/10.1187/cbe.18-07-0117

Duckworth, A. L., Quinn, P. D., Lynam, D. R., Loeber, R., & Stouthamer-Loeber, M. (2011). Role of test motivation in intelligence testing. Proceedings of the National Academy of Sciences USA, 108(19), 7716–7720. https://doi.org/10.1073/pnas.1018601108

Eccles, J. S., Adler, T. F., Futterman, R., Goff, S. B., Kaczala, C. M., Meece, J. L., & Midgley, C. (1983). Expectancies, values, and academic behaviors. In Spence, J. T. (Ed.), Achievement and achievement motivation (pp. 75–146). San Francisco, CA: W. H. Freeman.

Ferguson, C. J. (2009). An effect size primer: A guide for clinicians and researchers. Professional Psychology: Research and Practice, 40(5), 532–538. https://doi.org/10.1037/a0015808

Fisher, R. J., & Katz, J. E. (2000). Social-desirability bias and the validity of self-reported values. Psychology & Marketing, 17(2), 105–120. https://doi.org/10.1002/(SICI)1520-6793(200002)17:2<105::AID-MAR3>3.0.CO;2-9

Flesch, R. (1948). A new readability yardstick. Journal of Applied Psychology, 32(3), 221–40. https://doi.org/10.1037/h0057532

Hoyt, J. E. (2001). Performance funding in higher education: The effects of student motivation on the use of outcomes tests to measure institutional effectiveness. Research in Higher Education, 42(1), 71–85. https://doi.org/10.1023/A:1018716627932

Huffman, L., Adamopoulos, A., Murdock, G., Cole, A., & McDermid, R. (2011). Strategies to motivate students for program assessment. Educational Assessment, 16(2), 90–103. https://doi.org/10.1080/10627197.2011.582771

Kincaid, J. P., Fishburne, R. P., Rogers, R. L., & Chissom, B. S. (1975). Derivation of new readability formulas (Automated Readability Index, Fog Count, and Flesch Reading Ease Formula) for Navy enlisted personnel. Research Branch Report 8-75. Millington, TN: Naval Air Station Memphis.

Kong, X. J., Wise, S. L., & Bhola, D. S. (2007). Setting the response time threshold parameter to differentiate solution behavior from rapid-guessing behavior. Educational and Psychological Measurement, 67(4), 606–619. https://doi.org/10.1177/0013164406294779

Lee, Y.-H., & Jia, Y. (2014). Using response time to investigate students' test-taking behaviors in a NAEP computer-based study. Large-Scale Assessments in Education, 2(1), 8. https://doi.org/10.1186/s40536-014-0008-1

Lepper, M. R., Greene, D., & Nisbett, R. E. (1973). Undermining children's intrinsic interest with extrinsic reward: A test of the "overjustification" hypothesis. Journal of Personality and Social Psychology, 28(1), 129–137. https://doi.org/10.1037/h0035519

Liu, O. L., Rios, J. A., & Borden, V. (2015). The effects of motivational instruction on college students' performance on low-stakes assessment. Educational Assessment, 20(2), 79–94. https://doi.org/10.1080/10627197.2015.1028618

Napoli, A. R., & Raymond, L. A. (2004). How reliable are our assessment data? A comparison of the reliability of data produced in graded and un-graded conditions. Research in Higher Education, 45(8), 921–929. https://doi.org/10.1007/s11162-004-5954-y

Pintrich, P. R., & Schrauben, B. (1992). Students' motivational beliefs and their cognitive engagement in classroom academic tasks. In Schunk, D. H.Meece, J. L. (Eds.), Student perceptions in the classroom (pp. 149–183). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Qualtrics. (2019). Qualtrics. Provo, UT. https://www.qualtrics.com

Salomon, G. (1984). Television is "easy" and print is "tough": The differential investment of mental effort in learning as a function of perceptions and attributions. Journal of Educational Psychology, 76(4), 647–658. https://doi.org/10.1037/0022-0663.76.4.647

SAS Institute Inc. (2019). JMP (Version 15). Cary, NC. https://www.jmp.com

Schiel, J. (1996). Student effort and performance on a measure of postsecondary educational development (ACT research report series 96-9). Retrieved September 2, 2020, from https://eric.ed.gov/?id=ED405380

Schnipke, D. L., & Scrams, D. J. (1997). Modeling item response times with a two-state mixture model: A new method of measuring speededness. Journal of Educational Measurement, 34(3), 213–232. JSTOR.

Schunk, D. H. (1995). Self-efficacy and education and instruction. In Self-efficacy, adaptation, and adjustment: Theory, research, and application (pp. 281–303). New York: Plenum. https://doi.org/10.1007/978-1-4419-6868-5_10

Schunk, D. H., Pintrich, P. R., & Meece, J. L. (2008). Motivation in education: Theory, research, and applications (3rd ed.). Upper Saddle River, NJ: Pearson Prentice Hall.

Segarra, V. A., Hughes, N. M., Ackerman, K. M., Grider, M. H., Lyda, T., & Vigueira, P. A. (2018). Student performance on the Test of Scientific Literacy Skills (TOSLS) does not change with assignment of a low-stakes grade. BMC Research Notes, 11(1), 422. https://doi.org/10.1186/s13104-018-3545-9

Semsar, K., Brownell, S., Couch, B. A., Crowe, A. J., Smith, M. K., Summers, M. M., ... & Knight, J. K. (2019). Phys-MAPS: A programmatic physiology assessment for introductory and advanced undergraduates. Advances in Physiology Education, 43(1), 15–27. https://doi.org/10.1152/advan.00128.2018 ,

Setzer, J. C., Wise, S. L., van den Heuvel, J. R., & Ling, G. (2013). An investigation of examinee test-taking effort on a large-scale assessment. Applied Measurement in Education, 26(1), 34–49. https://doi.org/10.1080/08957347.2013.739453

Smith, M. K., Brownell, S. E., Crowe, A. J., Holmes, N. G., Knight, J. K., Semsar, K., ... & Couch, B. A. (2019). Tools for change: Measuring student conceptual understanding across undergraduate biology programs using Bio-MAPS assessments. Journal of Microbiology & Biology Education, 20(2). https://doi.org/10.1128/jmbe.v20i2.1787

Steedle, J. T. (2014). Motivation filtering on a multi-institution assessment of general college outcomes. Applied Measurement in Education, 27(1), 58–76. https://doi.org/10.1080/08957347.2013.853072

Summers, M. M., Couch, B. A., Knight, J. K., Brownell, S. E., Crowe, A. J., Semsar, K., ... & Smith, M. K. (2018). EcoEvo-MAPS: An ecology and evolution assessment for introductory through advanced undergraduates. CBE—Life Sciences Education, 17(2), ar18. https://doi.org/10.1187/cbe.17-02-0037

Sundre, D. L. (2007). The Student Opinion Scale (SOS): A measure of examinee motivation, Test manual. Harrison, VA: Center for Assessment & Research Studies.

Sundre, D. L., & Moore, D. L. (2002). The Student Opinion Scale: A measure of examinee motivation. Assessment Update, 14(1), 8–9. https://doi.org/10.1002/au.141

Sundre, D. L., & Wise, S. L. (2003). Motivation filtering: An exploration of the impact of low examinee motivation on the psychometric quality of tests. Chicago: National Council on Measurement in Education.

Swerdzewski, P. J., Harmes, J. C., & Finney, S. J. (2011). Two approaches for identifying low-motivated students in a low-stakes assessment context. Applied Measurement in Education, 24(2), 162–188. https://doi.org/10.1080/08957347.2011.555217

Thelk, A. D., Sundre, D. L., Horst, S. J., & Finney, S. J. (2009). Motivation matters: Using the Student Opinion Scale to make valid inferences about student performance. Journal of General Education, 58(3), 129–151.

Waskiewicz, R. A. (2011). Pharmacy students' test-taking motivation-effort on a low-stakes standardized test. American Journal of Pharmaceutical Education, 75(3), 41. https://doi.org/10.5688/ajpe75341

Wigfield, A., & Eccles, J. S. (2000). Expectancy–value theory of achievement motivation. Contemporary Educational Psychology, 25(1), 68–81. https://doi.org/10.1006/ceps.1999.1015

Wise, S. L. (2006). An investigation of the differential effort received by items on a low-stakes computer-based test. Applied Measurement in Education, 19(2), 95–114. https://doi.org/10.1207/s15324818ame1902_2

Wise, S. L., Bhola, D. S., & Yang, S.-T. (2006a). Taking the time to improve the validity of low-stakes tests: The effort-monitoring CBT. Educational Measurement: Issues and Practice, 25(2), 21–30. https://doi.org/10.1111/j.1745-3992.2006.00054.x

Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. Educational Assessment, 10(1), 1–17. https://doi.org/10.1207/s15326977ea1001_1

Wise, S. L., & DeMars, C. E. (2010). Examinee noneffort and the validity of program assessment results. Educational Assessment, 15(1), 27–41. https://doi.org/10.1080/10627191003673216

Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. Applied Measurement in Education, 18(2), 163–183. https://doi.org/10.1207/s15324818ame1802_2

Wise, S. L., Pastor, D. A., & Kong, X. J. (2009). Correlates of rapid-guessing behavior in low-stakes testing: Implications for test development and measurement practice. Applied Measurement in Education, 22(2), 185–205. https://doi.org/10.1080/08957340902754650

Wise, V., Wise, S., & Bhola, D. (2006b). The generalizability of motivation filtering in improving test score validity. Educational Assessment, 11(1), 65–83. https://doi.org/10.1207/s15326977ea1101_3

Wolf, L. F., & Smith, J. K. (1995). The consequence of consequence: Motivation, anxiety, and test performance. Applied Measurement in Education, 8, 341–351. https://doi.org/10.1207/s15324818ame0803_3

Zimmerman, B. J., & Ringle, J. (1981). Effects of model persistence and statements of confidence on children's self-efficacy and problem solving. Journal of Educational Psychology, 73(4), 485–493. https://doi.org/10.1037/0022-0663.73.4.485

SUPPLEMENTAL MATERIAL FOR CHAPTER 1

Supplemental Table 1.1: Institution and course demographics

Institution characteristic	n	%
Control		
Public	15	75
Private	5	25
Region[b]		
Mid-Atlantic	2	10
Midwest	10	50
Northwest	3	15
Southwest	5	25
Carnegie basic classification		
Associate's Colleges: Mixed Transfer/Career & Technical-High Nontraditional	2	10
Baccalaureate Colleges: Arts & Sciences Focus	3	15
Master's Colleges & Universities: Larger or Medium Programs	7	35
Doctoral Universities: Higher or Moderate Research Activity	3	15
Doctoral Universities: Highest Research Activity	5	25
Course time point		
Beginning of introductory series	58	38
End of introductory series	45	30
Advanced	49	32

[a] Data originally collected and reported in Couch et al. (2019).
[b] Region designations are based on PULSE regional boundaries. No institutions from the Northeast or Southeast regions are represented in the data set.

Supplemental Table 1.2: Number of page times replaced

Number of pages replaced	Percent of students
0	88.4
1 - 5	11.5
6 - 10	0.1
10 -15	0

Note: Students saw one multiple-true-false question containing 4-5 T/F statements per page. Response times to individual questions exceeding 15 minutes were replaced with the average page time of 1.5 minutes

Supplemental Table 1.3: Response time thresholds for GenBio-MAPS questions

GenBio-MAPS question	Number of characters in question	Response threshold
BM-01	855	8.55
BM-02	633	6.33
BM-03	875	8.75
BM-04	758	7.58
BM-07	717	7.17
BM-08	1299	12.99
BM-12	1036	10.36
BM-13	937	9.37
BM-14	418	4.18
BM-15	1163	11.63
BM-16	895	8.95
BM-18	762	7.62
BM-19	700	7.00
BM-20	954	9.54
BM-21	1172	11.72
BM-22	1083	10.83
BM-23	462	4.62
BM-24	825	8.25
BM-27	920	9.20
BM-28	973	9.73
BM-30	904	9.04
BM-31	466	4.66
BM-32	840	8.40
BM-33	912	9.12
BM-35	970	9.70
BM-36	777	7.77
BM-37	988	9.88
BM-38	866	8.66
BM-40	749	7.49
BM-43	726	7.26
BM-44	858	8.58
BM-45	737	7.37
BM-49	618	6.18
BM-50	938	9.38
BM-54	1069	10.69
BM-55	1480	14.80
BM-59	1344	13.44
BM-60	1188	11.88
BM-61	733	7.33

Supplemental Table 1.4: Correlations between demographic variables, test-taking effort, and GenBio-MAPS score

	Gender	Race/ Ethnicity	Parental education	Language	Major	Self-reported effort	Test time	Solution behavior	GenBio-MAPS score
Gender	–	-0.01	-0.02	-0.02	0.01	-0.00	0.03**	0.04***	-0.12***
Race/ethnicity	-0.01	–	0.19***	0.14***	-0.04***	0.04***	-0.04***	0.03**	0.14***
Parental education	-0.02	0.19***	–	0.17***	-0.02	-0.00	-0.01	0.04***	0.18***
Language	-0.02	0.14***	0.17***	–	-0.02*	0.06***	-0.02*	0.05***	0.09***
Major	0.01	-0.04***	-0.02	-0.02*	–	0.05***	0.07***	0.07***	0.12***
Self-reported effort	-0.00	0.04***	-0.00	0.06***	0.05***	–	0.23***	0.33***	0.29***
Test time	0.03**	-0.04***	-0.01	-0.02*	0.07***	0.23***	–	0.51***	0.30***
Solution behavior	0.04***	0.03**	0.04***	0.05***	0.07***	0.33***	0.51***	–	0.42***
GenBio-MAPS score	-0.12***	0.14***	0.18***	0.09***	0.12***	0.29***	0.30***	0.42***	–

*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$

Supplemental Table 1.5: Standard least squares linear regression model[a] of the effects of student demographic characteristics on self-reported effort

Parameter[b]	Estimate	SE	*df*	*t*	*p*
Intercept	3.203	0.033	23.69	96.95	<0.0001
Gender: male (ref: female)	0.003	0.009	7536	0.37	0.7086[c]
Race/ethnicity: underserved (ref: non-underserved)	-0.031	0.011	7366	-2.92	0.0035
Parental education: did not complete bachelor's degree (ref: completed bachelor's degree)	0.020	0.009	7483	2.22	0.0263
Language: English not spoken at home (ref: English spoken at home)	-0.064	0.012	7534	-5.18	<0.0001
Major: not majoring in biology (ref: majoring in biology)	-0.046	0.009	7518	-4.81	<0.0001

[a] Self-reported effort ~ institution + gender + race/ethnicity + parental education + language + major + time point. Only variables that passed model selection are listed.

[b] Estimates for nominal variables indicate the effect based on being a member of the focal group in comparison to the reference (ref) group.

[c] Removing this non-significant term raised AIC above the threshold for exclusion.

Supplemental Table 1.6: Standard least squares linear regression models[a] of the effects of student demographic characteristics and self-reported effort on observed test-taking behaviors

	Solution behavior					Test completion time (minutes)				
Parameter[b]	**Estimate**	**SE**	***df***	***t***	***p***	**Estimate**	**SE**	***df***	***t***	***p***
Intercept	0.612	0.013	133.3	47.69	<.0001	13.86	0.977	120.1	14.18	<.0001
Gender: male (ref: female)	-0.011	0.002	7656	-4.89	<.0001	-0.568	0.174	7531	-3.26	0.0011
Race/ethnicity: underserved (ref: non-underserved)						1.081	0.217	7146	4.98	<.0001
Parental education: did not complete bachelor's degree (ref: completed bachelor's degree)	-0.008	0.002	7550	-3.30	0.0010	-0.029	0.182	7406	-0.16	0.8746[c]
Language: English not spoken at home (ref: English spoken at home)	-0.009	0.003	7636	-2.64	0.0084	0.551	0.253	7502	2.18	0.0294
Major: not majoring in biology (ref: majoring in biology)	-0.009	0.003	7608	-3.68	0.0002	-0.905	0.196	7428	-4.61	<.0001
Time point: end of introductory series (ref: beginning of introductory series)	-0.016	0.005	5818	-3.09	0.0020	-2.741	0.400	4854	-6.86	<.0001
Time point: advanced series (ref: end of introductory series)	0.040	0.007	7426	5.65	<.0001	4.147	0.552	7093	7.51	<.0001
Self-reported effort	0.087	0.003	7666	28.81	<.0001	4.564	0.234	7542	19.47	<.0001

[a] Solution behavior ~ institution + gender + race/ethnicity + parental education + language + major + time point + self-reported effort; Test completion time ~ institution + gender + race/ethnicity + parental education + language + major + time point + self-reported effort. Only variables that passed model selection are listed.
[b] Estimates for nominal variables indicate the effect based on being a member of the focal group in comparison to the reference (ref) group.
[c] Removing this non-significant term raised AIC above the threshold for exclusion.

Supplemental Table 1.7: Standard least squares linear regression model of the effects of student demographic characteristics and test-taking effort on GenBio-MAPS score

Parameter	Self-reported effort					Solution behavior					Test completion time (minutes)				
	Estimate	SE	*df*	*t*	*p*	Estimate	SE	*df*	*t*	*p*	Estimate	SE	*df*	*t*	*p*
Intercept	0.459	0.011	65.49	40.78	<.0001	0.415	0.010	81.39	39.61	<.0001	0.530	0.009	27.05	60.69	<.0001
Gender: male (ref: female)	0.013	0.001	7525	10.73	<.0001	0.016	0.001	7664	14.09	<.0001	0.014	0.001	7663	12.20	<.0001
Race/ethnicity: underserved (ref: non-underserved)	-0.011	0.001	7542	-7.36	<.0001	-0.012	0.001	7679	-8.32	<.0001	-0.014	0.001	7680	-9.88	<.0001
Parental education: did not complete bachelor's degree (ref: parent completed bachelor's degree)	-0.013	0.001	7541	-10.79	<.0001	-0.012	0.001	7680	-9.99	<.0001	-0.013	0.001	7680	-10.51	<.0001
Language: English not spoken at home (ref: English spoken at home)	-0.014	0.002	7536	-8.23	<.0001	-0.014	0.002	7676	-8.55	<.0001	-0.017	0.002	7675	-10.06	<.0001
Major: not majoring in biology (ref: majoring in biology)	-0.009	0.001	7540	-6.54	<.0001	-0.007	0.001	7679	-5.92	<.0001	-0.008	0.001	7678	-5.89	<.0001
Time point [2-1]: end of introductory series (ref: beginning of introductory series)	0.054	0.003	7444	19.39	<.0001	0.057	0.003	7551	22.15	<.0001	0.061	0.003	7576	22.32	<.0001
Time point [3-2]: advanced series (ref: end of introductory series)	0.064	0.004	7542	16.88	<.0001	0.052	0.004	7679	14.37	<.0001	0.055	0.004	7680	14.53	<.0001
Self-reported effort	0.038	0.002	7530	16.39	<.0001	.	.	.	.	.	.	.	.	.	.
Time point [2-1]*self-reported effort	0.009	0.004	7532	2.60	0.0093	.	.	.	.	.	.	.	.	.	.
Time point [3-2]*self-reported effort	0.022	0.005	7527	4.48	<.0001	.	.	.	.	.	.	.	.	.	.
Solution behavior	.	.	.	.	.	0.185	0.008	7670	22.71	<.0001	.	.	.	.	.
Time point [2-1]*solution behavior	.	.	.	.	.	0.067	0.011	7672	6.02	<.0001	.	.	.	.	.
Time point [3-2]*solution behavior	.	.	.	.	.	0.112	0.019	7665	5.83	<.0001	.	.	.	.	.
Test completion time	.	.	.	.	.	.	.	.	.	.	0.002	0.000	7675	16.14	<.0001
Time point [2-1]*test completion time	.	.	.	.	.	.	.	.	.	.	0.001	0.000	7672	7.48	<.0001
Time point [3-2]*test completion time	.	.	.	.	.	.	.	.	.	.	-0.000	0.000	7665	-1.54	0.1228

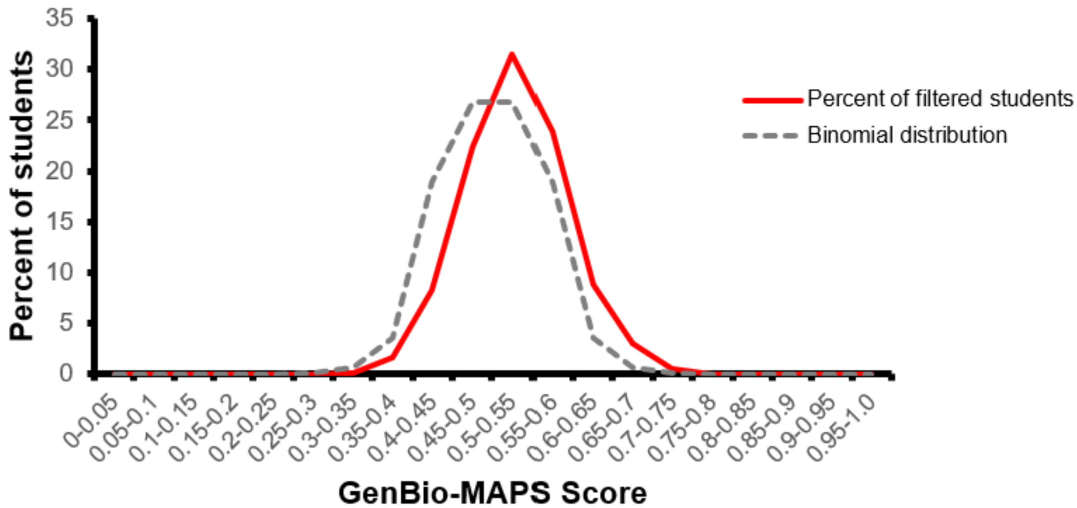
. Parameter not included in the model

Supplemental Table 1.8: Standard least squares linear regression models[a] on the effects of question display order on persistence behaviors and question score

Model outcome variable	Question display order estimate[b]	p	R^2
Question solution behavior	-0.0109	<0.0001	0.5072
Question time (minutes)	-0.0588	<0.0001	0.3676
Question score	-0.0034	<0.0001	0.3069

[a] Question solution behavior ~ student + GenBio-MAPS question + display order; Question time ~ student + GenBio-MAPS question + display order; Question score ~ student + GenBio-MAPS question + display order

[b] The estimate represents the change in the proportion of students using solution behavior on a question, the amount of time per question, or the proportion of correct responses as a student moves to each subsequent question.



Supplemental Figure 1.1: Distribution of GenBio-MAPS scores from students who were removed by our dual motivation filter compared to a binomial distribution arising from random responses. The red line represents the scores of students removed by the dual motivation filter who had demonstrated unmotivated behavior through low solution behavior or short test completion time. The gray dotted line represents a binomial distribution based on a 50% chance of correctly responding to 67 T/F statements, which represents the average number of statements seen by filtered students.

CHAPTER 2: HOW ADMINISTRATION STAKES AND SETTINGS AFFECT STUDENT BEHAVIOR AND PERFORMANCE ON A BIOLOGY CONCEPT ASSESSMENT[2]

ABSTRACT

Biology instructors use concept assessments in their courses to gauge student understanding of important disciplinary ideas. Instructors can choose to administer concept assessments based on participation (i.e., lower stakes) or the correctness of responses (i.e., higher stakes), and students can complete the assessment in an in-class or out-of-class setting. Different administration conditions may affect how students engage with and perform on concept assessments, thus influencing how instructors should interpret the resulting scores. Building on a validity framework, we collected data from 1578 undergraduate students over 5 years under five different administration conditions. We did not find significant differences in scores between lower-stakes in-class, higher-stakes in-class, and lower-stakes out-of-class conditions, indicating a degree of equivalence among these three options. We found that students were likely to spend more time and have higher scores in the higher-stakes out-of-class condition. However, we suggest that instructors cautiously interpret scores from this condition, as it may be associated with an increased use of external resources. Taken together, we highlight the lower-stakes out-of-class condition as a widely applicable option that produces outcomes

---

[2]This research was first published with minor formatting differences as Uminski, C., Hubbard, J. K., & Couch, B. A. (2023). CBE—Life Sciences Education*,* 22(2), ar27. https://www.lifescied.org/doi/10.1187/cbe.22-09-0181

similar to in-class conditions, while respecting the common desire to preserve classroom instructional time.

INTRODUCTION

Instructors and programs commonly use assessments to measure student performance and identify ways to improve student learning (National Research Council, 2003). Instructors can develop their own assessments or use publicly available instruments, such as published concept inventories or concept assessments. Concept assessments are constructed by a research team and designed to target common student misconceptions about important concepts within a topic or discipline (Adams and Wieman, 2011). The research that goes into developing a concept assessment allows instructors to use data from these instruments to diagnose student understanding of course content without requiring a large investment of time for assessment development or grading (Knight, 2010).

In deploying concept assessments, instructors need to identify administration conditions that fit within their course context while providing a valid reflection of student understanding. Administration conditions refer to how and where students complete a concept assessment and include the stakes assigned to student scores (i.e., the impact of the assessment on course grades) and the setting in which the testing session occurs, which often dictates the degree of associated proctoring. Differences in administration conditions can influence how students behave and perform on the assessment (American Educational Research Association et al., 2014). For example, lower-stakes grading in which students do not receive any course credit or receive participation credit may elicit lower test-taking effort, leading to lower scores (Wise and DeMars, 2005; Cole and Osterlind, 2008). Higher-stakes grading, such as when students are scored based on the correctness of their answers, may encourage greater test-taking effort and produce higher scores (Cole and Osterlind, 2008), but with the caveat that students may attain these higher scores by leveraging external resources (Munoz and Mackay, 2019). Disparities in scores between proctored and unproctored settings further indicate that students are likely using different test-taking behaviors under these different conditions (Carstairs and Myors, 2009; Alessio et al., 2017; Steger et al., 2020).

Concept assessment developers offer a variety of recommended administration conditions that they deem appropriate for maximizing student test-taking effort while minimizing threats to score validity. Some suggest administering instruments under lower-stakes in-class conditions (Kalas et al., 2013) or as in-class formative assessments (Bretz and Linenberger, 2012; McFarland et al., 2017). Other concept assessment developers recommend higher-stakes in-class conditions (Anderson et al., 2002; Smith et al., 2012). Several suggest lower-stakes out-of-class conditions (Bowling et al., 2008; Marbach-Ad et al., 2009; Couch et al., 2015), and a few indicate that the instruments should be embedded within the final exam (Smith et al., 2008; Shi et al., 2010). Previous work in upper-division biology courses compared in-class and out-of-class performance under low-stakes conditions (Couch and Knight, 2015); however, this type of comparison has not occurred across the entire set of recommended administration conditions or in lower-division courses in which there may be less direct connection between course content and students' prospective careers. Given the wide range of recommendations and the associated lack of empirical comparisons, there remains a need to determine how different administration conditions influence student behaviors and performance on concept assessments (AERA et al., 2014).

Theoretical Framework

We use a validity framework (Messick, 1987, 1989) as a basis for evaluating and interpreting biology concept assessment scores across different administration conditions. In our study, we interpret student behavior and performance to make inferences about student understanding of foundational concepts in introductory molecular and cell biology. According to Messick (1987), score interpretation should account for the context of how the construct is measured (i.e., the assessment instrument), the situational context of the assessment (i.e., external environmental influences), and the interplay between those two contexts, and it should be aligned to a unified validity theory.

In our case, the measurement and situational contexts refer to the Introductory Molecular and Cell Biology Concept Assessment (IMCA; Shi et al., 2010) and the administration conditions for the concept assessment, respectively. We consider associated validity evidence with respect to six aspects of unified validity: content validity, substantive validity, structural validity, generalizability, external validity, and consequential aspects of construct validity (Messick, 1989). Some aspects of this theory, such as content validity (i.e., test content is relevant and covers the specified domain), substantive validity (i.e., respondents engage with the test items as theorized), and structural validity (i.e., scoring structure is aligned to the intended construct), are more related to the process of assessment development. In developing the IMCA, the researchers provided evidence of content, substantive, and structural validity through expert reviews, student interviews, and statistical analysis of student scores (Shi et al., 2010).

We focus here on evaluating evidence of generalizability, external validity, and consequential aspects of construct validity when the IMCA is administered under different stakes and settings. Generalizability reflects the extent to which measurement properties and score interpretations apply across settings. External validity refers to the relationship between a test and other methods of measuring the same construct. Consequential aspects of construct validity concern the implications of score interpretation as a basis for action, with particular attention to the potential for invalidity to propagate bias. In our conceptual model (Figure 2.1), we hypothesize that different administration conditions elicit different student behaviors, such as their test-taking effort and external resource use. We make inferences about how students engaged with the assessment based on test completion time, concept assessment score, and the relationship of concept assessment score to scores on course unit exams with similar learning goals. These behavioral indicators thereby provide evidence for score validity interpretation under the various conditions.

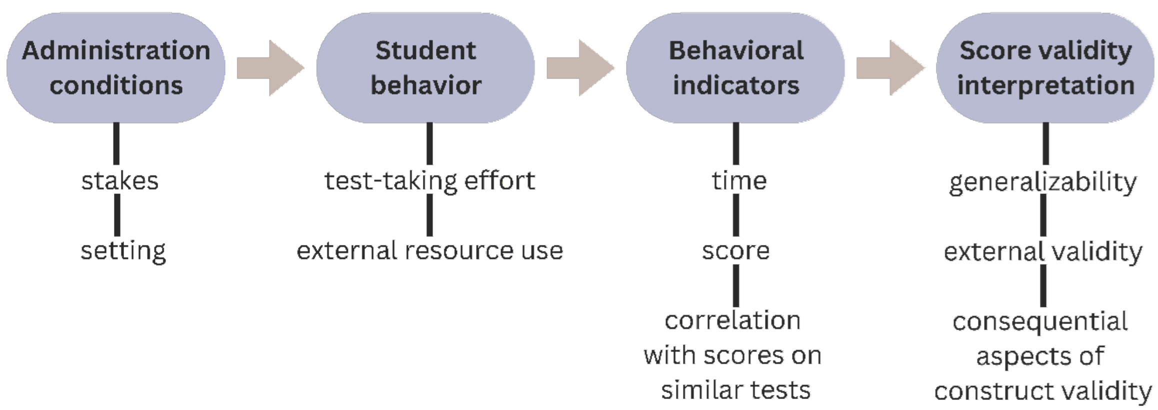


Figure 2.1: Conceptual model for score validity evidence and interpretation. This study aims to interpret how the situational context of an assessment (i.e., administration conditions) affects student behavior, indicated through test completion time, concept assessment score, and the correlation of concept assessment score to scores on course unit exams that assess similar learning goals. We use these behavioral indicators as evidence for interpreting score validity in each administration condition.

The administration conditions in this study vary systematically in the stakes and setting under which students complete the concept assessment, which we predict will elicit certain student behaviors (Figure 2.2). Given the desire for students to achieve high grades in their courses, we anticipate that increasing the assessment stakes leads students to expend greater effort, potentially reflected in students spending more time on the task (Wise and Kong, 2005). Higher stakes may also increase the tendency for students to seek external resources (e.g., peers, course materials, Internet resources) as a means to boost their scores, but this behavior also depends on the extent to which students perceive they will be penalized (Murdock and Anderman, 2006). In this way, the proctored in-class and unproctored out-of-class settings principally shape whether students can access and use external resources.

In our study, we examined five administration conditions: four “pre-final” conditions that took place during the last week of a course and one condition in which the concept assessment was embedded in the final exam. The four pre-final conditions (i.e., lower-stakes in-class, higher-stakes in-class, lower-stakes out-of-class, and higher-stakes out-of-class) differed substantively from the final exam condition, which was administered later in the course schedule, was delivered on paper rather than an electronic survey, was embedded within an exam, and had a higher point value in the overall course grade. For these reasons, we primarily consider the pre-final conditions and use the final exam condition as a comparative reference group. In the following sections, we apply our validity framework to describe how the pre-final and final exam conditions may influence student behavior and concept assessment score interpretation.

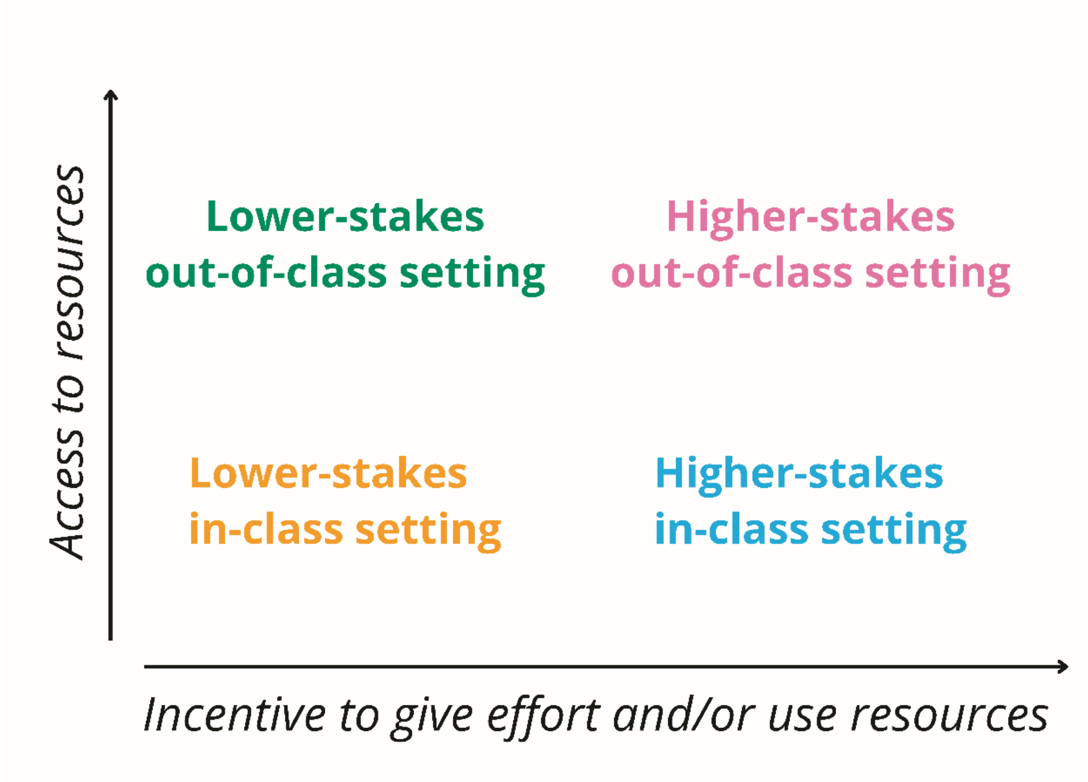


Figure 2.2: Administration conditions within our theoretical framework. We designed concept assessment administration conditions to reflect the various dimensions with our underlying theoretical framework. Compared with the lower-stakes (participation-graded) conditions, the higher-stakes (correctness-graded) conditions provide students with a greater impetus to give effort as well as an increased incentive to use external resources. Compared with the proctored in-class setting, the unproctored out-of-class setting provides students with greater access to external resources. We view student behavior as the product of a student's test-taking effort and associated incentive to use and access to use external resources.

Lower-Stakes In-Class: Because students receive credit based on participation, the lower stakes generate little extrinsic incentive for students to achieve a high score. Although this minimizes the incentive to use external resources, it may also result in low test-taking effort (Wise and DeMars, 2005). Low test-taking effort threatens valid score interpretation, because it may underestimate student knowledge, and it can be detected in assessments by identifying characteristically low completion times (Wise and Kong, 2005; Uminski and Couch, 2021). Research associating lower stakes with decreased effort has mostly been conducted with general education tests (Schiel, 1996; Hoyt, 2001; Sundre and Wise, 2003; Wise and Kong, 2005; Thelk et al., 2009), but this pattern may not hold for disciplinary assessments with more relevance or meaning to the test-taker. As effort partially arises from the importance an individual assigns to a task (Eccles et al., 1983; Wigfield and Eccles, 2000), when the content falls within students' disciplinary domain and they perceive completing the assessment to support their learning, students may place a higher importance on achieving a high score. Thus, they may not exhibit the lower-effort behavior traditionally associated with this condition.

Higher-Stakes In-Class: The higher stakes created by grading students based on answer correctness give students an extrinsic goal that can lead to higher scores (Wolf and Smith, 1995; Cole and Osterlind, 2008). While extrinsic goals may elicit greater effort and higher scores (Wise and DeMars, 2005; Liu et al., 2012), the increased score in this administration condition may also stem from students using external resources as a strategy for attaining their extrinsic goals. However, the in-class setting enables proctors (e.g., instructors, teaching assistants) to limit this strategy (Cizek, 1999), thus mitigating score increases due to external resource use.

Lower-Stakes Out-of-Class: Because students receive participation credit, their effort primarily depends on their intrinsic desire to do well on the assessment. Students who place a high intrinsic value on a task may be more cognitively engaged while performing the task (Pintrich and de Groot, 1990). The intrinsic value of a lower-stakes assessment given outside class time may also depend on whether the instructor encourages students to see the task as useful and important to their learning (Cole et al., 2008). In this lower-stakes out-of-class condition, students are likely to have low extrinsic incentive to use external resources despite having access in this unproctored condition. These features mirror the lower-stakes in-class condition, but the out-of-class setting may present additional time constraints or other challenges that prevent students from giving a full effort. In upper-division courses, we found that concept assessment scores under lower stakes were similar across in-class and out-of-class settings (Couch and Knight, 2015), but we do not know whether this similarity occurs for introductory courses.

Higher-Stakes Out-of-Class: The increased incentive to use and access resources potentially spurs notable differences in student behavior. This condition pairs an extrinsic incentive to achieve a high score with a low risk that external resource use will be detected, thereby presenting students with a relevant cause and potential means to improve their scores. Students using external resources may be spending additional time locating relevant information, which may be reflected in longer amounts of time spent on the assessment. While using external resources represents an important skill for students to develop, instructors often seek to measure unaided student knowledge under conditions without access to peers, textbooks, websites, or other information. Student use of external resources is of particular concern, because it may artificially inflate scores relative to what students would have achieved on their own (Tippins et al., 2006; Carstairs and Myors, 2009). These inflated scores threaten score validity, because they cannot be easily interpreted for their intended purposes of diagnosing student learning, may mask areas of student misunderstanding, and may not provide accurate feedback to instructors about their teaching and curricula (Munoz and Mackay, 2019).

Final Exam: Instructors may choose to administer concept assessments on the final exam to encourage students to take the assessment seriously and maximize participation rates (Smith et al., 2012). Concept assessments embedded within final exams represent a form of summative assessment. Students view the summative assessment as a culminating evaluation of their individual learning, rather than as a formative tool to identify knowledge gaps for personal or course improvement. While the final exam condition is similar to the higher-stakes in-class condition in that they both present an extrinsic incentive for students to achieve a high score in a proctored setting, the final exam carries a much higher importance to students in terms of its influence on overall course grade. Given the summative role of the final exam and its weight in course grades, students will be incentivized to spend time studying, and the scores from concept assessments administered in this condition likely reflect that additional test preparation.

Research Question: To date, there has been little empirical work to determine the impact of concept assessment administration conditions in the context of an undergraduate science course. Thus, we studied the effects of stakes and settings by systematically varying administration conditions over consecutive semesters. By comparing across administration conditions, we sought to address one overarching research question: How do administration stakes and settings affect student test-taking behavior and performance and influence interpretation of student scores on a biology concept assessment?

METHODS

Experimental Context

We compared five administration conditions over 5 years in a high-enrollment introductory molecular and cell biology course at a large midwestern research university.

The course included preclass homework, in-class formative assessments using an audience response system (i.e., clickers), and postclass homework quizzes. In addition to the final exam, the course had four unit exams that were administered on paper during class time and contained a mix of multiple-choice, multiple true-false, and open-ended questions. The unit exams demonstrated evidence of acceptable reliability, with Cronbach's alpha values above 0.75. A total of 1799 students were enrolled during the study period. After data processing, our sample contained responses from 1578 students who consented to share their data for research purposes, representing 88% of the total enrollment (see Table 2.1 for demographic information). While demographic information is provided to represent the study sample, our study did not seek to explore additional

Table 2.1: Demographic characteristics of students in the study

Demographic categories[a]	**n**	**%**[b]
Gender		
Female	916	61.7
Male	568	38.3
Race/ethnicity[c]		
Non-underrepresented	1229	83.5
Underrepresented	242	16.5
Generation status[d]		
Continuing-generation	940	68.7
First-generation	429	31.3
Class rank		
First-year	858	57.9
Sophomore	358	24.1
Junior	198	13.4
Senior	63	4.2
Non-degree seeking	6	0.4

[a] Information was obtained from the institution research office. Information was not available for every student.
[b] Percentages are calculated from the available demographic information.
[c] We use the term "underrepresented" to reflect racial/ethnic groups that have faced disproportionate challenges within STEM disciplines, including Black/African American, Hispanic/Latinx, American Indian/Alaskan Native, and Native Hawaiian/Pacific Islander. This grouping is not intended to obscure the unique histories and identities of any group.
[d] Students were considered first-generation if neither of their parents received a bachelor's degree, while continuing-generation students had one or both parents with a bachelor's degree.

associations with demographic characteristics. This research was given exempt status by the University of Nebraska–Lincoln (protocol 14314).

Preliminary Item Metrics and Development of Half-Length Instruments

We first embedded and scored the full-length IMCA instrument as part of the final exam in 2014, which students completed on paper in a proctored classroom setting (Figure 2.3). The IMCA consists of 24 multiple-choice items aligned with course learning objectives and unit exams. We calculated score as the proportion of items answered correctly. We calculated item difficulty (i.e., the proportion of students answering the question correctly) as the total number of correct responses divided by the total number of responses to the item, and item discrimination (i.e., a measure of how well a question

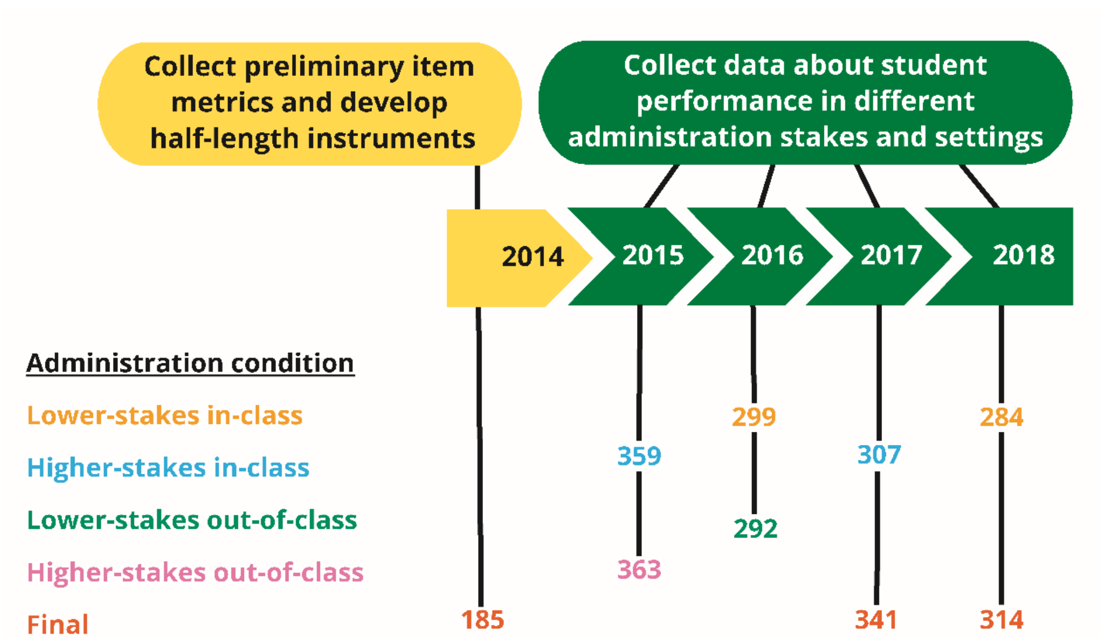


Figure 2.3: Experimental design and sample size for each administration condition. We collected data over the course of 5 years. The first-year (2014) data informed the development of half-length instruments. For the next 4 years (2015–2018), we administered the instruments in two different conditions per year and collected data about student behavior and performance. In a given year, each student saw a different instrument version in the two respective conditions.

distinguishes the highest-scoring and lowest-scoring students) as the difference in difficulty between the upper third of respondents and the lower third of respondents. The mean IMCA score was 0.67 ± 0.01 SEM. The difficulty and discrimination values for each item on the IMCA are reported in Supplemental Table 2.1. Student IMCA score was correlated with their average score on the four unit exams from the course ($r = 0.75$, $p < 0.001$), which provides evidence of convergent external validity for the IMCA regarding its ability to assess student knowledge in the given course context. Cronbach's alpha for the full-length IMCA was 0.84, which indicates acceptable reliability (Downing, 2004).

The 2014 administration informed our development of half-length IMCA instruments, henceforth referred to as version A and version B. Based on the original item-naming scheme and associated learning goals (Shi et al., 2010), version A contained items 1, 3, 9, 11, 13, 15, 17, 19, 20, 21, 23, and 24. Version B contained items 2, 4, 5, 6, 7, 8, 10, 12, 14, 16, 18, and 22. Both instruments contained items aligned with learning goals related to features of microorganisms, properties of water, thermodynamics of reactions, solubility, flow of matter and energy, and gene expression. Version A additionally assessed concepts related to evolution and information storage, and version B had a set of items assessing macromolecular structure. This distribution ensured that each instrument assessed content from across the course. Within the 2014 data, scores on the two instruments were correlated ($r = 0.70$, $p < 0.001$), and the average scores on the two instruments were similar (version A mean = 0.66 ± 0.02 SEM, version B mean = 0.68 ± 0.02 SEM, paired *t* test $p = 0.10$). Cronbach's alpha values were 0.63 and 0.80 for versions A and B, respectively. Version B contained items 4, 5, 6, 7, and 8, all sharing a common stem, which likely explains the higher internal consistency.

Administration of Half-Length Instruments

For the pre-final administration conditions, students completed the half-length instruments via Qualtrics survey during the last week of the course. The instructor informed students during class time that the task(s) would serve as practice for the final exam, told students that the activity would be credited with up to a 5% bonus on the final exam grade, explained how the assessments would be graded (i.e., lower-stakes participation grading or higher-stakes grading based on response correctness), and asked students not to consult peers or other external resources. This message was reiterated accordingly on the first page of the Qualtrics surveys. The lower-stakes conditions contained the text: “The following survey contains practice questions for the cumulative portion of the final exam. You can earn up to 5% points extra credit for the cumulative final by completing the practice questions. You will not be graded based on the correctness of your responses. Please use only the information in your own head and do not consult your peers or any other external resources.” The higher-stakes administrations had identical text, except the second and third sentences were changed to: “You can earn up to 5% points extra credit for the cumulative final based on how many questions you answer correctly.”

Students saw the items in a random order and could not return to questions once an answer was submitted. For the in-class administrations, the instructor provided students with as much time as they needed to complete the concept assessment, and the instructor and teaching assistants proctored while students completed the instrument. For the out-of-class administrations, students completed the instrument at a time and location of their choosing within 3 days after the activity was announced during class time. For the final exam condition, the instrument was embedded as the first 12 items on the exam, and students completed the exam on paper in the proctored classroom setting. Students could complete the questions on the final exam in any order and return to previous questions. The embedded IMCA instrument comprised 40% of the final exam points.

We implemented two different administration conditions each year (Figure 2.3), taking advantage of the course being taught as two separate sections (i.e., two class meeting times) during these 4 years. Each year, students in the first section completed one half-length instrument (e.g., version A) in the in-class setting and the other half-length instrument (e.g., version B) in the out-of-class setting or on the final exam, depending on the year. Students in the second section completed the reciprocal instrument in the same respective settings (e.g., they completed version B in the in-class setting and version A in either the out-of-class setting or on the final exam). The grading stakes were alternately varied by year to achieve the full range of conditions across the 4 years.

Data Processing and Statistical Analysis

Our data set contained responses from students who consented to release survey data, completed at least 80% of the instrument, and submitted during the intended time window. We recorded page-level response times for pre-final surveys. All items appeared on separate survey pages, except for items 4–8 and 19 and 20, which needed to appear as item groups. Approximately 0.07% of page times exceeded 15 minutes and were replaced with the mean time for that page. Total test completion time was calculated by summing the individual item page times for each student. We could not record time data when the instrument was administered on paper in the final exam condition.

We conducted linear mixed-effects models to analyze concept assessment completion time and score with student as a random effect. When tested as main effects, demographic variables (gender, race/ethnicity, and first-generation status) were excluded during model selection based on Akaike information criterion (AIC) values or were not significant predictors ($p > 0.05$), so these variables were not retained as covariates. To account for student biology proficiency, we included the average of the four unit exam scores for each student as a covariate in models predicting score. Full models are included in the footnotes of the corresponding results tables (Table 2.2; Supplemental Table 2.2). We calculated Pearson correlation coefficients between student IMCA scores and average unit exam scores, followed by pairwise Fisher's *z*-tests to evaluate the statistical significance of differences between correlation values.

Table 2.2: Linear mixed effects model[a] on the effects of administration condition on concept assessment score

Parameter	**Sum Sq**	**Mean Sq**	***df***	***F***	***p***
Administration condition	4.561	1.140	2175.3	42.716	<.001
Average exam score	41.738	41.738	1	1563.470	<.001
Post-hoc comparisons					
Contrast	**Estimate**	**SE**	***df***	***t***	***p***
Final Exam – Higher In	0.085	0.01	2060	9.16	<.001
Final Exam – Higher Out	-0.014	0.01	2542	-1.27	.711
Final Exam – Lower In	0.069	0.01	2065	7.12	<.001
Final Exam – Lower Out	0.098	0.01	2541	8.04	<.001
Higher In – Higher Out	-0.099	0.01	1751	-9.12	<.001
Higher In – Lower In	-0.016	0.01	2552	-1.68	.448
Higher In – Lower Out	0.013	0.01	2553	1.04	.837
Higher Out – Lower In	0.083	0.01	2553	7.17	<.001
Higher Out – Lower Out	0.112	0.01	2553	8.24	<.001
Lower In – Lower Out	0.029	0.01	1756	2.42	.109
Model $R^2 = 0.49$					
[a]Score ~ administration condition + average unit exam score + (1 | ID)					

Data processing and statistical analysis was completed using R v. 4.1.1 (R Core Team, 2021) and several packages: tidyverse (Wickham et al., 2019), rstatix (Kassambara, 2021), psych (Revelle, 2021), lmerTest (Kuznetsova et al., 2017), performance (Lüdecke et al., 2021), ShinyItemAnalysis (Martinkova and Drabinova, 2018), emmeans (Lenth, 2022), and diffcor (Blötner, 2022).

RESULTS

The Higher-Stakes Out-of-Class Condition Produced the Longest Completion Times

We observed a few patterns in the distributions of assessment completion times (represented as violin plots in Figure 4) across administration conditions. For the in-class settings, the bulk of students (89%) completed the instrument in roughly 3–20 minutes. For the out-of-class settings, many students (70%) fell within this same range, but a small proportion (9%) took longer than 20 minutes, creating a noticeable skew in the distributions. This skew may reflect students who multitasked during the activity, thereby conflating their completion time with time dedicated to extraneous tasks. The lower-

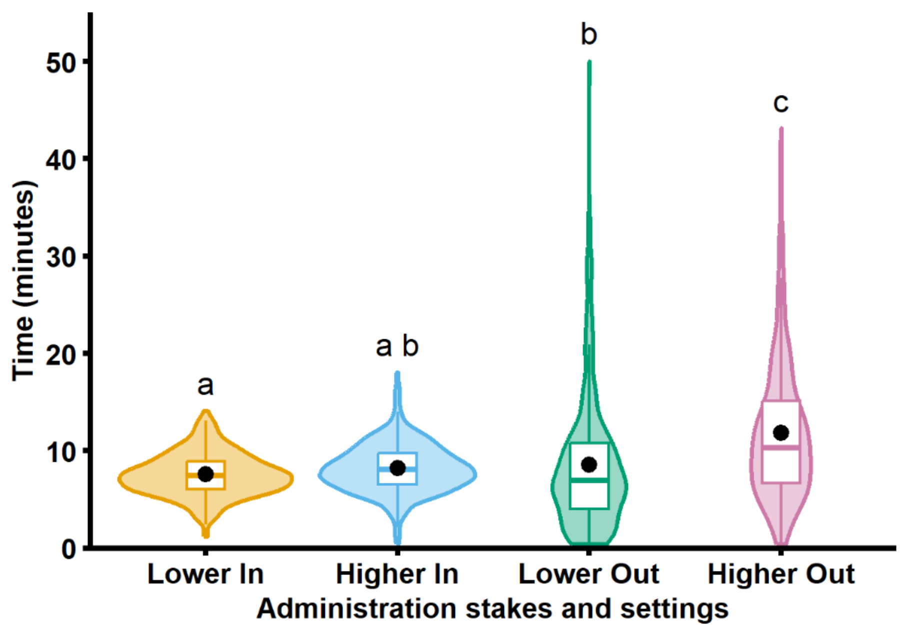


Figure 2.4: Test completion time in each administration condition. Completion times represent the sum of time spent on each page of the concept assessment. Completion time data were not collected when the concept assessment was administered on paper in the final exam condition. Violin plots show the distribution of completion times in each administration condition. Boxes represent the 25th, 50th, and 75th percentiles. Whiskers represent 5th and 95th percentiles. The dot represents the mean times. Conditions sharing the same letters were not significantly different ($p \geq 0.05$), as determined by the post hoc tests shown in Supplemental Table 2.3. Lower In, lower-stakes in-class; Higher In, higher-stakes in-class; Lower Out, lower-stakes out-of-class; Higher Out, higher-stakes out-of-class.

stakes out-of-class distribution also included 17% of students who completed the instrument in less than 3 minutes, likely an inadequate amount of time to read and thoughtfully respond to the items. Meanwhile, the higher-stakes out-of-class distribution was shifted noticeably upward relative to the other pre-final conditions.

We used a linear mixed-effects model to analyze completion times across administration conditions (Supplemental Table 2.2). We detected an effect of administration condition, so we conducted post hoc pairwise comparisons. We found that the two in-class conditions had similar completion times (lower-stakes in-class mean = 7.6 minutes ± 0.1 SEM, higher-stakes in-class mean = 8.2 minutes ± 0.1 SEM, p = 0.053). The lower-stakes out-of-class condition (mean = 8.6 minutes ± 0.4 SEM) was increased relative to the lower-stakes in-class condition ($p < 0.01$) but not different from the higher-stakes in-class condition ($p = 0.73$). Finally, the higher-stakes out-of-class condition (mean = 11.8 minutes ± 0.3 SEM) yielded longer completion times than all the other pre-final conditions ($p < 0.001$).

The Higher-Stakes Out-of-Class Condition Led to the Highest Scores

Students displayed a broad distribution of assessment scores (represented as violin plots in Figure 2.5) across the administration conditions. The lower-stakes in-class, higher-stakes in-class, and lower-stakes out-of-class distributions appeared similar, with the bulk of scores (71%) falling between 0.25 and 0.75. Conversely, the higher-stakes out-of-class score distribution was shifted upward. The majority of scores in this condition (50%) fell between 0.50 and 0.90, with an additional 12% of students achieving scores between 0.90 and 1.0. Scores in the final exam condition exhibited a similar upward shift, but also presented a noticeable proportion of scores in the 0.25 and 0.50 range.

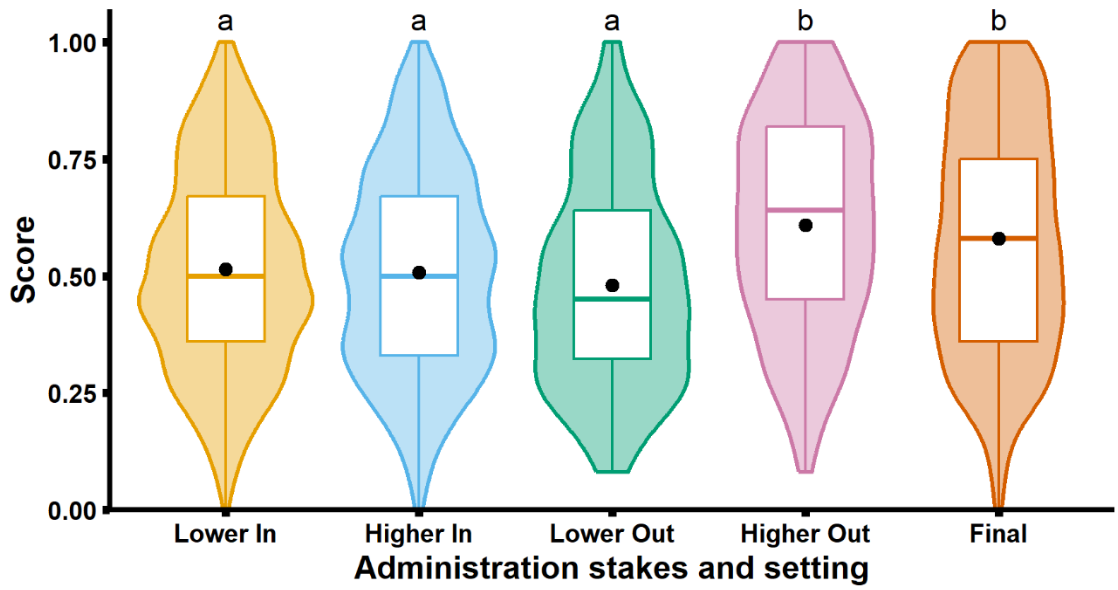


Figure 2.5. Concept assessment scores in each administration condition. Violin plots show the distribution of scores in each administration condition. Boxes represent the 25th, 50th, and 75th percentiles. Whiskers represent 5th and 95th percentiles. The dot represents the mean scores. Conditions sharing the same letters were not significantly different ($p \geq 0.05$), as determined by the post hoc tests shown in Table 2.2. Lower In, lower-stakes in-class; Higher In, higher-stakes in-class; Lower Out, lower-stakes out-of-class; Higher Out, higher-stakes out-of-class.

We used a linear mixed-effects model to analyze scores across administration conditions (Table 2.2). In this case, we included student average score on the other four unit exams as a covariate. Thus, the model enabled us to estimate how well students performed in a given condition, relative to how they would have been expected to score based on their broader exam performance. We detected an effect of administration condition and average exam score. Post hoc comparisons revealed no differences between the lower-stakes in-class (mean = 0.51 ± 0.01 SEM), higher-stakes in-class (mean = 0.51 ± 0.01 SEM), and lower-stakes out-of-class (mean = 0.48 ± 0.01 SEM) conditions ($p >$ 0.05). The higher-stakes out-of-class condition (mean = 0.61 ± 0.01 SEM) produced the highest scores, with the model estimating that scores in this condition were 8–11% above the other pre-final conditions (p < 0.001). Meanwhile, the final exam (mean = 0.58 ± 0.01 SEM) was estimated to produce scores 7–10% above these other pre-final conditions ($p < 0.001$) for all but the higher-stakes out-of-class condition ($p = 0.71$).

Higher-Stakes Out-of-Class Scores Correlated the Least with Unit Exam Performance

As part of exploring assessment properties, scores on a particular instrument are often compared with performance on a separate task or instrument (i.e., convergent validity). Stronger correlations between scores serve as an indication that the two activities measure similar attributes, whereas weaker correlations suggest that the two activities capture different constructs or processes (AERA et al., 2014). Within the course, the four unit exams represented additional measures of student biology proficiency. Students likely expended considerable effort to prepare for and complete the unit exams, which comprised a large proportion of the course grading scheme. Furthermore, because the unit exams occurred during class time under proctored conditions, the resulting scores should reflect each student's independent proficiency (i.e., students were prohibited from using external resources).

Thus, we examined correlations between student IMCA scores in the various administration conditions and average unit exam scores (Figure 2.6). All four pre-final conditions yielded scores that correlated with unit exam scores to a moderate degree, with correlation coefficients ranging from 0.54 to 0.71. Fisher's z-tests revealed nuanced differences in the extent to which the various concept assessment administration conditions aligned with unit exam performance (Supplemental Table 2.3). We first consider the impact of stakes within each setting. The two in-class conditions each correlated with unit exam performance to the same degree (lower-stakes in-class $r = 0.63$, higher-stakes in-class $r = 0.64$, $p = 0.41$), and the two out-of-class conditions each correlated with unit exam performance to the same degree (lower-stakes out-of-class $r = 0.59$, higher-stakes out-of-class r = 0.54, $p = 0.16$). We next consider the impact of setting for the given stakes. Under lower stakes, we did not see a difference in correlation with unit exam performance when moving from in-class to out-of-class settings ($p = 0.19$). However, under higher stakes, we observed a higher correlation with unit exam performance when the concept assessment was administered in the in-class setting than in the out-of-class setting ($p < 0.01$). Finally, we observed the highest correlation between concept assessment score and average exam score in the final exam condition ($r = 0.71$, $p < 0.01$).

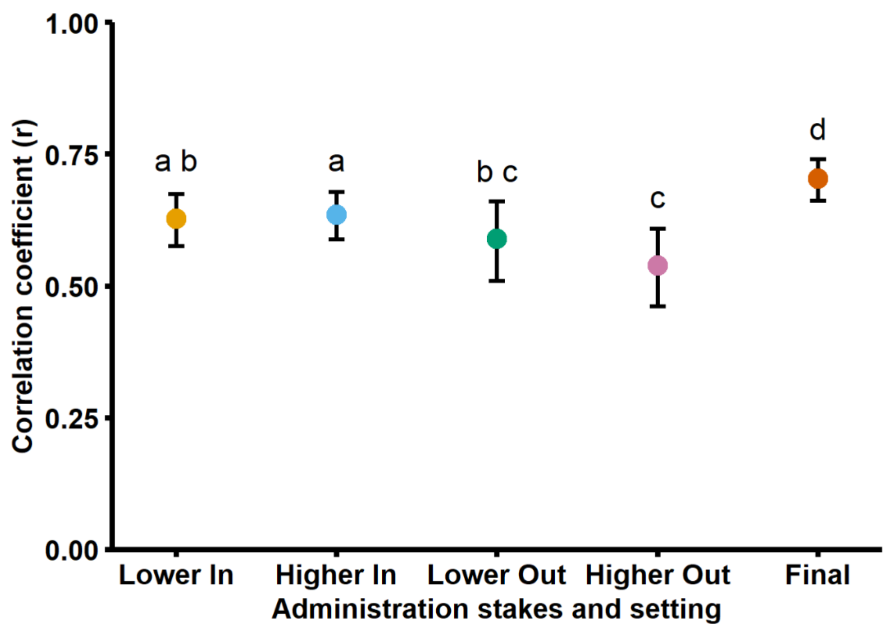


Figure 2.6: Correlation between concept assessment score and average course exam score for each administration condition. Dots represent correlation coefficients and whiskers represent the 95% confidence interval. Conditions sharing the same letters did not have significantly different correlation values ($p \geq 0.05$), as determined by the Fisher's *z* transformations shown in Supplemental Table 2.3. Lower In, lower-stakes in-class; Higher In, higher-stakes in-class; Lower Out, lower-stakes out-of-class; Higher Out, higher-stakes out-of-class.

Item Difficulty and Discrimination

Across administration conditions, the IMCA items had adequate values for item difficulty and discrimination (Ebel and Frisbie, 1986; Supplemental Figure 2.1). The exceptions were items 15 and 20, which were the most difficult for students (0.20–0.31 and 0.13–0.17, respectively) and had the lowest discrimination values (0.06–0.12 and 0.12–0.23, respectively). Items 15 and 20 also had low difficulty and discrimination values in the initial IMCA publication but were retained because they reflected that students struggle with particular concepts (Shi et al., 2010). The greatest variation in item difficulty and discrimination across conditions occurred for items 4–8, a set of matching items that addressed one learning goal related to recognition of monomer structures. These items shared a common question stem and answer options that all appeared on a single test page, which can explain why these items tended to vary similarly across the administration conditions.

DISCUSSION

Biology instructors have options for how they administer concept assessments in their courses, and each administration condition has the potential to affect student behavior and performance in ways that affect score interpretation. According to our theoretical framework, administration stakes and settings have the potential to influence test-taking effort and external resource use, behaviors that can shape the extent to which assessment scores accurately reflect student understanding of biology concepts. Because instructors and researchers use data from concept assessments to make decisions about course effectiveness, it is important for them to select optimal administration conditions and to account for potential impacts of these conditions. Our study aimed to provide empirical data about student behavior and performance in different conditions to inform associated score interpretations.

The Two In-Class Conditions Produce Similar Student Behaviors and Performance

The lower-stakes in-class and higher-stakes in-class conditions were equivalent with respect to completion time, test score, and correlation with unit exam performance, suggesting a certain degree of generalizability across these conditions. For these conditions, we note that students were given as much time as they needed at the beginning of class to complete the instrument. The resulting completion times and test scores thus provide a baseline of how students behave and perform under conditions where they have been given time and space for the task.

Our finding that there was no difference in scores between lower-stakes and higher-stakes in-class assessments differs from previous work reporting higher scores for higher-stakes proctored assessments (Wolf and Smith, 1995; Wise and DeMars, 2005; Cole and Osterlind, 2008). This discrepancy may stem from these earlier studies using general education assessments, whereas our study used a discipline-specific instrument. Students enrolled in a course intended for life sciences majors may have placed a higher value on a discipline-specific concept assessment and may have been incentivized to perform well even under the lower-stakes conditions. These ideas resonate with another study finding that incentive structure (i.e., regular vs. extra credit) did not affect biology student performance on a natural selection instrument (Sbeglia and Nehm, 2022). Students in our lower-stakes condition may have derived additional incentive to achieve a high score from our framing of the IMCA questions as practice for the final exam. The lack of alignment with previous findings may also be linked to the small sample of existing studies in higher education that compare student performance on the same assessment instrument administered under both lower and higher stakes (Cole and Osterlind, 2008).

The Lower-Stakes Out-of-Class Condition Represents a Practical Alternative to In-Class Conditions

Class time represents a limited resource, and instructors often feel pressure to cover a wide breadth of content in biology courses (Wright et al., 2018). Instructors may also have legitimate concerns about using class time to administer an instrument that is being given for research purposes or that does not completely align with their course content, such as a program-level assessment (Couch et al., 2015, 2019; Summers et al., 2018; Semsar et al., 2019; Smith et al., 2019; Branchaw et al., 2020). As a result of these factors, they may choose to administer concept assessments outside class time to conserve instructional time. Our results suggest that instructors may see similar results outside class time as compared with the in-class setting, so long as they use lower-stakes participation grading. Indeed, we found that student scores in the lower-stakes out-of-class condition did not differ from either of the two in-class conditions. Furthermore, the lower-stakes out-of-class condition correlated with unit exam performance to a similar degree as the lower-stakes in-class condition. These results agree with our previous work in upper-division courses (Couch and Knight, 2015) and suggest that similarity in performance occurs across course levels for a low-stakes concept assessment administered in-class versus out-of-class. The similar student performance between lower-stakes in-class and lower-stakes out-of-class conditions could also stem from broader course experiences. Students in our study had extensive experience with other in-class and out-of-class assignments, which may have led them to develop habits that were manifested when they completed the concept assessment in the last week of class.

One potential limitation of the lower-stakes out-of-class condition lies in its association with low test-taking effort, as students may devote less outside time to this task graded based on participation. Despite these concerns, we observed that the distribution of lower-stakes out-of-class completion times overlapped considerably with the in-class settings, suggesting that many students gave roughly equivalent efforts across these conditions. However, we did observe that 17% of students did not take what we would consider an adequate time to answer the questions in the lower-stakes out-of-class condition, indicating that they likely rushed through the task. This finding adds an important caveat that this condition should not be considered completely generalizable with or equivalent to the in-class conditions. This behavior may explain the lower-stakes out-of-class scores having a slightly lower correlation and external validity with unit exam performance than the higher-stakes in-class scores, for which very few students took less than 3 minutes. Instructors and researchers may want to apply motivation-filtering processes to identify and remove scores from low-effort test takers (Wise and Kong, 2005; Uminski and Couch, 2021). Another potential challenge associated with out-of-class conditions comes from students having increased opportunity to leverage external resources, which undermines the validity of the assessment as a measure of independent proficiency (AERA et al., 2014). The similarity in score distributions compared with the in-class settings results suggests that students did not gain significant advantage from external resources in the lower-stakes out-of-class condition. While this remains an area for further exploration, we anticipate that external resource use is minimized when students are not graded based on answer correctness.

Higher-Stakes Out-of-Class Conditions May Produce Artificially High Scores

Students behaved and performed differently in the higher-stakes out-of-class condition, for which they had both the incentive to use and access to external resources. Indeed, students spent more time and had the highest scores in this condition. While these differences could have reflected students operating in a more relaxed environment or taking more time to individually think through the assessment questions, we hypothesize that the increased times and scores more likely stemmed from students finding and using external resources to answer the assessment questions. This hypothesis is supported by the comparatively lower completion times and scores in the higher-stakes in-class condition, in which students were given as much time as they needed but proctoring mitigated the opportunity to use external resources. Compared with the other pre-final conditions, the lower correlation and external validity with unit exam scores also provided evidence that the higher-stakes out-of-class condition led to the concept assessment measuring somewhat different cognitive processes or attributes, such as the willingness or ability to extract information from external resources. Our results align with previous research finding that students had inflated scores and spent longer amounts of time on assessments completed in higher-stakes unproctored conditions (Alessio et al., 2017) and provide additional support for the argument that proctored and unproctored assessments should not be deemed equivalent under higher-stakes conditions (Carstairs and Myors, 2009).

Understanding test-taking behaviors in out-of-class conditions remains an important area for investigation. While students may have cause and opportunity to use external resources in an unproctored high-stakes setting, the extent of such behaviors is not well understood (Tippins et al., 2006; Steger et al., 2020) and detecting the use of external resources is logistically difficult (Fisher and Katz, 2000). Test-takers are likely to have higher scores when the tasks on unproctored assessments are easy to find using Internet searches (Steger et al., 2020), due to being posted on online answer-sharing platforms (e.g., Chegg, Course Hero) or having content amenable to online answer discovery (Munoz and Mackay, 2019). While all of the IMCA answers can be readily found online, the higher scores for some of the IMCA questions, such as items 4–8 assessing identification of common monomer structures, suggests that the answers to some items might be easier to find online than others. Altogether, we caution against administering concept assessments under the higher-stakes out-of-class condition, because this condition likely overestimates independent student proficiency and creates an unfair advantage for students who use unapproved resources. These consequential aspects of construct validity can shape instructional choices and lead to students maintaining misunderstandings about foundational biology concepts. We also note that this finding calls important attention to the fairness of other homework assignments graded based on answer correctness.

Interpreting Concept Assessment Scores from Final Exam Administrations

The final exam represents an additional vehicle to administer a course-level concept assessment (Smith et al., 2008; Shi et al., 2010), but this option might not be appropriate in situations in which the instrument covers a narrow topic or does not align fully with the course content (e.g., program assessment). The instructor may also wish to use the final exam for other purposes or to give the final exam back to students after the semester. In our case, the final exam differed in several ways from the pre-final conditions (e.g., summative nature, preparation time, paper administration format, grade weight). Given these caveats, we interpret the final exam condition as a reference group providing a comparative basis for student performance, but we consider it to substantially differ in its applicability.

We found that scores from the final exam condition were higher than three of the pre-final conditions (i.e., lower-stakes in-class, higher-stakes in-class, lower-stakes out-of-class) but on par with the higher-stakes out-of-class condition. We speculate that the higher scores in the final exam condition likely reflected additional time that students spent preparing for the high-stakes summative exam. The IMCA and the course's final exam represent broad cumulative assessments of introductory molecular and cell biology concepts, so effective studying for the final exam would likely have increased student scores on the IMCA as well. In contrast, students were not expected to spend extensive time studying for the pre-final concept assessments. These results echo previous studies highlighting the potential effects of incentives and time frames for concepts assessments given toward the end of a term, a period when students may engage in particularly focused studying (Ding et al., 2008). While not tested in our study, student performance may remain stable for at least 2 weeks after the final exam (Sbeglia and Nehm, 2022). Student study behaviors and final exam performance may also have been affected by the experience of completing a half-length IMCA instrument in-class during the week before the final exam. Ideally, this experience of completing a short set of cumulative questions helped encourage students to begin studying and gave them a sense of the question types they might see on the final, even though no student saw the exact same questions (because they had the alternate version on the final).

Scores from the final exam condition also had the highest correlation with unit exam scores. This correspondence likely stemmed from the marked similarity between unit exams and the final exam. Given their high weight in the course grading scheme and timing throughout the course calendar, students would have made roughly the same types of preparations for each of these exams. These exams were all completed on paper in the same proctored setting, thereby standardizing any potential sources of construct-irrelevant variance, such as technology issues or environmental distractions. Finally, we note that the final exam condition and the higher-stakes out-of-class condition had the largest discrepancy in their correlations with unit exam performance ($r = 0.71$ vs. $r = 0.54$, $p < 0.001$), suggesting that their similar score distributions resulted from markedly different underlying processes.

CONCLUSIONS

Based on our theoretical framework, every concept assessment administration condition has the potential to alter student behavior in ways that affect score interpretation. We view optimal administration conditions as eliciting sufficient student effort while minimizing the incentive to use external resources or the opportunity to use external resources. We gathered evidence in the form of assessment time, score, and correlation with scores on course exams to inform our interpretations of student behaviors and performance in each administration condition. We discovered that the two in-class conditions yielded similar results, suggesting that either way represents a roughly equivalent approach to collect information about student understanding. The lower-stakes out-of-class condition produced scores similar to the in-class administration conditions while preserving instructional time and potentially minimizing external resource use. However, this condition may prompt lower effort from a small proportion of students, so instructors and researchers can decide if this downside outweighs the costs of using class time and can apply motivation filtering to remove responses that did not take sufficient time (Wise and Kong, 2005; Uminski and Couch, 2021). Our results suggest that instructors should avoid the higher-stakes out-of-class condition, as these scores may reflect external resource use. Artificially inflated scores from this condition may contribute to overestimates of student understanding with potential consequences for instruction and fairness in assessment practices. The final exam condition led to high scores and represents a potential option for gauging student understanding after a period of focused studying, although instructors need to consider the appropriateness of the assessment content and the degree to which it can be kept secure across sections and semesters. Instructors and researchers will have different needs and constraints depending on their course contexts and intended use of assessment scores, but they should carefully consider how their administration conditions might affect student performance and strive to keep their approach as similar as possible across course sections, academic years, or experimental groups.

Acknowledgements

We thank Kati Brazeal for her contributions and thank the students who participated in the research. This material is based upon work supported by the National Science Foundation (NSF) Graduate Research Fellowship (DGE-1610400) and Improving Undergraduate STEM Education Grant (DUE-1610621). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

REFERENCES FOR CHAPTER 2

Adams, W. K., & Wieman, C. E. (2011). Development and validation of instruments to measure learning of expert-like thinking. International Journal of Science Education, 33(9, 1289–1312. https://doi.org/10.1080/09500693.2010.512369

Alessio, H. M., Malay, N., Maurer, K., Bailer, A. J., & Rubin, B. (2017). Examining the effect of proctoring on online test scores. Online Learning, 21(1). https://doi.org/10.24059/olj.v21i1.885

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (2014). Standards for educational and psychological testing. Washington, DC: AERA.

Anderson, D. L., Fisher, K. M., & Norman, G. J. (2002). Development and evaluation of the Conceptual Inventory of Natural Selection. Journal of Research in Science Teaching, 39(10), 952–978. https://doi.org/10.1002/tea.10053

Blötner, C. (2022). diffcor: Fisher's z-tests concerning difference of correlations (R Package Version 0.7.1). Retrieved May 24, 2022, from https://CRAN.R-project.org/package=diffcor

Bowling, B. V., Acra, E. E., Wang, L., Myers, M. F., Dean, G. E., Markle, G. C., ... & Huether, C. A. (2008). Development and evaluation of a genetics literacy assessment instrument for undergraduates. Genetics, 178(1), 15–22. https://doi.org/10.1534/genetics.107.079533

Branchaw, J. L., Pape-Lindstrom, P. A., Tanner, K. D., Bissonnette, S. A., Cary, T. L., Couch, B. A., ... & Brownell, S. E. (2020). Resources for teaching and assessing the Vision and Change biology core concepts. CBE—Life Sciences Education, 19(2), es1. https://doi.org/10.1187/cbe.19-11-0243

Bretz, S. L., & Linenberger, K. J. (2012). Development of the enzyme–substrate interactions concept inventory. Biochemistry and Molecular Biology Education, 40(4), 229–233. https://doi.org/10.1002/bmb.20622

Carstairs, J., & Myors, B. (2009). Internet testing: A natural experiment reveals test score inflation on a high-stakes, unproctored cognitive test. Computers in Human Behavior, 25(3), 738–742. https://doi.org/10.1016/j.chb.2009.01.011

Cizek, G. J. (1999). Cheating on tests: how to do it, detect it, and prevent it. New York, NY: Routledge. https://doi.org/10.4324/9781410601520

Cole, J. S., Bergin, D. A., & Whittaker, T. A. (2008). Predicting student achievement for low stakes tests with effort and task value. Contemporary Educational Psychology, 33(4), 609–624. https://doi.org/10.1016/j.cedpsych.2007.10.002

Cole, J. S., & Osterlind, S. J. (2008). Investigating differences between low- and high-stakes test performance on a general education exam. Journal of General Education, 57(2), 119–130.

Couch, B. A., & Knight, J. K. (2015). A comparison of two low-stakes methods for administering a program-level biology concept assessment. Journal of Microbiology & Biology Education, 16(2), 178–185. https://doi.org/10.1128/jmbe.v16i2.953

Couch, B. A., Wood, W. B., & Knight, J. K. (2015). The Molecular Biology Capstone Assessment: A concept assessment for upper-division molecular biology students. CBE—Life Sciences Education, 14(1), ar10. https://doi.org/10.1187/cbe.14-04-0071

Couch, B. A., Wright, C. D., Freeman, S., Knight, J. K., Semsar, K., Smith, M. K., ... & Brownell, S. E. (2019). GenBio-MAPS: A programmatic assessment to measure student understanding of Vision and Change core concepts across general biology programs. CBE—Life Sciences Education, 18(1), ar1. https://doi.org/10.1187/cbe.18-07-0117

Ding, L., Reay, N. W., Lee, A., & Bao, L. (2008). Effects of testing conditions on conceptual survey results. Physical Review Special Topics—Physics Education Research, 4(1), 010112. https://doi.org/10.1103/PhysRevSTPER.4.010112

Downing, S. M. (2004). Reliability: On the reproducibility of assessment data. Medical Education, 38(9), 1006–1012. https://doi.org/10.1111/j.1365-2929.2004.01932.x

Ebel, R. L., & Frisbie, D. A. (1986). Essentials of educational measurement (4th ed.). Englewood Cliffs, NJ: Prentice-Hall.

Eccles, J. S., Adler, T. F., Futterman, R., Goff, S. B., Kaczala, C. M., Meece, J. L., & Midgley, C. (1983). Expectancies, values, and academic behaviors. In Achievement and achievement motivation (pp. 75–146). San Francisco, CA: Freeman.

Fisher, R. J., & Katz, J. E. (2000). Social-desirability bias and the validity of self-reported values. Psychology & Marketing, 17(2), 105–120. https://doi.org/10.1002/(SICI)1520-6793(200002)17:2<105::AID-MAR3>3.0.CO;2-9

Hoyt, J. E. (2001). Performance funding in higher education: The effects of student motivation on the use of outcomes tests to measure institutional effectiveness. Research in Higher Education, 42(1), 71–85. https://doi.org/10.1023/A:1018716627932

Kalas, P., O'Neill, A., Pollock, C., & Birol, G. (2013). Development of a meiosis concept inventory. CBE—Life Sciences Education, 12(4), 655–664. https://doi.org/10.1187/cbe.12-10-0174

Kassambara, A. (2021). rstatix: Pipe-friendly framework for basic statistical tests (0.7.0). Retrieved November 14, 2021, from https://CRAN.R-project.org/package=rstatix

Knight, J. (2010). Biology concept assessment tools: Design and use. Microbiology Australia, 31(1), 5–8. https://doi.org/10.1071/ma10005

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in linear mixed effects models. Journal of Statistical Software, 82(13), 1–26. https://doi.org/10.18637/jss.v082.i13

Lenth, R. V. (2022). emmeans: Estimated marginal means, aka least-squares means (R Package Version 1.7.4-1). Retrieved May 24, 2022, from https://CRAN.R-project.org/package=emmeans

Liu, O. L., Bridgeman, B., & Adler, R. M. (2012). Measuring learning outcomes in higher education: Motivation matters. Educational Researcher, 41(9), ar9. https://doi.org/10.3102/0013189X12459679

Lüdecke, D., Ben-Shachar, M. S., Patil, I., Waggoner, P., & Makowski, D. (2021). performance: An R package for assessment, comparison and testing of statistical models. Journal of Open Source Software, 6(60), 3139. https://doi.org/10.21105/joss.03139

Madsen, A., McKagan, S. B., & Sayre, E. C. (2017). Best practices for administering concept inventories. Physics Teacher, 55(9), 530–536. https://doi.org/10.1119/1.5011826

Marbach-Ad, G., Briken, V., El-Sayed, N. M., Frauwirth, K., Fredericksen, B., Hutcheson, S., ... & Smith, A. C. (2009). Assessing student understanding of host pathogen interactions using a concept inventory. Journal of Microbiology & Biology Education, 10(1), 43–50.

Martinkova, P., & Drabinova, A. (2018). ShinyItemAnalysis for teaching psychometrics and to enforce routine analysis of educational tests. R Journal, 10(2), 503–515. https://doi.org/10.32614/RJ-2018-074

McFarland, J. L., Price, R. M., Wenderoth, M. P., Martinková, P., Cliff, W., Michael, J., ... & Wright, A. (2017). Development and validation of the Homeostasis Concept Inventory. CBE—Life Sciences Education, 16(2), ar35. https://doi.org/10.1187/cbe.16-10-0305

Messick, S. (1987). Validity. ETS Research Report Series, 1987(2), i–208. https://doi.org/10.1002/j.2330-8516.1987.tb00244.x

Messick, S. (1989). Validity. In Educational measurement (3rd ed., pp. 13–103). New York, NY: American Council on Education.

Munoz, A., & Mackay, J. (2019). An online testing design choice typology towards cheating threat minimisation. Journal of University Teaching & Learning Practice, 16(3). https://doi.org/10.53761/1.16.3.5

Murdock, T. B., & Anderman, E. M. (2006). Motivational perspectives on student cheating: Toward an integrated model of academic dishonesty. Educational Psychologist, 41(3), 129–145. https://doi.org/10.1207/s15326985ep4103_1

National Research Council. (2003). Assessment in support of instruction and learning: Bridging the gap between large-scale and classroom assessment—Workshop report. Washington, DC: National Academies Press. https://doi.org/10.17226/10802

Pintrich, P. R., & de Groot, E. V. (1990). Motivational and self-regulated learning components of classroom academic performance. Journal of Educational Psychology, 82(1), 33–40. https://doi.org/10.1037/0022-0663.82.1.33

R Core Team. (2021). R: A language and environment for statistical computing (4.1.1). Vienna: R Foundation for Statistical Computing. Retrieved November 14, 2021, from https://www.R-project.org/

Revelle, W. (2021). psych: Procedures for personality and psychological research (2.1.6). Evanston, IL: Northwestern University. Retrieved November 14, 2021, from https://CRAN.R-project.org/package=psych

Sbeglia, G. C., & Nehm, R. H. (2022). Measuring evolution learning: Impacts of student participation incentives and test timing. Evolution: Education and Outreach, 15(1), 9. https://doi.org/10.1186/s12052-022-00166-2

Schiel, J. (1996). Student effort and performance on a measure of postsecondary educational development (96-9) (ACT research report). Retrieved September 2, 2020, from https://eric.ed.gov/?id=ED405380

Semsar, K., Brownell, S., Couch, B. A., Crowe, A. J., Smith, M. K., Summers, M. M., ... & Knight, J. K. (2019). Phys-MAPS: A programmatic physiology assessment for introductory and advanced undergraduates. Advances in Physiology Education, 43(1), 15–27. https://doi.org/10.1152/advan.00128.2018

Shi, J., Wood, W. B., Martin, J. M., Guild, N. A., Vicens, Q., & Knight, J. K. (2010). A diagnostic assessment for introductory molecular and cell biology. CBE—Life Sciences Education,, 9(4), 453–461. https://doi.org/10.1187/cbe.10-04-0055

Smith, M. K., Brownell, S. E., Crowe, A. J., Holmes, N. G., Knight, J. K., Semsar, K., ... & Couch, B. A. (2019). Tools for change: Measuring student conceptual understanding across undergraduate biology programs using Bio-MAPS assessments. Journal of Microbiology & Biology Education, 20(2). https://doi.org/10.1128/jmbe.v20i2.1787

Smith, M. K., Thomas, K., & Dunham, M. (2012). In-class incentives that encourage students to take concept assessments seriously. Journal of College Science Teaching, 42(2), 57–61.

Smith, M. K., Wood, W. B., & Knight, J. K. (2008). The genetics concept assessment: A new concept inventory for gauging student understanding of genetics. CBE—Life Sciences Education, 7(4), 422–430. https://doi.org/10.1187/cbe.08-08-0045

Steger, D., Schroeders, U., & Gnambs, T. (2020). A meta-analysis of test scores in proctored and unproctored ability assessments. European Journal of Psychological Assessment, 36(1), 174–184. https://doi.org/10.1027/1015-5759/a000494

Summers, M. M., Couch, B. A., Knight, J. K., Brownell, S. E., Crowe, A. J., Semsar, K., ... & Smith, M. K. (2018). EcoEvo-MAPS: An ecology and evolution assessment for introductory through advanced undergraduates. CBE—Life Sciences Education, 17(2), ar18. https://doi.org/10.1187/cbe.17-02-0037

Sundre, D. L., & Wise, S. L. (2003). Motivation filtering: An exploration of the impact of low examinee motivation on the psychometric quality of tests. Chicago: National Council on Measurement in Education.

Thelk, A. D., Sundre, D. L., Horst, S. J., & Finney, S. J. (2009). Motivation matters: Using the Student Opinion Scale to make valid inferences about student performance. Journal of General Education, 58(3), 129–151. JSTOR.

Tippins, N. T., Beaty, J., Drasgow, F., Gibson, W. M., Pearlman, K., Segall, D. O., & Shepherd, W. (2006). Unproctored internet testing in employment settings. Personnel Psychology, 59(1), 189–225. https://doi.org/10.1111/j.1744-6570.2006.00909.x

Uminski, C., & Couch, B. A. (2021). GenBio-MAPS as a case study to understand and address the effects of test-taking motivation in low-stakes program assessments. CBE—Life Sciences Education, 20(2), ar20. https://doi.org/10.1187/cbe.20-10-0243

Wendy, K. A., & Wieman, C. E. (2011). Development and validation of instruments to measure learning of expert-like thinking. International Journal of Science Education, 33(9), 1289–1312. https://doi.org/10.1080/09500693.2010.512369

Wickham, H., Averick, M., Bryan, J., Chang, W., D'Agostino McGowan, L., François, R., ... & Hiroaki, Y. (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686. https://doi.org/10.21105/joss.01686

Wigfield, A., & Eccles, J. S. (2000). Expectancy–value theory of achievement motivation. Contemporary Educational Psychology, 25(1), 68–81. https://doi.org/10.1006/ceps.1999.1015

Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. Educational Assessment, 10(1), 1–17. https://doi.org/10.1207/s15326977ea1001_1

Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. Applied Measurement in Education, 18(2), 163–183. https://doi.org/10.1207/s15324818ame1802_2

Wolf, L. F., & Smith, J. K. (1995). The consequence of consequence: Motivation, anxiety, and test performance. Applied Measurement in Education. 8(3), 227–242. https://doi.org/10.1207/s15324818ame0803_3

Wright, C. D., Huang, A., Cooper, K., & Brownell, S. (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://doi.org/10.20429/ijsotl.2018.120214

SUPPLEMENTAL MATERIAL FOR CHAPTER 2

Supplemental Table 2.1: Item difficulty and discrimination for the full-length IMCA instrument administered in 2014

Item	Difficulty	Discrimination
1	0.85	0.28
2	0.86	0.24
3	0.62	0.45
4	0.6	0.81
5	0.66	0.74
6	0.69	0.74
7	0.62	0.80
8	0.69	0.62
9	0.89	0.28
10	0.86	0.27
11	0.81	0.41
12	0.62	0.50
13	0.58	0.36
14	0.52	0.54
15	0.34	0.33
16	0.45	0.46
17	0.82	0.30
18	0.72	0.45
19	0.55	0.54
20	0.24	0.31
21	0.52	0.59
22	0.86	0.30
23	0.56	0.60
24	0.69	0.50

Supplemental Table 2.2: Linear mixed effects model[a] on the effects of administration stakes and setting on concept assessment completion time

Parameter	Sum Sq	Mean Sq	*df*	*F*	*p*
Administration condition	14807266	4935755	3	76.768	<.001
Post-hoc comparisons					
Contrast	**Estimate**	**SE**	***df***	***t***	***p***
Higher In: Higher Out	-218.4	16.9	1097	-12.88	<.001
Higher In: Lower In	39.1	15.4	1888	2.55	.053
Higher In: Lower Out	-19.6	19.0	1899	-1.03	.730
Higher Out: Lower In	257.5	18.1	1900	14.24	<.001
Higher Out: Lower Out	198.8	21.2	1895	9.36	<.001
Lower In: Lower Out	-58.7	18.6	1101	-3.15	.009

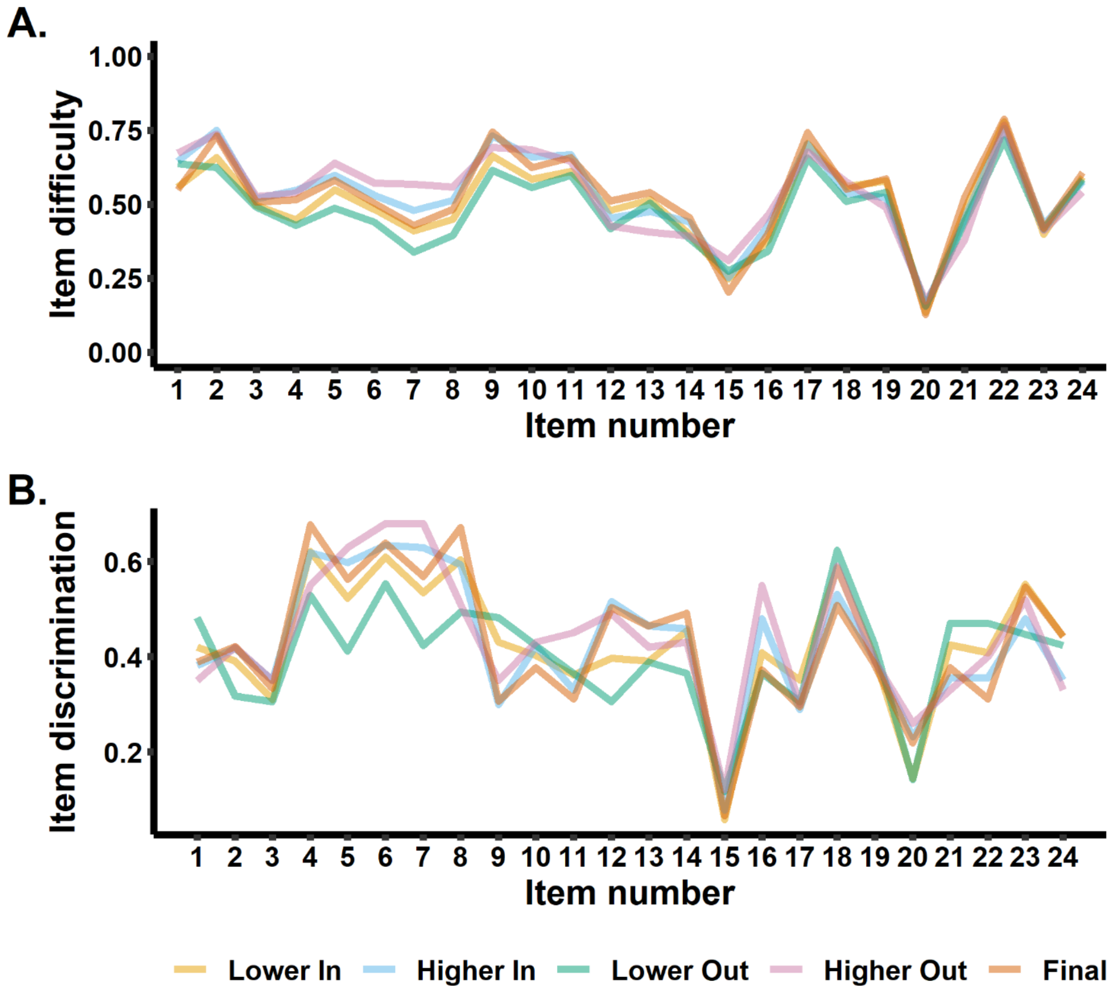
Model R^2 = 0.22

[a] Completion time ~ administration condition + (1 | ID)

Supplemental Table 2.3: Computations of Fisher's *z*-tests concerning differences between correlations of concept assessment score and average unit exam score

	Lower In	Higher In	Lower Out	Higher Out	Final
Lower In	-	0.405	0.190	0.017	0.007
Higher In	0.405	-	0.013	0.009	0.010
Lower Out	0.190	0.013	-	0.160	0.000
Higher Out	0.017	0.009	0.160	-	0.000
Final	0.007	0.010	0.000	0.000	-

Red text indicates significant differences in correlation values.



Supplemental Figure 2.1: Item difficulty and discrimination values for each question on the IMCA in the different administration conditions. The item number corresponds to the numbering scheme used in Shi et al., (2010). (A) Items with higher difficulty values indicate a higher proportion of students responded to the item correctly. (B) Items with a high discrimination value indicate that the item differentiated well between high- and low-performing students.

CHAPTER 3: TESTING SCIENTIFIC PRACTICES: A NATIONWIDE ANALYSIS OF UNDERGRADUATE BIOLOGY EXAMS

ABSTRACT

Scientific practices are the skills used to develop scientific knowledge and are essential across careers in science disciplines. Despite calls from education and government agencies to cultivate scientific practices, there remains little evidence of how often students are asked to apply them in undergraduate courses. We analyzed exams from 111 lower-division biology courses at 100 institutions across the United States and found that only 7% of exam questions addressed a scientific practice. Exams that incorporated scientific practices tended to have a higher average Bloom's Taxonomy level, indicating that scientific practices elicit higher-order cognitive skills. The low occurrence of scientific practices on exams signals that undergraduate courses may not be integrating foundational scientific skills throughout their curriculum in the manner envisioned by recent national frameworks. However, the close association with higher-order cognitive skills suggests that scientific practices represent a primary means to help students develop critical thinking skills.

INTRODUCTION

To address the demands of increasingly interdisciplinary science fields and solve emerging global challenges, education and government agencies have called for undergraduate science courses to emphasize scientific practices (National Research Council [NRC], 2007a; American Association for the Advancement of Science [AAAS], 2011; National Academies of Sciences, Engineering, and Medicine [NASEM], 2022). Scientific practices, such as planning investigations, analyzing data, and evaluating information, represent essential skills for establishing, extending, and refining scientific knowledge (NRC, 2007b).

A robust research synthesis highlighted the importance of scientific practices by naming them as one of the dimensions in a three-dimensional framework for science education (NRC, 2012). These three dimensions consist of scientific practices (i.e., the skills students use to engage in science), crosscutting concepts (i.e., interdisciplinary ways of thinking about scientific processes), and disciplinary core ideas (i.e., concepts central to each science discipline). While previous frameworks have featured elements of scientific practices through their emphasis on inquiry (AAAS, 1993; NRC, 1996), these aspects tended to focus on designing investigations and testing hypotheses. The scientific practices included within the three-dimensional framework present a more complete articulation of inquiry and more fully represent the range of actions scientists take to make sense of phenomena (Schwarz et al., 2017). The three-dimensional framework also explicitly stresses that students develop deep understanding of science when their learning integrates the three-dimensions, rather than approaching them as separate entities.

The scientific practices of the three-dimensional framework address the common instructional goal of improving student "critical thinking" abilities (Stowe & Cooper, 2017; Yuretich, 2003). While definitions of critical thinking vary, researchers agree that it represents an essential part of inquiry and involves interpretation, analysis, evaluation, making inferences, and constructing explanations based on evidence (Facione, 1990). Within undergraduate biology education (Crowe et al. 2008), critical thinking has often been identified through Bloom's Taxonomy (Anderson et al., 2001; Bloom et al., 1956).

While limited in its ability to capture the full spectrum of knowledge types (Blumberg, 2009), Bloom's Taxonomy provides a useful tool for classifying cognitive skills that students use when working through a task. The taxonomy is commonly divided into lower-order skills (remember and understand) and higher-order skills (apply, analyze, evaluate, and create). Biology education researchers often equate critical thinking with the higher-order skills (Allen & Tanner, 2002; Bissell & Lemons, 2006; Moon et al., 2021; Zheng et al., 2008), and the higher-order skills have considerable parallels to the scientific practices of the three-dimensional framework (Larsen et al., 2022), with some of the same verbs (e.g., analyze, evaluate) appearing in both frameworks. While they contain considerable overlap, there has not yet been an empirical comparison of scientific practices and Bloom's Taxonomy at the undergraduate level.

The three-dimensional framework serves as the foundation for K-12 science education in the United States, with 44 of the U.S. states currently using the framework as the basis for their statewide science standards (NASEM, 2021; NGSS Lead States, 2013). Despite this widespread adoption at K-12 levels, there is little evidence indicating to what degree undergraduate biology courses incorporate the three dimensions, particularly with respect to scientific practices. Achieving a smooth transition from high school to undergraduate coursework may depend on the degree to which instruction maintains continuity in three-dimensional language, terminology, and expectations (Clemmons et al., 2020b). Previous efforts have adapted the three-dimensional framework for undergraduate courses (Bain et al., 2020; Laverty et al., 2016), marking an important step for further curriculum development and associated research at the college level.

In light of ongoing national calls, there remains a need to determine the extent to which students in undergraduate courses apply the scientific practices outlined in the three-dimensional framework, particularly within the lower-division courses that serve as gateways—and often gatekeepers—to science degree programs (NASEM, 2016). One way to gauge the frequency of scientific practices in a course is to examine course assessments, such as tests and exams. Instructors in lower-division STEM courses often rely heavily on exams as the primary summative method to measure student learning (Goubeaud, 2010). Since the content of exams inherently reflects the knowledge and skills that instructors value and intend for students to learn (Scouller 1998, NRC 2003), an exam including scientific practices signifies that they represent a prioritized learning outcome. This approach of using assessments to gauge the extent of three-dimensional learning in a course has been applied in previous work (Matz et al. 2018, Stowe et al. 2021); however, these studies were conducted using courses taught at a single institution or within organic chemistry.

Our study aims to provide the first large-scale, nationwide portrait of how the three-dimensional framework is incorporated into undergraduate biology courses. We use exams as a window into the skills and knowledge instructors prioritize (NRC, 2003), and we analyze exam alignment to the three-dimensional framework, with a particular focus on the incorporation of scientific practices. We also analyze exam alignment to Bloom's Taxonomy given its overlap with the science practices of the three-dimensional framework (Larsen et al., 2022) and its wide use in biology education (Allen & Tanner, 2002; Crowe et al., 2008). Our analysis of course exams addresses two research questions: (1) To what extent do exams align to the three-dimensional framework with particular reference to the scientific practices? (2) What is the relationship between an exam's alignment to the three-dimensional framework and to Bloom's Taxonomy of cognitive skills?

METHODS

Survey Development and Administration

We developed an online survey through Qualtrics to collect course artifacts (e.g., an exam document, the associated exam answer key, and a syllabus) along with demographic and institutional information from instructors of undergraduate lower-division biology courses. We define lower-division courses as 100- and 200-level courses and their equivalents. To participate in the survey, instructors had to confirm that they were located at a 2- or 4-year institution of higher education in the United States, were currently teaching or had taught a lecture-based lower-division biology course within the past three years, and had administered graded tests or exams in their course. We provided instructors in this study with $75 USD in compensation for the approximately half-hour of time spent completing the survey. This research was classified as exempt from human-subjects review by the University of Nebraska–Lincoln (protocol 21082).

We distributed the survey between May–August 2021 through listservs for professional societies, including the Society for the Advancement of Biology Education Research (SABER), Ecological Society of America (ESA) EcoEd, Ecological Research as Education Network (EREN), Quantitative Undergraduate Biology Education and Synthesis (QUBES), and National Association of Biology Teachers (NABT). Because of expected overlap in these email lists, we cannot estimate the total number of biology instructors who received a survey invitation. We wanted to sample from instructors who may not subscribe to education-related listservs, so we randomly selected institutions from a complete list of United States Associate's, Baccalaureate, Master's, and Doctoral institutions. We randomly selected five institutions from each institution type and distributed the survey to all biology instructors at each institution via the email address provided on institution websites. We emailed 384 instructors using this method and had a response rate of 2%.

In this study, we collected one summative exam from each instructor from a lecture (i.e., non-lab) course. We focus here on summative assessments, but we recognized that instructors may also be utilizing formative assessments and other summative assessments (e.g., projects, papers, presentations) within their courses. Given the variation in the design, format, and grading of these other assessments, we excluded them from this study.

Data Sources

The final dataset contained responses from 111 instructors at 100 unique institutions across the United States, including broad representation from each undergraduate institution type (Table 3.1). Our sample included instructors across career stages (Table 3.2) and from different categories of lower-division courses (Table 3.3). The majority of the courses (80%) were introductory-level, and the remaining courses spanned a variety of lower-division biology topics such as anatomy and physiology, environmental science, and microbiology. Class sizes ranged from 4 to 600 students ($M = 83.8 \pm 10.6$ SEM).

Table 3.1: Institutional Carnegie classifications and geographic regions

Institution region	Associate's	Baccalaureate	Master's	Doctoral	Total
Northeast	4	4	7	6	21
Midwest and Great Plains	6	10	6	7	29
Pacific Northwest	3	2	0	2	7
Southeast	7	9	4	9	29
Southwest	6	0	2	6	14
Total	26	25	19	30	100

Note: Institutional categories are based on Carnegie classifications (Indiana University Center for Postsecondary Research, 2021). Institution regions are based on the PULSE regional network classifications (Partnership for Undergraduate Life Sciences Education, 2019).

Table 3.2: Self-reported demographic information of undergraduate biology instructors

Characteristic	n	%
Gender		
Female	67	60
Male	42	38
Preferred not to disclose	2	2
Race/ethnicity[a]		
Non-underrepresented	97	87
Underrepresented	11	10
Self-described	1	1
Preferred not to disclose	2	2
Teaching experience		
0-1 year	5	5
2-5 years	20	18
6-10 years	30	27
11-15 years	29	26
16-20 years	10	9
21-25 years	11	10
> 25 years	6	5

[a]We use the term "underrepresented" here to convey our focus on racial/ethnic groups that have faced disproportionate challenges within STEM disciplines, including Black/African American, Hispanic/Latinx, American Indian/Alaska Native, and Native Hawaiian/Pacific Islander. This grouping is not intended to obscure the unique histories and identities of any group.

Table 3.3: Categories of lower-division biology courses included in the sample

Course category[a]	**n**	**%**
Introductory – Cell/Molecular	32	29
Introductory – Organismal	31	28
Introductory – General Biology	26	23
Ecology/Evolution	6	5
Genetics	3	3
Microbiology	3	3
Anatomy/Physiology	3	3
Cell/Molecular Biology	2	2
Environmental Science	2	2
Plant Biology	2	2
Zoology	1	< 1
Lab courses		
Course has an associated lab component	95	86
Course does not have an associated lab component	16	14

[a]If course category was not evident based on the title of the course, we used the content in the course syllabus to designate the categories. We categorized introductory-series courses that primarily deal with molecules, cells, and genetics as "Introductory – Cell/Molecular," introductory-level courses that primarily deal with animal systems, biodiversity, ecology, and evolution topics as "Introductory – Organismal," and courses that broadly span both cell/molecular biology and ecology/evolution topics as "Introductory – General Biology."

Codebook Development

We assembled our modified three-dimensional framework (Table 3.4) and associated codebook (Supplemental Table 3.1) from existing protocols and tools for characterizing assessments in undergraduate science courses. We used the codebook from the Three-Dimensional Learning Assessment Protocol (3D-LAP; Laverty et al., 2016) to characterize scientific practices and crosscutting concepts and used the *Vision and Change* core concepts (AAAS, 2011), as delineated in the BioCore Guide (Brownell et al., 2014), for core ideas. We note that the 3D-LAP includes a protocol for coding core ideas that overlaps considerably with the *Vision and Change* core concepts of evolution, information flow, energy and matter, structure and function, and systems. We chose to use the Vision and Change core concepts and associated BioCore Guide because they provided a more comprehensive portrait of these topics across biological scales. We used the protocol from Bloom's Dichotomous Key (Semsar and Casagrand, 2017) to assign levels of Bloom's Taxonomy to exam items. Each Bloom's level was assigned an ordinal

Table 3.4: Modified dimensions of the three-dimensional framework

Scientific Practices[a] 1. Asking Questions 2. Developing and Using Models 3. Planning Investigations 4. Analyzing and Interpreting Data 5. Using Mathematics and Computational Thinking 6. Constructing Explanations and Engaging in Argument from Evidence 7. Evaluating Information **Crosscutting Concepts**[b] 1. Patterns 2. Cause and Effect: Mechanism and Explanation 3. Scale 4. Proportion, and Quantity 5. Systems and System Models 6. Energy and Matter: Flows, Cycles, and Conservation 7. Structure and Function 8. Stability and Change **Biology Core Ideas**[c] 1. Evolution 2. Information Flow, Exchange, and Storage 3. Structure and Function 4. Pathways and Transformations of Energy and Matter 5. Systems
Sources: NRC, 2012; Laverty et al., 2016; AAAS, 2011. [a] The *Framework for K-12 Science Education* includes both scientific and engineering practices. For the purposes of this research based in biology courses, we focus exclusively on the scientific practices as presented in the Three-Dimensional Learning Assessment Protocol (3D-LAP). Note that the 3D-LAP differs from the K-12 practices in that it combines “Constructing Explanations” and “Engaging in Argument from Evidence” into a single scientific practice and narrows the focus of the practice “Obtaining, Evaluating, and Communicating Information” to only evaluating information. [b] The 3D-LAP separates “Scale” and “Proportion and Quantity” into two separate crosscutting concepts, where in the K-12 framework, these are combined into a single concept. [c] We use the biology core ideas that are articulated in *Vision and Change* core concepts but note that there are similar biology core ideas outlined in the 3D-LAP and within the K-12 framework. There are separate sets of core ideas for chemistry and physics disciplines.

numeric value between 1 and 6, where 1 = remember, 2 = understand, 3 = apply, 4 = analyze, 5 = evaluate, and 6 = create.

We note that the mental processes that a student engages in when responding to assessment items is context-dependent and may be affected by previous instruction or experiences within a course. As such, we coded items based on the potential of the item to elicit specific dimensions or cognitive skills, but as we did not have insight into the course content or structure, this coding only captures the apparent cognitive processes targeted by a given item.

Item Coding

We used the point values and numbering schemes set by the instructor to determine the boundaries of individual items (i.e., test questions). Items that shared a common stem and/or used a sub-part numbering scheme (e.g., 4a, 4b, 4c) were coded as a single clustered item. Our sample of 111 exams contained a total of 4337 items. Exams ranged from 1 to 120 items ($M = 39.1 \pm 2.0$ SEM).

We used instructor-provided answer keys to inform our coding of individual items. In certain cases, particularly for constructed-response items, the answer key informed us that the instructor expected students to include explanations or reasoning in their response, which may not have been evident in exact wording of the item stem or prompt. For such items, we defaulted to the student performance expectations contained in the answer key.

The 3D-LAP delineates each scientific practice as consisting of nested criteria statements describing different levels within the practice. Similar to previous studies (Laverty et al., 2016; Laverty and Caballero, 2018; Matz et al., 2018; Underwood et al., 2018; Carmel et al., 2019; Stowe et al., 2021), we coded an item as eliciting a scientific practice when it satisfied all of the criteria statements for the corresponding constructed-response or selected-response item type. We coded an item as addressing a crosscutting concept or core idea if the item aligned with any of the criteria statements within the code. Items may have met multiple scientific practices, crosscutting concepts, or core ideas. We coded only the highest Bloom's Taxonomy level that the item was capable of eliciting.

Interrater Reliability

Two members of the research team used the codebook (Supplemental Material 3.1) to independently code a total of 48 items that were randomly selected from the entire item pool. The team members coded the items in iterative sets of 12, and any disagreements from a set of items were discussed until consensus before beginning coding the next set of items. There was an average of 93% agreement across all codes and ≥ 75% agreement for each individual code (Supplemental Table 3.2). We calculated percent agreement using the arsenal package [v. 3.6.3] (Heinzen et al., 2021) in R statistical software. For the items that the two raters discussed, the consensus values were used in the final dataset. The remaining items in the dataset were coded by only one member of the research team.

Item Normalization and Weighting

Given that exams use different point schemes across courses, for some analyses, we calculated a normalized item point value by dividing the individual item point value by the total number of points on the exam. For other analyses, we determined the percentage of exam points aligned with the three-dimensional framework by multiplying each normalized item point value by either 1 or 0 based on whether the item met or did not meet a dimension, respectively, and summed the values for each exam. We determined the weighted Bloom's value for each exam by multiplying each normalized item point value by the coded Bloom's level [1, 2, 3, 4, 5, or 6] and summed the values for each exam.

Correlating Percentage of Exam Points Aligned with Each Dimension and Weighted Bloom's Level

Considering the ordinal nature of Bloom's Taxonomy levels, we used Spearman rank order correlations to determine the relationship between the percentage of exam points aligned with each dimension and the weighted Bloom's levels of each exam. We used Fisher's *z* transformations to compare the correlation coefficients with respect to each dimension (Supplemental Table 3.3). We calculated Spearman correlations using the stats package [v 4.1.1] (R Core Team, 2021) and calculated Fisher's *z* transformations using the diffcor package [v 0.7.1] (Blötner, 2022) in R statistical software.

Exam Weighting in Course Grade

Out of the 111 instructors in the sample, 104 (94%) included a grading scheme that revealed the overall weight of exam grades in their course syllabus. For each course, we determined the total percentage of the course grade that came from exam grades. We included unit, midterm, and final exams in our value for weight of exam grades but did not include formative assessments or other summative assessments.

RESULTS

Across our sample of 111 exams with a total of 4337 items (i.e., test questions), only 5% of items achieved the principal goal of the three-dimensional framework by simultaneously incorporating a scientific practice, crosscutting concept, and core idea (Figure 3.1). This lack of three-dimensional alignment was driven by the small percentage of items that met the criteria for a scientific practice. Only 7% of items incorporated a scientific practice, but the majority of those items were three-dimensional (Figure 3.2). Despite the abundance of items that included a crosscutting concept (47%) or core idea (59%), only a small proportion of those items qualified as three-dimensional.

Strikingly, over a third of items on the exams did not align with any of the three dimensions.

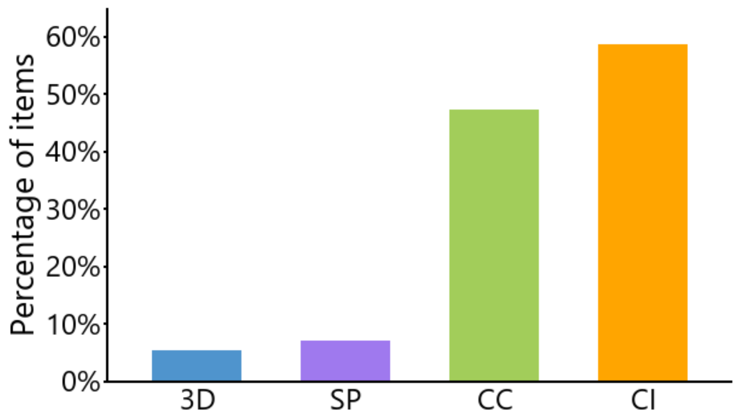


Figure 3.1: Percentage of undergraduate biology exam items aligned to each dimension of three-dimensional framework. Exam items (n = 4337) are represented only once in each bar even if they may align with multiple scientific practices, crosscutting concepts, or core ideas within that dimension. Abbreviations: 3D = three-dimensional; SP = scientific practice; CC = crosscutting concept; CI = core idea.

When items did align to a scientific practice, the practice was most commonly "Analyzing Data," "Engaging in Argument," or to a lesser extent "Using Models" (Figure 3.3). While all the scientific practices were represented in the sample, there were notably few items meeting the practices of "Evaluating Information," "Asking Questions," "Planning Investigations," and "Using Mathematics and Computational Thinking." Each crosscutting concept and core idea was represented across the range of items in the sample. In both the crosscutting concepts and core ideas, "Structure and Function" was the most common code applied to items. The codes for "Structure and Function" as a crosscutting concept and as a core idea can be coded independently but given the considerable overlap in the code criteria (Supplemental Table 3.1), these codes were often applied together.

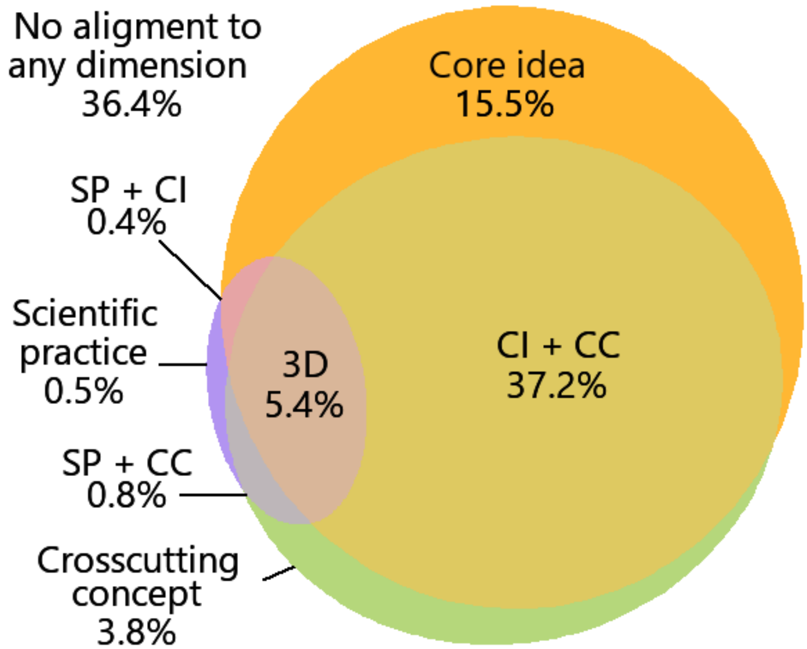


Figure 3.2: Intersections of the three-dimensional alignment of undergraduate biology exam items. The size of the ellipses for scientific practices, crosscutting concepts, and core ideas are proportional to the number of items in the sample aligned with each dimension(s). Approximately 36% of items in the sample did not align with any dimension and are not included within an ellipse. Abbreviations: 3D = three-dimensional; SP = scientific practice; CC = crosscutting concept; CI = core idea.

While the exams contained few items addressing scientific practices overall, these items could have been more involved or taken students more time to complete, thus constituting a larger portion of the exam experience. To address this possibility, we analyzed exam content based on normalized item point values, since instructors tend to assign more points to more substantial items. When accounting for item point value, we found that most exams still had fewer than 10% of points aligned with scientific practices (Figure 3.4). Thus, items targeting scientific practices had higher point values than other exam items, but scientific practices represented a relatively small proportion of the overall exam content.

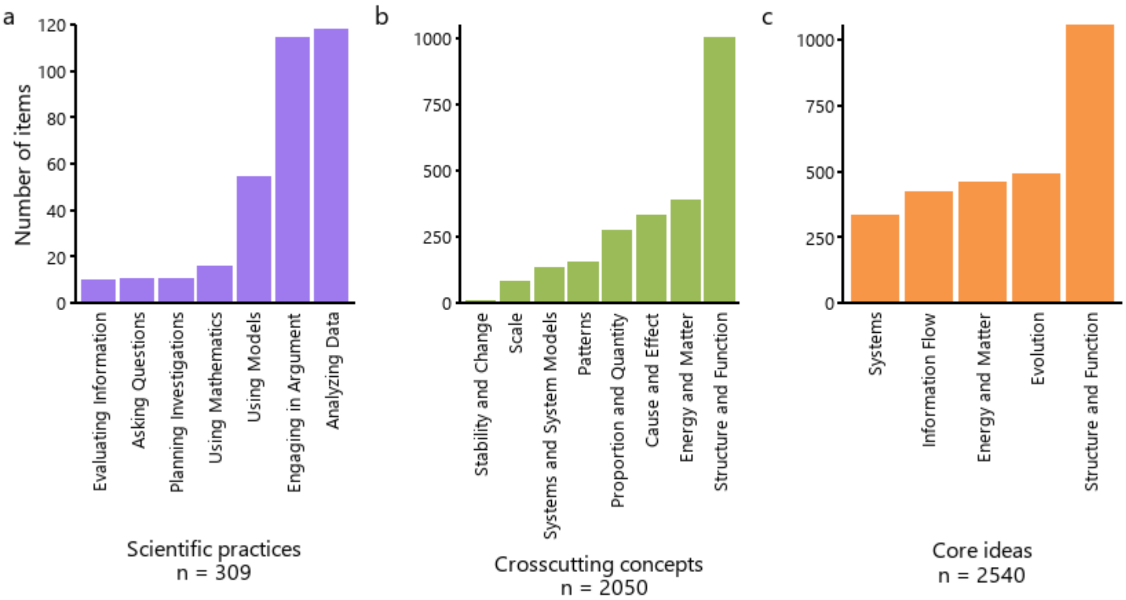


Figure 3.3: Alignment of undergraduate biology exam items to each of the scientific practices, crosscutting concepts, and core ideas of the three-dimensional framework. Individual items may have addressed more than one scientific practice (a), crosscutting concept (b), or core idea (c), thus the sum of the bars in each plot may exceed the total number of items aligned to the dimension.

We applied Bloom's Taxonomy to see which cognitive skills predominated in undergraduate biology exams. We found that the majority of items (86%) aligned with the lower-order skills remember (level 1) or understand (level 2), with just 14% of items aligning to the higher-order skills apply, analyze, evaluate, or create (levels 3-6; Figure 3.5). We also considered Bloom's at the exam level by computing a weighted average accounting for item point values. The mean of the weighted Bloom's level across exams was 2.02 ± 0.09 SEM. Even after accounting for the tendency for instructors to place more points on higher-level Bloom's items, we found that the overall exam tends toward lower-order cognitive skills.

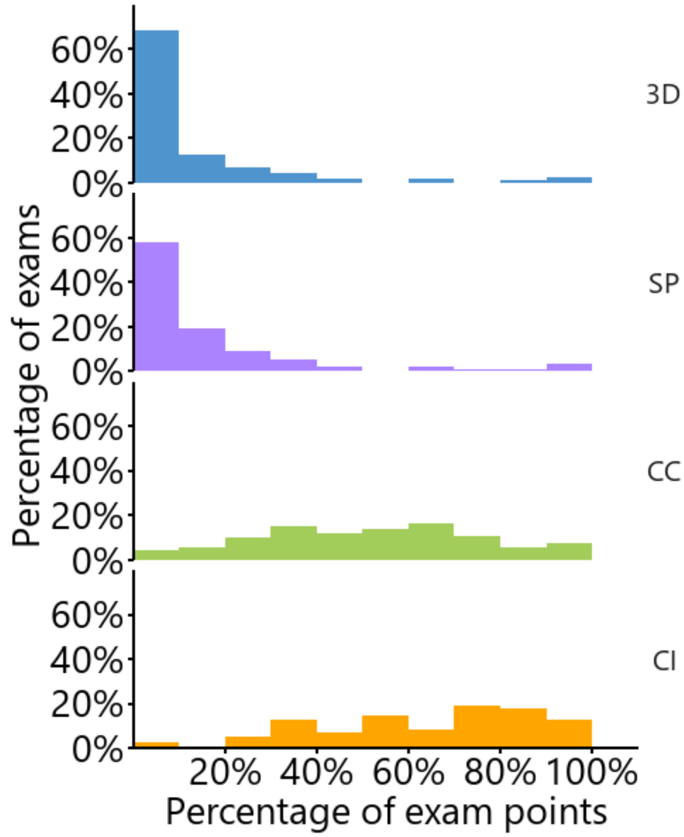


Figure 3.4: Percentage of exam points aligned to the three-dimensional framework. An exam from each course (n = 111) is represented once within each dimension. Abbreviations: 3D = three-dimensional; SP = scientific practice; CC = crosscutting concept; CI = core idea.

There was a considerable correlation between the percentage of three-dimensional points on an exam and its weighted Bloom's level (ρ = 0.75; Figure 3.6). This strong positive relationship was driven by scientific practices, which had the highest correlation with Bloom's Taxonomy of any of the three dimensions (ρ = .83). Crosscutting concepts and core ideas were also correlated with the Bloom's level of exams (ρ = .48; .61), albeit to a lesser extent (Supplemental Material 3.3).

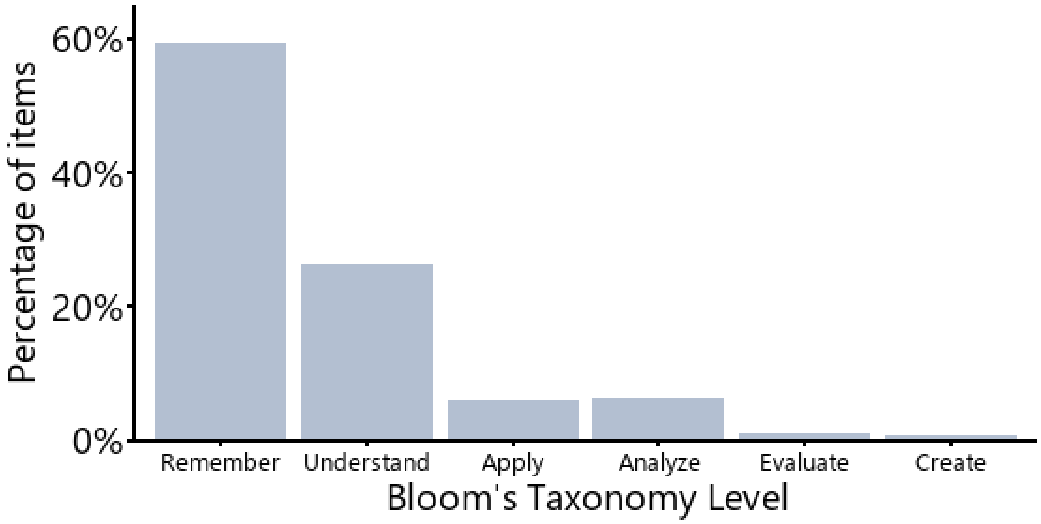


Figure 3.5: Alignment of undergraduate biology exam items to levels of Bloom's Taxonomy. Of the 4337 exam items in the sample, 86% align to the lower-order cognitive skills (remember and understand) and 14% align to the higher-order cognitive skills (apply, analyze, evaluate, and create).

Instructors can use other activities to target scientific practices or focus on scientific practices in associated lab courses. However, within our sample, exam grades comprised half of total course grade (M = 49.7 ± 1.5 SEM), and we observed no difference in the extent to which scientific practices (Welch's ANOVA, $F(1, 18.7) = 0.15$, $p = 0.71$) or Bloom's levels (Welch's ANOVA, $F(1, 22.4) = 0.09$, $p = 0.77$) were assessed in courses with or without associated labs.

We note that approximately 65% of the exams in our sample (n = 72) were administered during the semesters affected by the COVID-19 pandemic. While this period of time was marked by changes in instructional modality, with many courses shifting into a partially or fully online format, we did not find notable differences in the content of the assessments administered during the global pandemic. When comparing exams administered to students before and after March 2020, we found no significant differences in the percentage of three-dimensional points (*t*-test, *df* = 65.1, *p* = 0.14) nor in the weighted Bloom's level of the exams (*t*-test, *df* = 68.8, *p* = 0.58).

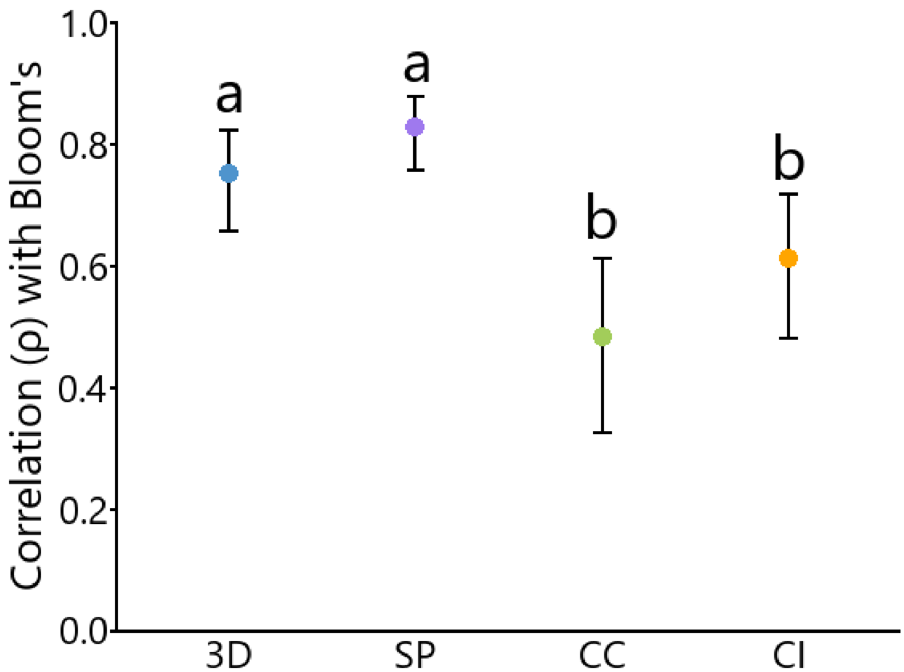


Figure 3.6: Spearman correlation coefficients and 95% confidence intervals representing the relationship between the percentage of exam points in each dimension and the weighted Bloom's level of the exam. Letters represent differences in significance between correlation coefficients as determined by Fisher's *z*-tests (Supplemental Table 3.3). Abbreviations: 3D = three-dimensional; SP = scientific practice; CC = crosscutting concept; CI = core idea.

DISCUSSION

Taken together, our results highlight a disconnect between what educational reports propose as optimal science assessment (NRC, 2014) and what undergraduate biology courses actually assess. These reports indicate that integrating scientific practices with the crosscutting concepts and core ideas is needed for students to reason through how scientific ideas form and to view science as a dynamic and ongoing process (NRC, 2012), but we found that scientific practices are largely missing from biology exams. The low frequency of science practices paired with the high frequency of items only addressing lower-order cognitive skills means students are more often assessed on conceptual knowledge rather than their ability to apply that information to conduct science. This exclusion of scientific practices may unintentionally reinforce the perception of science as a collection of discrete facts (NRC, 2012, 2014), which may have negative consequences for retention and persistence of students in STEM majors (Olson & Riordan, 2012).

Despite calls for scientific practices to be taught and assessed throughout undergraduate course sequences (AAAS, 2011; NASEM, 2022), our analysis of exam content suggests that these critical skills remain a minor part of lower-division lecture courses. While this study necessarily focused on biology, previous work indicates that this phenomenon may be the norm in gateway courses across science disciplines (Matz et al., 2018; Stowe & Cooper, 2017). The underrepresentation of scientific practices likely reflects constraints placed on instructors who lack the time, resources, and support for implementing three-dimensional lessons and assessments (NRC, 2014) and who may feel pressured to cover broad ranges of content knowledge (Wright et al., 2018). Another possible explanation for the low frequency is that instructors may be reserving instruction and assessment of scientific practices for upper-division courses, yet our previous work found that the extent to which instruction focuses on scientific practices does not differ between course levels (Durham et al., 2017).

Our current findings highlight a need to shift instruction and assessment toward incorporating scientific practices. Many instructors share the goal of teaching and assessing critical thinking and higher-order cognitive skills (Yuretich, 2003), but our findings echo previous studies (Momsen et al., 2010, 2013) and indicate that many instructors may not be meeting that goal. We found that most exam items were only capable of assessing lower-order cognitive skills on Bloom's Taxonomy. This abundance of lower-order skills may be in part attributed to a common interpretations of Bloom's Taxonomy in which a high level of difficulty associated with answering the item is conflated with higher-order Bloom's levels (Lemons & Lemons, 2013; Monrad et al., 2021; Wright et al., 2018). The scientific practices offer a way to navigate around this tendency. We found that the extent to which an exam engages students in higher-order cognitive skills associated with critical thinking is closely aligned with the inclusion of scientific practices. This provides additional support for the idea that incorporating scientific practices may be a more specific way to target the higher-order cognitive skills and associated critical thinking intended by instructors (Stowe & Cooper, 2017).

Although there were few scientific practices in our sample overall, we found that scientific practices rarely occurred in isolation and were typically paired with crosscutting concepts and/or core ideas. The instructors who did incorporate scientific practices into their exams usually situated them within a disciplinary context as intended by the three-dimensional framework (NRC, 2014). Instructors wishing to incorporate scientific practices into their exams may also find it helpful to consult the Three-Dimensional Learning Assessment Protocol (Laverty et al., 2016). The 3D-LAP provides detailed criteria that can be used to determine if an exam item has the potential to engage students in scientific practices, and there are guides for using the 3D-LAP to adapt existing exam items (Underwood et al., 2018). Like other calls for greater adoption of three-dimensional assessment at the undergraduate level, we are not suggesting that every item on an exam needs to be three-dimensional (Laverty et al., 2016). The transition into three-dimensional learning and assessment can be challenging and time-intensive for instructors (Furtak, 2017), but it is a task that may lead to more equitable science assessments (Bang et al. 2017, Ralph et al. 2022).

Each of the scientific practices was represented in our sample, indicating that exams are capable of assessing each practice, but not all the practices were represented equally. The practices “Asking Questions,” “Evaluating Information,” “Planning Investigations,” and “Using Mathematics and Computational Thinking” occurred least frequently on exams. These practices associated with traditional definitions of inquiry and the scientific method may see more prominent implementation in the curriculum of lab courses (Carmel et al., 2019). This raises the possibility that instructors are carrying out instruction and assessment of these and other scientific practices within the associated lab course. However, courses without associated labs did not assess more science practices, suggesting that the assessment content of the lecture portion of a course may be fairly independent from the presence or absence of associated lab sections. Instructors may also have targeted scientific practices through other course activities, such as formative assessments or other summative assessments (e.g., projects, papers, presentations). Even if this is the case, the three-dimension framework contends that scientific practices should be incorporated throughout lecture courses because they help students to develop a robust understanding of disciplinary knowledge as the dynamic product of a scientific process.

The majority of items in our sample met the criteria for a core idea or crosscutting concept; however, most of these items did not elicit a scientific practice. Furthermore, although not true for every case, many of these one- or two-dimensional items tended to ask students to recall only definitions or discrete pieces of memorized information (i.e., lower-order cognitive skills). While it is important for students to remember and understand these foundational ideas, the goal of the three-dimensional framework is to have students apply their knowledge and understanding using the scientific practices (Cooper et al., 2015). Our work highlights that many exams tend to lend credence to the longstanding criticism that lower-division STEM courses, particularly in biology, overemphasize the memorization of factual information (Momsen et al., 2010, 2013; Sundberg et al., 1994). Such a finding has consequences for student learning, as memorization-based exams may not be as effective at promoting long-term retention of course content compared to exams that encourage deeper understanding and application of the material (Jensen et al., 2014).

We applied the three-dimensional framework because of our focus on lower-division courses. The three-dimensional framework is used extensively in K-12 science education and adopting this framework in lower-division courses can help provide a familiar scaffold for students to aid their learning of skills and concepts expected at the undergraduate level. While we use the three-dimensional framework here, we acknowledge that other frameworks can be used similarly to characterize important skills and concepts in undergraduate science courses. The Advanced Placement (AP) Biology Course Framework (College Board, 2020) provides a guide for skills and concepts, but its application may be limited to introductory biology courses. The *Vision and Change* framework (AAAS, 2011) provides a wider lens for program-level learning outcomes that can be applied across all levels of undergraduate biology and are intended to be completed by the end of a four-year degree. Although there are slight differences in terminology, there is substantial overlap between the scientific practices in the three-dimensional framework and the *Vision and Change* core competencies and their articulation within the more delineated BioSkills Guide (Clemmons et al., 2020a, 2020b). For biology courses focused on ecological concepts, instructors may choose to use the 4-Dimensional Ecology Education framework (4DEE; Berkowitz et al., 2018, Prevost et al., 2019), which in addition to practices, core concepts, and crosscutting themes features an additional dimension examining human-environment interactions. While we use the three-dimensional framework for this study, each of these aforementioned frameworks may be used to help center curriculum, instruction, and assessments around foundational ideas and skills that are important for scientific literacy, understanding, and participation.

CONCLUSION

The three-dimensional framework represents a major educational advancement because it presents science proficiency as integrating science practices, crosscutting concepts, and core ideas (NRC, 2012). Indeed, scientific knowledge arises from research investigations, so curriculum reform efforts should seek to engage students with conceptual models as evolving products of the science process, rather than invariant truths (Matz et al., 2018; Passmore et al., 2009; Zagallo et al., 2016). Our research suggests that a more direct incorporation of scientific practices represents a key avenue to helping students develop the envisioned integrative proficiency. By focusing on scientific practices within instruction and assessment, we can help cultivate the types of critical thinking needed by scientifically literate citizens and science professionals to tackle global challenges that require both knowledge and action.

Acknowledgements

We thank participating instructors, and Karli Workman and Sara Burbach for their contributions to project development This material is based upon work supported by the National Science Foundation (NSF) Graduate Research Fellowship (DGE-1610400). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

REFERENCES FOR CHAPTER 3

Allen, D., & Tanner, K. (2002). Approaches to cell biology teaching: Questions about questions. CBE—Life Sciences Education, 1(3), 63–67. https://doi.org/10.1187/cbe.02-07-0021

American Association for the Advancement of Science. (1993). Benchmarks for Science Literacy. Oxford University Press.

American Association for the Advancement of Science. (2011). Vision and Change in Undergraduate Biology Education: A Call to Action. AAAS. https://live-visionandchange.pantheonsite.io/wp-content/uploads/2011/03/Revised-Vision-and-Change-Final-Report.pdf

Anderson LW, Krathwohl DR, Bloom BS. 2001. A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives. Longman.

Bain, K., Bender, L., Bergeron, P., Caballero, M. D., Carmel, J. H., Duffy, E. M., Ebert-May, D., Fata-Hartley, C. L., Herrington, D. G., Laverty, J. T., Matz, R. L., Nelson, P. C., Posey, L. A., Stoltzfus, J. R., Stowe, R. L., Sweeder, R. D., Tessmer, S. H., Underwood, S. M., Urban-Lurain, M., & Cooper, M. M. (2020). Characterizing college science instruction: The Three-Dimensional Learning Observation Protocol. PLOS ONE, 15(6), e0234640. https://doi.org/10.1371/journal.pone.0234640

Berkowitz, A. R., Cid, C., Doherty, J., Ebert-May, D., Klemow, K., Middendorf, G., Mourad, T., & Pohlad, B. (2018). The 4-Dimensional Ecology Education Framework. Ecological Society of America.

Bissell, A. N., & Lemons, P. P. (2006). A New Method for Assessing Critical Thinking in the Classroom. BioScience, 56(1), 66–72. https://doi.org/10.1641/0006-3568(2006)056[0066:ANMFAC]2.0.CO;2

Bloom, B. S., Krathwohl, D. R., & Masia, B. B. (1956). Taxonomy of Educational Objectives: The Classification of Educational Goals. McKay.

Blötner, C. (2022). diffcor: Fisher's z-tests concerning difference of correlations (R package version 0.7.1). https://CRAN.R-project.org/package=diffcor

Blumberg, P. (2009). Maximizing learning through course alignment and experience with different types of knowledge. Innovative Higher Education, 34(2), 93–103. https://doi.org/10.1007/s10755-009-9095-2

Brownell, S. E., Freeman, S., Wenderoth, M. P., & Crowe, A. J. (2014). BioCore Guide: A tool for interpreting the Core Concepts of Vision and Change for biology majors. CBE—Life Sciences Education, 13(2), 200–211. https://doi.org/10.1187/cbe.13-12-0233

Carmel, J. H., Herrington, D. G., Posey, L. A., Ward, J. S., Pollock, A. M., & Cooper, M. M. (2019). Helping students to "do science": Characterizing scientific practices in general chemistry laboratory curricula. Journal of Chemical Education, 96(3), 423–434. https://doi.org/10.1021/acs.jchemed.8b00912

Clemmons, A. W., Timbrook, J., Herron, J. C., & Crowe, A. J. (2020a). BioSkills Guide. QUBES Educational Resources. https://doi.org/10.25334/156H-T617

Clemmons, A. W., Timbrook, J., Herron, J. C., & Crowe, A. J. (2020b). BioSkills Guide: Development and national validation of a tool for interpreting the Vision and Change core competencies. CBE—Life Sciences Education, 19(4), ar53. https://doi.org/10.1187/cbe.19-11-0259

College Board. (2020). AP Biology Course and Exam Description, Effective Fall 2020.

Cooper, M. M., Caballero, M. D., Ebert-May, D., Fata-Hartley, C. L., Jardeleza, S. E., Krajcik, J. S., Laverty, J. T., Matz, R. L., Posey, L. A., & Underwood, S. M. (2015). Challenge faculty to transform STEM learning. Science, 350(6258), 281–282. https://doi.org/10.1126/science.aab0933

Crowe, A., Dirks, C., & Wenderoth, M. P. (2008). Biology in Bloom: Implementing Bloom's Taxonomy to Enhance Student Learning in Biology. CBE—Life Sciences Education, 7(4), 368–381. https://doi.org/10.1187/cbe.08-05-0024

Durham, M. F., Knight, J. K., & Couch, B. A. (2017). Measurement Instrument for Scientific Teaching (MIST): A tool to measure the frequencies of research-based reaching practices in undergraduate science courses. CBE—Life Sciences Education, 16(4). https://doi.org/10.1187/cbe.17-02-0033

Facione, P. A. (1990). Critical thinking: A statement of expert consensus for purposes of educational assessment and instruction. American Philosophical Association; ED315423.

Furtak, E. M. (2017). Confronting dilemmas posed by three-dimensional classroom assessment: Introduction to a virtual issue of Science Education. Science Education, 101(5), 854–867. https://doi.org/10.1002/sce.21283

Goubeaud, K. (2010). How is science learning assessed at the postsecondary level? Assessment and grading practices in college biology, chemistry and physics. Journal of Science Education and Technology, 19(3), 237–245. https://doi.org/10.1007/s10956-009-9196-9

Heinzen, E., Sinnwell, J., Atkinson, E., Gunderson, T., & Dougherty, G. (2021). arsenal: An Arsenal of "R" Functions for Large-Scale Statistical Summaries (3.6.3). https://CRAN.R-project.org/package=arsenal

Indiana University Center for Postsecondary Research. (2021). The Carnegie Classification of Institutions of Higher Education (2021 edition).

Jensen, J. L., McDaniel, M. A., Woodard, S. M., & Kummer, T. A. (2014). Teaching to the test…or testing to teach: Exams requiring higher order thinking skills encourage greater conceptual understanding. Educational Psychology Review, 26(2), Article 2. https://doi.org/10.1007/s10648-013-9248-9

Larsen, T. M., Endo, B. H., Yee, A. T., Do, T., & Lo, S. M. (2022). Probing internal assumptions of the revised Bloom's Taxonomy. CBE—Life Sciences Education, 21(4), ar66. https://doi.org/10.1187/cbe.20-08-0170

Laverty, J. T., & Caballero, M. D. (2018). Analysis of the most common concept inventories in physics: What are we assessing? Physical Review Physics Education Research, 14(1), 010123. https://doi.org/10.1103/PhysRevPhysEducRes.14.010123

Laverty, J. T., Underwood, S. M., Matz, R. L., Posey, L. A., Carmel, J. H., Caballero, M. D., Fata-Hartley, C. L., Ebert-May, D., Jardeleza, S. E., & Cooper, M. M. (2016). Characterizing college science assessments: The Three-Dimensional Learning Assessment Protocol. PLOS ONE, 11(9), e0162333. https://doi.org/10.1371/journal.pone.0162333

Lemons, P. P., & Lemons, J. D. (2013). Questions for assessing higher-order cognitive skills: It's not just Bloom's. CBE—Life Sciences Education, 12(1), 47–58. https://doi.org/10.1187/cbe.12-03-0024

Matz, R. L., Fata-Hartley, C. L., Posey, L. A., Laverty, J. T., Underwood, S. M., Carmel, J. H., Herrington, D. G., Stowe, R. L., Caballero, M. D., Ebert-May, D., & Cooper, M. M. (2018). Evaluating the extent of a large-scale transformation in gateway science courses. Science Advances, 4(10), eaau0554. https://doi.org/10.1126/sciadv.aau0554

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. https://doi.org/10.1187/cbe.10-01-0001

Momsen, J. L., 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. https://doi.org/10.1187/cbe.12-08-0130

Monrad, S. U., Zaidi, N. L. B., Grob, K. L., Kurtz, J. B., Tai, A. W., Hortsch, M., Gruppen, L. D., & Santen, S. A. (2021). What faculty write versus what students see? Perspectives on multiple-choice questions using Bloom's Taxonomy. Medical Teacher, 43(5), 575–582. https://doi.org/10.1080/0142159X.2021.1879376

Moon, S., Jackson, M. A., Doherty, J. H., & Wenderoth, M. P. (2021). Evidence-based teaching practices correlate with increased exam performance in biology. PLOS ONE, 16(11), e0260789. https://doi.org/10.1371/journal.pone.0260789

National Academies of Sciences, Engineering, and Medicine. (2016). Barriers and Opportunities for 2-Year and 4-Year STEM Degrees: Systemic Change to Support Students' Diverse Pathways (S. Malcom & M. Feder, Eds.). National Academies Press. https://doi.org/10.17226/21739

National Academies of Sciences, Engineering, and Medicine. (2021). Call to Action for Science Education: Building Opportunity for the Future. National Academies Press. https://doi.org/10.17226/26152

National Academies of Sciences, Engineering, and Medicine. (2022). Imagining the Future of Undergraduate STEM Education: Proceedings of a Virtual Symposium (K. Brenner, A. Beatty, & J. Alper, Eds.). National Academies Press. https://doi.org/10.17226/26314

National Research Council. (1996). National Science Education Standards. National Academies Press. https://doi.org/10.17226/4962

National Research Council. (2003). Assessment in Support of Instruction and Learning: Bridging the Gap Between Large-Scale and Classroom Assessment - Workshop Report. National Academies Press. https://doi.org/10.17226/10802

National Research Council. (2007a). Rising Above the Gathering Storm: Energizing and Employing America for a Brighter Economic Future. National Academies Press. https://doi.org/10.17226/11463

National Research Council. (2007b). Taking Science to School: Learning and Teaching Science in Grades K-8. National Academies Press. https://doi.org/10.17226/11625

National Research Council. (2012). A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas. National Academies Press. https://doi.org/10.17226/13165

National Research Council. (2014). Developing Assessments for the Next Generation Science Standards. National Academies Press. https://doi.org/10.17226/18409

NGSS Lead States. (2013). Next Generation Science Standards: For States, By States. The National Academies Press.

Olson, S., & Riordan, D. G. (2012). Engage to Excel: Producing One Million Additional College Graduates with Degrees in Science, Technology, Engineering, and Mathematics. Report to the President. In Executive Office of the President. Executive Office of the President. https://eric.ed.gov/?id=ED541511

Partnership for Undergraduate Life Sciences Education. 2019. PULSE Regional Network. (https://pulse-community.org/regions).

Passmore, C., Stewart, J., & Cartier, J. (2009). Model-based inquiry and school science: Creating connections. School Science and Mathematics, 109(7), 394–402. https://doi.org/10.1111/j.1949-8594.2009.tb17870.x

Prevost, L., Sorensen, A. E., Doherty, J. H., Ebert-May, D., & Pohlad, B. (2019). 4DEE—What's next? Designing instruction and assessing student learning. Bulletin of the Ecological Society of America, 100(3), 1–6.

R Core Team. (2021). R: A language and environment for statistical computing (4.1.1). R Foundation for Statistical Computing. https://www.R-project.org/

Schwarz, C. V., Passmore, C., & Reiser, B. J. (2017). Moving beyond "knowing about" science to making sense of the world. In Helping Students Make Sense of the World Using Next Generation Science and Engineering Practices (pp. 3–21). National Science Teachers Association Press.

Scouller, K. (1998). The influence of assessment method on students' learning approaches: Multiple choice question examination versus assignment essay. Higher Education, 35(4), 453–472. https://doi.org/10.1023/A:1003196224280

Semsar, K., & Casagrand, J. (2017). Bloom's dichotomous key: A new tool for evaluating the cognitive difficulty of assessments. Advances in Physiology Education, 41(1), 170–177. https://doi.org/10.1152/advan.00101.2016

Stowe, R. L., & Cooper, M. M. (2017). Practicing what we preach: Assessing "critical thinking" in organic chemistry. Journal of Chemical Education, 94(12), 1852–1859. https://doi.org/10.1021/acs.jchemed.7b00335

Stowe, R. L., Scharlott, L. J., Ralph, V. R., Becker, N. M., & Cooper, M. M. (2021). You are what you assess: The case for emphasizing chemistry on chemistry assessments. Journal of Chemical Education, acs.jchemed.1c00532. https://doi.org/10.1021/acs.jchemed.1c00532

Sundberg, M. D., Dini, M. L., & Li, E. (1994). Decreasing course content improves student comprehension of science and attitudes towards science in freshman biology. Journal of Research in Science Teaching, 31(6), 679–693. https://doi.org/10.1002/tea.3660310608

Underwood, S. M., Posey, L. A., Herrington, D. G., Carmel, J. H., & Cooper, M. M. (2018). Adapting assessment tasks to support three-dimensional learning. Journal of Chemical Education, 95(2), 207–217. https://doi.org/10.1021/acs.jchemed.7b00645

Wright, C. D., Huang, A., Cooper, K., & Brownell, S. (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://doi.org/10.20429/ijsotl.2018.120214

Yuretich, R. F. (2003). Encouraging critical thinking: Measuring skills in large introductory science classes. Journal of College Science Teaching, 33(3), 40–45.

Zagallo, P., Meddleton, S., & Bolger, M. S. (2016). Teaching Real Data Interpretation with Models (TRIM): Analysis of student dialogue in a large-enrollment cell and developmental biology course. CBE—Life Sciences Education, 15(2), ar17. https://doi.org/10.1187/cbe.15-11-0239

Zheng, A. Y., Lawhorn, J. K., Lumley, T., & Freeman, S. (2008). Application of Bloom's Taxonomy debunks the "MCAT myth." Science, 319(5862), 414–415. https://doi.org/10.1126/science.1147852

SUPPLEMENTAL MATERIAL FOR CHAPTER 3

Supplemental Table 3.1: Codebook adapted from the Three-Dimensional Learning Assessment Protocol, BioCore Guide, and Bloom's Dichotomous Key

Code name	**Code criteria**
Scientific Practice	Indicates that the item does (1) or does not (0) assess a Science Practice (as defined by the 3D-LAP protocol). To code a 1, the item must meet the highest criteria for at least one of the following, "Asking Questions," "Developing and Using Models," "Planning Investigations," "Analyzing and Interpreting Data," "Using Mathematics and Computational Thinking," "Constructing Explanations and Engaging in Argument from Evidence" or "Evaluating Information."
Science Practice: Asking Questions	This code only applies to constructed response items. Student is asked to generate a scientific question about a real world event, observation, phenomenon, data, scenario, or model. 1. Question gives an event, observation, phenomenon, data, scenario, or model. 2. Question asks student to generate an empirically testable question about the given event, observation, phenomenon, data, scenario, or model.
Science Practice: Developing and Using Models	Constructed Response: Student is given or asked to construct a mathematical, graphical, computational, symbolic, or pictorial representation and use it to explain or predict an event, observation, or phenomenon. 1. Question gives an event, observation, or phenomenon for the student to explain or make a prediction about. 2. Question gives a representation or asks student to construct a representation. 3. Question asks student to explain or make a prediction about the event, observation, or phenomenon. 4. Question asks student to provide the reasoning that links the representation to their explanation or prediction. Selected Response: Student is given or asked to select a mathematical, graphical, computational, symbolic, or pictorial representation and select an appropriate explanation or prediction about an event, observation, or phenomenon based on the representation. 1. Question gives an event, observation, or phenomenon for the student to explain or make a prediction about. 2. Question gives a representation or asks student to select a representation. 3. Question asks student to select an explanation for or prediction about the event, observation, or phenomenon. 4. Question asks student to select the reasoning that links the representation to their explanation or prediction.

Science Practice: Planning Investigations	Constructed Response: Student is asked to design an experimental method or identify a set of observations that can be used to answer a scientific question or test a claim or hypothesis. 1. Question poses a scientific question, claim, or hypothesis to be investigated. 2. Question asks student to describe or design an investigation, or identify the observations required to answer the question or test the claim or hypothesis. 3. Question asks student to justify how their description, design, or observations can be used to answer the question or test the claim or hypothesis. Selected Response: Student is asked to select an appropriate design of an experimental method or an observation that can be used to answer a scientific question or test a claim or hypothesis. 1. Question poses a scientific question, claim, or hypothesis to be investigated. 2. Question asks student to select a description of or a design for an investigation or select the observations that could be used to answer the question or test the hypothesis. 3. Question asks student to select a justification of how the description, design, or observations can be used to answer the question or test the claim or hypothesis.
Science Practice: Analyzing and Interpreting Data	Constructed Response: Student is given a question, claim, or hypothesis and data collected from an experiment or observation and is asked to analyze the resulting data and interpret their meaning. 1. Question gives a scientific question, claim, or hypothesis to be investigated. 2. Question gives a representation of the data (e.g., table or graph, or list of observations) provided to answer the question or test the claim or hypothesis. 3. Question gives an analysis of the data or asks student to analyze the data. 4. Question asks student to interpret the results or assess the validity of the conclusions in the context of the scientific question, claim, or hypothesis. Selected Response: Student is given a question, claim, or hypothesis and data collected from an experiment or observation and is asked to select an interpretation of its meaning. 1. Question gives a scientific question, claim, or a hypothesis to be investigated. 2. Question gives a representation of data (table, graph, list of observations, etc.) provided to answer the question or test the claim or hypothesis. 3. Question asks student to select an interpretation of the results or an assessment of the validity of the conclusions in the context of the scientific question, claim, or hypothesis.

Science Practice: Using Mathematics and Computational Thinking	Constructed Response: Student is asked to use mathematical reasoning or a calculation and interpret the results within the context of the given event, observation, or phenomenon. 1. Question gives an event, observation, or phenomenon. 2. Question asks student to perform a calculation or statistical test, generate a mathematical representation, or demonstrate a relationship between parameters. 3. Question asks student to give a consequence or an interpretation (not a restatement) in words, diagrams, symbols, or graphs of their results in the context of the given event, observation, or phenomenon. Selected Response: Student is expected to perform a mathematical manipulation and asked to select an interpretation of the results within the context of a given event, observation, or phenomenon. 1. Question gives an event, observation, or phenomenon. 2. Question asks student to perform a calculation or statistical test, use a mathematical representation, or derive a relationship between parameters in order to obtain the correct answer. 3. Question asks student to select a consequence or an interpretation (not a restatement) in words, diagrams, symbols, or graphs of their results in the context of the given event, observation, or phenomenon.
Science Practice: Constructing Explanations and Engaging in Argument from Evidence	Constructed Response: Student is asked to provide reasoning based on evidence to support a claim. 1. Question gives an event, observation, or phenomenon. 2. Question gives or asks student to make a claim based on the given event, observation, or phenomenon. 3. Question asks student to provide scientific principles or evidence in the form of data or observations to support the claim. 4. Question asks student to provide reasoning about why the scientific principles or evidence support the claim. Selected Response: Student is asked to select reasoning and evidence to support a claim. 1. Question gives an event, observation, or phenomenon. 2. Question gives or asks student to select a claim based on the given event, observation, or phenomenon. 3. Question asks student to select scientific principles or evidence in the form of data or observations to support the claim. 4. Question asks student to select the reasoning about why the scientific principles or evidence support the claim.

Science Practice: Evaluating Information	Constructed Response: Student is asked to make sense of information or ideas presented to them. 1. Question gives an excerpt from a conversation, article, student solution, or video (or similar form of communication) that makes one or more assertions. 2. Question gives a conclusion about the validity of the assertion(s) made or asks student to make a conclusion about the validity of the assertion(s) or reconcile multiple assertions with each other. 3. Question asks student to provide reasoning to support their conclusion(s) about the validity of the assertion(s) or reconciliation with data, observations, or scientific principles. Selected Response: Student is asked to make sense of information or ideas presented to them. 1. Question gives an excerpt from a conversation, article, student solution, or video (or similar form of communication) that makes one or more assertions. 2. Question gives a conclusion about the validity of the assertion(s) or asks student to select a conclusion about the validity of the assertion(s) or reconciliation of multiple assertions. 3. Question asks student to select reasoning to support their conclusion(s) about the validity of the assertion(s) or reconciliation with data, observations, or scientific principles.
Crosscutting Concept	Indicates that the item does (1) or does not (0) assess a Crosscutting Concept (as defined by the 3D-LAP protocol). To code a 1, the item must meet the criteria for at least one of the following, “Patterns,” “Cause and Effect: Mechanism and Explanation,” “Scale,” “Proportion and Quantity,” “Systems and System Models,” “Energy and Matter: Flows, Cycles, and Conservation,” “Structure and Function,” and “Stability and Change.”
Crosscutting Concept: Patterns	To code an assessment task with Patterns, the question asks the student to identify patterns or trends emerging from three or more events, observations, or data.
Crosscutting Concept: Cause and Effect: Mechanism and Explanation	To code an assessment task with Cause and Effect: Mechanism and Explanation, the question provides at most two of the following: 1) a cause, 2) an effect, and 3) the mechanism that links the cause and effect, and the student is asked to provide the other(s).
Crosscutting Concept: Scale	To code an assessment task with Scale, the question asks the student 1) to compare objects, processes, or properties across size, time, or energy scales, or to dimensions of familiar objects, timescales, or energies or 2) to identify non-negligible/relevant interactions at various scales.
Crosscutting Concept: Proportion and Quantity	To code an assessment task with Proportion and Quantity, the question asks the student to predict the response of one variable to changes in another or identify the relationship between two or more variables from data.
Crosscutting Concept: Systems and System Models	To code an assessment task with Systems and System Models, the question asks the student to identify a system (by defining its components or boundaries), any assumptions made, and the surroundings (if necessary), and how the system and surroundings interact with each other.

Crosscutting Concept: Energy and Matter: Flows, Cycles, and Conservation	To code an assessment task with Energy and Matter: Flows, Cycles, and Conservation, the question asks the student to describe the transfer or transformation of energy or matter within or across systems, or between a system and its surroundings~~, with explicit recognition that energy and/or matter are conserved~~. *The phrase "with explicit recognition that energy and/or matter are conserved" is restrictive, and as a result, few items meet this crosscutting concept. We removed this phrase from our operational definition of the crosscutting concept "Energy and Matter: Flows, Cycles, and Conservation."*
Crosscutting Concept: Structure and Function	To code an assessment task with Structure and Function, the question asks the student to predict or explain a function or property based on a structure, or to describe what structure could lead to a given function or property. *To meet this crosscutting concept, the item needs to clearly address both structure and function. The function does not have to be immediate and may be either proximal or distal. Items that only ask to identify a structure do not meet this crosscutting concept.*
Crosscutting Concept: Stability and Change	To code an assessment task with Stability and Change, the question asks the student to determine 1) if a system is stable and provide the evidence for this, or 2) what forces, rates, or processes make a system stable (static, dynamic, or steady state), or 3) under what conditions a system remains stable, or 4) under what conditions a system is destabilized and the resulting state.
Core Idea	Indicates the that the item does (1) or does not (0) assess a Core Idea (as defined by the BioCore Guide). To code a 1, the item must meet the criteria for at least one of the following, "Evolution," "Information Flow," "Structure Function," "Transformations of Energy and Matter," and "Systems."

Core Idea: Evolution	To meet the Core Idea, the exam item must align with at least one of the following criteria: • Overarching Principle: All living organisms share a common ancestor. • Overarching Principle: Species evolve over time, and new species can arise, when allele frequencies change due to mutation, natural selection, gene flow, and genetic drift. • Molecular: Multiple molecular mechanisms, including DNA damage and errors in replication, lead to the generation of random mutations. These mutations create new alleles that can be inherited via mitosis, meiosis, or cell division. • Molecular: Mutations and epigenetic modifications can impact the regulation of gene expression and/or the structure and function of the gene product. If mutations affect phenotype and lead to increased reproductive success, the frequency of those alleles will tend to increase in the population. • Physiology: Mutations that change protein structure and/or regulation can impact anatomy and physiological function at all levels of organization. • Physiology: Most organisms have anatomical and physiological traits that tend to increase their fitness for a particular environment. • Physiology: Physiological systems are constrained by ancestral structures, physical limits, and the requirements of other physiological systems, leading to trade-offs that affect fitness. • Ecology/Evolutionary Biology: The characteristics of populations change over time due to changes in allele frequencies. Changes in allele frequencies are caused by random and nonrandom processes – specifically mutation, natural selection, gene flow, and genetic drift. Not all of these changes are adaptive. • Ecology/Evolutionary Biology: All species alive today are derived from the same common ancestor. New species arise when populations become genetically isolated and diverge due to mutation, natural selection, and genetic drift. Phylogenetic trees depict relationships among ancestral and descendant species, and are estimated based on data. • Ecology/Evolutionary Biology: Fitness is an individual's ability to survive and reproduce. It is environment-specific and depends on both abiotic and biotic factors. Evolution of optimal fitness is constrained by existing variation, trade-offs and other factors.

Core Idea: Information Flow	To meet the Core Idea, the exam item must align with at least one of the following criteria: • Overarching Principle: Organisms inherit genetic and epigenetic information that influences the location, timing, and intensity of gene expression. • Overarching Principle: Cells/organs/organisms have multiple mechanisms to perceive and respond to changing environmental conditions. • Molecular: In most cases, genetic information flows from DNA to mRNA to protein, but there are important exceptions. • Molecular: Gene expression and protein activity are regulated by intracellular and extracellular signaling molecules. Signal transduction pathways are crucial in relaying these signals. • Molecular: The signals that a cell receives depend on its location, and may change through time. As a result, different types of cells express different genes, even though they contain the same DNA. • Physiology: Information stored in DNA is expressed as RNA and proteins. These gene products impact anatomical structures and physiological function. • Physiology: Organisms have sophisticated mechanisms for sensing changes in the internal or external environment. They use chemical, electrical, or other forms of signaling to coordinate responses at the cellular, tissue, organ, and/or system level. • Ecology/Evolutionary Biology: Individuals transmit genetic information to their offspring; some alleles confer higher fitness than others in a particular environment. • Ecology/Evolutionary Biology: A genotype influences the range of possible phenotypes in an individual; the actual phenotype results from interactions between alleles and the environment.

Core Idea: Structure Function	To meet the Core Idea, the exam item must align with at least one of the following criteria: • Overarching Principle: Biological structures exist at all levels of organization, from molecules to ecosystems. A structure's physical and chemical characteristics influence its interactions with other structures, and therefore its function. • Overarching Principle: Natural selection leads to evolution of structures that tend to increase fitness within the context of evolutionary, developmental, and environmental constraints. • Molecular: The structure of a cell – its shape, membrane, organelles, cytoskeleton, and polarity – impacts its function. • Molecular: The three dimensional structure of a molecule and its subcellular localization impact its function, including the ability to catalyze reactions or interact with other molecules. Function can be regulated through reversible alterations of structure e.g. phosphorylation. • Molecular: The structure of molecules or organisms may be similar due to common ancestry or selection for similar function. • Physiology: Physiological functions are often compartmentalized into different cells, tissues, organs, and systems, which have structures that support specialized activities. • Physiology: The size, shape, and physical properties of organs and organisms all affect function. The ratio of surface area to volume is particularly critical for structures that function in transport or exchange of materials and heat. • Physiology: Structure constrains function in physiology; specialization for one function may limit a structure's ability to perform another function. • Ecology/Evolutionary Biology: Natural selection has favored structures whose shape and composition contribute to their ecological function. • Ecology/Evolutionary Biology: Competition, mutualism, and other interactions are mediated by each species' morphological, physiological, and behavioral traits.

Core Idea: Transformations of Energy and Matter	To meet the Core Idea, the exam item must align with at least one of the following criteria: • Overarching Principle: Energy and matter cannot be created or destroyed, but can be changed from one form to another. • Overarching Principle: Energy captured by primary producers is necessary to support the maintenance, growth and reproduction of all organisms. • Overarching Principle: Natural selection leads to the evolution of efficient use of resources within constraints. • Molecular: Energy captured by primary producers is stored as chemical energy. This stored energy can be converted through a series of biochemical reactions into ATP for immediate use in the cell. • Molecular: In cells, the synthesis and breakdown of molecules is highly regulated. Biochemical pathways usually involve multiple reactions catalyzed by enzymes that lower activation energies. Energetically unfavorable reactions are driven by coupling to energetically favorable reactions such as ATP hydrolysis. • Molecular: Intracellular and intercellular movement of molecules occurs via 1) energy-demanding transport processes and 2) random motion. A molecule's movement is affected by its thermal energy, size, electrochemical gradient, and biochemical properties. • Physiology: Energy captured by primary producers is stored as chemical energy. This stored energy can be converted into ATP, which is required for energetically demanding activities necessary for life, including synthesis, transport, and movement. • Physiology: Due to the inefficiency of biochemical reactions and other constraints, physiological processes are never 100% efficient. • Physiology: Organisms have limited energetic and material resources which must be distributed across competing functional demands. These include movement of material across gradients, growth, maintenance, and reproduction, inevitably leading to trade-offs. • Ecology/Evolutionary Biology: Energy captured by primary producers is stored as chemical energy. At each trophic level, most of this energy is used for maintenance, with a relatively small fraction available for growth and reproduction. As a consequence, each trophic level in an ecosystem has less energy available than the preceding level. • Ecology/Evolutionary Biology: Chemical elements are transferred among the abiotic and biotic components of an ecosystem; changes in the amount and distribution of chemical elements can impact the ecosystem.

Core Idea: Systems	To meet the Core Idea, the exam item must align with at least one of the following criteria: • Overarching Principle: Biological molecules, genes, cells, tissues, organs, individuals, and ecosystems interact to form complex networks. A change in one component of the network can affect many other components. • Overarching Principle: Organisms have complex systems that integrate internal and external information, incorporate feedback control, and allow them to respond to changes in the environment. • Molecular: Cells receive a complex array of chemical and physical signals that vary in time, location, and intensity over the lifespan of the organism; a cell's response depends on integration and coordination of these various signals. • Molecular: During development the signals a cell receives depend on its spatial orientation within the embryo and its intercellular interactions. As a consequence, cells adopt different cell fates depending on their local environment and/or cell lineage. • Molecular: Alteration of a single gene or molecule in a signaling network may have complex impacts at the cell, tissue or whole-organism level. • Physiology: Organ systems are not isolated but interact with each other through chemical and physical signals at the level of cells, tissues, and organs. • Physiology: An individual's physiological traits affect its interactions with other organisms and with its physical environment. • Physiology: In the face of environmental changes, organisms may maintain homeostasis through control mechanisms that often use negative feedback; others have adaptations that allow them to acclimate to environmental variation. • Ecology/Evolutionary Biology: The size and structure of a population is dynamic. A species' abundance and distribution are limited by available resources and by interactions between biotic and abiotic factors. • Ecology/Evolutionary Biology: Ecosystems are not isolated and static – they respond to change, both as a result of intrinsic changes to networks of species and as a result of extrinsic environmental drivers. Within an ecosystem, interactions among individuals form networks; changes in one node of a network can cause changes in other nodes – directly or indirectly. • Ecology/Evolutionary Biology: Biodiversity impacts many aspects of ecosystems.

Bloom's Taxonomy	Only apply the code for the highest level of Bloom's Taxonomy that the item is capable of assessing. 1 = Remember • To code for Remember, students could memorize the answer to the question and students are repeating nearly exactly what they have heard or seen in class materials (including lecture, textbook, laboratory, homework, clicker, etc.). 2 = Understand/Comprehend • To code for Understand/Comprehend, students demonstrate a conceptual understanding by putting the answer in their own words, matching examples to concepts, representing a concept in a new form (words to graph, etc.), etc., or demonstrate that they understand a concept by putting it into a different form (new example, analogy, comparison, etc.) than they have seen in class. 3 = Apply • To code for Apply, students are using data to calculate the value of a variable or are predicting the outcome of a trend of a fairly simple change to a scenario. 4 = Analyze • To code for Analyze, students are asked to compare/contrast information, or have to interpret data (graph, table, figure, story problem, etc.) and come to a conclusion about the data mean (they may or may not be required to explain the conclusion) and/or have to decide what data are important to solve the problem (i.e., picking out relevant from irrelevant information). 5 = Evaluate • To code for Evaluate, students have to interpret data (graph, table, figure, story, problem, etc.) then determine whether the data are consistent with a given scenario or whether conclusions are consistent with the data, critique validity, quality, or experimental data/methods, or make a judgment and/or justifying their answer. 6 = Create/Synthesize • To code for Create/Synthesize, students must be synthesizing information into a bigger picture (coherent whole) or creating something they haven't seen before (a novel hypothesis, a novel model, etc.), building up a model or novel hypothesis from data, or putting information from several areas together to create a new pattern/structure/model/etc.
Sources: NRC, 2012; Laverty et al., 2016; AAAS, 2011; Brownell et al., 2014; Semsar & Casagrand, 2017	

Supplemental Table 3.2: Percent agreement between two raters

Code name	Percent agreement
Scientific Practice	98
Scientific Practice: Asking Questions	100
Scientific Practice: Developing and Using Models	98
Scientific Practice: Planning Investigations	100
Scientific Practice: Analyzing and Interpreting Data	96
Scientific Practice: Using Mathematics and Computational Thinking	100
Scientific Practice: Constructing Explanations and Engaging in Argument from Evidence	98
Scientific Practice: Evaluating Information	100
Crosscutting Concept	75
Crosscutting Concept: Patterns	98
Crosscutting Concept: Cause and Effect	92
Crosscutting Concept: Scale	100
Crosscutting Concept: Proportion	100
Crosscutting Concept: Systems and System Models	85
Crosscutting Concept: Energy and Matter	92
Crosscutting Concept: Structure and Function	88
Crosscutting Concept: Stability and Change	100
Biology Core Idea	79
Biology Core Idea: Evolution	98
Biology Core Idea: Information Flow	90
Biology Core Idea: Structure Function	88
Biology Core Idea: Transformations of Energy and Matter	94
Biology Core Idea: Systems	83
Bloom's Taxonomy Level	79

Note: Percent agreement was calculated based on two rater's coding of 48 randomly selected exam items.

Supplemental Table 3.3: Computations of Fisher's *z*-tests concerning differences between correlations of weighted Bloom's Taxonomy level and the percentage of exam points in each dimension

	3D	Scientific practices	Crosscutting concepts	Core ideas
3D	-	1.52	-3.31	-1.93
Scientific practices	1.52	-	4.83	3.45
Crosscutting concepts	-3.31	4.83	-	-1.38
Core ideas	-1.93	3.45	-1.38	-

Note: Red text indicates significant differences ($p < 0.05$) between correlation coefficients.

CHAPTER 4: IDENTIFYING FACTORS ASSOCIATED WITH INSTRUCTOR IMPLEMENTATION OF THREE-DIMENSIONAL ASSESSMENTS IN UNDERGRADUATE BIOLOGY COURSES

ABSTRACT

Recent national calls to reform undergraduate science education have centered on engaging students in scientific practices as a means of helping them develop deeper and more robust understandings of foundational disciplinary concepts. A three-dimensional framework encapsulates the goals of these national calls, and we used alignment of course exams to this framework as a way to measure the progress of reform efforts in undergraduate biology. As very few biology exams were three-dimensionally aligned, we hypothesized that there are likely to be barriers or challenges that biology instructors face in meeting the goals of national calls. We sought to better understand these challenges and we used a generalized linear mixed model to predict what factors may be associated with three-dimensional alignment of course exams. Our model indicated that instructors who used three-dimensional items on their exams were more likely to write the items using a constructed-response format and were more likely to use Bloom's Taxonomy as a tool when designing their exams. We also found that professional development opportunities did not necessarily change the likelihood an instructor would have three-dimensional assessments. Based on these results, we suggest that institutions and departments consider supporting instructors with the time and resources needed to grade constructed-response assessments and that further refining of professional development offerings may be an important step in meeting the goals of national calls.

INTRODUCTION

For the past several decades, the landscape of science education has been defined by national calls for rich and contextualized teaching that engages in students in scientific processes to help them better understand foundational disciplinary concepts (American Association for the Advancement of Science [AAAS], 1989, 1990, 1993, 2011; National Academies of Sciences, Engineering, and Medicine [NASEM], 2016b, 2021, 2022; National Commission on Excellence in Education [NCEE], 1983; National Research Council [NRC], 1996, 2003, 2007, 2012a). Over the years, the focus of these calls has centered around different aspects of science education, such as scientific literacy (AAAS, 1989), inquiry (NRC, 1996, 2000), career preparation (NASEM, 2016b; NCEE, 1983; NRC, 2007), and integrating scientific concepts and competencies (AAAS, 2011; NRC, 2012a). Within the K-12 education system, public school districts are often held accountable for achieving the goals outlined in these calls through standardized assessments, accountability-based policies, and federal intervention programs (Hardy & Campbell, 2020; U.S. Department of Education et al., 2019); however, there are few analogous assessments, policies, and programs in postsecondary education to measure progress in meeting these national calls (NASEM, 2016a). Thus, the extent to which national calls have percolated through the undergraduate biology education system remains an area of active research. Recent research in this area tends to examine the impact of national calls on discrete levels of the education system, focusing on national-level discourse (Vasaly et al., 2014), department-level initiatives (Clark & Hsu, 2023; Peteroy-Kelly et al., 2019), and classroom-level implementation (Matz et al., 2018;

Uminski & Couch, in revision)[3]. Yet there still remain unanswered questions about how well these levels interact to support learning aligned with national priorities (NASEM, 2016a) and what factors in this system may help or hinder the implementation of national calls (Matz et al., 2018).

The undergraduate biology education system is composed of many levels, spanning from federal agencies, policymakers, and professional organizations to undergraduate institutions, science departments, and biology instructors. The conceptual model of coherence can help to shape our thinking about how national calls get translated across these levels of the education system. Coherence refers to a congruous alignment of the levels of the education system in ways that promote a common vision and reinforce norms for teaching and learning (Fuhrman, 1993; NRC, 2006, 2015; Webb, 1997). When biology education is coherent with the priorities outlined in national calls, learning outcomes that integrate scientific content and scientific practices are emphasized by institutions, supported by departments, and enacted within biology classrooms.

Coherence can be difficult to achieve, however, as there are often conflicting priorities at different levels within education systems (Cherbow et al., 2020) which may be reflected in the resources and types of supports that are provided to instructors to improve their classroom practice (Bradforth et al., 2015). Such resources and supports may be in the forms of providing professional development opportunities (Smith et al., 2014; Sunal et al., 2001), incorporating Learning Assistants or Teaching Assistants in high-enrollment courses (Biswas et al., 2022; Matz et al., 2018), and having faculty with discipline-based education research experience within the department (NRC, 2012b;

[3] The citation Uminski & Couch (in revision) refers to the text within Chapter 3, which was submitted as a manuscript and is currently under revision.

Wieman et al., 2010). The decisions they make about their classroom may be linked to the degree to which their institutions and departments provide such resources and supports that enhance their capacity to implement instruction in line with that envisioned in national calls (Austin, 2011; Stepans et al., 2001). Thus, when local practice does not reflect national calls, we can use the concept of coherence as a lens for identifying potential constraints or barriers in the education system to determine where instructors may need additional support.

Previous studies help to inform our understanding of how national calls to improve science education have permeated into undergraduate biology (Bain et al., 2020; Clark & Hsu, 2023; Clemmons et al., 2022; Crowe et al., 2008; Durham et al., 2017, 2018; Ebert-May et al., 2011; Matz et al., 2018; Momsen et al., 2010, 2013; Peteroy-Kelly et al., 2019; Vasaly et al., 2014). These studies often rely on validated instruments and protocols that can be used as tools for examining the current state of biology departments and classrooms through the lens of pedagogical frameworks including Bloom's Taxonomy (Anderson et al., 2001; Bloom et al., 1956), Scientific Teaching (Couch, Brown, et al., 2015; Handelsman et al., 2004, 2007), *Vision and Change* (AAAS, 2011; Brownell et al., 2014; Clemmons et al., 2020), and Three-Dimensional Learning (NRC, 2012b; NGSS Lead States, 2013). While the frameworks in these studies may center on different facets of undergraduate biology education, each framework encapsulates the main goals of the national calls by emphasizing student engagement in science through evidence-based instructional practices. Across these studies, a common finding was that the current biology education system may not be consistently or effectively meeting the main goals of the national calls. From national-level discourse to department-level learning objectives to classroom-level instruction and assessment, these studies indicate that there remain gaps between what is envisioned for undergraduate biology education and what is actually enacted.

We aim to better understand this gap between envisioned and enacted educational goals through the lens of a three-dimensional framework (NRC, 2012a) which developed from a robust synthesis of educational research in response to national calls (e.g., NCEE, 2008; NRC, 2007; Schmidt et al., 1997). This framework suggests that students develop deep understanding of science when their learning integrates scientific practices (e.g., skills and processes used by scientists) with both crosscutting concepts (i.e., interdisciplinary approaches to thinking about scientific phenomena) and disciplinary core ideas (i.e., foundational concepts central to each science discipline). While the three-dimensional framework was intended for K-12 science education and is widely used in statewide science education standards (NASEM, 2021; NGSS Lead States, 2013), this framework is easily translated to the undergraduate level and is particularly relevant for gateway introductory-level courses that bridge many students' high school science experiences (Bain et al., 2020; Cooper et al., 2015; Laverty et al., 2016; Matz et al., 2018; Radloff et al., 2022).

The three-dimensional framework scaffolds science curriculum, instruction, and assessment to align with national priorities, but in this work here, we narrow the focus of our study to only assessment. Guided by principles of backward design (Wiggins & McTighe, 2005), we can use the content and skills on assessments to make inferences about the learning objectives that were included in curriculum and incorporated into instruction. Our study of assessments specifically looked at three-dimensional alignment in course exams. Exams are types of summative assessments that tend to carry a significant weight in course grades and are a common assessment strategy in undergraduate science courses (Gibbons et al., 2022; Goubeaud, 2010; Hurtado et al., 2012; Stanger-Hall, 2012; Wright et al., 2018). Since what is included on exams reflects what instructors intend for students to learn, the content and skills targeted on exams can be used to gauge the extent that these same content and skills have also been taught to students during instruction (Scouller, 1998; NRC, 2003). Hence, if an exam is three-dimensional, we can assume that students have encountered the associated scientific practice, crosscutting concept, and core idea in their biology class.

The approach of using assessments as a proxy for course alignment to the three-dimensional framework has been used in several studies (Matz et al., 2018; Stowe et al., 2020, 2021; Stowe & Cooper, 2017; Uminski & Couch, in revision). These studies used the Three-Dimensional Learning Assessment Protocol (Laverty et al., 2016) as a tool for characterizing the three-dimensional alignment of assessment items (i.e., exam questions). A common finding across these studies was that the majority of items in undergraduate science courses were not three-dimensionally aligned. Given the low frequency of three-dimensional assessment items, Matz et al. (2018) raised a question about which supports and barriers help or hinder the use of the three-dimensional framework in undergraduate science. To date, this question remains unanswered, and we still know very little about what factors affect how undergraduate science instructors implement the three-dimensional framework in their courses.

Our work here seeks to answer the question posed by Matz et al. (2018) in the context of undergraduate biology courses and builds off of our past work looking at three-dimensional alignment of biology exams. We previously found that only 5% of the items in our nationwide sample of undergraduate biology exams were three-dimensional—a finding that was largely driven by the small number of scientific practices we observed (Uminski & Couch, in revision). Scientific practices occurred in less than 10% of biology exam items, as compared to crosscutting concepts and core ideas, which were present in approximately half and two-thirds of items, respectively. Based on the infrequency of three-dimensional items we observed, our past work suggests that lower-division biology instructors likely encounter barriers to implementing the goals of national calls in their courses. We infer that such barriers to three-dimensional assessment were most likely related to the challenges of assessing scientific practices in an exam format, particularly when the exams mainly use a closed-ended or selected-response format like multiple choice. We also hypothesize that three-dimensional alignment can be challenging because eliciting explicit evidence that students have engaged in a scientific practice can be a daunting task for instructors, especially when there is a lack of training, a lack of resources, or a lack of support for implementing three-dimensional assessments (Furtak, 2017; Laverty et al., 2016; National Research Council, 2014; Siebert & McIntosh, 2001). The purpose of our current research is to contextualize our previous findings about the low frequency of scientific practices in biology exams and to better understand what barriers may exist to three-dimensional assessment in undergraduate science education. We aim to answer the following research question: What constraints and challenges are undergraduate biology instructors facing in implementing three-dimensional assessments in their courses and where may they need additional support?

METHODS

Survey Development and Administration

Our methods in this study expand upon the methods and data collection reported in Uminski & Couch (in revision). Briefly, we developed an online survey through Qualtrics intended to collect course artifacts (e.g., a course syllabus, a summative exam, the exam answer key) along with demographic and institutional information from instructors of lower-division undergraduate biology courses. We define lower-division courses as 100- and 200-level courses and their equivalents. Our final dataset contained responses from 111 lower-division biology instructors at 100 unique undergraduate institutions across the United States. Our sample includes broad representation from each undergraduate institution type as defined by Carnegie classifications (see Table 3.1) and from instructors across career stages (see Table 3.2). The majority of the courses in this study were introductory-level (80%), and the remaining courses spanned a variety of lower-division biology topics including anatomy and physiology, environmental science, and microbiology (see Table 3.3).

In our survey, we asked instructors to self-report on a series of factors we anticipated might be related to the structure and design of their assessments. These factors ranged from instructional practices (e.g., Scientific Teaching methods) to department-level policies (e.g., providing support for professional development). Brief descriptions of these factors and how they were measured are outlined in Table 4.1. The survey items and additional descriptions of how these factors were measured are in Supplemental Material 4.1.

This research was classified as exempt from human-subjects review by the University of Nebraska–Lincoln (protocol 21082).

Item Coding

Our dataset contained 111 exams consisting of 4337 items (i.e., questions). We used the point values and numbering schemes specified by the instructor to determine the boundaries of individual items. In line with recommendations from (Laverty et al., 2016), we coded items that shared a common stem and/or used a sub-part numbering scheme (e.g., 2a, 2b, 2c) as a single clustered item. As exams use different grading point schemes across courses, we calculated a normalized item point value by dividing individual item point value by the total number of points on the exam and multiplying it by 100.

We coded individual exam items for three-dimensional alignment using existing protocols and tools for characterizing assessments in undergraduate science courses. Briefly, we coded scientific practices and crosscutting concepts based on the Three-Dimensional Learning Assessment Protocol (Laverty et al., 2016). We coded core ideas from the *Vision and Change* core concepts (AAAS, 2011), as delineated in the BioCore Guide (Brownell et al., 2014). We coded for Bloom's Taxonomy levels using the Bloom's Dichotomous Key (Semsar & Casagrand, 2017). We assigned Bloom's levels ordinal numeric values between 1 and 6, where 1 = remember, 2 = understand, 3 = apply, 4 = analyze, 5 = evaluate, and 6 = create, and only coded the highest Bloom's value the item was capable of eliciting. There was 93% agreement between two raters across this set of codes and ≥ 75% agreement for each individual code. For full details on coding procedures and calculation of interrater reliability, please see Uminski & Couch (in revision).

Table 4.1: Factors we anticipated might be related to how undergraduate biology instructors design their course exams

Factor	Measurement
Authorship	Self-reported data about whether the instructor wrote original exam items, sourced the exam items from other materials, or had a combination of both original and sourced items.
Course audience	Self-reported data about whether the course was intended for students with STEM majors, non-STEM majors, or both STEM and non-STEM majors.
Course lab	Self-reported data about whether the course had an associated lab component.
Course setting	Self-reported data about whether the course was taught in-person, online, online (because of the COVID-19 pandemic but had previously been taught in-person), or hybrid (both in-person and online).
Department DBER faculty	Self-reported data about whether the instructor's department contains any faculty who identify as discipline-based education researchers (including the instructor themselves, if applicable).
Department professional development	Self-reported data about whether the instructor's department has allocated resources (e.g., time or money) for faculty professional development.
Exam weight	The percentage of the final course grade that was attributed to summative exams (including midterm and final exams if applicable). Data was collected from course syllabus documents.
Institution type	Institutions were classified as Associate's, Baccalaureate, Master's or Doctoral based on the 2018 Carnegie classifications (Indiana University Center for Postsecondary Research, 2021).
Instructor professional development	Self-reported data about the extent to which the instructor completed professional development about assessment (reported in 4-hour time increments).
Item point value	The point value of individual exam items was collected from either the exam document, the associated answer key, or instructor-provided text description of their exam. Item point values were normalized across each instructor's exam by dividing the point value of the item by the total number of points on the exam and multiplying by 100.
Item response format	Individual exam items were classified as selected-response or constructed-response based on whether students were provided a list of options to pick from or had to generate a response to the item. See Supplemental Table 4.1 for additional details.
Scientific Teaching	Self-reported data about the degree to which instructional practices aligned with Scientific Teaching principles related to active learning, data analysis and interpretation, and experimental design. Data was collected using subscales of the Measurement Instrument for Scientific Teaching (MIST; Durham et al., 2017).
Teaching years	Self-reported data about the number of years of teaching experience (reported in 5-year time increments).
Use of 3D-LAP	Self-reported data about the degree to which instructors used the Three-Dimensional Learning Assessment Protocol (3D-LAP; Laverty et al., 2016) when writing their exams. Reported using a Likert scale ranging from Never to Almost Always.
Use of Bloom's Taxonomy	Self-reported data about the degree to which instructors used Bloom's Taxonomy (Bloom et al., 1956) when writing their exams. Reported using a Likert scale ranging from Never to Almost Always.
Use of Vision and Change	Self-reported data about the degree to which instructors used *Vision and Change* (AAAS, 2011) when writing their exams. Reported using a Likert scale ranging from Never to Almost Always.

Coding for Item Format

Using the coding protocols described in Uminski & Couch (in revision), we coded 13 different item formats that were classified into either the constructed-response or selected-response item type and there was 98% agreement between the two authors. We consider constructed-response items (i.e., open-ended) items those that required students to generate an original response and selected-response (i.e., closed-ended) items those that asked students to select from a predetermined or provided set of responses. Constructed-response item types included fill-in-the-blank, short answer, and essay, which were determined by the relative length of the expected student response (a single word or phrase, up to a paragraph, or multiple paragraphs, respectively). Constructed-response items also included clusters (a series of constructed-response items that shared a common stimulus or prompt), math manipulation (involving an algorithmic calculation), modeling (test taker creates or modifies a model), and discipline-specific items (procedures, algorithms, or processes specific to biological sciences, such as complementary base pairing, completing Punnett squares). Selected-response items included multiple-choice, multiple select (a multiple-choice item in which more than one option is selected), true-false, multiple-true-false, matching, and reorder. Full descriptions of the item types coded are in Supplemental Table 4.1.

Recoding for Partial Alignment to Scientific Practices

The 3D-LAP coding protocol (Laverty et al., 2016) provides a set of 2-4 criteria statements for each scientific practice. Strictly following the protocol recommended by the 3D-LAP, scientific practices are coded in a binary manner based on whether or not the item meets all the criteria statements for a given scientific practice. There is value in using the binary approach to scientific practices, but we found that few instructors were meeting the standards for full alignment to the practices. To better represent the variation underlying this binary coding, we recoded our data into consistent ordinal scale based on the number of scientific practice criteria statements to which each item aligned. This scale included the categories: not aligned, partially aligned, mostly aligned, or fully aligned to a scientific practice. Briefly, items that were not aligned did not meet any of the criteria statements for a scientific practice. Items that were partially aligned met surface-level criteria, such as including a real-world biological phenomenon described in text or presented as a visual model. Items that were mostly aligned met the majority of the scientific practice criteria but lacked a prompt for students to explicitly engage in the scientific practice by providing reasoning or justification of their thought processes. Items that were fully aligned met each criteria statement for the scientific practice. When items met criteria for multiple scientific practices, we coded the item at the highest level of alignment. For further details on the translation of the 3D-LAP protocol into the partial alignment coding scheme, see Supplemental Material 4.2.

MIST Instrument

Our survey contained an abbreviated version of the Measurement Instrument for Scientific Teaching (MIST; Durham et al., 2017, 2018), consisting of the items within the subcategories of Active Learning Strategies, Data Analysis and Interpretation, and Experimental Design and Communication. We applied the methods outlined in Durham et al. (2017) for normalizing the three MIST subcategories into a single MIST scale in which the responses from the MIST items were summed and divided by the number of contributing questions and multiplying by 100. The resulting MIST scores were on a 0-100 scale with higher MIST scores indicate the instructor reported using a greater amount of Scientific Teaching practices in their classroom instruction.

Statistical Analysis

We categorized three-dimensional alignment of items as a binary variable (i.e., items were either three-dimensional or not three-dimensional); thus, when three-dimensional alignment was the response variable, we used a generalized linear mixed model (GLMM) with a logit link. As we had multiple items per instructor in the sample, we included instructor as a random effect in the GLMM. We used forward stepwise model selection procedures that based on Akaike Information Criterion (AIC) to determine the subset of variables that explain variability in three-dimensional alignment while avoiding overfitting. Variables were individually tested for retention in the model and were only retained if the new model had an AIC value more than two units lower than the prior model. We conducted statistical analysis with R statistical software [v 4.2.3] (R Core Team, 2023) using tidyverse (Wickham et al., 2019) for data processing and figure generation and lme4 (Bates et al., 2015) for our GLMM.

RESULTS

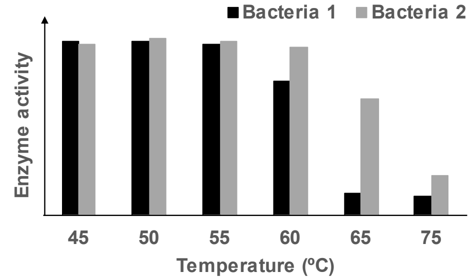
Identifying Three-Dimensional Items

Three-dimensional items were those that elicited evidence of student engagement with a scientific practice, crosscutting concept, and core idea. Three-dimensional items may have met the criteria for multiple scientific practices, crosscutting concepts, or core ideas within the same item. As there are few examples of three-dimensional exam items in the literature for undergraduate biology education, we provide a few examples of three-dimensional items in Figure 4.1. These examples are adapted from items in our sample and we pair each adaptation of a three-dimensional item with an adaptation of zero-dimensional item that was administered on the same exam. Zero-dimensional items did not meet the criteria for any of the scientific practices, crosscutting concepts,

(a)

A student compared the enzyme activity of two different bacteria across a range of temperatures and created a graph of their results (pictured at the left). Which conclusion is supported by the student's data?

A) The enzyme activity of both bacteria increased as the temperature increased because higher temperatures generally enhance enzymatic activity.
B) Bacteria 1 showed higher enzyme activity at all tested temperatures because it possesses a more thermally stable enzyme.
C) The enzyme activity of both bacteria decreased as the temperature increased because excessive heat can disrupt molecular structure of proteins leading to a loss of enzyme activity.
D) The two bacteria exhibited similar enzyme activity across all tested temperatures because the bacteria species are likely adapted to similar environmental conditions.

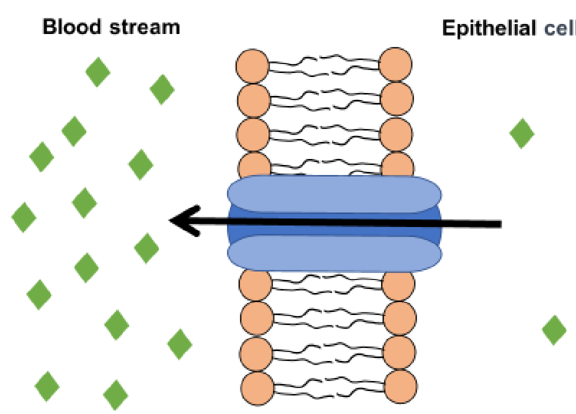


(b)

Which method would be most suitable for observing cilia on a bacterium's surface in great detail ?
A) Electron microscopy
B) Light microscopy
C) Scanning probe microscopy
D) Fluorescence microscopy

(c)

A student created a model (pictured at the right) to illustrate the how epithelial cells in the small intestines release glucose into the blood stream via facilitated diffusion. The model illustrates glucose (represented by green diamonds) moving in the direction of the black arrow through a membrane protein (in blue). How can this model be improved to be more accurate? Explain your reasoning.



(d)

Sporophytes are haploid.
A) True
B) False

Figure 4.1: Example three-dimensional and zero-dimensional items. Items (a) and (a) were adapted from one instructor, and items (c) and (d) were adapted from a second instructor in the sample. The three-dimensional item (a) is aligned to the scientific practice "Analyzing and Interpreting Data," the crosscutting concepts "Cause and Effect" and "Structure and Function" and the Core Idea "Structure Function." The three-dimensional item (c) is aligned to the scientific practice "Developing and Using Models," the crosscutting concepts "Patterns" and "Transformations of Energy and Matter," and the Core Idea "Energy Flow." The zero-dimensional items (b) and (d) are not aligned with any scientific practices, crosscutting concepts, or core ideas.

or core ideas. We note a few features of these sets of items are reflective of other items in our sample. The zero-dimensional items tend to focus on singular pieces of discrete factual information that are important in biology but fall outside the purview of the core ideas. In contrast, the three-dimensional items ask students to draw upon a more robust understanding of biological phenomena and often incorporate small datasets, graphs, or models into the item stimulus.

Identifying Challenges and Constraints in Implementing Three-Dimensional Items

We used a generalized linear mixed-effects model with a logit link function to identify the most salient factors affecting the likelihood that an item fully aligns to the three-dimensional framework. After model selection, our model retained the following predictors: institution type, use of Bloom's Taxonomy, item point value, and item response format. While it is important to consider the factors that are associated with three-dimensional items, it is also important to consider which predictors were excluded from the model. All factors related to course format (e.g., course setting, courses audience, courses labs) were excluded during model selection. We similarly saw little effect of instructor teaching methods and experience, and our best-fit model excluded factors such as years of teaching experience, amount of professional development related to assessment, instructional practices related to Scientific Teaching, and instructor use of educational frameworks and tools such as *Vision and Change* and the 3D-LAP. Our best-fit model also excluded factors at the department level, and we saw no effect of department support for professional development or departments that contain faculty with discipline-based education research expertise on the likelihood of three-dimensional alignment.

Our model indicated that odds of an item being three-dimensional increased when the item had a higher point value, when the item used a constructed-response format, or when the item was written by an instructor who more frequently used Bloom's Taxonomy (Table 4.2). When holding all other factors constant, our model predicted that item response format would have the greatest effect on the likelihood of an item being three-dimensional. Constructed-response items were 11.75 times more likely to be three-dimensional compared to selected-response items. Our model also indicated that, when controlling for other factors, each one percent increase in the normalized item point value

Table 4.2: Generalized linear mixed model[a] with binomial logit link predicting whether an item was likely to be three-dimensionally aligned

Term	Estimate	Standard Error	Odds Ratio	Confidence Interval
Item point value	0.05	0.02	1.06	[1.01, 1.10]
Item response format: Constructed response	2.46	0.23	11.75	[7.55, 18.30]
Institution type: Baccalaureate	-0.80	0.51	0.45	[0.17, 1.21]
Institution type: Master's	-0.22	0.53	0.80	[0.28, 2.29]
Institution type: Doctoral	0.68	0.47	1.97	[0.78, 4.95]
Use of Bloom's Taxonomy	0.38	0.17	1.46	[1.05, 2.03]
$R^2 = 0.496$				

[a]Model: Three-dimensional alignment ~ item point value + item response format + institution type + use of Bloom's Taxonomy + (1|instructor), family = binomial(link = logit)

increased the likelihood of an item being three-dimensional by 6%. In addition to how instructors wrote and assigned point values to individual items, we found an effect from instructors who reported using Bloom's Taxonomy more frequently when they were constructing their exam. We used a five-point Likert scale to measure the frequency of using Bloom's Taxonomy and instructors' items were 1.46 times more likely to be three dimensional for each additional one-unit increase they reported on this scale. Institution type was retained in the best-fit model.

Identifying Generalizable Characteristics of Three-Dimensional Items

We used the factors identified in our best-fit model as a lens for examining the generalizable characteristics of three-dimensional items; thus, we narrowed our analysis to the response format, point value, and Bloom's Taxonomy levels of the items in our sample. We found that over half of three-dimensional items (55%, n = 130) used a constructed-response format compared to only 10% (n = 436) that were not three-dimensional (Figure 4.2). Among the three-dimensional items, short answer and clusters were the most commonly used constructed-response item type (Table 4.3). Of the three-dimensional items, nearly all of the selected response items were multiple choice, but this was a trend common to both three-dimensional and non-three-dimensional items.

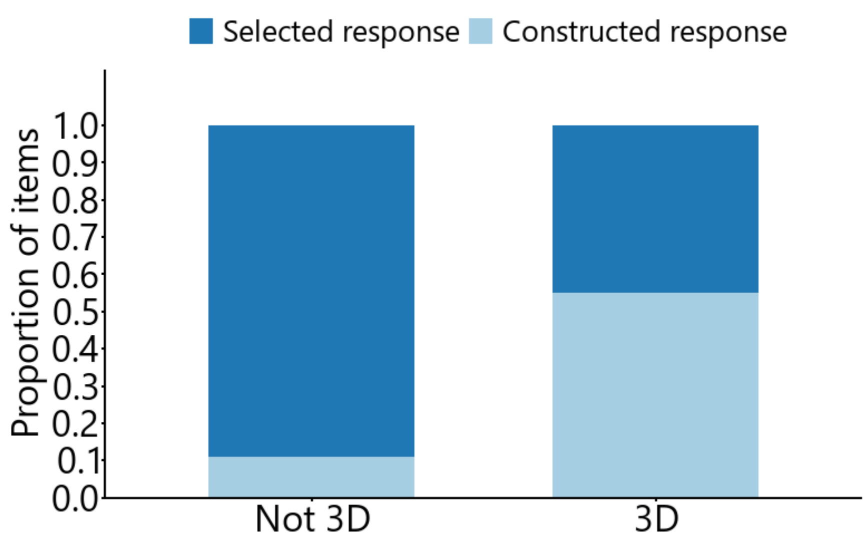


Figure 4.2: Proportion of three-dimensional and non-three-dimensional items using selected-response and constructed-response item types. Out of the entire sample of items (n = 4337), there were 236 items that were three-dimensional and 4101 items that were not three-dimensional.

Table 4.3: Item types of three-dimensional and non-three-dimensional items

Item type[a]	Non-three-dimensional items	Percent[b]	Three-dimensional items	Percent[b]
Multiple choice	3145	72.52	95	2.19
Matching	240	5.53	10	0.23
Short answer	216	4.98	71	1.64
True-False	173	3.99	0	0.00
Fill-in-the-blank	100	2.31	1	0.02
Multiple select	64	1.48	1	0.02
Cluster	41	0.95	35	0.81
Multiple-True-False	30	0.69	0	0.00
Model	29	0.67	6	0.14
Essay	20	0.46	16	0.37
Discipline-specific	15	0.35	1	0.02
Math manipulation	15	0.35	0	0.00
Reorder	13	0.30	0	0.00

[a]Multiple choice, matching, True-False, Multiple True-False, and reorder items use a selected-response format. Short answer, fill-in-the-blank, cluster, model, essay, discipline-specific, and math manipulation items use a constructed-response format. See Supplemental Table 4.1 for additional details about the classification of these item types.
[b]Percentage was calculated based on the total item pool (n = 4337 items)

Multiple choice was the most common item type, representing almost three-quarters of the items within our entire sample. There were no three-dimensional items that used the selected-response true-false, multiple-true-false, or reorder item types. There were also no three-dimensional items that used the constructed-response math manipulation item type, which is characterized by students writing out their mathematical computations.

Across our sample, three-dimensional items were worth more points on exams (Welch ANOVA, $F(1, 238.2) = 65.7$, $p < .001$). On average, three-dimensional items were assigned 5.70 ± 0.41 SE points compared to 2.35 ± 0.034 points for non-three-dimensional items. We found that the higher item point value of three-dimensional items was associated with the response format of the item (Figure 4.3) and that there was a significant interaction between response format and three-dimensional alignment (Supplemental Table 4.2). On average, constructed-response items tended to be

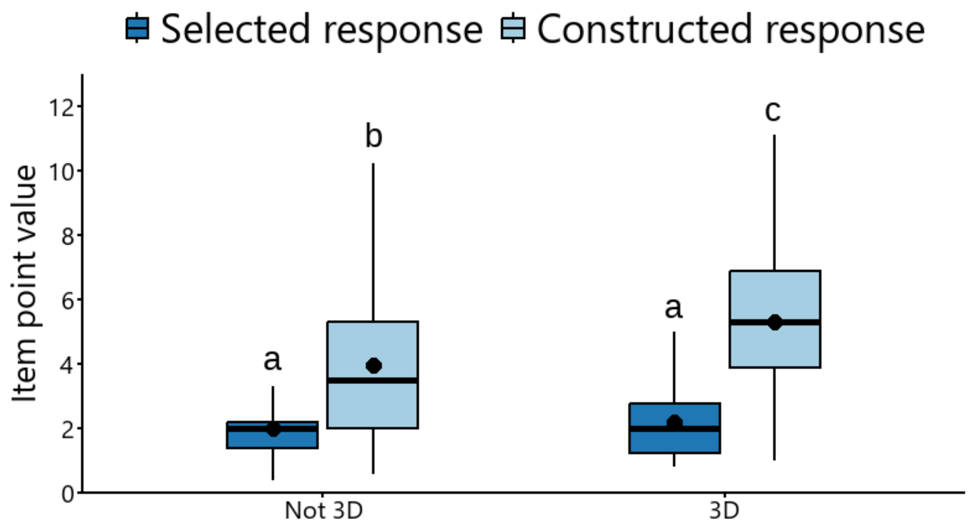


Figure 4.3: Normalized item point value by item response format and three-dimensional alignment. Boxes represent the interquartile range and whiskers represent the fifth and ninety-fifth percentile. Dots represent the mean value. Letters indicate a statistically significant difference ($p < 0.05$) between groups.

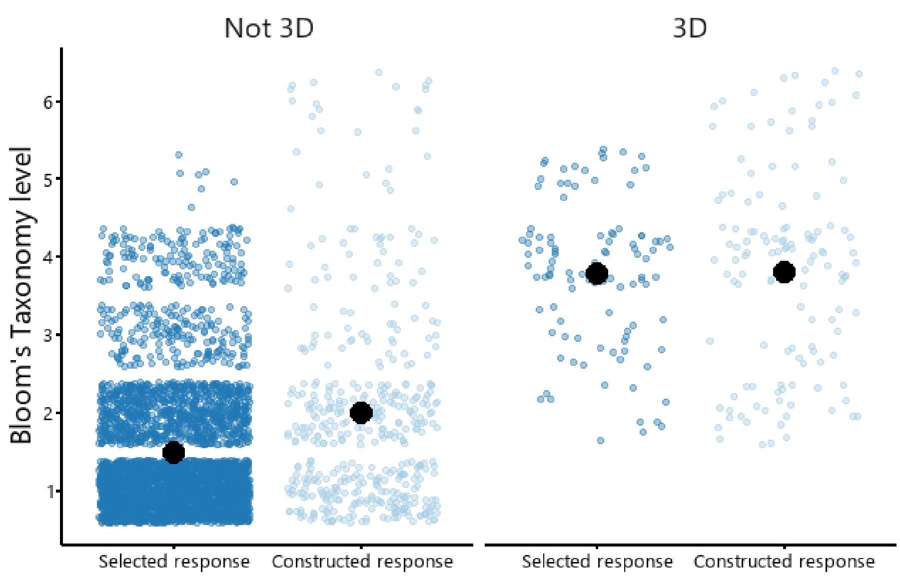


Figure 4.4: Bloom's Taxonomy level by item response format and three-dimensional alignment. Smaller points represent individual exam items and are jittered to better illustrate overlapping points. Larger black dots represent the mean value for each group.

worth more points than selected-response items (M = 5.93 ± 0.25 SE; M = 2.03 ± 0.017 SE, respectively), and the average point value tended to be even higher when we examine the subset of constructed-response items that are three-dimensional (M = 8.56 ± 0.64 SE).

Our best-fit model suggested that instructors who use Bloom's Taxonomy more frequently had a greater likelihood of having three-dimensional items, so we investigated the relationship between the Bloom's Taxonomy levels and the three-dimensional alignment of items. We found that the average Bloom's level of three-dimensional items (M = 3.80 ± 0.073 SE) was greater than that of items that are not three-dimensional (M = 1.54 ± 0.013 SE) and we observed an interacting effect with response format (Figure 4.4; Supplemental Table 4.3). Constructed-response items tended to have a higher average Bloom's level than selected response items (M = 2.42 ± 0.062 SE; M = 1.55 ± 0.014 SE, respectively) and when accounting for variation between instructors, three-dimensional constructed response items had a higher average Bloom's level than three-dimensional selected-response items, although this difference is small and marginally significant (p = 0.044).

Reevaluating Alignment to Scientific Practices

We previously found that the low three-dimensional alignment was driven by the small number of items fully meeting the 3D-LAP criteria for scientific practices (Uminski & Couch, in revision). To fully meet the 3D-LAP criteria for scientific practices, the item had to explicitly ask students to indicate their reasoning or to justify their thinking about a scientific phenomenon. We hypothesized that the low number of three-dimensional items could be in-part attributed to the stringent coding scheme of the 3D-LAP for scientific practices rather than a lack of scientific practices being incorporated into undergraduate biology education. To test this hypothesis, we analyzed our data to illustrate degrees of alignment to the 3D-LAP scientific practice criteria statements (Figure 4.5). We found that even when accounting for partial alignment, most items (61%, n = 2666) still did not meet any of the criteria for scientific practices (Figure 4.5a). Approximately 19% of items were partially aligned to a scientific practice because they met surface-level criteria by including a biological phenomenon. About 12% of items were mostly aligned to a scientific practice but failed to meet full alignment because they did not ask students to explicitly engage in the practice using reasoning or justification. Together, these partially- and mostly-aligned items suggest that about a third of the items in our sample have the potential to be transformed into three-dimensional items.

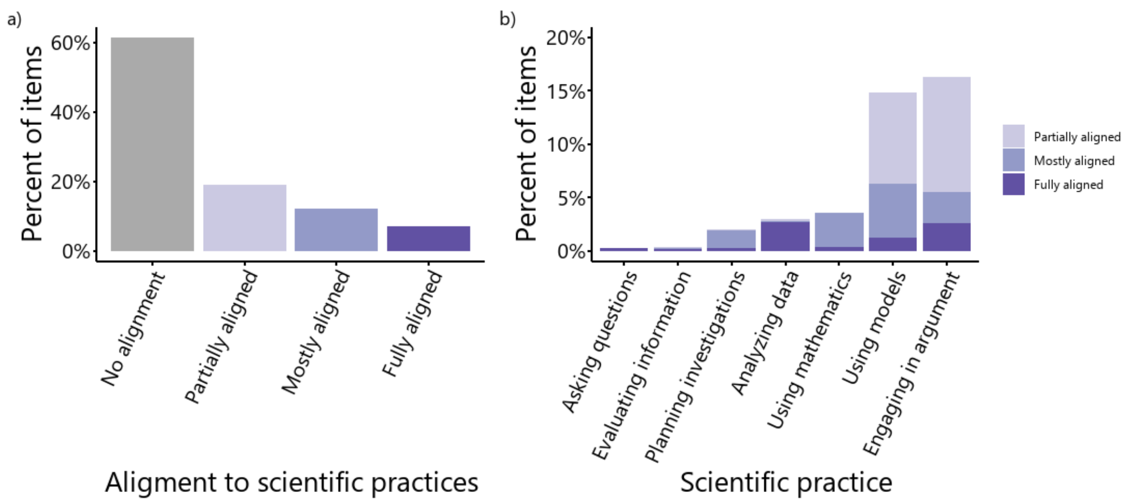


Figure 4.5: Partial alignment of biology exam items to 3D-LAP criteria for scientific practices. a) The highest level of alignment to scientific practices out of the entire item pool (n = 4337). One or more scientific practices may have been present within the item, but only the highest level of alignment to any of the scientific practices was recorded. b) Alignment of items to each scientific practice. Percent of items is calculated out of the entire item pool. Items may have aligned to different scientific practices and may be represented in multiple columns.

Looking at this subset of items that were aligned to scientific practices criteria statements, we found that the occurrence of partial alignment was not evenly distributed across the scientific practices (Figure 4.5b). “Constructing Explanations and Engaging in Argument” and “Developing and Using Models” were represented most frequently, likely reflecting the low bar for partial alignment which could be reached by including a real-world phenomenon in text or in model, respectively. Our partial alignment coding also allows us to see that instructors were incorporating elements of the practices “Using Mathematics and Computational Thinking” and “Planning Investigations,” but were missing the criteria for assessing reasoning that is required for full alignment to these scientific practices. Interestingly, we did not see many items partially aligned to the practice “Analyzing and Interpreting Data.” When instructors had exam items that involved data analysis, they were often fully meeting the associated scientific practice. There were no instances of partial alignment for the practice “Asking Questions,” but this

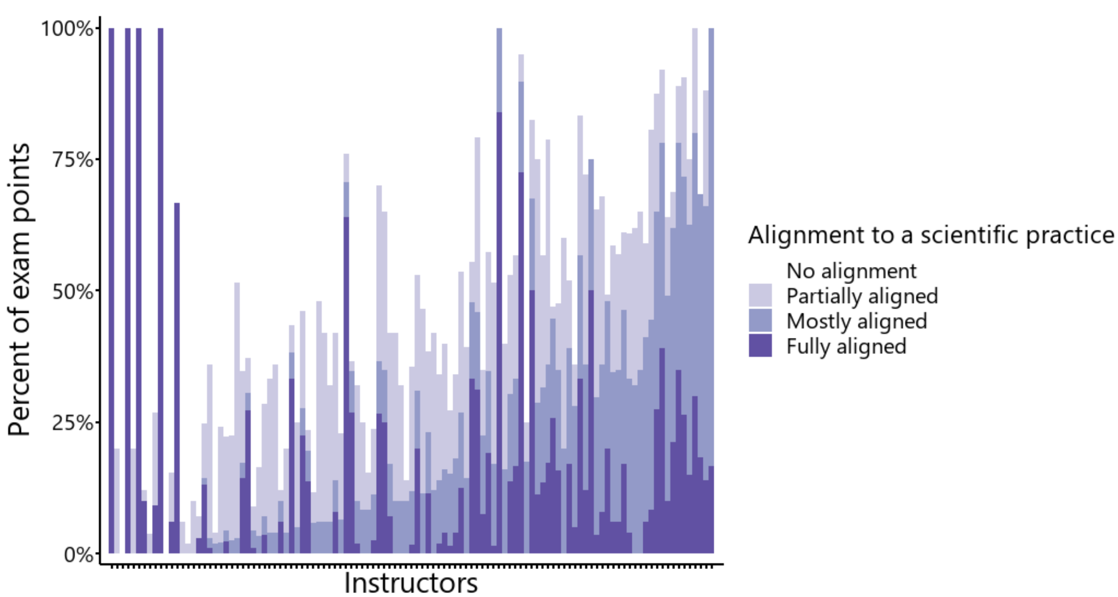


Figure 4.6: Alignment of each instructor’s exam to scientific practices. Each instructor is represented as a bar in this graph. The instructors are sorted by increasing percentage of exam points that were mostly aligned to scientific practices.

is an artifact of the coding scheme for this scientific practice which only contained two criteria statements (as compared to the other scientific practices which all had either three or four criteria statements).

When we look at the characteristics of instructors' exams as a whole, it becomes clear that the majority of instructors had items that incorporated important components of scientific practices (Figure 4.6). Excluding the four instructors who had all of their exam points fully aligned to scientific practices, 98% of instructors (n = 105) had at least one item that was partially or mostly aligned to a scientific practice and had the potential to be transformed into a three-dimensional item. Within this set of instructors that had room to incorporate more fully-aligned scientific practices into their exams, there was on average about 15% of points mostly aligned and 18% of points partially aligned to a scientific practice, but as Figure 4.6 indicates, there is a large amount of variation that underlies these averages. The percentages of instructor's exam points that were mostly aligned and partially aligned to scientific practices ranged from 0–83% and 0–48%, respectively.

DISCUSSION

Undergraduate biology education is a complex and interconnected system that spans from instructors to their institutions to the national landscape of STEM education and we sought to examine which factors in this system might help or hinder the use of assessments that reflect the educational priorities outlined in national calls. Using the conceptual model of coherence, we anticipated that national-level and institutional-level factors may provide support or place constraints on biology instructors in ways that affect their implementation of assessments that incorporate scientific practices, crosscutting concepts, and core ideas. While our work sought to identify these supports and constraints, we found very few significant relationships between these national- and institutional-level factors in terms of how they are related to three-dimensional alignment of instructor's exams. Thus, we conclude that these factors in the undergraduate biology education system are not necessarily hindering three-dimensional assessment, but they are not necessarily helping instructors implement the three-dimensional framework in their courses either. We found that challenges and constraints of three-dimensional assessment may be occurring mostly at the instructor level, with the most notable barriers likely being the time and resources required to grade constructed-response items that assess higher-order cognitive skills. Our research highlights the need for future work to better understand how instructors are meeting these national-level goals in their courses and what additional resources may be important for instructors to fully align their assessments, instruction, and curriculum to the three-dimensional framework.

Not Necessarily Barriers to Three-Dimensional Assessment at the Institutional and Department Levels

We sought to answer the question posed by Matz et al. (2018) to determine the supports and barriers to adopting the three-dimensional framework in undergraduate science courses. We narrowed our analysis to just the supports and barriers to assessment, with the assumption that the supports and barriers to three-dimensional assessment would reflect supports and barriers to integrating the three-dimensional framework throughout course instruction and curriculum. Based on the results from our generalized linear mixed model (Table 4.2), we did not necessarily find any specific barriers at the institutional and department levels. While our model did not indicate barriers to three-dimensional assessment, we did not necessarily find institutional- or department-level supports either. Our best fit model excluded all department-level variables as they did not provide any additional explanatory power. These excluded variables did not significantly increase or decrease the likelihood of three-dimensional alignment, and for categorical variables, such as course setting, this may signal a degree of equivalence across categories. Our model retained institution type as a predictor, but overlapping confidence intervals between the institutional categories indicate no statistical difference between Associate's, Baccalaureate, Master's, and Doctoral institutions. Hence, we can interpret our results to mean that three-dimensional assessments can be used in biology courses with small class sizes, such as those typical of Associate's and Baccalaureate colleges, as well as in high-enrollment courses, like those commonly seen in Master's and Doctoral universities. This finding supports previous work which indicates that three-dimensional assessments can be effectively administered even in high-enrollment courses (Matz et al., 2018; Stowe et al., 2021). Similarly, the lack of a significant difference between course settings suggests that three-dimensional assessments may be used with a degree of equivalency in courses with in-person, online, and hybrid instructional modalities, which corroborates past work suggesting that three-dimensional learning and assessment can be implemented online without adding an appreciable burden on instructors (Stowe et al., 2020). Our finding that there were not necessarily barriers to three-dimensional assessment is encouraging and emphasizes the wide applicability of this framework across diverse educational contexts in undergraduate biology education.

Identifying Where Institutions and Departments May Provide Additional Support

Another interpretation of the variables that were excluded from our best-fit model is that these may be areas where instructors could benefit from additional targeted support to help facilitate three-dimensional alignment of their assessments. We can extrapolate that professional development is one such area in need of support. Our finding that neither department-level professional development opportunities nor the amount of instructor-level professional development increased the likelihood of three-dimensional items indicates that the presence of professional development alone may not be enough to initiate and sustain adoption of the three-dimensional framework in undergraduate biology courses. This finding may be explained by previous qualitative research conducted in high school biology classrooms. Heredia (2020) found that incoherence between district expectations for student learning, the school's goals for classroom practice, and the information presented in professional development sessions created a source of uncertainty and ambiguity among biology teachers that hampered the degree to which they leveraged ideas and resources from professional development in their teaching. Biology teachers were less likely to use the content from professional development if they were unsure if that content was aligned to the metrics that would be rewarded in their teacher evaluation rubrics (Heredia, 2020). Such findings from the K-12 system are likely to generalize to the levels of the undergraduate biology educations system, which often operates with similar expectations and evaluations of undergraduate teaching. Although professional development is crucial for three-dimensional adoption (NRC, 2014), our research may provide additional evidence that just being exposed to professional development alone may not be sufficient to create long-lasting and sustainable changes in undergraduate biology education (Derting et al., 2016).

For professional development to be effective and sustained, we recommend that the content be coherent with clear department expectations about the educational goals (Sunal et al., 2001), and we encourage departments to align their expectations with the educational priorities outlined in national calls. Institutions and departments interested in increasing in meeting the goals of national calls may consider gearing their professional development offerings toward using frameworks like *Vision and Change* and using pedagogical tools such as the 3D-LAP. These frameworks and tools may be especially important areas for professional development as our best-fit model indicated that instructors who reported more frequently using *Vision and Change* or the 3D-LAP when writing their exams were no more likely to have three-dimensional items. Instructors may be familiar with these frameworks and tools but may face barriers to using them in ways that are fully aligned with the goals of the national calls. Interestingly, we did not find the same relationship with Bloom's Taxonomy, and instructors who reported using Bloom's more frequently were more likely to have three-dimensional assessments. Based on this result, we hypothesize that professional development related to assessing higher-order cognitive skills of Bloom's Taxonomy may be effective as a means of achieving goals related to three-dimensional alignment, but this is an area that will need further study. As professional development is an important agent of change in department teaching culture, we recommend that departments align their professional development with the metrics used for teaching evaluation, as such a congruous alignment between educational goals may prevent uncertainty and ambiguity about evaluation that hampers change. Change around teaching culture is a slow process, so the long-term effectiveness of professional development in terms of its ability to increase three-dimensional alignment in undergraduate courses is an area where future research is necessary.

Another support that institutions and departments can provide to instructors is facilitating purposeful and meaningful interactions with DBER faculty. Our best-fit model excluded the variable which indicated if there were DBER faculty within the department, but we do not intend this finding to minimize the impact of DBER within the field of biology education. Research demonstrates the positive role of DBER faculty in creating positive cultures around teaching (NRC, 2012b). Our finding may largely reflect the wide array of sub-disciplines within DBER that have wide reach beyond the realm of three-dimensional learning. DBER faculty represent a valuable resource within departments and we encourage institutions and departments to consider ways to facilitate conversations and bridge connections with DBER faculty as such conversations and connections are important avenues for promoting evidence-based teaching practices (Lane et al., 2022).

Teaching Practices May Not Reflect Assessment Practices

We asked instructors to self-report on their instructional practices that aligned with the principles of Scientific Teaching (Couch, Brown, et al., 2015; Handelsman et al., 2004, 2007) using an abbreviated version of the Measurement Instrument for Scientific Teaching (MIST; Durham et al., 2017) which included the subcategories Active Learning Strategies, Data Analysis and Interpretation, and Experimental Design and Communication. These three subcategories reflect many of the components of the three-dimensional framework. After model selection, MIST score was excluded from the best-fit model, suggesting that Scientific Teaching methods do not provide any additional significant explanatory power in predicting the likelihood of using three-dimensional assessments. This null result is surprising, as it indicates a potential misalignment between teaching and assessment practices. Instructors who had higher MIST scores and reported teaching content relevant to scientific practices did not necessarily have a greater number of scientific practices embedded in three-dimensional items on their exams.

We propose that this misalignment between teaching and assessment can arise within more traditional courses where science content and science practices are often taught and assessed separately (Pellegrino, 2013; Pruitt, 2014). For example, in traditional courses, scientific practices are often introduced to students as rote procedures, such as in the ritualized and singular "scientific method" (NRC, 2012a). Instructors who themselves were taught using this traditional pedagogical method may feel unprepared for three-dimensional teaching in which scientific practices and scientific content are taught in conjunction (Krajcik, 2015). Additional research suggests that many instructors across STEM courses still largely rely on instructor-centered teaching practice (Stains et al., 2018), and such teaching styles do not facilitate active student engagement in scientific practices (Bain et al., 2020). Misalignment at the instructional level can also occur if there is confusion or misinterpretation of how students are engaging in learning. Instructors can have best intentions to create highly active classrooms with frequent formative assessments yet may only facilitate student learning of discrete pieces of factual information (Cooper et al., 2015). This potential area of misalignment between teaching and assessment of the three-dimensional framework remains an area where additional research is necessary. The Three-Dimensional Learning Observation Protocol (Bain et al., 2020) may be a useful tool for this type of research, as it does not rely on self-reported data and allows a more direct comparison of three-dimensional assessments to observable three-dimensional teaching practices.

Instructors May Need More Time and Resources For Grading Three-Dimensional Exams

There are constraints on the amount of time that instructors have for writing and grading exams, which may affect their choices in what types of exams they are administering in their courses (Wright et al., 2018). As constructed-response items usually need to be graded manually by the instructor or by a paid assistant, instructors may choose not to use these items because of the associated time and/or resources needed to grade them. We found that the majority of three-dimensional items in our sample used a constructed-response format, from which we can extrapolate that three-dimensional assessments are more time- and resource-intensive to use in a classroom context. There are efforts to use machine learning to grade student responses to constructed-response items, but this approach requires a large sample of student responses that is usually beyond the scope of what can be collected in a single classroom setting (Moharreri et al., 2014; Nehm et al., 2012). While most three-dimensional items were constructed-response, we want to emphasize that three-dimensional items can certainly be multiple-choice or use other types of selected-response formats (Laverty et al., 2016; Underwood et al., 2018). We encourage instructors who write three-dimensional items in the selected response format to carefully consider how students are engaging with the scientific practices, particularly when the practice calls for reasoning about a phenomenon (Figure 4.5). Institutions and departments that want to support their instructors in incorporating three-dimensional assessments into their courses may need to provide instructors with time (e.g., teaching releases or decreased service apportionment), and with resources for grading (e.g., assigning teaching or learning assistants to the course). We issue our recommendations here, but we recognize that these recommendations involve financial considerations for institutions and departments that may not be feasible under all budgets.

Scientific Practices as a Target for Three-Dimensional Alignment

To better understand potential barriers to three-dimensional alignment in undergraduate biology courses, we focused on the dimension of the framework that was the least represented in our sample—the scientific practices. One of the reasons we hypothesized that there were so few scientific practices was because of the necessary stringency of the coding scheme for scientific practices in the 3D-LAP which includes the important criteria of including explicit prompts for student reasoning. Such prompts encourage students to justify and explain their logic about scientific phenomena and provide evidence that they have appropriately engaged in a scientific practice (Cooper & Stowe, 2018; Laverty et al., 2016, 2017; Stowe & Cooper, 2017). When these prompts are missing and the assessment does not explicitly ask students to provide reasoning, it is possible for students to respond without actually engaging in a scientific practice. In these cases, instructors run the risk of making assumptions about student thinking processes that do not mirror the actual processes students engaged with to answer the item (Stowe & Cooper, 2017). The 3D-LAP avoids the risk of making such assumptions by requiring items to ask for explicit evidence that students engaged in the scientific practice. While we agree with the authors of the 3D-LAP and concur that assessment items targeting scientific practices should elicit explicit evidence that students are using appropriate reasoning about scientific phenomena (Cooper & Stowe, 2018; Laverty et al., 2016, 2017; Stowe & Cooper, 2017), our results suggest that this approach of coding assessment items may have systematically underestimated instructors' attempts to incorporate scientific practices into their assessments. It is possible that instructors may be attempting to include three-dimensional items in their assessments, but these attempts may not have been detected with the strict interpretation of the coding protocol. Overall, very few biology exam items explicitly met all the 3D-LAP criteria for engaging in scientific practices. However, when we account for items that met some, but not all of the criteria, for scientific practices, we see that almost all instructors had some of the basic components of scientific practices in their exams (Figure 4.6).

In our sample, there were a notable number of items that met the majority of the scientific practice criteria but were just missing the key final component of student reasoning (Figure 4.5). This finding is not unique to biology, and previous work in chemistry has suggested that the reasoning component is often missing from typical assessment tasks (Laverty et al., 2017; Reed et al., 2017). In our sample, instructors were most commonly missing reasoning from the practices “Developing and Using Models,” “Using Mathematics and Computational Thinking,” “Constructing Explanations and Engaging in Argument,” and “Planning Investigations.” We highlight such items where instructors were mostly aligned to scientific practices as starting places to build upon existing items and make small modifications that would bring the item into full alignment with the 3D-LAP criteria for scientific practices. Our sample also contained many items that were partially aligned to the scientific practices “Developing and Using Models” and “Constructing Explanations and Engaging in Argument.” These partially aligned items met surface-level criteria for the practices, such as introducing a visual or verbal representation of a biological phenomenon, but these items will need major revisions to engage students in a scientific practice. We encourage instructors to carefully review the criteria of the 3D-LAP and to consult publications on adapting assessment tasks to the three-dimensional framework (Laverty et al., 2016; Underwood et al., 2018). We present our findings here not as a critique of the 3D-LAP, but as a way to showcase the work of biology instructors that may have been masked by a stringent coding scheme and to highlight the areas where instructors can build upon their existing assessments to fully align with the intent of the three-dimensional framework.

Limitations

We acknowledge several limitations of our study that should be considered in the interpretation of our findings. Our work took a broad quantitative approach that may not have captured individual perspectives about challenges and constraints of three-dimensional assessment. We suggest future qualitative research to more deeply explore how instructors are perceiving and implementing the three-dimensional framework within their courses.

We focused on exams as a summative assessment, as this is a common assessment strategy among undergraduate science courses (Gibbons et al., 2022; Goubeaud, 2010; Hurtado et al., 2012; Stanger-Hall, 2012; Wright et al., 2016, 2018), but there are other types of summative assessments, such as projects, presentations, essays, and reports, that instructors may be using to assess scientific practices. Instructors may also be engaging students in scientific practices during formative assessments, such as in-class activities and homework assignments. Given the anticipated variability in these other types of summative and formative assessments, we limited the scope of our study to exams, which tend to have a more similar format and structure between instructors.

We present the levels of Bloom's Taxonomy as ordinal in our analysis, which is in line with previous research and interpretations of Bloom's Taxonomy in biology education research (Freeman et al., 2011; Momsen et al., 2010, 2013; Zheng et al., 2008). However, we acknowledge that there are different interpretations of Bloom's Taxonomy within the field of biology education (Arneson & Offerdahl, 2018; Crowe et al., 2008; Lemons & Lemons, 2013; Semsar & Casagrand, 2017; Thompson & O'Loughlin, 2015), that Bloom's Taxonomy does not capture the full spectrum of knowledge types (Blumberg, 2009; Larsen et al., 2022), and there is not a consensus on the ordinal nature of the levels (Anderson et al., 2001; Furst, 1981; Lo et al., 2016; Lord & Baviskar, 2007).

We focused this research on lower-division courses, which face a unique set of challenges, including high enrollment and the pressure to cover a wide range of topics, that may be barriers to evidence-based instructional strategies, such as those aligned with the three-dimensional framework (Ebert-May et al., 2011; Henderson & Dancy, 2007; Wright et al., 2018). It is possible that our findings are not generalizable to upper-division courses which may not feel these challenges to the same extent.

Our work is by no means meant to be prescriptive of three-dimensional items or how they are used in undergraduate biology assessments. Instead, our work is meant to characterize instructor exams using broad strokes to form an abstract portrait of the current landscape of three-dimensional assessment in biology. While we found that three-dimensional assessments tended to use constructed response formats, be worth more points, and assess higher levels of Bloom's Taxonomy, a large amount of variation underlies these findings, and we provide these statistics as a way to help instructors conceptualize how other instructors have approached three-dimensional assessments in their courses.

CONCLUSION

For decades, national-level reports (e.g., AAAS, 1989, 2011; NASEM, 2021, 2022; NRC, 2003) have called for contextualized science education that engages students in scientific practices. The three-dimensional framework (NRC, 2012a) encapsulates many of the principles of these national calls and provides a lens for studying how national priorities are integrated across levels of the undergraduate biology education system. Institutions and departments can work towards meeting the three-dimensional alignment by setting clear and coherent expectations for undergraduate education aligned with national priorities and by providing supports for instructors in ways that enable and encourage them to exceed those expectations. We suggest that institutions and departments consider offering professional development on teaching and assessment that is aligned to both national priorities and institutional expectations. This professional development may be more impactful when instructors have the time, resources, and support to enact three-dimensional curriculum, instruction, and assessments in their courses. Institutions and departments may want to consider ways to structure courses and teaching appointments in ways that provide the time and resources to accommodate three-dimensional assessments which may take longer to grade compared to multiple-choice assessments that mainly test recall of facts. Our work highlights a need for a broader qualitative approach to better understand the nuances of how instructors are perceiving the existing support structures for three-dimensional education provided by institutions and departments and future research is needed to determine what additional supports instructors may need to facilitate instruction aligned with national priorities.

Acknowledgements

We thank participating instructors for their contributions to project development. This material is based upon work supported by the National Science Foundation (NSF) Graduate Research Fellowship (DGE-1610400). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

REFERENCES FOR CHAPTER 4

American Association for the Advancement of Science. (1989). Science for All Americans: A Project 2061 Report on Literacy Goals in Science, Mathematics, and Technology. American Association for the Advancement of Science.

American Association for the Advancement of Science. (1990). The Liberal Art of Science: Agenda for Action. American Association for the Advancement of Science. https://www.aaas.org/sites/default/files/the_liberal_art_of_science.pdf

American Association for the Advancement of Science. (1993). Benchmarks for Science Literacy. Oxford University Press.

American Association for the Advancement of Science. (2011). Vision and Change in Undergraduate Biology Education: A Call to Action. AAAS. https://live-visionandchange.pantheonsite.io/wp-content/uploads/2011/03/Revised-Vision-and-Change-Final-Report.pdf

Anderson, L. W., Krathwohl, D. R., & Bloom, B. S. (2001). A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives. Longman.

Arneson, J. B., & Offerdahl, E. G. (2018). Visual literacy in Bloom: Using Bloom's Taxonomy to support visual learning skills. CBE—Life Sciences Education, 17(1), ar7. https://doi.org/10.1187/cbe.17-08-0178

Austin, A. E. (2011). Promoting evidence-based change in undergraduate science education: A paper commissioned by the National Academies National Research Council. https://sites.nationalacademies.org/cs/groups/dbassesite/documents/webpage/dbasse_072578.pdf

Bain, K., Bender, L., Bergeron, P., Caballero, M. D., Carmel, J. H., Duffy, E. M., Ebert-May, D., Fata-Hartley, C. L., Herrington, D. G., Laverty, J. T., Matz, R. L., Nelson, P. C., Posey, L. A., Stoltzfus, J. R., Stowe, R. L., Sweeder, R. D., Tessmer, S. H., Underwood, S. M., Urban-Lurain, M., & Cooper, M. M. (2020). Characterizing college science instruction: The Three-Dimensional Learning Observation Protocol. PLOS ONE, 15(6), e0234640. https://doi.org/10.1371/journal.pone.0234640

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. Journal of Statistical Software, 67(1), 1–48. https://doi.org/doi:10.18637/jss.v067.i01

Biswas, S., Benabentos, R., Brewe, E., Potvin, G., Edward, J., Kravec, M., & Kramer, L. (2022). Institutionalizing evidence-based STEM reform through faculty professional development and support structures. International Journal of STEM Education, 9(1), 36. https://doi.org/10.1186/s40594-022-00353-z

Bloom, B. S., Krathwohl, D. R., & Masia, B. B. (1956). Taxonomy of Educational Objectives: The Classification of Educational Goals. McKay.

Blumberg, P. (2009). Maximizing learning through course alignment and experience with different types of knowledge. Innovative Higher Education, 34(2), 93–103. https://doi.org/10.1007/s10755-009-9095-2

Bradforth, S. E., Miller, E. R., Dichtel, W. R., Leibovich, A. K., Feig, A. L., Martin, J. D., Bjorkman, K. S., Schultz, Z. D., & Smith, T. L. (2015). University learning: Improve undergraduate science education. Nature, 523(7560), Article 7560. https://doi.org/10.1038/523282a

Brownell, S. E., Freeman, S., Wenderoth, M. P., & Crowe, A. J. (2014). BioCore Guide: A tool for interpreting the Core Concepts of Vision and Change for biology majors. CBE—Life Sciences Education, 13(2), 200–211. https://doi.org/10.1187/cbe.13-12-0233

Cherbow, K., McKinley, M. T., McNeill, K. L., & Lowenhaupt, R. (2020). An analysis of science instruction for the science practices: Examining coherence across system levels and components in current systems of science education in K-8 schools. Science Education, 104(3), 446–478. https://doi.org/10.1002/sce.21573

Clark, N., & Hsu, J. L. (2023). Insight from biology program learning outcomes: Implications for teaching, learning, and assessment. CBE—Life Sciences Education, 22(1), ar5. https://doi.org/10.1187/cbe.22-09-0177

Clemmons, A. W., Donovan, D. A., Theobald, E. J., & Crowe, A. J. (2022). Using the Intended–Enacted–Experienced curriculum model to map the Vision and Change core competencies in undergraduate biology programs and courses. CBE—Life Sciences Education, 21(1), ar6. https://doi.org/10.1187/cbe.21-02-0054

Clemmons, A. W., Timbrook, J., Herron, J. C., & Crowe, A. J. (2020). BioSkills Guide: Development and national validation of a tool for interpreting the Vision and Change core competencies. CBE—Life Sciences Education, 19(4), ar53. https://doi.org/10.1187/cbe.19-11-0259

Cooper, M. M., Caballero, M. D., Ebert-May, D., Fata-Hartley, C. L., Jardeleza, S. E., Krajcik, J. S., Laverty, J. T., Matz, R. L., Posey, L. A., & Underwood, S. M. (2015). Challenge faculty to transform STEM learning. Science, 350(6258), 281–282. https://doi.org/10.1126/science.aab0933

Cooper, M. M., & Stowe, R. L. (2018). Chemistry education research—From personal empiricism to evidence, theory, and informed practice. Chemical Reviews, 118(12), 6053–6087. https://doi.org/10.1021/acs.chemrev.8b00020

Couch, B. A., Brown, T. L., Schelpat, T. J., Graham, M. J., & Knight, J. K. (2015). Scientific Teaching: Defining a taxonomy of observable practices. CBE—Life Sciences Education, 14(1), ar9. https://doi.org/10.1187/cbe.14-01-0002

Couch, B. A., Prevost, L. B., Stains, M., Whitt, B., Marcy, A. E., Apkarian, N., Dancy, M. H., Henderson, C., Johnson, E., Raker, J. R., Yik, B. J., Earl, B., Shadle, S. E., Skvoretz, J., & Ziker, J. P. (2023). Examining whether and how instructional coordination occurs within introductory undergraduate STEM courses. Frontiers in Education, 8. https://www.frontiersin.org/articles/10.3389/feduc.2023.1156781

Crowe, A., Dirks, C., & Wenderoth, M. P. (2008). Biology in Bloom: Implementing Bloom's Taxonomy to enhance student learning in biology. CBE—Life Sciences Education, 7(4), 368–381. https://doi.org/10.1187/cbe.08-05-0024

Durham, M. F., Knight, J. K., Bremers, E. K., DeFreece, J. D., Paine, A. R., & Couch, B. A. (2018). Student, instructor, and observer agreement regarding frequencies of scientific teaching practices using the Measurement Instrument for Scientific Teaching-Observable (MISTO). International Journal of STEM Education, 5(1), 31. https://doi.org/10.1186/s40594-018-0128-1

Durham, M. F., Knight, J. K., & Couch, B. A. (2017). Measurement Instrument for Scientific Teaching (MIST): A tool to measure the frequencies of research-based reaching practices in undergraduate science courses. CBE—Life Sciences Education, 16(4). https://doi.org/10.1187/cbe.17-02-0033

Ebert-May, D., Derting, T. L., Hodder, J., Momsen, J. L., Long, T. M., & Jardeleza, S. E. (2011). What we say is not what we do: Effective evaluation of faculty professional development programs. BioScience, 61(7), 550–558. https://doi.org/10.1525/bio.2011.61.7.9

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

Fuhrman, S. (Ed.). (1993). Designing Coherent Education Policy: Improving the System (1st ed). Jossey-Bass.

Furst, E. J. (1981). Bloom's Taxonomy of educational objectives for the cognitive domain: Philosophical and educational issues. Review of Educational Research, 51(4), 441–453. https://doi.org/10.3102/00346543051004441

Furtak, E. M. (2017). Confronting dilemmas posed by three-dimensional classroom assessment: Introduction to a virtual issue of Science Education. Science Education, 101(5), 854–867. https://doi.org/10.1002/sce.21283

Gibbons, R. E., Reed, J. J., Srinivasan, S., Murphy, K. L., & Raker, J. R. (2022). Assessment tools in context: Results from a national survey of postsecondary chemistry faculty. Journal of Chemical Education, 99(8), 2843–2852. https://doi.org/10.1021/acs.jchemed.2c00269

Goubeaud, K. (2010). How is science learning assessed at the postsecondary level? Assessment and grading practices in college biology, chemistry and physics.

Journal of Science Education and Technology, 19(3), 237–245. https://doi.org/10.1007/s10956-009-9196-9

Handelsman, J., Ebert-May, D., Beichner, R., Bruns, P., Chang, A., DeHaan, R., Gentile, J., Lauffer, S., Stewart, J., Tilghman, S. M., & Wood, W. B. (2004). Scientific Teaching. Science, 304(5670), 521–522. https://doi.org/10.1126/science.1096022

Handelsman, J., Miller, S., & Pfund, C. (2007). Scientific Teaching. Macmillan.

Hardy, I., & Campbell, T. (2020). Developing and supporting the Next Generation Science Standards: The role of policy entrepreneurs. Science Education, 104(3), 479–499. https://doi.org/10.1002/sce.21566

Henderson, C., & Dancy, M. H. (2007). Barriers to the use of research-based instructional strategies: The influence of both individual and situational characteristics. Physical Review Special Topics - Physics Education Research, 3(2), 020102. https://doi.org/10.1103/PhysRevSTPER.3.020102

Heredia, S. C. (2020). Exploring the role of coherence in science teachers' sensemaking of science-specific formative assessment in professional development. Science Education, 104(3), 581–604. https://doi.org/10.1002/sce.21561

Hurtado, S., Eagan, K., Pryor, Whang, H., & Tran, S. (2012). Undergraduate teaching faculty: The 2010–2011 HERI Faculty Survey. Higher Education Research Institute, UCLA. https://www.heri.ucla.edu/monographs/HERI-FAC2011-Monograph-Expanded.pdf

Indiana University Center for Postsecondary Research. (2021). The Carnegie Classification of Institutions of Higher Education (2021 edition).

Lane, A. K., Earl, B., Feola, S., Lewis, J. E., McAlpin, J. D., Mertens, K., Shadle, S. E., Skvoretz, J., Ziker, J. P., Stains, M., Couch, B. A., & Prevost, L. B. (2022). Context and content of teaching conversations: Exploring how to promote sharing of innovative teaching knowledge between science faculty. International Journal of STEM Education, 9(1), 53. https://doi.org/10.1186/s40594-022-00369-5

Larsen, T. M., Endo, B. H., Yee, A. T., Do, T., & Lo, S. M. (2022). Probing internal assumptions of the revised Bloom's Taxonomy. CBE—Life Sciences Education, 21(4), ar66. https://doi.org/10.1187/cbe.20-08-0170

Laverty, J. T., Underwood, S. M., Matz, R. L., Posey, L. A., Carmel, J. H., Caballero, M. D., Fata-Hartley, C. L., Ebert-May, D., Jardeleza, S. E., & Cooper, M. M. (2016). Characterizing college science assessments: The Three-Dimensional Learning Assessment Protocol. PLOS ONE, 11(9), e0162333. https://doi.org/10.1371/journal.pone.0162333

Laverty, J. T., Underwood, S. M., Matz, R. L., Posey, L. A., Carmel, J. H., Caballero, M. D., Fata-Hartley, C. L., Ebert-May, D., Jardeleza, S. E., & Cooper, M. M. (2017).

Comment on “Analyzing the Role of Science Practices in ACS Exam Items.” Journal of Chemical Education, 94(6), 673–674. https://doi.org/10.1021/acs.jchemed.7b00170

Lemons, P. P., & Lemons, J. D. (2013). Questions for assessing higher-order cognitive skills: It’s not just Bloom’s. CBE—Life Sciences Education, 12(1), 47–58. https://doi.org/10.1187/cbe.12-03-0024

Lo, S. M., Larsen, V. M., & Yee, A. T. (2016). A two-dimensional and non-hierarchical framework of Bloom’s taxonomy for biology. The FASEB Journal, 30(S1), 662.14-662.14. https://doi.org/10.1096/fasebj.30.1_supplement.662.14

Lord, T., & Baviskar, S. (2007). Moving students from information recitation to information understanding: Exploiting Bloom’s Taxonomy in creating science questions. Journal of College Science Teaching, 36(5), 40–44.

Matz, R. L., Fata-Hartley, C. L., Posey, L. A., Laverty, J. T., Underwood, S. M., Carmel, J. H., Herrington, D. G., Stowe, R. L., Caballero, M. D., Ebert-May, D., & Cooper, M. M. (2018). Evaluating the extent of a large-scale transformation in gateway science courses. Science Advances, 4(10), eaau0554. https://doi.org/10.1126/sciadv.aau0554

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

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. https://doi.org/10.1187/cbe.10-01-0001

Momsen, J. L., 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. https://doi.org/10.1187/cbe.12-08-0130

National Academies of Sciences, Engineering, and Medicine. (2016a). Barriers and Opportunities for 2-Year and 4-Year STEM Degrees: Systemic Change to Support Students’ Diverse Pathways (S. Malcom & M. Feder, Eds.). National Academies Press. https://doi.org/10.17226/21739

National Academies of Sciences, Engineering, and Medicine. (2016b). Developing a National STEM Workforce Strategy: A Workshop Summary. The National Academies Press. https://doi.org/10.17226/21900

National Academies of Sciences, Engineering, and Medicine. (2021). Call to Action for Science Education: Building Opportunity for the Future. National Academies Press. https://doi.org/10.17226/26152

National Academies of Sciences, Engineering, and Medicine. (2022). Imagining the Future of Undergraduate STEM Education: Proceedings of a Virtual Symposium (K. Brenner, A. Beatty, & J. Alper, Eds). National Academies Press. https://doi.org/10.17226/26314

National Center on Education and the Economy. (2008). Tough Choices or Tough Times: The Report of the New Commission on the Skills of the American Workforce (Revised and Expanded edition). Jossey-Bass.

National Commission on Excellence in Education. (1983). A Nation At Risk: The Imperative For Educational Reform. National Commission on Excellence in Education. https://eric.ed.gov/?id=ED226006

National Research Council. (1996). National Science Education Standards (p. 4962). National Academies Press. https://doi.org/10.17226/4962

National Research Council. (2000). Inquiry and the National Science Education Standards: A Guide for Teaching and Learning (S. Olson & S. Loucks-Horsley, Eds.). National Academies Press. https://doi.org/10.17226/9596

National Research Council. (2003). BIO2010: Transforming Undergraduate Education for Future Research Biologists. National Academies Press. https://doi.org/10.17226/10497

National Research Council. (2006). Systems for State Science Assessment. National Academies Press. https://doi.org/10.17226/11312

National Research Council. (2007). Rising Above the Gathering Storm: Energizing and Employing America for a Brighter Economic Future. National Academies Press. https://doi.org/10.17226/11463

National Research Council. (2012a). A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas. National Academies Press. https://doi.org/10.17226/13165

National Research Council. (2012b). Discipline-Based Education Research: Understanding and Improving Learning in Undergraduate Science and Engineering. National Academies Press. https://doi.org/10.17226/13362

National Research Council. (2014). Developing Assessments for the Next Generation Science Standards. National Academies Press. https://doi.org/10.17226/18409

National Research Council. (2015). Guide to Implementing the Next Generation Science Standards. National Academies Press. https://doi.org/10.17226/18802

Nehm, R. H., Ha, M., & Mayfield, E. (2012). Transforming biology assessment with machine learning: Automated scoring of written evolutionary explanations. Journal of Science Education and Technology, 21(1), 183–196. https://doi.org/10.1007/s10956-011-9300-9

NGSS Lead States. (2013). Next Generation Science Standards: For States, By States. The National Academies Press.

Peteroy-Kelly, M., Brancaccio-Taras, L., Awong-Taylor, J., Balser, T., Jack, T., Lindsay, S., Marley, K., Romano, S., Uzman, J. A., & Pape-Lindstrom, P. (2019). A qualitative analysis to identify the elements that support department level change in the life sciences: The PULSE Vision & Change recognition program. PLOS ONE, 14(5), e0217088. https://doi.org/10.1371/journal.pone.0217088

R Core Team. (2023). R: A language and environment for statistical computing (4.2.3). R Foundation for Statistical Computing. https://www.R-project.org/

Radloff, J., Capobianco, B., Weller, J., Rebello, S., Eichinger, D., & Erk, K. (2022). Aligning undergraduate science curricula with three-dimensional learning. Journal of College Science Teaching, 52(1), 35–42.

Reed, J. J., Brandriet, A. R., & Holme, T. A. (2017). Analyzing the role of science practices in ACS exam items. Journal of Chemical Education, 94(1), 3–10. https://doi.org/10.1021/acs.jchemed.6b00659

Schmidt, W. H., McKnight, C. C., & Raizen, S. (Eds.). (1997). A Splintered Vision: An Investigation of U.S. Science and Mathematics Education. Springer.

Semsar, K., & Casagrand, J. (2017). Bloom's dichotomous key: A new tool for evaluating the cognitive difficulty of assessments. Advances in Physiology Education, 41(1), 170–177. https://doi.org/10.1152/advan.00101.2016

Siebert, E. D., & McIntosh, W. J. (Eds.). (2001). College Pathways to the Science Education Standards. NSTA Press.

Smith, M. K., Vinson, E. L., Smith, J. A., Lewin, J. D., & Stetzer, M. R. (2014). A campus-wide study of STEM courses: New perspectives on teaching practices and perceptions. CBE—Life Sciences Education, 13(4), 624–635. https://doi.org/10.1187/cbe.14-06-0108

Stains, M., Harshman, J., Barker, M. K., Chasteen, S. V., Cole, R., DeChenne-Peters, S. E., Eagan, M. K., Esson, J. M., Knight, J. K., Laski, F. A., Levis-Fitzgerald, M., Lee, C. J., Lo, S. M., McDonnell, L. M., McKay, T. A., Michelotti, N., Musgrove, A., Palmer, M. S., Plank, K. M., … Young, A. M. (2018). Anatomy of STEM teaching in North American universities. Science, 359(6383), 1468–1470. https://doi.org/10.1126/science.aap8892

Stanger-Hall, K. F. (2012). Multiple-choice exams: An obstacle for higher-level thinking in introductory science classes. CBE—Life Sciences Education, 11(3), 294–306. https://doi.org/10.1187/cbe.11-11-0100

Stepans, J. I., Shiflett, M., Yager, R. E., & Saigo, B. W. (2001). Professional Development Standards. In College Pathways to the Science Education Standards (pp. 25–56). National Science Teachers Association Press.

Stowe, R. L., & Cooper, M. M. (2017). Practicing what we preach: Assessing "critical thinking" in organic chemistry. Journal of Chemical Education, 94(12), 1852–1859. https://doi.org/10.1021/acs.jchemed.7b00335

Stowe, R. L., Esselman, B. J., Ralph, V. R., Ellison, A. J., Martell, J. D., DeGlopper, K. S., & Schwarz, C. E. (2020). Impact of maintaining assessment emphasis on three-dimensional learning as organic chemistry moved online. Journal of Chemical Education, 97(9), 2408–2420. https://doi.org/10.1021/acs.jchemed.0c00757

Stowe, R. L., Scharlott, L. J., Ralph, V. R., Becker, N. M., & Cooper, M. M. (2021). You are what you assess: The case for emphasizing chemistry on chemistry assessments. Journal of Chemical Education, acs.jchemed.1c00532. https://doi.org/10.1021/acs.jchemed.1c00532

Sunal, D. W., Hodges, J., Sunal, C. S., Whitaker, K. W., Freeman, L. M., Edwards, L., Johnston, R. A., & Odell, M. (2001). Teaching science in higher education: Faculty professional development and barriers to change. School Science and Mathematics, 101(5), 246–257. https://doi.org/10.1111/j.1949-8594.2001.tb18027.x

Thompson, A. R., & O'Loughlin, V. D. (2015). The Blooming Anatomy Tool (BAT): A discipline-specific rubric for utilizing Bloom's taxonomy in the design and evaluation of assessments in the anatomical sciences. Anatomical Sciences Education, 8(6), 493–501. https://doi.org/10.1002/ase.1507

Uminski & Couch (in revision). Testing Scientific Practices: A Nationwide Analysis of Undergraduate Biology Exams. *BioScience.*

Underwood, S. M., Posey, L. A., Herrington, D. G., Carmel, J. H., & Cooper, M. M. (2018). Adapting assessment tasks to support three-dimensional learning. Journal of Chemical Education, 95(2), 207–217. https://doi.org/10.1021/acs.jchemed.7b00645

U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, & National Assessment of Educational Progress. (2019). NAEP Report Card: 2019 NAEP Science Assessment. U.S. Department of Education. https://www.nationsreportcard.gov/science/supporting_files/2019_infographic_science.pdf

Vasaly, H. L., Feser, J., Lettrich, M. D., Correa, K., & Denniston, K. J. (2014). Vision and Change in the biology community: Snapshots of change. CBE—Life Sciences Education, 13(1), 16–20. https://doi.org/10.1187/cbe.13-12-0234

Webb, N. L. (1997). Criteria for Alignment of Expectations and Assessments in Mathematics and Science Education. Research Monograph No. 6. Council of Chief State School Officers, Attn: Publications, One Massachusetts Avenue, NW, Ste. https://eric.ed.gov/?id=ED414305

Wickham, H., Averick, M., Bryan, J., Chang, W., D'Agostino McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Lin Pedersen, T., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., … Hiroaki, Y. (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686. https://doi.org/10.21105/joss.01686

Wieman, C., Perkins, K., & Gilbert, S. (2010). Transforming Science Education at Large Research Universities: A case study in progress. Change: The Magazine of Higher Learning, 42(2), 6-14. Retrieved from http://www.cwsei.ubc.ca/resources/files/WiemanPerkinsGilbert_SEI-Model_Change_Mar-10.pdf

Wright, C. D., Huang, A., Cooper, K., & Brownell, S. (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://doi.org/10.20429/ijsotl.2018.120214

Zheng, A. Y., Lawhorn, J. K., Lumley, T., & Freeman, S. (2008). Application of Bloom's Taxonomy debunks the "MCAT myth." Science, 319(5862), 414–415. https://doi.org/10.1126/science.1147852

SUPPLEMENTAL MATERIAL FOR CHAPTER 4

Supplemental Material 4.1: Additional details on how factors were collected, measured, and analyzed.

This supplemental material provides the survey items and additional information regarding how the survey data was processed and analyzed. Factors are listed alphabetically here, which may not reflect the order that instructors saw the items as they were presented in the original survey. Parenthetical numbers at the end of options were not seen by instructors and indicate how survey item responses were recorded. Instructor responses were retained as-is unless additional data processing is noted.

4.1.1 Authorship

Survey item:

Did you write the majority of exam questions yourself?

- ◯ Yes, all by myself (1)
- ◯ Yes, by myself and with colleagues teaching the same course (2)
- ◯ No, exam questions were modified from other materials (3)
- ◯ No, exam questions were straight from other materials (4)
- ◯ Other (5) ____

Additional data processing: Instructor responses were recoded into three categories representing original authorship (options 1 and 2) and authorship that drew from other materials (options 3 and 4) and mixed authorship indicating a combination of both original items and items from other sources. Fifteen instructors indicated option 5 (Other) and provided a text description of their authorship process, which were reviewed and recoded as “mixed authorship” or “original authorship.”

4.1.2 Course audience
Survey item:

This course was intended for:

- ◯ STEM majors (1)
- ◯ Non-STEM majors (2)
- ◯ Both STEM majors and non-STEM majors (3)
- ◯ Other (4) ____

Additional data processing: Instructor responses to “other” included courses intended for pre-health science students, which were recoded to “Both STEM majors and non-STEM majors.”

4.1.3 Course lab
Survey item:

Was there a required laboratory component to this course?

- ◯ Yes (1)
- ◯ No (2)

4.1.4 Course setting
Survey item:

At the time the exam was administered, this course was taught:

- ◯ In-person only (1)
- ◯ Online only, but previous semesters of this course were in-person (2)
- ◯ Online only and previous semesters of this course were taught online (3)
- ◯ Hybrid (i.e., contained both in-person and online components) (4)
- ◯ Other (5) ____

Additional data processing: Six instructors selected option 5 and based on their text clarifications, these responses were re-assigned to options 1, 3, and 4.

4.1.5 Department DBER faculty
Survey item:

Including yourself, does the department contain any faculty who identify as discipline-based education researchers (i.e., DBER faculty)?

- ◯ Yes (1)
- ◯ No (2)
- ◯ Unsure (3)

4.1.6 Department professional development

Survey item:

Has the department allocated resources (e.g., time or money) for faculty professional development?

- ○ Yes (1)
- ○ No (2)
- ○ Unsure (3)

4.1.7 Instructor professional development

Survey item:

Approximately how many hours of professional development sessions (e.g., conference presentations, courses, workshops, or other forms of training) on the **topic of assessments** have you attended in the **past 10 years**?

- ○ Zero hours (i.e., no professional development specific to assessments) (1)
- ○ 1-3 hours (e.g., attending a conference presentation on assessment) (2)
- ○ 4-8 hours (e.g., participating in a half- or full-day assessment-focused workshop) (3)
- ○ 8-12 hours (i.e., several conference presentations, workshops, or trainings) (4)
- ○ Greater than 12 hours (i.e., many conference presentations, workshops, or trainings) (5)

Additional data processing: The options were recoded to an ordinal scale.

Note: The bolding in this item was also in the original item presented to instructors.

4.1.8 Teaching years

Survey item:

How many years of teaching experience do you have as an instructor of record?

- ◯ 0-1 yea (1)
- ◯ 2-5 years (2)
- ◯ 6-10 years (3)
- ◯ 11-15 years (4)
- ◯ 16-20 years (5)
- ◯ 21-25 years (6)
- ◯ Greater than 25 years (7)

Additional data processing: The options were recoded to an ordinal scale.

4.1.9 Uses 3D-LAP

Survey item:

To what degree do you refer to, consider, or use the following when you are constructing assessments?

	Never (1)	Rarely (2)	Sometimes (3)	Often (4)	Almost Always (5)
Three-Dimensional Learning Assessment Protocol (3D-LAP)	◯	◯	◯	◯	◯

Additional data processing: The options were recoded to an ordinal scale.

4.1.10 Uses Bloom's Taxonomy
Survey item:

To what degree do you refer to, consider, or use the following when you are constructing assessments?

	Never (1)	Rarely (2)	Sometimes (3)	Often (4)	Almost Always (5)
Bloom's Taxonomy	○	○	○	○	○

Additional data processing: The options were recoded to an ordinal scale.

4.1.11 Uses Vision and Change
Survey item:

To what degree do you refer to, consider, or use the following when you are constructing assessments?

	Never (1)	Rarely (2)	Sometimes (3)	Often (4)	Almost Always (5)
Recommendations made by the *Vision and Change* report	○	○	○	○	○

Additional data processing: The options were recoded to an ordinal scale.

Supplemental Table 4.1: Descriptions of item types

Item type	**Item response**	**Description**
Cluster	Constructed response	Test-takers respond to a series of items that share a common stimulus. The series of items are designed as sub-parts or sub-items, which may or may not be scored independently. Cluster items often differ from essay items in that test-takers are provided a bulleted or numbered list of discrete sub-parts to respond to rather than a single paragraph of text directions.
Discipline-specific	Constructed response	Test-takers use procedures, algorithms, or other processes that are specific to biological sciences but are not easily categorized as strictly modeling or mathematical manipulation of information. Examples include matching complementary nucleotide base pairs or completing Punnett squares.
Essay	Constructed response	Test-taker responses to an essay item typically require more than one paragraph. Essay items often use verbs such as “explain” or “justify” to elicit longer responses from test-takers.
Fill-in-the-blank	Constructed response	Test-takers fill in a word or a short phrase that is missing from the stimulus and there is not a list of responses (i.e., a “word bank”) provided.
Matching	Selected response	For each option in one list, the test-taker selects the correct match from a second list. Matching options may be presented as a series of items where each item in the series has the same set of common options.
Math manipulation	Constructed response	Test-takers manipulate information to solve mathematical or algorithmic problems.
Model	Constructed response	Test-takers respond to the item by creating a model of a biological phenomenon or by adding to, contributing to, or otherwise modifying an existing model.
Multiple choice	Selected response	The test-taker selects one option from a list of two or more provided options.
Multiple select	Selected response	A multiple-choice item where more than one option can be selected as correct.
Multiple True-False	Selected response	A form of multiple select where the options consist of binary factual statements and are preceded by a prompt or question statement linking the options together.
Reorder	Selected response	Test-takers put a series of provided options into a sequence or specified order.
Short answer	Constructed response	Test-takers response to the item with a word, phrase, or response that does not exceed one paragraph (approximately 3-4 sentences).
True-False	Selected response	The test-taker selects whether a single statement is true or false. Unlike multiple-true false, there is no preceding prompt or question linking multiple statements together.

Supplemental Material 4.2: Coding for partial alignment to scientific practices

Our coding for partial alignment to the 3D-LAP scientific practices criteria accounts for the inconsistent number of criteria statements between the different scientific practices. When the scientific practice had three criteria statements (e.g., “Planning Investigations,” “Using Mathematics and Computational Thinking,” or “Evaluating Information”), we coded alignment to each statement, where meeting zero, one, two, or three statements was coded as no alignment, partially aligned, mostly aligned, and fully aligned to the scientific practice, respectively.

The 3D-LAP criteria for the scientific practice “Asking Questions” only contained two criteria statements, so we coded alignment as either no alignment or fully aligned. The first criterion (“Question gives an event, observation, phenomenon, data, scenario, or model”) was similar to the first criterion of multiple practices and could not accurately be coded at the level of that statement.

In cases where the 3D-LAP contained four criteria statements for the scientific practice (e.g., “Developing and Using Models”, “Analyzing and Interpreting Data”, “Constructing Explanations and Engaging in Argument from Evidence”), we similarly disregarded the first criterion as we did for “Asking Questions.” In each case when there were four criteria statements, the first criterion could be met by providing an event, observation, phenomenon, or hypothesis, and as such could not be distinguished between scientific practices that shared the same or similar criteria. When there were four criteria for the scientific practice, we coded alignment to zero, two, three, or four statements as no alignment, partially aligned, mostly aligned, and fully aligned to a scientific practice, respectively.

Supplemental Table 4.2: Linear mixed model[a] predicting item point value with an interacting effect of item response and three-dimensional alignment

Term	Estimate	Standard error	*t*	*p*
(Intercept)	4.44	0.93	4.78	< 0.001
Item response: Constructed response	2.62	0.09	28.65	< 0.001
Alignment: Three-dimensional	-0.02	0.16	-0.11	0.91
Item response*alignment	0.90	0.23	3.92	< 0.001
$R^2 = 0.977$				
[a]Model: item point value ~ item response format*alignment + (1|instructor)				

Supplemental Table 4.3: Linear mixed model[a] predicting Bloom's Taxonomy level with interaction effect of item response and three-dimensional alignment

Term	**Estimate**	**Standard Error**	***t***	***p***
(Intercept)	1.48	0.05	32.60	< 0.001
Item response: Constructed response	0.57	0.05	12.31	< 0.001
Alignment: Three-dimensional	1.79	0.08	21.98	< 0.001
Item response*alignment	-0.28	0.12	-2.39	0.017
$R^2 = 0.423$				
[a]Model: Bloom's Taxonomy level ~ item response format*alignment + (1|instructor)				

CONCLUSION

Reform efforts in undergraduate biology, guided by landmark documents such as *Vision and Change* (American Association for the Advancement of Science, 2011) and *A Framework for K-12 Science Education* (National Research Council [NRC], 2012a), encourage educators to eschew a mile-wide and inch-deep approach to teaching and instead focus on core concepts and engage students in scientific practices that will help them more deeply understand and contribute to the discipline. Assessments can provide important information on how instructors and departments are making progress in meeting the goals of these reform efforts. In this dissertation, I studied how programmatic assessments and concept assessments can provide data about student learning aligned to core concepts, and I conducted a nation-wide survey of biology instructors to determine how their exams integrate scientific practices and foundational concepts and to investigate what additional factors are associated with incorporating three-dimensional assessments into their courses. The major findings of these studies are summarized below.

1) **Departments and instructors using programmatic and concept assessments to determine their progress in meeting reform efforts should carefully evaluate student performance in light of assessment administration conditions to optimize score validity.**

Programmatic assessments, such as GenBio-MAPS (Couch et al., 2019), and concept assessments, such as the IMCA (Shi et al., 2010), provide departments and instructors a way to measure student learning of foundational concepts in undergraduate biology. These measures of student learning can illustrate which concepts students have mastered and which parts of the curriculum may need to be reevaluated to improve student learning outcomes. In Chapters 1 and 2 (Uminski & Couch, 2021; Uminski et al., 2023), I examined student performance on GenBio-MAPS and the IMCA to illustrate that departments and instructors should consider the available evidence for score validity before using assessment scores to make changes to curriculum and instruction. Self-reports of test-taking effort, measurements of test-taking behaviors, such as the amount of time spent on test questions or the entire test overall, and correlations of assessment score with previous scores on course exams testing similar content provide lines of validity evidence we can use to interpret programmatic and concept assessment scores.

Using these lines of validity evidence, I found that content knowledge for some students may be underestimated in lower-stakes out-of-class contexts in which students are not graded on the correctness of their responses. In lower-stakes conditions, a small portion of students may be more likely to demonstrate low test-taking effort, such as rapid selection of test answers or short test completion times, and these behaviors may yield assessment scores that do not accurately reflect what students know about biology concepts. In cases where students demonstrate these behaviors, departments and instructors that take the assessment scores at face value may be prompted to make unnecessary changes to curriculum and instruction, which can be a costly error in terms of time and resources.

I also found that student scores may be higher compared to performance when the assessment is completed in a higher-stakes out-of-class context in which students have both access to external resources and the incentive to use them. In these cases where scores potentially overestimate student knowledge, departments and instructors may unintentionally overlook areas of curriculum or instruction where students are struggling to grasp foundational concepts.

My research indicates that lower-stakes in-class and higher-stakes in-class conditions provide reasonable information about student understanding, and these may be appropriate for administering programmatic or concept assessments; however, class time is a limited resource and instructors may wish to use out-of-class administrations to preserve time for instruction. If departments or instructors choose to use programmatic or concept assessments in lower-stakes out-of-class or higher-stakes out-of-class contexts, my findings suggest that they should collect evidence of score validity and carefully evaluate assessment data in light of the administration conditions.

2) Course exams indicate that there is still progress to be made to fully align undergraduate biology with broader curriculum reform calls.

The majority of undergraduate biology courses use high point-value summative exams as a way to measure student learning. What is assessed on these exams provides a window into the prioritized learning goals in a course. In Chapters 3 and 4, I analyzed the content of exams from a nationwide sample of biology courses for alignment to the scientific practices, crosscutting concepts, and core ideas of the three-dimensional framework (NRC, 2012a). I found that the overwhelming majority of exam items were not testing scientific practices, and as such, these items were not three-dimensionally aligned and were not fully meeting the goals of reform efforts in biology education.

Although instructors were often assessing biology core ideas, which is an important component of the aligning to national calls for reform, most of these items were only capable of engaging students in lower-order cognitive skills associated with recall of memorized facts. This overrepresentation of lower-order cognitive skills mirrors findings from over a decade ago (Momsen et al., 2010, 2013), indicating that there is still much to be done in undergraduate biology courses to fully meet the national calls to integrate conceptual knowledge and scientific practices.

My work here suggests that the format of exam items may be a potential barrier to this integration of scientific practices. Three-quarters of the items in our sample used a selected-response multiple-choice format, but three-dimensional items were disproportionately constructed-response items. These constructed-response items can be difficult to implement because they are typically time consuming or resource-intensive to grade. Institutions and departments seeking to better align with the goals of reform efforts by increasing the number of three-dimensional assessments may want to consider ways of providing adequate time and resources for grading constructed-response exams. My work also indicates that existing professional development opportunities may not necessarily be helping instructors in meeting the goals of three-dimensional alignment. Departments may consider offering professional development opportunities specifically targeting the three-dimensional framework. My work highlights paths for institutions, departments, and instructors to more closely align their undergraduate biology education with reform efforts.

Future directions

Programmatic and concept assessments have an important role in measuring progress in reform efforts, yet there are currently no programmatic or concept assessments for undergraduate biology that are specifically aligned to the three-dimensional framework. There is a need for a validated three-dimensional assessment instrument. As there are few published examples of three-dimensional exam items in undergraduate biology, a three-dimensional instrument can be a useful reference or serve as inspiration for course instructors aiming to incorporate more scientific practices into their exams. As programmatic and concept assessments often take many months to years of development, a more immediate solution to the lack of three-dimensional exam items would be the creation of a publicly-accessible database or repository to which instructors can submit their own three-dimensional exam items.

My work in Chapter 3 and 4 was mostly quantitative, and there remains a need for a qualitative investigation to explore instructor decision making about three-dimensional assessments as well as instructors' perceived challenges and barriers to three-dimensional alignment in their courses. In addition to instructor perspectives, there is also a gap in the literature about undergraduate students' engagement with three-dimensional assessment items. Student interviews and think-aloud protocols can better uncover whether items that have the potential to elicit scientific practices are actually engaging students in those practices. Future work is also needed to create professional development opportunities related to the three-dimensional framework and to study the short- and long-term effectiveness of this professional development in terms of advancing the goals of reform efforts.

REFERENCES FOR THE CONCLUSION

American Association for the Advancement of Science. (2011). Vision and Change in Undergraduate Biology Education: A Call to Action. AAAS. https://live-visionandchange.pantheonsite.io/wp-content/uploads/2011/03/Revised-Vision-and-Change-Final-Report.pdf

Couch, B. A., Wright, C. D., Freeman, S., Knight, J. K., Semsar, K., Smith, M. K., Summers, M. M., Zheng, Y., Crowe, A. J., & Brownell, S. E. (2019). GenBio-MAPS: A programmatic assessment to measure student understanding of *Vision and Change* core concepts across general biology programs. CBE—Life Sciences Education, 18(1), ar1. https://doi.org/10.1187/cbe.18-07-0117

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. https://doi.org/10.1187/cbe.10-01-0001

Momsen, J. L., 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. https://doi.org/10.1187/cbe.12-08-0130

National Research Council. (2012). A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas. National Academies Press. https://doi.org/10.17226/13165

Shi, J., Wood, W. B., Martin, J. M., Guild, N. A., Vicens, Q., & Knight, J. K. (2010). A diagnostic assessment for introductory molecular and cell biology. CBE — Life Sciences Education, 9(4), 453–461. https://doi.org/10.1187/cbe.10-04-0055

Uminski, C., & Couch, B. A. (2021). GenBio-MAPS as a case study to understand and address the effects of test-taking motivation in low-stakes program assessments. CBE—Life Sciences Education, 20(2), ar20. https://doi.org/10.1187/cbe.20-10-0243

Uminski, C., Hubbard, J. K., & Couch, B. A. (2023). How administration stakes and settings affect student behavior and performance on a biology concept assessment. CBE—Life Sciences Education, 22(2), ar27. https://doi.org/10.1187/cbe.22-09-0181