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

Crystal Uminski,[†] Joanna K. Hubbard,[‡] and Brian A. Couch[†]*

[†]School of Biological Sciences, University of Nebraska–Lincoln, Lincoln, NE 68588; [‡]Biology Department, Truman State University, Kirksville, MO 63501

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 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

Abdi Warfa, *Monitoring Editor*

Submitted Sep 15, 2022; Revised Feb 8, 2023; Accepted Feb 23, 2023

CBE Life Sci Educ June 1, 2023 22:ar27

DOI:10.1187/cbe.22-09-0181

*Address correspondence to: Brian Couch (bcouch2@unl.edu).

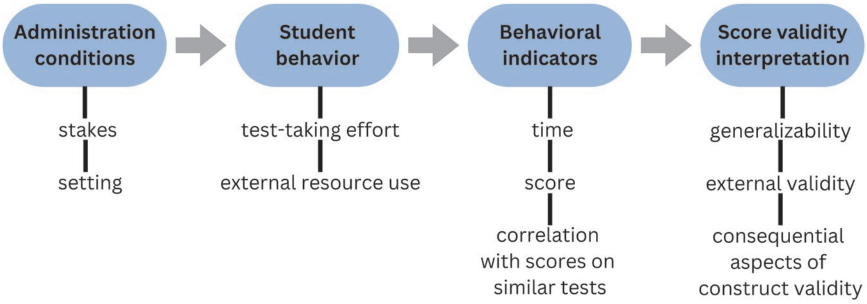
Administration conditions
Student behavior
Behavioral indicators
Score validity interpretation
stakes
setting
test-taking effort
external resource use
time
score
correlation with scores on similar tests
generalizability
external validity
consequential aspects of construct validity

FIGURE 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.

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 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.

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). 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

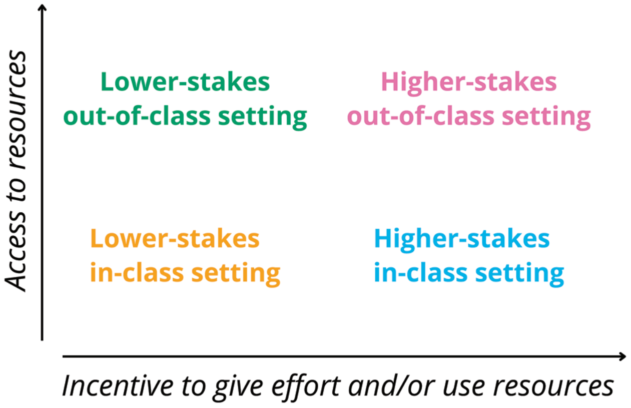
Access to resources
Lower-stakes out-of-class setting
Higher-stakes out-of-class setting
Lower-stakes in-class setting
Higher-stakes in-class setting
Incentive to give effort and/or use resources

FIGURE 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.

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.

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 1 for demographic information). While demographic information is provided to represent the study sample, our study did not seek to explore additional 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 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 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 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

TABLE 1. Demographic characteristics of students in the study[a]

Demographic categories	*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.

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 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; Supplemental Table 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.

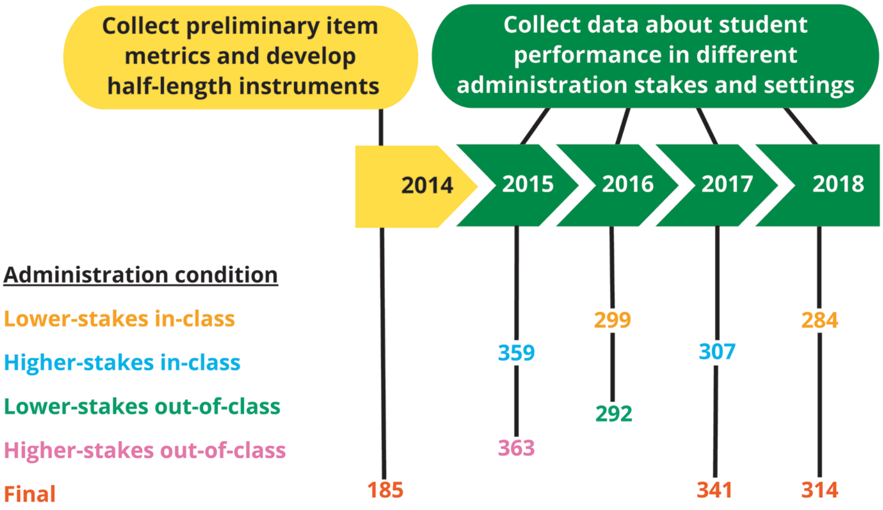
Collect preliminary item metrics and develop half-length instruments
Collect data about student performance in different administration stakes and settings
2014
2015
2016
2017
2018
Administration condition
Lower-stakes in-class
Higher-stakes in-class
Lower-stakes out-of-class
Higher-stakes out-of-class
Final
359
363
299
292
307
341
284
314
185

FIGURE 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.

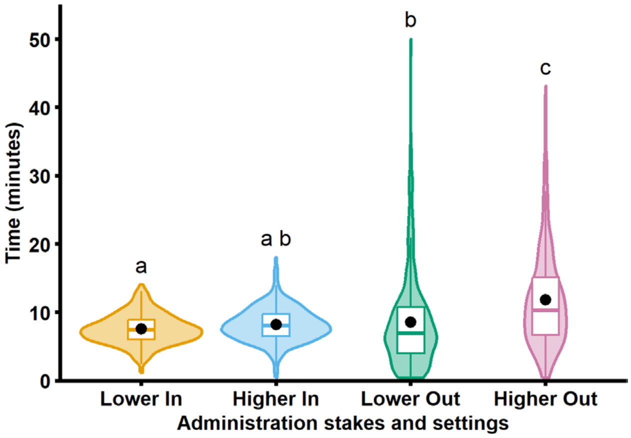

FIGURE 4. Test completion time in each administration condition. Completion times represent the sum of time spent on each page of the concept assessment. Completion time data were not collected when the concept assessment was administered on paper in the final exam condition. Violin plots show the distribution 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 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.

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-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). 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 5) across the administration conditions. The lower-stakes in-class, higher-stakes

TABLE 2. Linear mixed-effects model on the effects of administration condition on concept assessment score[a]

Parameter	Sum of squares	Mean squares	*df*	*F*	*p*
Administration condition	4.561	1.140	2175.3	42.716	<0.001
Average exam score	41.738	41.738	1	1563.470	<0.001
Post hoc comparisons					
Contrast[c]	**Estimate**	**SE**	***df***	***t***	***p***
Final Exam – Higher In	0.085	0.01	2060	9.16	<0.001
Final Exam – Higher Out	–0.014	0.01	2542	–1.27	0.711
Final Exam – Lower In	0.069	0.01	2065	7.12	<0.001
Final Exam – Lower Out	0.098	0.01	2541	8.04	<0.001
Higher In – Higher Out	–0.099	0.01	1751	–9.12	<0.001
Higher In – Lower In	–0.016	0.01	2552	–1.68	0.448
Higher In – Lower Out	0.013	0.01	2553	1.04	0.837
Higher Out – Lower In	0.083	0.01	2553	7.17	<0.001
Higher Out – Lower Out	0.112	0.01	2553	8.24	<0.001
Lower In – Lower Out	0.029	0.01	1756	2.42	0.109

[a] Score ~ administration condition + average unit exam score + (1 | ID)
[b] Model $R^2 = 0.49$.
[c] 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.

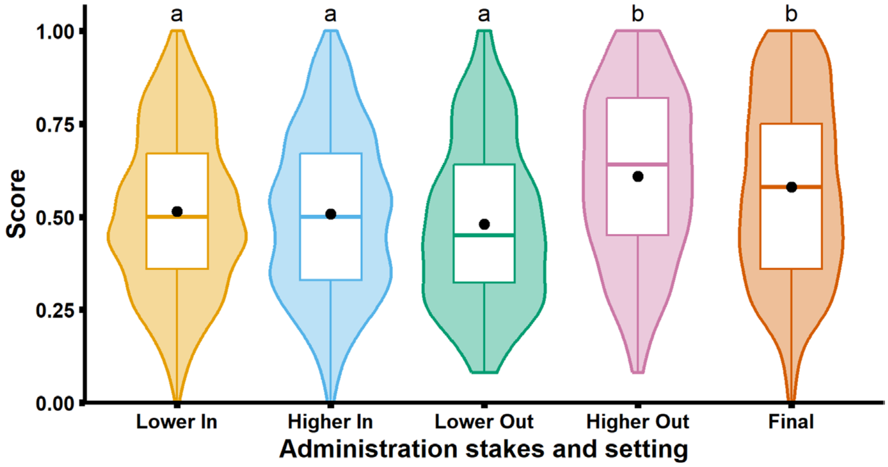

FIGURE 5. Concept assessment scores in each administration condition. Violin plots show the distribution of scores in each administration condition. Boxes represent the 25th, 50th, and 75th percentiles. Whiskers represent 5th and 95th percentiles. The dot represents the mean scores. Conditions sharing the same letters were not significantly different ($p \geq 0.05$), as determined by the post hoc tests shown in Table 2. 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.

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.

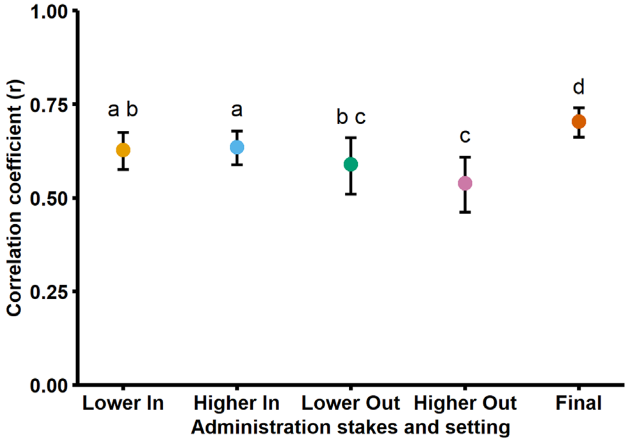

FIGURE 6. Correlation between concept assessment score and average course exam score for each administration condition. Dots represent correlation coefficients and whiskers represent the 95% confidence interval. Conditions sharing the same letters did not have significantly different correlation values ($p \geq 0.05$), as determined by the Fisher's z transformations shown in Supplemental Table 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.

We used a linear mixed-effects model to analyze scores across administration conditions (Table 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 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 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 ($r = 0.71$, $p < 0.01$).

Item Difficulty and Discrimination

Across administration conditions, the IMCA items had adequate values for item difficulty and discrimination (Ebel and Frisbie, 1986; Supplemental Figure 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.

ACKNOWLEDGMENTS

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

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.2018.120214Adams