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

Crystal Uminski and Brian A. Couch*
School of Biological Sciences, University of Nebraska–Lincoln, Lincoln, NE 68588

ABSTRACT

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

INTRODUCTION

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

GenBio-MAPS is administered to undergraduate students outside class time as an online survey. The online out-of-class format does not take time from class instruction and allows the instrument to be administered and scored consistently and efficiently across different courses and institutions. While the online out-of-class administration may be convenient for test administrators, this format necessitates low-stakes testing

Tammy Long, *Monitoring Editor*
Submitted Oct 22, 2020; Revised Jan 20, 2021; Accepted Jan 29, 2021
CBE Life Sci Educ June 1, 2021 20:ar20
DOI:10.1187/cbe.20-10-0243
*Address correspondence to: Brian A. Couch (bcouch2@unl.edu).

conditions in which students are not graded based on test performance. If GenBio-MAPS had higher stakes, there might be greater incentive for students to access external resources, and maintaining test security to prevent academic dishonesty in the out-of-class context would be difficult for departments to achieve. Under low-stakes testing conditions, prior research on a similar instrument (Couch *et al.,* 2015) found that student performance in the out-of-class context does not differ significantly from an in-class administration, suggesting that students engage with the assignment to roughly the same degree as they would for an in-class activity (Couch and Knight, 2015).

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

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

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

Theoretical Framework

The literature on motivation is vast, and the term "motivation" can have different meanings depending on context. For this research, "motivation" is defined as "the process whereby goal-directed activity is instigated and sustained" (Schunk *et al*., 2008, p. 4), and we refer to motivation specifically in the context of low-stakes testing. In this work, we studied motivation by examining students' test-taking behaviors related to the intended goal of students performing to the best of their abilities on GenBio-MAPS. Motivation can be inferred when student behavior aligns with the four indexes of motivation: choice of tasks, effort, persistence, and achievement (Lepper *et al*., 1973; Zimmerman and Ringle, 1981; Salomon, 1984; Pintrich and Schrauben, 1992; Schunk, 1995). Specific test-taking behaviors align with each index of motivation (Table 1). Choice of tasks would be evidenced by students initiating the assessment, but we will not study this here, as we have no information from students who chose not to complete GenBio-MAPS. In the current study, we will focus on test-taking effort (inferred by the three behavioral indicators of self-reported effort, solution behavior, and test completion time), persistence behavior (determined by the amount of time spent on each question as the test progresses), and achievement (measured by GenBio-MAPS score). Each of these indexes of motivation will be discussed in more detail in the following paragraphs.

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

Effort can also be inferred based on timing data from students as they progress through a test, and these data are readily collected by computer-based testing platforms. The amount of time spent per question can be processed to determine the proportion of questions on which students exceed a minimal threshold time (i.e., solution behavior) or to quantify the

TABLE 1. Behavioral indicators associated with test-taking motivation.

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

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

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

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

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

Research Questions

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

METHODS

GenBio-MAPS administration

GenBio-MAPS consists of 39 question stems with four to five true-false (T/F) statements each for a total of 175 accompanying T/F statements that assess *Vision and Change* core concepts (AAAS, 2011). Each student was administered a random subset of 15 question stems and their associated T/F statements. The order of the question stems and T/F statements within each question stem were randomized for each student. Full details regarding the development and administration of the GenBio-MAPS instrument can be found in Couch *et al*. (2019).

Our analyses used the final data set from the instrument development process (Couch *et al*., 2019). These cross-sectional data were collected during the 2016 calendar year from students in 152 biology courses at 20 institutions (Supplemental Table 1). Each student responded at only a single time point and thus is only represented once in this data set. Students completed GenBio-MAPS as part of normal course or program requirements and received course credit or extra credit for completing the instrument. Credit was determined by course instructors, and there was no additional benefit to students based on correctness of responses or the decision to release their responses for research purposes.

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

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

Data Processing, Participation Rates, and Student Demographics

We applied initial and minimal data processing to remove responses that were incomplete, duplicated, or unusable. Note that, although we used the same data set as Couch *et al.* (2019), we targeted a broader range of students in our study and accordingly used less-restrictive data-processing procedures. We first removed submissions from individuals who did not reach the end of the survey, reported being under 18 years of age, did not consent to release survey data, or had already submitted complete survey data in the same course. We also excluded data from individuals who had responded to fewer than 60 T/F statements, a cutoff selected because it represents the minimum number of statements that students could encounter in an administration of 15 GenBio-MAPS question stems. Our final data set contained 8185 responses (Table 2). Roughly 3% of students who remained in the data set did not complete the SOS instrument; these students were only excluded from analyses that involved SOS scores. Response times for individual questions that exceeded 15 minutes represented 1% of the response times recorded, and the data for those pages were replaced with the average page time of 1.5 minutes (Supplemental Table 2).

Identifying Solution Behavior and Persistence Behaviors

We set response time thresholds based on the number of characters in the text of each GenBio-MAPS MTF question, including spaces. The standardized directions in each question and text within figures, graphs, or tables were excluded from the character count. We calculated thresholds based on a rate of 100 characters per second (Supplemental Table 3), which approximates threshold rates used in prior studies (Wise and Kong, 2005; Kong *et al.*, 2007). Response times above the threshold (i.e., solution behavior) were assigned a value of 1, and response times below the threshold (i.e., rapid-selection behavior) were assigned a value of 0. We used the methods established by Wise and Kong (2005) and calculated the sum of the values for solution behavior, then divided by the number of questions on the assessment. The resulting value represented the proportion of test questions for which the student used solution behavior. Consistent with previous studies (Wise and Kong, 2005; Kong *et al.*, 2007), we did not consider the readability of the text (e.g., Flesch reading ease or Flesch-Kincaid level [Flesch, 1948; Kincaid *et al.*, 1975]) when setting the response time thresholds. We determined persistence behaviors by examining changes to the proportion of students using solution behavior and the length of response times for each page in the survey.

Statistical Analyses

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

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

TABLE 2. Student self-reported demographics

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

[a] Numbers do not add to full sample size because some students left the given item blank.

[b] Due to low numbers, responses in this group were excluded from analyses.

[c] Underserved racial/ethnic groups included students who self-identified as African American/Black, Filipino, Hispanic/Latinx, Native American/Alaska Native, Native Hawaiian, and Pacific Islander. This grouping is not intended to obscure the unique histories and identities of any group.

We used JMP (SAS Institute Inc., 2019) to calculate Cronbach's alpha to determine the estimated reliability of the items on the SOS instrument and to estimate standard least squares linear regression models to understand how different variables explained student effort, persistence, and overall score. Predictor variables were tested based on whether they had previously shown significant effects in Couch *et al.* (2019) or were hypothesized to explain variance in the outcome variable. We included self-reported demographic variables as fixed effects and institution as a random effect in our models predicting effort and overall score. Reference groups were selected based on the group having the larger sample size. We included student and question as random effects in our models for test-taking persistence. A correlation matrix for variables is provided in Supplemental Table 4. Given the correlations between predictor variables, we applied a backward stepwise model-selection procedure to address potential issues with multicollinearity (Akaike, 1973). Starting with the highest p-values, nonsignificant variables were individually tested for retention in the model and were only retained if the new model had an Akaike information criterion (AIC) value more than two units greater than the prior model.

Institutional Review Board Approval

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

RESULTS

How Are Students Engaging with the GenBio-MAPS Instrument?

We examined student engagement with GenBio-MAPS based on self-reported effort, solution behavior, and test completion time (Figure 1). The estimated reliability of the SOS effort subscale (using Cronbach's alpha) was 0.81. Most students (86%) reported a score on the effort subscale greater than or equal to 2.5. The mean score on the effort subscale was 3.26, with an SD of 0.72. Most students (90%) used solution behavior on greater than 60% of GenBio-MAPS questions, and 64% of students used solution behavior on every question. Approximately 90% of students had test completion times longer than 10 minutes. The mean test completion time was 27.78 minutes with an SD of 15.11.

We found that the different measures of effort generally correlated with each other (Supplemental Table 4). To understand differences in student motivation classifications, we analyzed how commonly students received the same classification of either "motivated" or "unmotivated" across measures. There was a 72% agreement between self-reported effort and solution behavior. Self-reported effort and test completion time agreed 69% of the time. The two time-based indicators of effort, solution behavior and test completion time, had the largest agreement at 93%. Agreement across all three indicators of effort was 66%. Thus, while there is correspondence across these three indicators of test-taking effort, they each capture slightly different subsets of student behaviors.

Most of the demographic variables that we included in our models significantly predicted scores on the SOS effort subscale (Supplemental Table 5); however, the effect size for each variable was small and the adjusted R^2 for our model was low (0.033). Our results suggest that student demographic characteristics had negligible effects on self-reported effort, which provides further evidence that the SOS effort subscale consistently measures test-taking effort across diverse student populations.

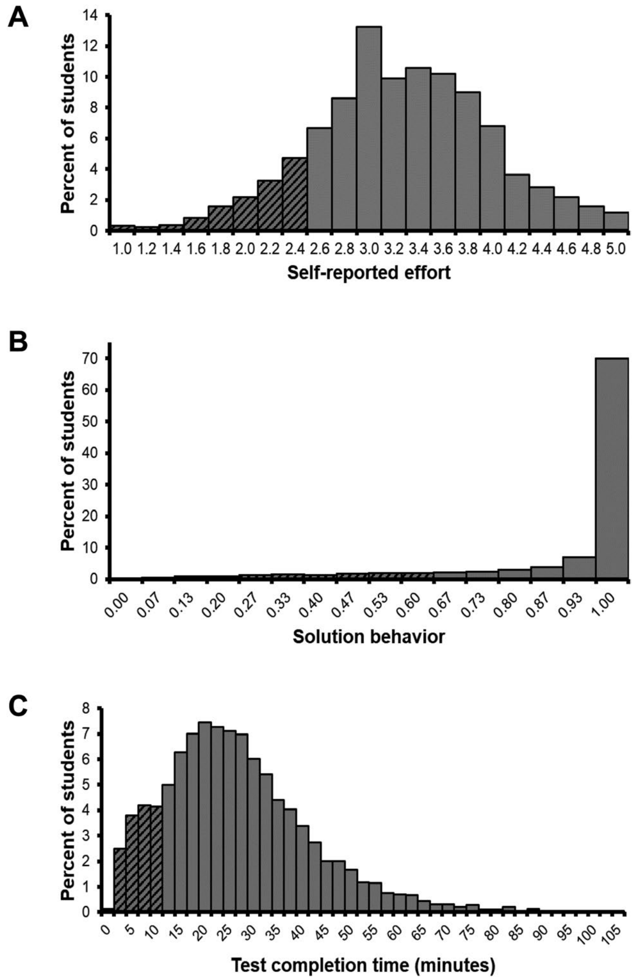

FIGURE 1. Distribution of (A) self-reported effort, (B) solution behavior, and (C) test completion time. The striped portion of each distribution represents the students considered to be demonstrating unmotivated test-taking behavior. (A) Self-reported effort was determined using the average of students' responses to the effort subscale of the SOS instrument. Higher average scores reflect student perception of using a greater amount of effort on GenBio-MAPS. (B) Solution behavior represents the proportion of questions for which a student did not use rapid-selection behavior. (C) The intended test completion time for GenBio-MAPS was 30 minutes.

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

We examined the degree to which students' self-reported effort predicted their observed time-based behaviors, using separate models to predict the effects of student demographics and self-reported effort on solution behavior and test completion time (Supplemental Table 6). We found that most demographic variables had significant ($p < 0.05$) but weak effects on solution behavior and test completion time. These findings suggest that variation in observed time-based behavior cannot be largely attributed to differences in student demographic characteristics.

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

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

We hypothesized that self-reported effort and observed time-based behaviors affect student performance on GenBio-MAPS. Given the correlations between the three indicators of effort, we used regression models to separately test for the effects of self-reported effort, solution behaviors, and test completion time (Supplemental Table 7). In each model, each demographic variable significantly ($p < 0.0001$) predicted score, as we have found previously (Couch *et al.*, 2019). We found that self-reported effort, solution behavior, and test completion time had positive effects on score, indicating that students who reported higher effort, used more solution behavior, or spent longer amounts of time on the test were likely to achieve higher scores. When considered separately, the model containing solution behavior explained more of the variance in score (adjusted $R^2 = 0.418$) compared with self-reported effort (adjusted $R^2 = 0.343$) or test completion time (adjusted $R^2 = 0.350$). When we added all three of these variables into one regression model to look at the combined effects of test-taking effort on score (Table 3), their effect sizes decreased, but the adjusted R^2 of the model increased to 0.452.

Our models indicated that time point in a degree program largely affects GenBio-MAPS performance. As expected, students at later time points in a degree program were predicted to have higher GenBio-MAPS scores than students at earlier points in a degree program. We also examined the interactions between test-taking effort and time point in a degree program. These interactions allow us to determine how effort affects scores at each time point (Figure 2). For self-reported effort, advanced students show a disproportionate benefit as they report increasing effort. For solution behavior, as students reach later time points, their engagement in solution behavior increasingly results in higher scores. Both of these results are consistent with the idea that effort has a greater impact on the performance of students at later time points. For test completion time, students at the end of the introductory series see a disproportionate benefit from taking more time than students at the beginning of the introductory series, but advanced students do not see any further benefit from taking more time to complete the test.

To What Extent Do Students Demonstrate Test-Taking Persistence?

Students used the SOS instrument to report their test-taking effort after completing GenBio-MAPS, but this single data point was not sufficient to capture subtle changes in test-taking effort that may have occurred as the test progressed. Our results

TABLE 3. Standard least-squares linear regression model[a] on the effects of student demographic characteristics and test-taking effort on GenBio-MAPS score

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

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

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

indicate that persistence behaviors generally decreased over the course of the test (Figure 3). When comparing the first and last questions on the test, the proportion of students using solution behavior decreased from 0.99 to 0.83, the average number of minutes per question decreased from 2.1 minutes to 1.3 minutes, and the proportion of students answering correctly decreased from 0.67 to 0.62. Regression models, which account for the difficulty of each randomly displayed question, confirm that the display order of questions had a significant ($p < 0.0001$) negative effect on solution behavior, the amount of time spent on the question, and the score that students achieved on the question (Supplemental Table 8).

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

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

Given that data should not be removed without sufficient cause, we established the criterion that data should only be filtered when there is a compelling indication that a student expended very little effort. Thus, we explored how various filters affect the data set before making recommendations about which filtering strategy is appropriate. First, we analyzed the score distributions of students excluded by each of the filters (Figure 4). We found that students who self-reported low effort on the SOS (<2.5) could still achieve reasonably high scores (i.e., 60–90% correct), suggesting that some high-performing students may not perceive or report themselves to be giving high effort. Conversely, students with low solution behavior (<0.6) or time (<10 minutes) mostly scored below 60% correct, indicating that these filters capture far fewer students with high scores. This pattern also remained when using a dual filter that removed students if they had either low solution behavior or low test completion time. The test scores of students who were removed by this dual filter mirrored but did not completely align with a binomial distribution arising from random responses (Supplemental Figure 1).

We next examined test metrics for the students remaining after application of each filter (Table 4). The filter based on self-reported effort was the most restrictive filter (excluding 16% of the data set) but resulted in the smallest change on the mean test score for the remaining sample. The separate filters based on solution behavior or test completion time performed

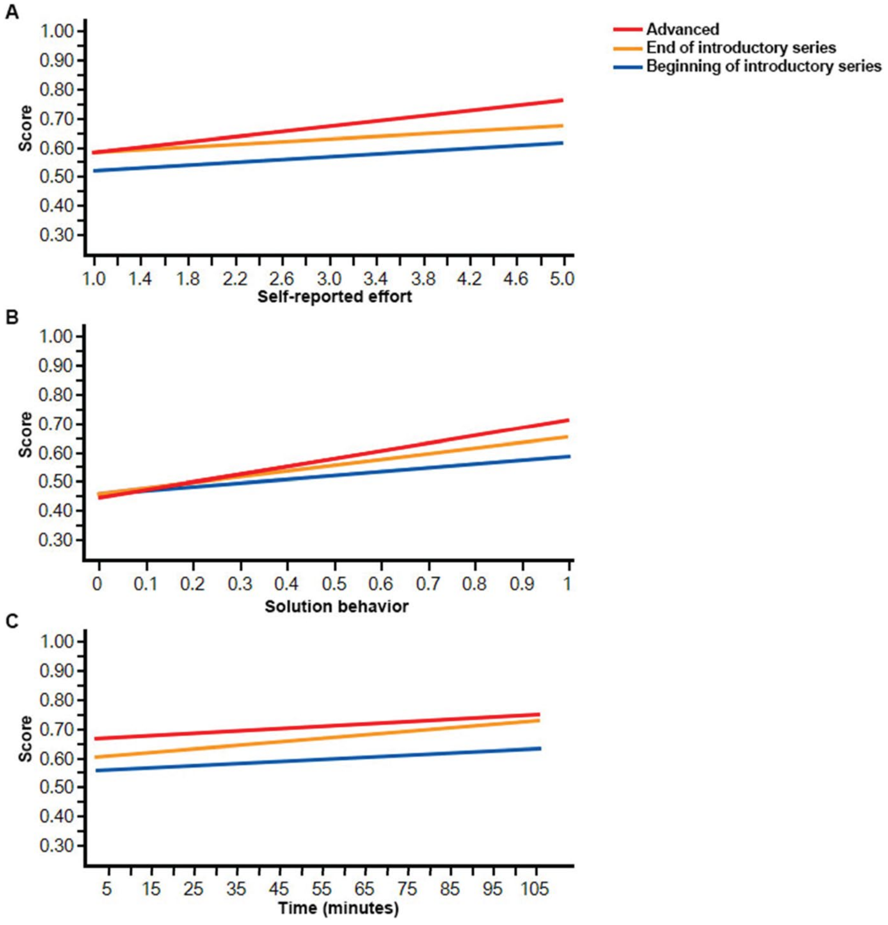

FIGURE 2. Modeled interaction effects between (A) self-reported effort, (B) solution behavior, and (C) test completion time and time point in a degree program on GenBio-MAPS score. Lines represent students enrolled in courses at the beginning of the introductory course series (blue), end of the introductory course series (orange), and end of advanced courses (red).

similarly, which can be attributed to the high agreement between the filters. However, these filters were not synonymous, as the dual filter removed a higher percentage of the sample and resulted in a slightly higher mean test score.

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

DISCUSSION

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

Most Students Used Motivated Behavior on GenBio-MAPS

While one of the goals of our work was to identify and remove scores from students with low test-taking effort, we want to emphasize that this group of students was only a small percentage of our data set. We found that the majority of students (>86%) reported and used motivated behavior when completing GenBio-MAPS and that there was a high degree of consistency across the self-reported and time-based effort measures (Figure 1). Student use of solution behavior on GenBio-MAPS was comparable to student behavior in other low-stake contexts (Wise *et al*., 2006a, 2009; Wise and DeMars, 2010); however, we observed a slightly higher percentage of students reporting motivated behavior on GenBio-MAPS compared with low-stakes general education tests (Schiel, 1996; Hoyt, 2001; Swerdzewski *et al*., 2011). The expectancy-value theory of achievement motivation (Eccles *et al*., 1983; Wigfield and Eccles, 2000) may provide an explanation for this result. This theory states that motivation to perform well on a task is influenced by expectancy for success on the task and the perception that the task is important or interesting. In our context, the task (GenBio-MAPS) is a discipline-specific test that was administered only to students enrolled in biology courses. Thus, the students may have had a greater expectancy to do well on a biology test and may have had greater interest in its biology content, which could have led them to report greater effort compared with a general education test outside the discipline. This interpretation also agrees with our finding that biology majors tended to have higher effort metrics than nonmajors (Supplemental Tables 5 and 6).

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

Although most students engaged in effortful behavior, we noticed a significant effect of question order on student

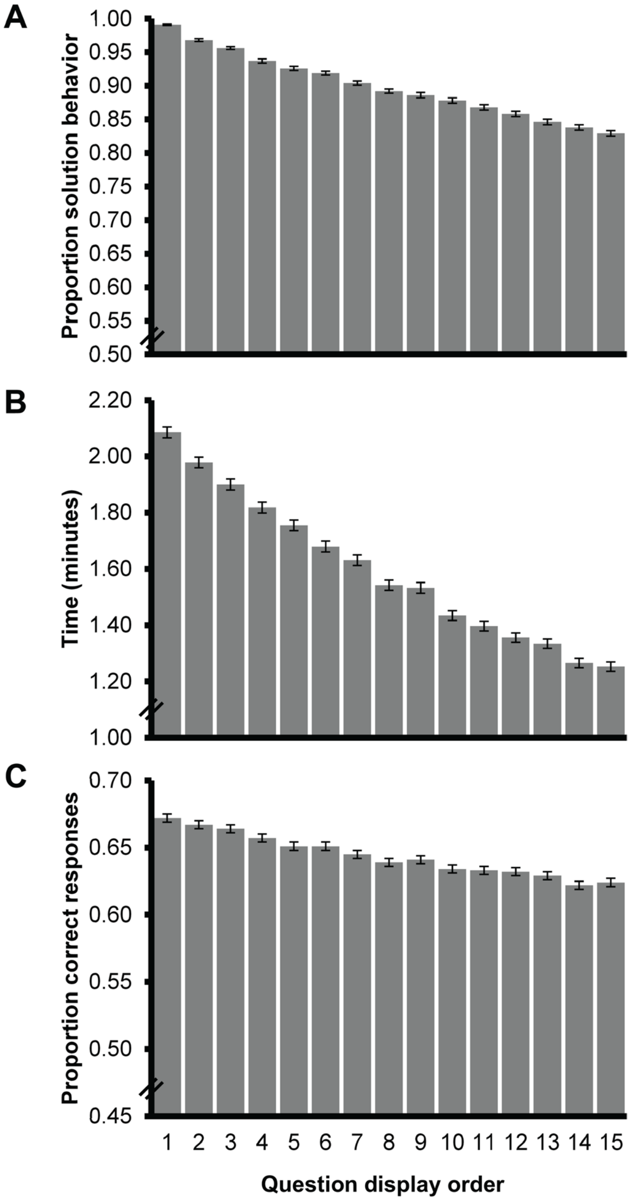

FIGURE 3. Effect of question display order on student test-taking behaviors and performance. Bars represent (A) the proportion of students using solution behavior, (B) the average minutes spent by each student, and (C) the proportion of correct responses for questions shown in each position on the test. Each student received a random subset of 15 GenBio-MAPS questions displayed in a random order, so differences between student behavior or performance on each question cannot be attributed to question characteristics. The *y*-axis for each graph was truncated for emphasis. Error bars represent standard errors.

behaviors. We found that test-taking persistence tended to decrease as students moved through the test (Figure 3; Supplemental Table 8). There was a decreasing proportion of solution behavior with increasing question position, which is a trend that has been documented in other low-stakes assessment contexts (Wise, 2006; Wise *et al*., 2009). The amount of time spent on a question as well as the percentage of correct responses also decreased as students moved through the test. The decrease in time spent on questions may be partially attributed to a growing familiarity with the test format. Each GenBio-MAPS question contains the same line of text providing instructions on how to respond to T/F statements, which students may have ignored later in the test. The decrease in solution behavior and decrease in time spent per question are closely related, because students who do not use solution behavior have inherently short question-response times. Changes in solution behavior and time spent per question both contribute to the decrease in the proportion of correct answers at the overall test level, but our results suggest that solution behavior has a greater influence on GenBio-MAPS score than time (Table 3; Supplemental Table 7).

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

Effortful Behavior Predicts Higher GenBio-MAPS Scores

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

Our models predicted that students who reported and used effortful behavior were likely to have higher scores (Table 3; Supplemental Table 7). This important result is consistent with motivation theory (Pintrich and Schrauben, 1992; Schunk, 1995) and aligns with previous findings in the literature on low-stakes assessments (Wolf and Smith, 1995; Schiel, 1996; Wise and DeMars, 2005; Cole *et al*., 2008; Thelk *et al*., 2009). Our work bolsters existing theory and matches findings from other low-stakes contexts, but we also contributed a new perspective to the field by examining how test-taking motivation is affected by students' time point in a degree program. We found that test-taking effort has a greater effect on student performance at later time points (Figure 2). Our findings suggest that, when students in upper-level courses have low test-taking effort,

TABLE 4. Comparison of filtered scores across methods of motivation filtering[a]

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

[a] Filters listed represent the population that is included in the sample.

[b] Standardized mean score change = $(\text{Mean}_{\text{filtered}} - \text{Mean}_{\text{original}})/\text{SD}_{\text{original}}$.

[c] GPA was self-reported on a scale where 5 = "A–" to "A+" (3.70–4.00); 4 = "B–" to "B+" (2.70–3.69); 3 = "C–" to "C+" (1.70–2.69); 2 = "D–" to "D+" (0.70–1.69); 1 = "E" or "F" (0.00–0.69).

there is likely to be a more pronounced discrepancy between their actual understanding of biology and the level of biology understanding that their low GenBio-MAPS score implies. This underestimation of students' skills and abilities threatens valid interpretation of GenBio-MAPS scores and provides support for the practice of motivation filtering to remove the scores of students with low test-taking effort.

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

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

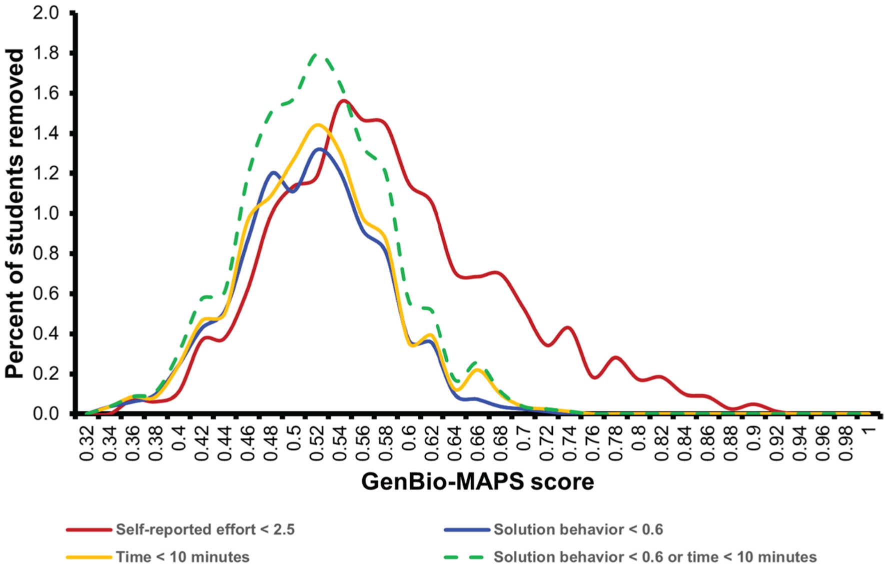

FIGURE 4. Distribution of student responses removed by each motivation filter. Lines represent the percentage of students who were removed by filters for self-reported effort (red), solution behavior (blue), and test completion time (yellow). The dashed green line represents the number of students removed by our recommended motivation filter, which removes students based on either low solution behavior or low test completion time.

While all the motivation filters addressed the effects of low test-taking effort, the filters did not address these effects equally, and they produced subtle differences in resulting scores (Table 4). Given that it is generally not ideal to remove responses from data sets, we sought to identify a filtering strategy that only eliminated data from students who clearly gave an insufficient effort. Based on our findings, we recommend using a dual filter that removes students who had either low solution behavior or short test completion time. While these individual filters largely overlap (93%), using the dual filter helps identify students who may have met one criterion, but who still gave an unsatisfactory effort. For example, a student may have spent just barely more than the threshold time on each question, or a student may have spent less than the threshold time on most questions and a considerable time on a few questions. This filter captures a range of low-effort behaviors that likely introduce construct-irrelevant variance, but it does not remove an excessive number of students from the data set.

Although the data from the SOS instrument are convenient to collect, we do not recommend using the data from the SOS effort subscale as a motivation filter. Compared with the time-based filters, we observed that the SOS filter captured a greater number of responses from students who achieved high scores (Figure 4), which also explains why there was a smaller effect on mean score with this filter. Steedle (2014) observed a similar trend in that many examinees who reported low effort using the SOS instrument actually performed well on the Collegiate Learning Assessment. Steedle proposed several explanations for this result and suggested that it may be attributed to students not accurately providing self-reported data, intentionally selecting inaccurate responses, or making errors when interpreting SOS item wording. Our recommended motivation filter avoids these potential problems with self-reported data and relies only on objective time-based behaviors. After applying the dual filter, departments may still incorporate SOS or time-based variables in their statistical models, although this option may not be viable at institutions with small student numbers.

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

Recommendations for GenBio-MAPS Administration

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

CONCLUSIONS

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

Accessing Instruments

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

ACKNOWLEDGMENTS

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

REFERENCES

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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