2015

A Comparison of Two Low-Stakes Methods for Administering a Program-Level Biology Concept Assessment

Brian A. Couch
University of Nebraska - Lincoln, bcouch2@unl.edu

Jennifer K. Knight
University of Colorado Boulder

JOURNAL OF MICROBIOLOGY & BIOLOGY EDUCATION, December 2015, p. 178-185
DOI: http://dx.doi.org/10.1128/jmbe.v16i2.953

Research

A Comparison of Two Low-Stakes Methods for Administering a Program-Level Biology Concept Assessment

Brian A. Couch[1]* and Jennifer K. Knight[2]
[1]*School of Biological Sciences, University of Nebraska, Lincoln, NE 68588,*
[2]*Department of Molecular, Cellular, and Developmental Biology, University of Colorado, Boulder, CO 80309*

Concept assessments are used commonly in undergraduate science courses to assess student learning and diagnose areas of student difficulty. While most concept assessments align with the content of individual courses or course topics, some concept assessments have been developed for use at the programmatic level to gauge student progress and achievement over a series of courses or an entire major. The broad scope of a program-level assessment, which exceeds the content of any single course, creates several test administration issues, including finding a suitable time for students to take the assessment and adequately incentivizing student participation. These logistical considerations must also be weighed against test security and the ability of students to use unauthorized resources that could compromise test validity. To understand how potential administration methods affect student outcomes, we administered the Molecular Biology Capstone Assessment (MBCA) to three pairs of matched upper-division courses in two ways: an online assessment taken by students outside of class and a paper-based assessment taken during class. We found that overall test scores were not significantly different and that individual item difficulties were highly correlated between these two administration methods. However, in-class administration resulted in reduced completion rates of items at the end of the assessment. Taken together, these results suggest that an online, outside-of-class administration produces scores that are comparable to a paper-based, in-class format and has the added advantages that instructors do not have to dedicate class time and students are more likely to complete the entire assessment.

INTRODUCTION

Over the past several decades, an increasing number of concept assessments have been generated for use within undergraduate science courses (16, 17). Concept assessments, also referred to as concept inventories, traditionally consist of a series of multiple-choice or other closed-ended questions where incorrect answer options represent common student misconceptions (1). Many concept assessments are geared toward introductory students, focusing on individual courses or course topics. For example, several concept assessments exist for introductory biology (10, 14, 26) as well as discrete topics taught within introductory biology, such as meiosis, diffusion and osmosis, respiration and photosynthesis, and natural selection (3, 12, 15, 25). Administering a concept assessment in a pre-post manner allows instructors to gauge the conceptual learning that results from a period of instruction. Concept assessments have been used widely to ascertain the prevalence of misconceptions, guide instructional decisions, and determine the effects of instructional interventions.

In a few cases, concept assessments have been developed to assess conceptual learning across a series of courses or an entire major. These program-level assessments share the qualities that their content exceeds that of any single course and they can be used to monitor cumulative achievement over a multiyear timescale. Standing outside any single course, these instruments are intended to guide discussions at the department level regarding student progression, curricular cohesion, and program deficiencies. Program-level assessments also provide information that departments can use to satisfy requests for documenting student achievement, including local institutional mandates as well as external requirements from accreditation agencies (4, 24).

Two recently developed concept assessments exemplify the scope and utility of program-level assessments. The Host-Pathogens Interaction (HPI) concept assessment was developed by a team of biology instructors to assess learning across a series of eight microbiology courses (21, 22). Administering the assessment in select courses

*Corresponding author. Mailing address: 204 Manter, Lincoln, NE 68588-0118. Phone: 402-472-8130. Fax: 402-472-2083. E-mail: bcouch2@unl.edu.

at the beginning, middle, and end of their local course series allowed this group of instructors to monitor learning and retention of key concepts within the broad domain of host-pathogen interactions and provided baseline data for continued improvement of their curriculum (23). The Molecular Biology Capstone Assessment (MBCA), developed by the authors of the present paper, was designed to assess conceptual learning within molecular biology programs (5). Administration to upper-division students in advanced molecular biology courses at seven institutions suggested that these students still retained a variety of incorrect conceptions regarding key disciplinary concepts.

In administering concept assessments to students, instructors must consider a variety of factors that could influence response rates and the degree to which students take the assessment seriously (2). These decisions must also be weighed against the extent to which the test can be secured and the likelihood that students will attempt to obtain a copy of the test or use external resources to answer test questions. One recommended administration method involves giving the assessment during class time, informing students that test results will help inform course practices, and awarding full credit to all participating students regardless of the correctness of their answers (1). This low-stakes method achieves high response rates and allows students to use class time for the purpose of providing feedback that will benefit their educational experience, while minimizing any temptation to use unauthorized resources.

While administering concept assessments in class under low-stakes conditions has several advantages, this method suffers from the possibility that students may not take the assessment as seriously as they would if it were given under higher stakes. To address this problem, several researchers have chosen to administer the instrument as part of the final exam to ensure that students are highly motivated to perform well (26, 27). A study from physics found that administering a concept assessment as part of the final exam led to improved performance relative to in-class administration with no or modest point incentives for correct answers (7). Conversely, a group of biologists found that administering a concept assessment as part of the final exam did not significantly affect student performance compared with a separate administration conducted in class the week before finals (28). These disparate findings could be attributed to the variety of incentives associated with the in-class format in the biology study: students correctly answering all questions earned five points on their final exam, the instructor encouraged students to take the assessment seriously and to use it as practice for the final exam, and the instructor added concepts from the hardest questions to the final exam study guide.

The broad scope of a program-level assessment presents additional administrative challenges. Since the assessment does not fully align with the content of any individual course, most instructors are unwilling to assign points based on correct responses or offer the assessment as part of the final exam. For this same reason, some instructors do not feel comfortable using class time for a program-level assessment, which would inevitably displace some amount of normal course content. As a result, in our initial large-scale pilot, we administered the MBCA in an online format outside of class time and gave students participation points for attempting the assessment. While this format was suitable for many instructors and expedient for data collection, questions remained regarding how this administration method affected student scores and whether students dedicated a sufficient amount of time to completing the assessment (20, 30).

To understand how different administration methods affect student performance, we compared two broadly feasible methods for administering program-level assessments: an online assessment taken by students outside of class and a paper-based assessment taken during class. We hypothesized that in-class scores would be higher than out-of-class scores, since we expected that students would be more willing to devote their time and attention to an in-class assessment. We hoped that the results of this study would help inform future decisions on how to administer program-level assessments in a way that maximizes student participation and performance, while working within the bounds of normal course structures and expectations.

METHODS

Assessment characteristics

The development and content of the Molecular Biology Capstone Assessment (MBCA) have been described previously (5). Briefly, this assessment was iteratively developed with extensive student and faculty input to ensure that the questions were clear and scientifically accurate. Each question is aligned with a specific concept and associated learning objective from molecular biology, cell biology, or genetics. The MBCA consists of 18 question stems followed by 4 true/false (T/F) statements each, resulting in a total of 72 T/F statements. A complete copy of the assessment (without correct answers) is available as supplementary material in the initial publication.

Assessment administration and data processing

To compare how different administration formats affect student outcomes, we administered the MBCA in three upper-division biology courses at different institutions (Table 1) that were taught in two separate semesters by the same instructor.[1] In both semesters, students were informed that the results of the assessment would help the department improve its undergraduate curriculum, asked

[1]For Course 1, the first semester was co-taught by an additional instructor, but the course structure and instructional materials remained nearly identical across terms.

TABLE 1.
Institution Carnegie classifications[a].

Course	Control	Research Activity[b]	Region
1	Public	RU/VH	Mountain West
2	Public	RU/VH	West Coast
3	Public	Master's/L	West Coast

[a]Institutions are ordered by participant numbers. All institutions offer doctoral degrees.
[b]RU = research university; VH = very high research activity; Master's/L = master's level, larger programs.

to give the assessment their best effort, and awarded participation points for attempting the assessment. Students self-reported their current class standing as the last question on the assessment. Students not wanting to participate in the research were allowed to submit a blank survey containing only their name and secondary identification and still receive full participation credit. This option was not exercised by any of the 287 students enrolled in the target courses.

In one semester, the MBCA was administered online to students via Qualtrics and completed by students outside of class time. The assignment was announced by the instructor during class, and students received an email with a link to complete the assessment, which remained open for a period of roughly one week. Students were asked to complete the online assessment in a single continuous session without consulting additional resources. Each multiple-T/F question was presented to students as an individual page, and the amount of time spent on each page was recorded by the survey software. In the other semester, the MBCA was administered in class. Students were informed during the week prior to the activity that they would be taking an in-class assessment on a particular day. Test questions were printed on paper, and students were given 30 to 45 minutes to record their answers on Scantron forms.

Prior to analyses, the data were processed to minimize the potential impact of invalid entries. One online submission was removed due to failure to complete at least half the assessment. To account for cases where students completed the assessment discontinuously, the time stamp for any question with a dwell time greater than 20 minutes was replaced with the class mean for that question. This substitution was made for roughly 1% of all survey pages collected (24 out of 2,457 total pages).

Data analyses

Participation rates were separately compared for each course pair using Fisher's exact test. Overall test scores were calculated using the fractional scoring method, where students received credit for each T/F statement answered correctly (5), expressed as the percent of total statements answered correctly (11, 29). Blank statement responses were omitted from the analysis. Overall scores for junior and senior students were compared using Student's *t*-test. Overall scores for each course under different administration formats were analyzed with a two-factor analysis of variance (ANOVA). Statement difficulties were calculated as the fraction of students answering each statement correctly; correlations between statement difficulties were determined by calculating Pearson's coefficient. Statement difficulties were also compared using the Mantel-Haenszel test (6), performed with Winsteps software (Version 3.81.0) (18). This differential item functioning (DIF) test is used to determine whether individual items show significant differences between two groups beyond what would be expected based on overall scores. For items with p values less than 0.05, the magnitude of the difference between the two groups was classified according to Educational Testing Services (ETS) criteria: Category B = slight to moderate difference, DIF ≥ 0.43 logits; Category C = moderate to large difference, DIF ≥ 0.64 logits (19, 33). Percent completion was calculated as the percent of students who marked an answer for each T/F statement. Overall completion times for each student were calculated as the sum of individual page dwell times. The correlation between test duration and overall score was calculated as Pearson's coefficient. Statistical analyses were conducted using SPSS software (Version 22) unless otherwise specified. This research was classified by the University of Colorado as exempt from Institutional Review Board (IRB) review, protocols 0603.081, 0108.9, and 12-0336.

RESULTS

The MBCA was taken by 117 students in the online, outside-of-class format and 144 students in the paper-based, in-class format (Table 2). Class participation ranged from 81% to 98%, and participation rates between semesters for each course did not differ significantly. The sample analyzed consisted almost entirely of upper-division students, including 25% juniors and 73% seniors. Overall test scores did not differ between juniors and seniors, so this variable was not considered for subsequent analyses (mean ± standard deviation: juniors = 71.4 ± 10.7, seniors = 72.4 ± 11.5; p = 0.55).

To assess how performance compared between administration formats, we analyzed student outcomes at the overall score and individual statement levels. Overall student scores were not significantly affected by administration format (Fig. 1). However, there was a main effect of course indicating that the MBCA has the capacity to distinguish between different groups of students, as previously demonstrated (5). Individual statement difficulties were highly correlated between the two formats (r = 0.92) (Fig. 2). While there was a high degree of similarity in statement difficulties between the

TABLE 2.
Assessment administration and participation.

Course	Term	Format, Location	*n*	Part.	*p* value[a]	Class Standing[b]			
						Fr.	So.	Jr.	Sr.
1	Fa13	online, outside class	56	98%	0.23	0	0	25	30
	Sp13	paper, in class	66	93%		1	2	14	45
2	Sp13	online, outside class	44	88%	0.99	0	0	12	32
	Sp14	paper, in class	52	90%		0	1	11	32
3	Sp14	online, outside class	17	81%	0.70	0	0	0	17
	Sp13	paper, in class	26	87%		0	0	0	23

Part. = participation; Fr. = freshman; So. = sophomore; Jr. = junior; Sr. = senior; Fa = fall; Sp = spring.
[a]*p* values are based on Fisher's exact test for each course pair.
[b]Class standings do not sum to total participant numbers because some students did not enter their class standing.

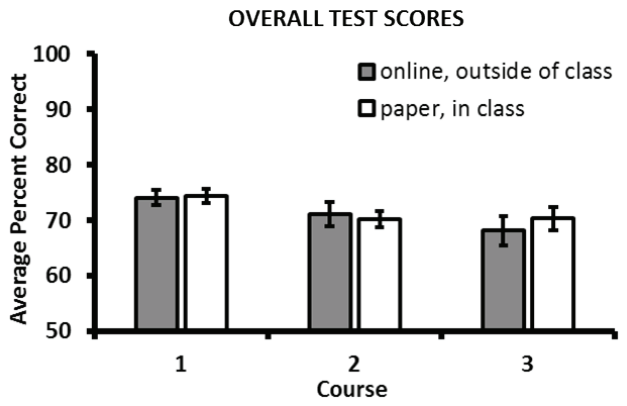

FIGURE 1. Effect of format on overall scores. Bars represent average percent correct for each class ± SEM. Filled bars indicate the semester in which the test was given online, outside of class. Unfilled bars indicate the semester in which the test was given on paper, in class. Two-factor ANOVA (format × course): main effect of format, $F_{(1,255)} = 0.10$, $p = 0.76$; main effect of course, $F_{(2,255)} = 4.62$, $p = 0.01$; interaction, $F_{(2,255)} = 0.26$, $p = 0.77$. SEM = standard error of the mean.

different formats, Mantel-Haenszel analysis revealed that three statements showed slight to moderate differences between formats and one statement showed a moderate to large difference.

We analyzed completion rates to determine how the different formats affected test completion (Fig. 3). Students given the assessment in the online, outside-of-class format attempted nearly every T/F statement, and this high completion rate remained constant across the entire assessment. Conversely, while nearly all students given the assessment in the paper-based, in-class format completed the first 12 questions, completion rates declined over the last six questions, reaching a roughly 90% completion rate for the final two questions.

We also estimated how long students spent on the assessment, which is designed to take roughly 30 minutes. For the online survey, we used time-stamp data to calculate the time it took students to complete the entire assessment. Students spent a wide range of times with an overall median completion time of 31.4 minutes (Fig. 4). A small fraction of students (8%) completed the assessment in less than 15 minutes, which we consider an inadequate amount of time for a typical student to read, contemplate, and answer each question on the test. The majority of students (64%) took 15 to 45 minutes, and the remaining students (28%) took 45 to 90 minutes. While we were unable to collect this same type of data for the paper-based, in-class administration, we did ask instructors to note how long it took students to complete the assessment. In this format, there were no students who completed the assessment in less than 15 minutes, and nearly all students turned in the assessment between 15 and 45 minutes, with longer durations being prevented by class time limits.

We also tracked the amount of time students spent on each online survey page, which corresponds to one multiple T/F question (Fig. 5). Most questions took students a median time of one to two minutes per question. Question 8, which involves using a codon table to translate a protein coding sequence, took noticeably longer. Importantly, the amount of time spent on each question remained stable over the entire assessment with similar values for the first and last six questions.

To understand the relationship between the amount of time spent completing the assessment and overall student performance, we calculated the Pearson's correlation between these values for the online, outside-of-class format (Fig. 6). Overall scores ranged widely at each time, with a modest correlation ($r = 0.24$), indicating that a small portion of the variance in overall scores is associated with variance in completion time ($r^2 = 0.06$).

DISCUSSION

Departments wishing to collect program-level information on student achievement must consider how instruments will be administered to students so that the resulting data present an accurate picture of student understanding. To determine whether administration format affects assessment outcomes, we compared student

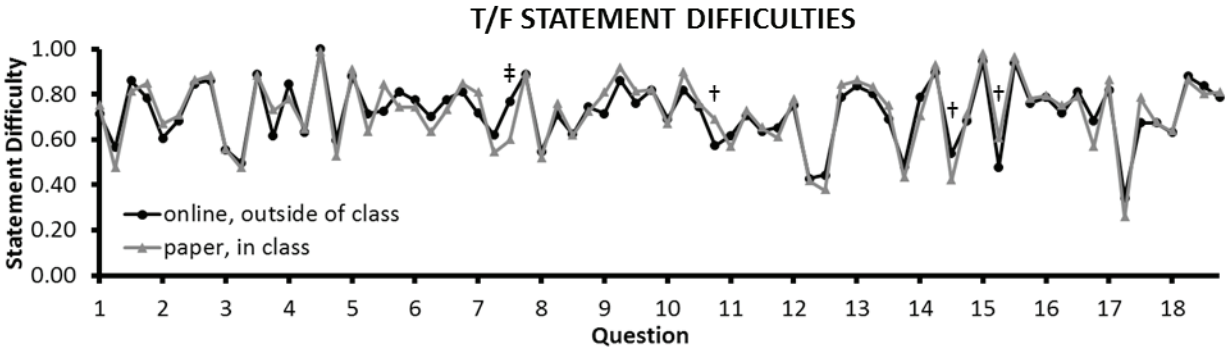

FIGURE 2. Comparison of T/F statement difficulties. Symbols represent difficulties for each T/F statement (4 per question) given either online, outside of class (black circles) or on paper, in class (gray triangles). Lines between data points are included to help visually trace the two administration formats. Note that a higher difficulty indicates a higher proportion of correct answers (i.e., an easier question). Correlation between statements: Pearson's r = 0.92. Mantel-Haenszel differential item functioning: † = Category B (10d, 14c, 15b); ‡ = Category C (7c). T/F = true/false.

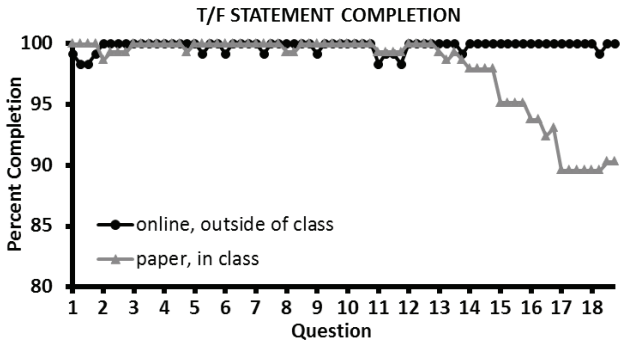

FIGURE 3. Individual T/F statement completion rates. Symbols represent the percent of students marking an answer for each T/F statement given either online, outside of class (black circles) or on paper, in class (gray triangles). T/F = true/false.

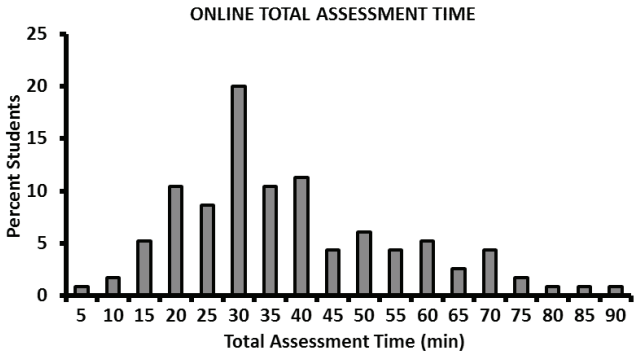

FIGURE 4. Total assessment completion time for the online, outside-of-class administration. Gray bars represent the percent of students taking the amount of time given for each bin. Labels indicate the upper threshold of each bin. For example, the rightmost bin contains students who took longer than 85 minutes and less than or equal to 90 minutes.

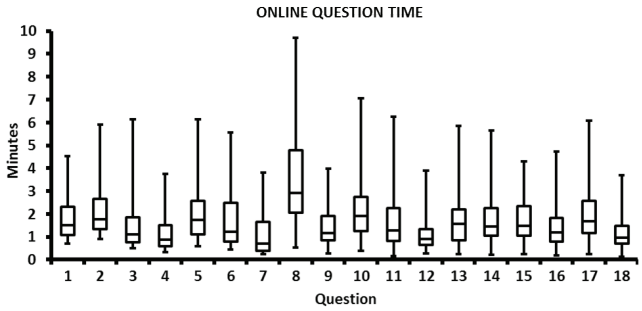

FIGURE 5. Time per question for the online, outside-of-class administration. Central bars represent median question time, boxes represent inner quartiles, and whiskers represent 5th and 95th percentiles.

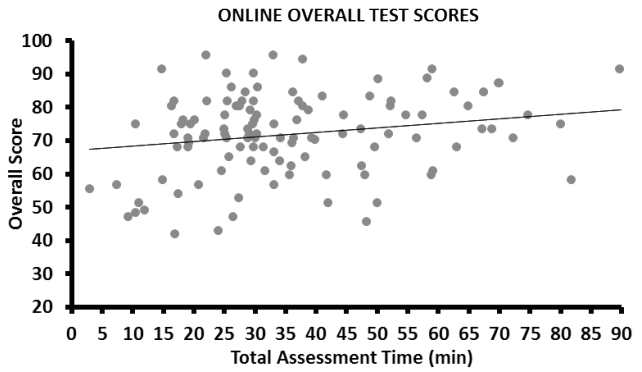

FIGURE 6. Relationship between total assessment time and overall test scores for the online, outside-of-class format. Each gray dot corresponds to the overall score for a single student taking the indicated amount of time. The black line shows the linear correlation between variables (r = 0.24).

performance and time spent under two low-stakes administration methods. The online, outside-of-class format is generally considered to be easier to implement because it does not require class time. However, the effect of this approach has been unclear: student scores might be lowered due to unwillingness to devote sufficient time to the assessment, or they may be raised as a result of consulting external resources, such as classmates or the internet. In this study, the assessment was given under low-stakes conditions where students were given credit for attempting the assessment but not penalized for incorrect responses, so we predicted that the effect of external resources would be minimal in either setting. We therefore hypothesized that an online, outside-of-class administration would yield lower scores than a paper-based assessment given during class, where students would be more likely to devote adequate time to the assessment and may be motivated by the presence of their course instructor.

This hypothesis was not supported: we found that the two administration formats produced comparable results in the three courses studied, both in participation rates and overall scores. The combined averages of overall student scores were nearly identical, although variation in scores was slightly larger for the online, outside-of-class format (mean ± standard deviation: online, outside of class = 72.1 ± 12.1; paper, in class = 72.1 ± 10.3). Importantly, the correlation between individual statement difficulties was consistent between formats (r = 0.92), and this correlation is equivalent to previous studies in which the MBCA was given in the same online, outside-of-class format in consecutive semesters (r = 0.91) (16). Thus, the differences observed between the two administration formats are no greater than semester-to-semester variation under a constant administration format.

Despite an overall high correlation, we detected significant differences in individual statement difficulties between administration formats. The magnitudes of the differences for these statements were further classified as Category B or C based on the degree of item bias between examinee groups. In testing development situations where the presence of item bias would disadvantage a particular group, Category B items may be reviewed for potential underlying issues, while Category C items are given much closer scrutiny and may be removed from the test (33, 34). In looking at the MBCA statements that showed bias, it is unclear why performance on these statements would significantly differ between formats. In cases where the online format showed higher performance, it is possible that these statements were particularly amenable to looking up on the internet (e.g., statements 7c and 14c). Conversely, one might expect that questions compelling written work would show higher scores in the paper-based format. Interestingly, there were no significant biases for the two questions where we predicted *a priori* that students would benefit from scratch paper (i.e., question 8 requiring translation of a coding sequence and question 18 involving use of a Punnett square).

We were surprised to find that nearly all students completed the online, outside-of-class assessment, while roughly 10 percent of students failed to complete the paper, in-class assessment. The latter finding illustrates the difficulty inherent in implementing program-based assessments: the three instructors were all enthusiastic about piloting the MBCA but, in each case, were unable to devote enough class time for every student to complete the assessment. We also observed differences in the amount of time students spent taking the assessment under the two conditions. The online, outside-of-class format showed a wider range of time values, including some students who completed the assessment faster than was ever observed with the in-class version. While these students likely did not give the assessment sufficient attention, this fraction was either not large enough to dramatically affect overall scores or may have been counterbalanced by students willing to invest more time outside of class than was available during class.

In light of these findings, we conclude that an online, outside-of-class administration produces results that are comparable to a paper-based, in-class administration for the given student population. Both administration conditions are sufficient to produce high participation rates and motivate a substantial fraction of the given student population to devote an adequate amount of time to taking the assessment. However, the online, outside-of-class administration has several additional advantages. It does not require class time, and thus instructors may be more willing to administer it in their courses. It also allows every student an adequate amount of time to complete the assessment and does not depend on class attendance. To account for students who engage in rapid guessing, programs may choose to remove submissions from students who did not spend an adequate amount of time on the assessment (31). Given the differences in completion rates and individual item biases, results from these two administration formats should not be considered interchangeable, and programs wishing to compare different cohorts will be best served by employing the same format and incentive system across administrations.

As an alternative to either of the course-based methods described here, some departments have successfully adopted a model where students complete assessments as part of their degree requirements, which provides a sufficient incentive to achieve high participation rates. In some cases, students attend a testing session organized by the department; in other cases students visit a testing center or complete an online assessment on their own time. While these approaches require a certain degree of administrative commitment, they have the benefit of being able to target students at defined points in the major (e.g., graduating seniors) rather than sampling from the subset of students enrolled in individual courses.

In administering a program-level assessment, a question remains regarding whether higher stakes would increase student performance and alter the overall interpretation

of student achievement (30). Increasing the stakes of an assessment by associating test scores with a course grade or incentivizing correct responses through monetary reward have each been shown to boost student motivation and test performance (8, 32). However, these methods seem impractical for departments administering a program-level assessment. Most instructors are unwilling to hold students accountable for content not explicitly covered in the current course, and departments generally lack the fiscal resources to compensate students for test performance. Raising the stakes of an assessment may also increase the likelihood that students seek out test answers through unauthorized means, and maintaining test security could be challenging for departments to manage. Test scores may improve under higher stakes, but unless test security can be guaranteed, these higher scores may be no more accurate than scores collected under lower stakes. Thus, when higher stakes are placed on program-level assessments, departments should administer the assessments under proctored conditions where students do not have access to external resources and cannot keep copies of the assessment questions.

Both administration formats described in this paper can provide useful information to departments engaged in curricular discussions at the programmatic level. The initial MBCA pilot revealed several areas in which advanced students still struggle, including specific concepts from evolution, development, cellular transport, thermodynamics, and genetics (5). Such information can guide discussions among faculty wishing to map the concepts onto the current curriculum and be helpful in deciding whether learning such concepts requires increased emphasis or alternative pedagogical approaches. In this manner, student data—rather than faculty opinions—serve as the starting point for discussions regarding curricular organization and implementation. Furthermore, by administering the same assessment across years, departments can determine whether their efforts have had measurable impacts on student performance.

Departments administering program-level assessments should consider the overarching purpose of the assessment and select conditions that meet their specific needs, while accounting for the limited time and resources available to support such efforts (13). In addition to providing specific feedback on student achievement, program-level assessments can communicate to students that the department values student learning and inspire departmental conversations regarding curriculum and pedagogy. Departments should resist falling for the "Single Indicator Fallacy," which holds that a single measurement instrument can suffice to capture the entirety of a complex system (9). Regardless of the administration format, program-level assessments must be interpreted in concert with other metrics of student achievement, including student coursework and ability to demonstrate competency in authentic disciplinary activities. While further research is warranted to understand the nuances of different administration methods, program-level assessments stand to provide useful information to identify potential areas for improvement and monitor student progress on a continual basis.

ACKNOWLEDGMENTS

We thank the instructors who administered the assessment as well as the students who completed the assessment. This work was supported by the University of Colorado-Boulder Science Education Initiative (SEI) and a NSF TUES II award (DUE-1322364) to BAC, JKK, and others. We are grateful to our collaborators on the TUES award, who provided critical insights and/or manuscript revisions: Sara Brownell, Alison Crowe, Scott Freeman, Kate Semsar, Michelle Smith, Mindi Summers, and Christian Wright. The authors declare that there are no conflicts of interest.

REFERENCES

1. **Adams, W. K., and C. E. Wieman.** 2011. Development and validation of instruments to measure learning of expert-like thinking. Int. J. Sci. Educ. **33**:1289–1312.
2. **American Educational Research Association (AERA), American Psychological Association (APA), National Council on Measurement and Education (NCME)**. 2014. The standards for educational and psychological testing. Washington, DC.
3. **Anderson, D. L., K. M. Fisher, and G. J. Norman.** 2002. Development and evaluation of the conceptual inventory of natural selection. J. Res. Sci. Teach. **39**:952–978.
4. **Beno, B. A.** 2004. The role of student learning outcomes in accreditation quality review. New Dir. Commun. Coll. **126**:65–72.
5. **Couch, B. A., W. B. Wood, and J. K. Knight.** 2015. The Molecular Biology Capstone Assessment: a concept assessment for upper-division molecular biology students. CBE Life Sci. Educ. **14**:ar10.
6. **Crocker, L., and J. Algina.** 2006. Introduction to classical and modern test theory. Wadsworth Pub. Co., Mason, OH.
7. **Ding, L., N. W. Reay, A. Lee, and L. Bao.** 2008. Effects of testing conditions on conceptual survey results. Phys. Rev. ST Phys. Educ. Res. **4**:010112.
8. **Duckworth, A. L., P. D. Quinn, D. R. Lynam, R. Loeber, and M. Stouthamer-Loeber.** 2011. Role of test motivation in intelligence testing. Proc. Natl. Acad. Sci. **108**:7716–7720.
9. **Ewell, P. T.** 1988. Implementing assessment: some organizational issues. New Dir. Institutional Res. **15**:15–28.
10. **Garvin-Doxas, K., and M. W. Klymkowsky.** 2008. Understanding randomness and its impact on student learning: lessons learned from building the Biology Concept Inventory (BCI). CBE Life Sci. Educ. **7**:227–233.
11. **Gross, L. J.** 1982. Scoring multiple true/false tests: some considerations. Eval. Health Prof. **5**:459–468.
12. **Haslam, F., and D. F. Treagust.** 1987. Diagnosing secondary students' misconceptions of photosynthesis

and respiration in plants using a two-tier multiple choice instrument. J. Biol. Educ. **21**:203–211.

13. **Higgerson, M. L.** 1993. Important components of an effective assessment program. J. Assoc. Commun. Adm. **2**:1–9.
14. **Howitt, S., T. Anderson, M. Costa, S. Hamilton, and T. Wright T.** 2008. A concept inventory for molecular life sciences: how will it help your teaching practice? Aust. Biochem. **39**:14–17.
15. **Kalas, P., A. O'Neill, C. Pollock, and G. Birol.** 2013. Development of a meiosis concept inventory. CBE Life Sci. Educ. **12**:655–664.
16. **Knight, J. K.** 2010. Biology concept assessment tools: design and use. Microbiol. Aust. **31**:5–8.
17. **Libarkin, J.** 2008. Concept inventories in higher education science. National Research Council, Washington, DC.
18. **Linacre, J. M.** 2014. Winsteps Rasch measurement computer program. Winsteps.com, Beaverton, OR.
19. **Linacre, J. M.** 2014. Winstep Rasch measurement computer program user's guide. Winsteps.com, Beaverton, OR.
20. **Liu, O. L., B. Bridgeman, and R. M. Adler.** 2012. Measuring learning outcomes in higher education: motivation matters. Educ. Res. **41**:352–362.
21. **Marbach-Ad, G., et al.** 2007. A faculty team works to create content linkages among various courses to increase meaningful learning of targeted concepts of microbiology. CBE Life Sci. Educ. **6**:155–162.
22. **Marbach-Ad, G., et al.** 2010. A model for using a concept inventory as a tool for students' assessment and faculty professional development. CBE Life Sci. Educ. **9**:408–416.
23. **Marbach-Ad, G., et al.** 2009. Assessing student understanding of host pathogen interactions using a concept inventory. J. Microbiol. Biol. Educ. **10**:43–50.
24. **Middaugh, M. F.** 2009. Planning and assessment in higher education: demonstrating institutional effectiveness. Jossey-Bass, Hoboken, NJ.
25. **Odom, A. L., and L. H. Barrow.** 1995. Development and application of a two-tier diagnostic test measuring college biology students' understanding of diffusion and osmosis after a course of instruction. J. Res. Sci. Teach. **32**:45–61.
26. **Shi, J., W. B. Wood, J. M. Martin, N. A. Guild, Q. Vicens, and J. K. Knight.** 2010. A diagnostic assessment for introductory molecular and cell biology. CBE Life Sci. Educ. **9**:453–461.
27. **Smith, M. K., W. B. Wood, and J. K. Knight.** 2008. The Genetics Concept Assessment: a new concept inventory for gauging student understanding of genetics. CBE Life Sci. Educ. **7**:422–430.
28. **Smith, M., K. Thomas, and M. Dunham.** 2012. In-class incentives that encourage students to take concept assessments seriously. J. Coll. Sci. Teach. **42**:57–61.
29. **Tsai, F.-J., and H. K. Suen.** 1993. A brief report on a comparison of six scoring methods for multiple true-false items. Educ. Psychol. Meas. **53**:399–404.
30. **Wise, S. L., and C. E. DeMars.** 2005. Low examinee effort in low-stakes assessment: problems and potential solutions. Educ. Assess. **10**:1–17.
31. **Wise, S. L., and C. E. DeMars.** 2010. Examinee noneffort and the validity of program assessment results. Educ. Assess. **15**:27–41.
32. **Wolf, L. F., and J. K. Smith.** 1995. The consequence of consequence: motivation, anxiety, and test performance. Appl. Meas. Educ. **8**:227–242.
33. **Zwick, R, D. T. Thayer, and C Lewis.** 1999. An empirical Bayes approach to Mantel-Haenszel DIF analysis. J. Educ. Meas. **36**:1–28.
34. **Zwick, R.** 2012. A review of ETS differential item functioning assessment procedures: flagging rules, minimum sample size requirements, and criterion refinement. ETS Res. Rep. Ser. **2012**:i–30.